Best Practices of CI/CD for High Performance Computation with Tapis Workflows API

Joe Stubbs[1] , Nathan Freeman[1], Anagha Jamthe[1], Steve Black[1], Sean Cleveland[2], Mike Packard[1]

[1]Texas Advanced Computing Center, Austin, TX, USA
{ jstubbs, nfreeman, ajamthe, scblack, mpackard}@tacc.utexas.edu
[2]University of Hawaii - System, Honolulu, HI, USA
seanbc@hawaii.edu

Abstract:

Developing research computing workflows often demands significant understanding of DevOps tooling and related software design patterns, requiring researchers to spend time learning skills that are often outside of the scope of their domain expertise. Tapis Workflows API, which is a production-grade service, provides researchers with a tool that simplifies the creation of their workflows. It simplifies the creation of workflows by abstracting away the complexities of the underlying technologies behind a user-friendly API that integrates with HPC resources available at any institution with a Tapis deployment. It offers users the ability to design and automate workflows in which they can build container images, run scientific computing jobs, and trigger subsequent workflow tasks on independent systems with webhook notifications as shown in figure below.

This tutorial will explain how to effectively create scientific workflows leveraging the NSF-funded Tapis v3 platform, an Application Program Interface (API) for distributed computation. Using techniques covered in the tutorial, attendees will be able to easily automate and validate their CI/CD pipelines for scientific computing applications. We will include several hands-on exercises that will allow attendees to build an automated CI/CD pipeline for a large scientific application. This workflow can be seamlessly moved to different execution environments, including a small virtual machine and a national-scale supercomputer.

Description and Format

We will set up a TACC project for all the registered attendees, which will have access to allocations on HPC resources such as TACC's Stampede2 supercomputers. Attendees will use their own TACC accounts, and they can register for a new account in a few minutes before the workshop. Course material including the slides and hands-on exercises for the tutorial will be published on Github pages and will remain available to the attendees during and after the tutorial.

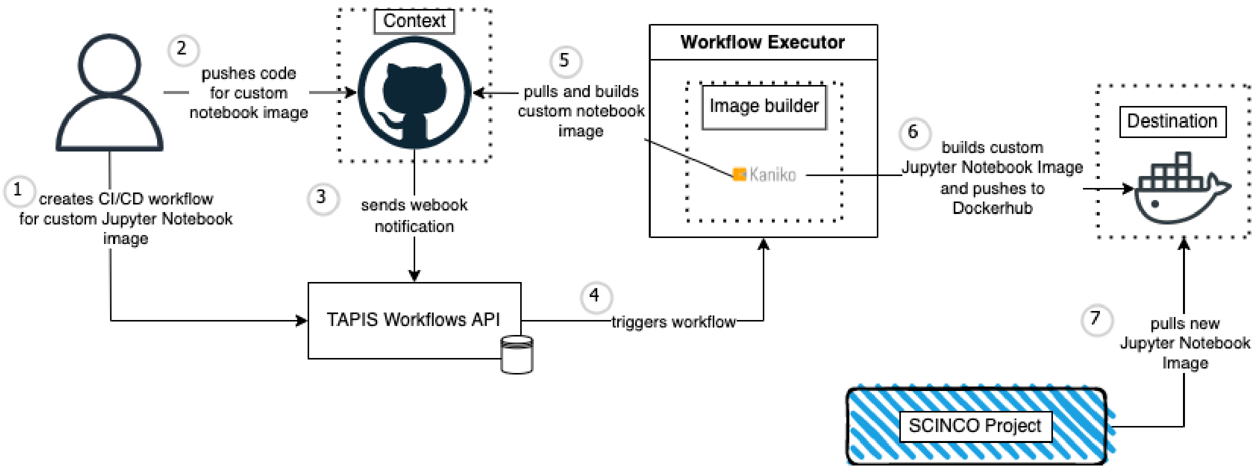
Context
1
creates CI/CD workflow for custom Jupyter Notebook image
2
pushes code for custom notebook image
3
sends webook notification
TAPIS Workflows API
4
triggers workflow
5
pulls and builds custom notebook image
Workflow Executor
Image builder
Kaniko
6
builds custom Jupyter Notebook Image and pushes to Dockerhub
Destination
7
pulls new Jupyter Notebook Image
SCINCO Project

Fig: Tapis Workflows architecture

Introduction to CI/CD for HPC codes and Tapis will include theoretical concepts followed by simple hands-on-exercises. Attendees will be able to run exercises in Jupyter Notebooks and TapisUI, a serverless interactive science gateway that runs entirely on GitHub pages to interact with different Tapis services. A pre-computed example Jupyter notebook, containing a description of building and running the scientific workflows will be available in case of internet outage. The tutorial will have a good mix of presentation and hands-on exercises for the attendees to learn and implement their scientific computing workflows. We will have enough proctors throughout the session, who will help attendees throughout the tutorial. Proposed tutorial schedule is as follows:

Time	**Block**	**Topics**
50 min	1	Introduction to Tapis and Workflows for CI/CD Tapis Resource Creation: Systems, Apps and Jobs
30 min	1	Introduction to CI/CD for HPC codes
10 min	1	Image build with Workflows using TapisUI
10 min		Break
15 min	2	Image build with Workflows
50 min	2	Automating Validation Steps Within CI/CD Workflows using TapisUI
15 min	2	Use cases and Q/A discussion

Tutorial Duration: 3 hours

Learning Outcomes:

By the end of this workshop attendees will be able to:

- Create pipelines that define workflows for a real-world scientific use-case.
- Have a basic understanding of Docker containers in relation to computational research.
- Use Tapis to access HPC storage and compute resources in a programmatic and reproducible way.
- Utilize Tapis UI to interact with Tapis services.

Target Audience

The audience for this workshop fits into three categories:

1) Researchers that utilize national, campus and local cyberinfrastructure resources and wish to do so in a reproducible, scalable and programmable manner.

2) Cyberinfrastructure specialists such as research software engineers (RSE), gateway providers/developers and infrastructure administrators. People in these roles can utilize open source technologies and state-of-the-art techniques to enable portable, reproducible computation.

3) Cyberinfrastructure directors, managers and facilitators that are looking for solutions to aid and educate their institutional researchers in order to better leverage local and distributed computational and cyberinfrastructure resources.

Content Level

Beginner 70%, Intermediate 30%

Audience Prerequisites

Attendees will be required to create a TACC account. They should have DockerHub and GitHub accounts. Attendees must use their own laptop for the hands-on part of the tutorial.

Acknowledgement

This workshop material is based upon work supported by the National Science Foundation Plant Cyberinfrastructure Program (DBI-0735191), the National Science Foundation Plant Genome Research Program (IOS-1237931 and IOS-1237931), the National Science Foundation Division of Biological Infrastructure (DBI-1262414), the National Science Foundation Division of Advanced CyberInfrastructure (1127210) and (1931439 and 1931575).

Presentation team

Dr. Joe Stubbs is a Research Associate and leads the Cloud and Interactive Computing (CIC) group at the Texas Advanced Computing Center at the University of Texas at Austin. Dr. Stubbs is currently the PI of two NSF-funded projects and has played a fundamental role in developing numerous national-scale cyberinfrastructure systems for various scientific and engineering communities used by thousands of researchers.
Previous trainings:

- PEARC 22: Building Portable, Scalable and Reproducible Scientific Workloads across Cloud and HPC for Gateways
- TACCster 2022: Tapis day at TACC.
- Gateways 21: Portable, Scalable, and Reproducible Scientific Computing: from Cloud to HPC
- PEARC 2020: Leveraging Tapis For Portable, Reproducible High Performance Computing In the Cloud
- Gateways 2019: Portable, Reproducible High Performance Computing In the Cloud.
- PEARC 2019: Portable, Reproducible High Performance Computing In the Cloud.

Nathan Freeman is an Engineering Scientist Associate in the Cloud and Interactive Computing (CIC) group at the Texas Advanced Computing Center at the University of Texas at Austin. Nathan manages the development of the Tapis Workflows API and related services, libraries, and UI.
Previous trainings:

- TACCster 2022: Tapis day at TACC.

Dr. Anagha Jamthe is a Research Associate in the Cloud and Interactive Computing group (CIC) at the Texas Advanced Computing Center at the University of Texas at Austin. Dr. Jamthe has played a key role in developing and validating Tapis APIs.
Previous trainings:

- TACCster 2022: Tapis day at TACC.
- PEARC 22: Building Portable, Scalable and Reproducible Scientific Workloads across Cloud and HPC for Gateways
- Gateways 21: Portable, Scalable, and Reproducible Scientific Computing: from Cloud to HPC
- PEARC 2020: Leveraging Tapis For Portable, Reproducible High Performance Computing In the Cloud
- Gateways 2019: Portable, Reproducible High Performance Computing In the Cloud.
- PEARC 2019: Portable, Reproducible High Performance Computing In the Cloud.

Dr. Sean Cleveland is a Cyberinfrastructure Research Scientist at the University of Hawaii in the Information Technology Services Cyberinfrastructure group and affiliate researcher at the Hawaii Data Science Institute(HI-DSI). In his current roles, he develops and deploys software and cyberinfrastructure solutions, such as science gateways, along with advanced cyberinfrastructure supporting scientific data management, computation and collaboration. As Co-PI on the Tapis framework grant and senior personnel on a number of NSF grants he also assists researchers in learning, adopting, and applying advanced cyberinfrastructure to accelerate and scale their research efforts.
Previous trainings:

- PEARC 22: Building Portable, Scalable and Reproducible Scientific Workloads across Cloud and HPC for Gateways
- Gateways 21: Portable, Scalable, and Reproducible Scientific Computing: from Cloud to HPC
- Leveraging Tapis For Portable, Reproducible High Performance Computing In the Cloud
- Gateways 2019: Portable, Reproducible High Performance Computing In the Cloud.
- PEARC 2019: Portable, Reproducible High Performance Computing In the Cloud.

Dr. Steve Black is an Engineering Scientist in the Cloud and Interactive Computing group at the Texas Advanced Computing Center at the University of Texas at Austin. Dr. Black has played a vital role in the development of Tapis APIs.
Previous trainings:

- Gateways 22: Building Portable, Scalable and Reproducible Scientific Workloads across Cloud and HPC for Gateways
- Gateways 21: Portable, Scalable, and Reproducible Scientific Computing: from Cloud to HPC
- PEARC 2020: Leveraging Tapis For Portable, Reproducible High Performance Computing In the Cloud.

Mike Packard is a Senior Systems Administrator in the Cloud and Interactive Computing group (CIC) at the Texas Advanced Computing Center at the University of Texas at Austin. He has 20 years of experience in Linux administration, scientific computing, devops, and cloud computing. He has helped build & support several generations of systems for NSF projects used by a variety of users across many scientific domains.
Previous trainings:

- Gateways 22: Building Portable, Scalable and Reproducible Scientific Workloads across Cloud and HPC for Gateways
- Gateways 21: Portable, Scalable, and Reproducible Scientific Computing: from Cloud to HPC
- PEARC 2020: Leveraging Tapis For Portable, Reproducible High Performance Computing In the Cloud
- Gateways 2019: Portable, Reproducible High Performance Computing In the Cloud.

- COE 332 Software Engineering and Design (Upper division course, Fall 2018, Spring 2020); Assistant course designer, Computational Engineering UT Austin.

Resources

1) Tapis Project: https://tapis-project.org
2) Tapis v3 documentation: https://tapis.readthedocs.io/en/latest/
3) Tapis Slack http://bit.ly/join-tapis