9th International Workshop on Science Gateways (IWSG 2017), 19-21 June 2017

Containerizing Conda

Experiences Building 2000 Applications As Portable Docker And Singularity Images

John Fonner
Texas Advanced Computing Center
University of Texas at Austin
Austin, TX USA
jfonner@tacc.utexas.edu

Rion Dooley
Texas Advanced Computing Center
University of Texas at Austin
Austin, TX USA
dooley@tacc.utexas.edu

ABSTRACT

Over the last 5 years, container technologies have gone from a novelty, to a niche, to a necessity for many research communities. As both infrastructure and cyberinfrastructure providers, the authors have worked to support their user communities in their adoption and use of these technologies. While Docker [1] is still the dominant mainstream container technology, the space has begun to fragment as containers move past the hype and into broad usage. Alternative image formats and container runtimes have emerged, providing compelling use cases and advantages over Docker for compute containers [2][3][4][5].

Compute containers have provided scientists a way to easily improve reproducibility and simplify the process of sharing apps with others in the field. While the concept of pre-packaged application images is similar to the way many scientists have leveraged virtualization technologies in create prepackaged machine images for use in cloud environments, the use of containers reduces the overhead of moving software across incompatible cloud providers and, ideally, enables portability between local, HPC, and cloud compute resources.

While several efforts are underway attempting to build up libraries of user-provided application codes, the Bioconda [6] project has taken a more aggressive approach by pairing Docker containers with the Bioconda repository of bioinformatics tools. The BioContainers project has generated containers for over 2000 scientifically relevant applications [7]. By building off an existing, popular polyglot dependency management system, they have taken a significant step forward in making containers approachable for anyone currently using Bioconda. Ironically, many of their academic users are still hampered by a lack of compatible container runtime environments on academic compute clusters and supercomputing systems. Those systems cannot support Docker because of issues such as existing security restrictions, incompatible versions of the Linux kernel, and hardware incompatibilities, just to name a few. These same systems are, however, increasingly choosing to support Singularity.

For the last year, the Texas Advanced Computing Center has worked closely with researchers in the CyVerse [8], Agave [9], and DesignSafe [10] projects to develop a library of containerized applications in both Docker and Singularity image formats that can be used pervasively across HPC, cloud, and local resources. Additionally, by creating a gateway that enables researchers to run these containers on academic cyberinfrastructure, we bridge the disconnect between app discovery and execution that remains unmet by existing registry services like [11-14], execution services like [15-18], and infrastructure services like [19-22]. In this talk, we present initial insights and progress adding the first 2000 images to our library, and the advantages of supporting dual container runtimes. As time permits, we will demonstrate how Agave's Application Exchange leverages the library to enable one click, reproducible execution of any code in the

BioConda repository on HPC, Cloud, and local resources.

Keywords—Docker; Singularity; Agave; App Exchange; reproducibility; portability; containers; BioConda

REFERENCES

[1] Docker, https://docker.com.

[2] Rocket, https://coreos.com/rkt

[3] Kurtzer, Gregory M., Vanessa Sochat, and Michael W. Bauer. "Singularity: Scientific Containers for Mobility of Compute." PLOS ONE 12, no. 5 (2017): 1–20. doi:10.1371/journal.pone.0177459.

[4] Canon, Richard Shane, and Douglas M. Jacobsen. "Shifter: Containers for HPC." In Proceedings of the 2016 Cray User Group Meeting. London, UK, 2016. https://cug.org/proceedings/cug2016_proceedings/includes/files/pap103.pdf.

[5] LXD, https://www.ubuntu.com/containers/lxd.

[6] Bioconda, https://bioconda.github.io/.

[7] Leprevost, F. D., et al. "BioContainers: An open-source and community-driven framework for software standardization." Bioinformatics (Oxford, England) (2017), doi: 10.1093/bioinformatics/btx192.

[8] Goff, Stephen A., et al., "The iPlant Collaborative: Cyberinfrastructure for Plant BIology," Frontiers in Plant Science 2 (2011), doi: 10.3389/fpls.2011.00034.

[9] Dooley, Rion, et al. "Software-as-a-Service: The iPlant Foundation API", 5th IEEE Workshop on Many-Task Computing on Grids and Supercomputers (MTAGS). IEEE, 2012.

[10] Rathje, E., Dawson, C. Padgett, J.E., Pinelli, J.-P., Stanzione, D., Adair, A., Arduino, P., Brandenberg, S.J., Cockerill, T., Dey, C., Esteva, M., Haan, Jr., F.L., Hanlon, M., Kareem, A., Lowes, L., Mock, S., and Mosqueda, G. 2017. "DesignSafe: A New Cyberinfrastructure for Natural Hazards Engineering," ASCE Natural Hazards Review, doi:10.1061/(ASCE)NH.1527-6996.0000246.

[11] Docker Store, https://store.docker.com/.

[12] Quay, https://quay.io.

[13] Nexus, https://nexus.com.

[14] Artifactory, https://artifactory.com.

[15] Mesos, https://mesos.apache.org.

[16] Kubernetes, https://kubernetes.io.

[17] Moab, http://www.adaptivecomputing.com/products/hpc-products/moab-hpc-basic-edition/.

[18] HTCondor, https://research.cs.wisc.edu/htcondor/.

[19] Enis Afgan, Dannon Baker, Marius van den Beek, Daniel Blankenberg, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Carl Eberhard, Björn Grüning, Aysam Guerler, Jennifer Hillman-Jackson, Greg Von Kuster, Eric Rasche, Nicola Soranzo, Nitesh Turaga, James Taylor, Anton Nekrutenko, and Jeremy Goecks. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Research (2016) 44(W1): W3-W10 doi:10.1093/nar/gkw343

[20] EGI Applications on Demand service, https://access.egi.eu/start.

[21] Nectar, https://nectar.org.au/about/.

[22] Fischer, Jeremy, Steven Tuecke, Ian Foster, and Craig A. Stewart. "Jetstream: A Distributed Cloud Infrastructure for Underresourced Higher Education Communities." In Proceedings of the 1st Workshop on The Science of Cyberinfrastructure: Research, Experience, Applications and Models, 53–61. SCREAM '15. New York, NY, USA: ACM, 2015. doi:10.1145/2753524.2753530.