Simplifying Scientific Application Access in Kubernetes with Push Button Deployments

Taylor Johnston
University of Washington Bothell
Bothell, USA

Victor Shaw
University of Washington Bothell
Bothell, USA

Brian Werts
Purdue University
West Lafayette, USA

Sam Weekly
Purdue University
West Lafayette, USA

Erik Gough
Purdue University
West Lafayette, USA

Preston Smith
Purdue University
West Lafayette, USA

ABSTRACT

The Geddes Composable Platform is a Kubernetes-based private cloud resource at Purdue University. To streamline adoption of the platform and lower the barrier to entry, we created push button deployments for some of the popular applications used by Purdue researchers and made them available via Geddes' Rancher web-based user interface using Helm and Rancher Charts. With little knowledge of the underlying system, a new user can use a web form to deploy custom applications, including JupyterHub instances, Alphafold, CryoSPARC and the Triton Inference Server.

CCS CONCEPTS

• **Information systems** → **Computing platforms**; • **Computer systems organization** → *Cloud computing*.

KEYWORDS

Kubernetes, containers, composable systems

ACM Reference Format:
Taylor Johnston, Victor Shaw, Brian Werts, Sam Weekly, Erik Gough, and Preston Smith. 2022. Simplifying Scientific Application Access in Kubernetes with Push Button Deployments. In *Practice and Experience in Advanced Research Computing (PEARC '22), July 10–14, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3491418.3535164

1 INTRODUCTION

The Geddes Composable Platform, a multi-tenant Kubernetes-based private cloud resource at Purdue University, is used to deploy containerized applications, data analysis tools and elastic software stacks for scientific computing. Geddes uses Rancher to provide web-based access to the Kubernetes cluster. Kubernetes has a steep learning curve and deployments can be very complex, requiring custom provisioning of compute, storage and networking resources. To reduce this complexity and lower the barrier to entry for deploying applications in Kubernetes, we designed custom push button deployments for popular applications and made them available to users via the Rancher UI.

PEARC '22, July 10–14, 2022, Boston, MA, USA

ACM ISBN 978-1-4503-9161-0/22/07.
https://doi.org/10.1145/3491418.3535164

2 COMPONENTS

The core components used to create push button applications are Helm and the Rancher Apps & Marketplace feature. Rancher supports deploying applications from both Helm charts and Rancher charts. Rancher charts are an extension of Helm charts, designed to simplify the user experience by automatically creating web-based forms for entering required Helm chart variables.

3 APPLICATIONS

Rancher provides access to several applications by default via the Apps & Marketplace feature. Examples include MySQL, MongoDB, Kafka and ELK stacks. Custom charts were created for the following additional applications.

JupyterHub: A Rancher chart was created based on the Zero-to-Jupyterhub Helm chart, allowing the user to integrate custom notebooks, create a custom URL for the deployment and configure authentication mechanisms. For example, a user can deploy a JupyterHub that uses Purdue's LDAP authentication with authorization based on a specific UNIX group.

Triton: The NVIDIA Triton Inference Server [4] provides an simple interface to deploy, access and scale AI models for inference. Helm and Rancher charts were created based off of the Triton container image from NVIDIA GPU Cloud (NGC) [3], allowing a user to serve models from a shared storage location and use Prometheus metrics with the horizontal pod autoscaler (HPA).

Alphafold: Alphafold is a neural network-based protein folding software used to predict the structure of a protein based on its amino acid sequence [2]. Helm and Rancher charts were created allowing users to enter a protein sequence in a web form and perform GPU-enabled searches from a shared database location using a customized version of the Alphafold Docker image [1].

CryoSPARC: CyroSPARC is software that processes cryo-electron microscopy (cryo-EM) data to solve cryo-EM structures such as membrane proteins, viruses, and small particles. Helm and Rancher charts were created to integrate users' license keys, GPUs and web-based access to the CryoSPARC interface via a custom Docker image.

4 CONCLUSIONS AND FUTURE WORK

This work has been successful in allowing users to easily deploy applications in Kubernetes, with the most popular deployment being JupyterHub instances for research groups and data science courses. We plan to continue this work by deploying additional charts for RStudio to support additional courses, BioContainers to

support life sciences researchers, and other NGC containers for AI/ML research. We will also generalize and publish Helm charts for use by the larger RCD community.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 2018926. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Alphafold. 2022. Alphafold Github Repository. Retrieved March 31, 2022 from https://github.com/deepmind/alphafold

[2] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 7873 (2021), 583–589. https://doi.org/10.1038/s41586-021-03819-2

[3] NVIDIA. 2022. GPU-optimized AI, Machine Learning, & HPC Software. Retrieved March 31, 2022 from https://catalog.ngc.nvidia.com

[4] NVIDIA. 2022. NVIDIA Triton Inference Server. Retrieved March 31, 2022 from https://developer.nvidia.com/nvidia-triton-inference-server