ACCESS Pegasus: Bringing Workflows to the ACCESS Masses

Mats Rynge
rynge@isi.edu
USC Information Sciences Institute
USA

Karan Vahi
vahi@isi.edu
USC Information Sciences Institute
USA

Mohammad Zaiyan Alam
mzalam@isi.edu
USC Information Sciences Institute
USA

Ewa Deelman
deelman@isi.edu
USC Information Sciences Institute
USA

Todd Miller
tlmiller@cs.wisc.edu
University of Wisconsin-Madison
USA

Miron Livny
miron@cs.wisc.edu
University of Wisconsin-Madison
USA

Shelley Knuth
Shelley.Knuth@colorado.edu
University of Colorado
USA

James Griffioen
griff@netlab.uky.edu
University of Kentucky
USA

John Goodhue
jtgoodhue@mghpcc.org
Massachusetts Green High
Performance Computing Center
USA

David Hudak
dhudak@osc.edu
Ohio Supercomputer Center
USA

Julie Ma
jma@mghpcc.org
Massachusetts Green High
Performance Computing Center
USA

Andrew Pasquale
andrew@elytra.net
Massachusetts Green High
Performance Computing Center
USA

Lissie Fein
lissie@sweetandfizzy.com
Massachusetts Green High
Performance Computing Center
USA

ACM Reference Format:
Mats Rynge, Karan Vahi, Mohammad Zaiyan Alam, Ewa Deelman, Todd Miller, Miron Livny, Shelley Knuth, James Griffioen, John Goodhue, David Hudak, Julie Ma, Andrew Pasquale, and Lissie Fein. 2023. ACCESS Pegasus: Bringing Workflows to the ACCESS Masses. In *Practice and Experience in Advanced Research Computing (PEARC '23), July 23–27, 2023, Portland, OR, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3569951.3597590

1 INTRODUCTION

ACCESS Support, the user support arm of ACCESS (Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support), has a novel multi-tiered support strategy that comprises 3 major themes: (1) leverage modern information delivery systems to simplify user interfaces; (2) leverage experts from the community to develop training materials and instructions that can dramatically reduce the user learning curve for several increasingly important Cyberinfrastructure (CI) computational techniques; and (3) employ a matchmaking service that will maintain a database of specialist consultants, mentors and student mentees that can be matched with projects to provide the domain-specific expertise needed to leverage ACCESS resources [5]. Both the *Pegasus Workflow Management System (WMS)* [3] and *Open OnDemand*[4] are integral to the first tier, offering user-friendly interfaces to ACCESS resources. Open OnDemand delivers a comprehensive web-based interface, while Pegasus WMS serves as a robust workflow management system capable of managing jobs across ACCESS resources.

PEARC '23, July 23–27, 2023, Portland, OR, USA

ACM ISBN 978-1-4503-9985-2/23/07.
https://doi.org/10.1145/3569951.3597590

2 ACCESS PEGASUS

This poster describes a new user-facing capability, *ACCESS Pegasus* that aims to provide ACCESS users with robust workflow management capabilities [1]. This poster describes this web-based capability that integrates a number of existing cyberinfrastructure (CI) components. The components include Open OnDemand[4], which provides the web-based environment, CILogon[2], which supports secure access to resources, Open Storage Network(OSN), which provides data management[7], and HTCondor[6], which provides workload and resource management, and Pegasus, which enables the definition and execution of user workflows across ACCESS CI. ACCESS Pegasus provides a centralized point of control for "bringing your own capacity" (BYOC), allowing researchers to provision ACCESS resources and to use them for executing their workflows. Designed with ease of use in mind, ACCESS Pegasus offers comprehensive self-guided training modules that guide users

through the process of composing, submitting, monitoring, and debugging their workflows on ACCESS resources.

The following briefly introduces the CI services used to create ACCESS Pegasus:

2.1 Open OnDemand

Open OnDemand[4] is a web-based platform designed to provide researchers with easy access to high-performance computing (HPC) resources. The platform enables users to manage their HPC jobs, files, and data directly through a user-friendly web interface, without requiring extensive knowledge of the command-line interface or the intricacies of HPC systems.

Within the ACCESS Pegasus system, Open OnDemand serves as the web interface, enabling users to employ Jupyter Notebooks and command-line interfaces directly from their web browser. Furthermore, Open OnDemand provides the integration to CILogon[2], enabling seamless login/authentication for any ACCESS user with a current allocation, thereby simplifying access to the ACCESS resources.

2.2 Pegasus

Our underlying strategy for promoting open science and the democratization of access to CI includes the use of workflows as models for specifying the computations that a user want to perform. Workflows provide the necessary structures that allow workflow management systems to provide:

- **Automation:** Workflows automate repetitive and time-consuming tasks, thereby reducing the workload of researchers, and avoiding many human errors;
- **Reusability:** Workflows can be used to build libraries of reusable code and tools that can be adapted by other researchers;
- **Reproducibility:** Workflows allow researchers to document and reproduce their analyses, ensuring their validity.
- **Scalability:** Workflows allow users to scale up their computations to handle large data sets and complex analyses, enabling scientists to tackle more challenging research problems.

Pegasus[3] as a choice of workflow system provides attractive advantages for users running on ACCESS resources namely

Data Management: Pegasus handles data transfers, input data selection and output registration by adding them as auxiliary jobs to the workflow.

Error Recovery: Pegasus handles errors by retrying tasks, workflow-level checkpointing, re-mapping and alternative data sources for data staging.

Pegasus also comes with powerful tools for tracking status of the workflow and debugging failed jobs in a workflow. This is critical in lowering the barrier of usage of ACCESS resources for science as errors often occur on nodes on ACCESS resources, to which the user does not have direct access

2.3 HTCondor Annex

The HTCondor access point (AP) underlies Pegasus and executes its workflows on the resources associated with the AP. Normally, only the administrator of an AP can associate resources to it, so empowering individual scientists means making it possible (and easy) to bring your own capacity to an AP. The HTCondor developers defined easy in this case to include the administrator of the AP: a feature that's hard to set up won't be widely available. Ideally, BYOC would require no additional effort from or user interaction with the AP administrator.

In ACCESS Pegasus, we use the new `htcondor annex` command-line tool, which supports a subset of ACCESS resources and offers a consistent method to manage annexes, named collections of capacity. (Annexes are implemented using the well-established concept and strategy of pilot jobs[8].) Management includes creating, monitoring, and shutting down annexes. An annex will shut itself down if it's idle for longer than a configurable amount of time, but it also allows users to manually stop using capacity when they no longer have jobs to run, minimizing waste.

The HTCondor developers implemented `htcondor annex` as a command-line tool to ensure as broad access as possible. Anyone with an allocation for an ACCESS resource can use SSH to log into that resource and make use of it. This implies the command-line because many of the supported systems require some form of user interaction to log in via SSH.

3 ACCESS PEGASUS SELF GUIDED JUPYTER-BASED TUTORIALS

ACCESS Pegasus also provides self-guided Jupyter Training Notebooks hosted in Open OnDemand that guides users through a complete Pegasus WMS tutorial navigates through various workflow concepts, and shows users how to compose, submit and monitor their workflows using the Pegasus Workflow API. The notebooks also provide real-world workflow examples such as a variant calling workflow (adapted from a popular data carpentry workshop[9] that is setup to run on multiple ACCESS sites and use Open Storage Network (OSN) for data management.

4 FUTURE PLANS/CONCLUSION

Users can explore ACCESS Pegasus by visiting https://support.access-ci.org/pegasus for detailed information about logging in and getting started. We plan to expand the number of Jupyter Workflow Training notebooks to include more AI and ML workflows in the near future.

REFERENCES

[1] ACCESS Pegasus Website [n. d.]. ACCESS Pegasus: Automate your Workflow . https://support.access-ci.org/pegasus.

[2] Jim Basney, Heather Flanagan, Terry Fleury, Jeff Gaynor, Scott Koranda, and Benn Oshrin. 2019. CILogon: Enabling Federated Identity and Access Management for Scientific Collaborations. *PoS* ISGC2019 (2019), 031. https://doi.org/10.22323/1.351.0031

[3] Ewa Deelman, Karan Vahi, Gideon Juve, Mats Rynge, Scott Callaghan, Philip J Maechling, Rajiv Mayani, Weiwei Chen, Rafael Ferreira da Silva, Miron Livny, and Kent Wenger. 2015. Pegasus: a Workflow Management System for Science Automation. *Future Generation Computer Systems* 46 (2015), 17–35. https://doi.org/10.1016/j.future.2014.10.008

[4] David E. Hudak, Douglas Johnson, Jeremy Nicklas, Eric Franz, Brian McMichael, and Basil Gohar. 2016. Open OnDemand: Transforming Computational Science Through Omnidisciplinary Software Cyberinfrastructure. In *Proceedings of the XSEDE16 Conference on Diversity, Big Data, and Science at Scale* (Miami, USA) *(XSEDE16)*. Association for Computing Machinery, New York, NY, USA, Article 43, 7 pages. https://doi.org/10.1145/2949550.2949644

[5] Shelley L Knuth, Julie Ma, Joel C Adams, Alan Chalker, Ewa Deelman, Layla Freeborn, Vikram Gazula, John Goodhue, James Griffioen, David Hudak, et al.

2022. The Multi-Tier Assistance, Training, and Computational Help (MATCH) Project, a Track 2 NSF ACCESS Initiative. *Journal of Computational Science* 13, 2 (2022).
[6] Douglas Thain, Todd Tannenbaum, and Miron Livny. 2005. Distributed computing in practice: the Condor experience. *Concurr. Comput.* 17, 2-4 (Feb. 2005), 323–356.
[7] The Open Storage Network [n. d.]. The Open Storage Network. https://www.openstoragenetwork.org/.
[8] Matteo Turilli, Mark Santcroos, and Shantenu Jha. 2018. A Comprehensive Perspective on Pilot-Job Systems. *ACM Comput. Surv.* 51, 2, Article 43 (apr 2018), 32 pages. https://doi.org/10.1145/3177851
[9] Variant Calling Workflow [n. d.]. Data Wrangling and Processing for Genomics: Variant Calling Workflow . https://datacarpentry.org/wrangling-genomics/04-variant_calling/index.html.