Introduction to Application Containerization Using Singularity [Tutorial]

William S. Sanders, Jason S. Macklin, Richard Yanicky, Aaron McDivitt

The Jackson Laboratory

www.jax.org

The 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB 2021)

August 1-4, 2021 (Virtual)

Course Content

https://github.com/TheJacksonLaboratory/acm-bcb-singularity2021

Overview

In this tutorial, participants will learn the following:

- What a software container is, and why adoption of containerized applications is increasing
- What are the existing, online resources for containers (DockerHub, Singularity-Hub, etc.)
- How to navigate and interact with Singularity containers from the Linux command line
- How to build their own containers both using container definition files and from scratch
- How to leverage containers to optimize their existing scientific workflows

Scientific Reproducibility

Reproducibility

- Rising concern about lack of repeatability, replicability and reproducibility in science and engineering
- A number of high-profile publications in a large number of domains have not been successfully replicated when efforts to reproduce the work have been attempted
- Variation in data collection methodologies, experimental environments, computational configuration, the lack of detailed and intricate documentation, and more are leading to a call for enhanced peer review and validation of experimentally produced artifacts

Artifacts

- Digital artifacts are produced during the course of research
  - Can be either inputs or outputs of a research project
  - Examples include input data sets, raw data, software systems, scripts to run experiments or analyze results
- ACM Digital Artifacts
  - ACM has initiated formal processes for artifact review to accompany publication

Artifacts Evaluated – Functional 

Artifacts Evaluated – Reusable 

Artifacts Available 

Results Replicated 

Results Reproduced 

Artifact Repositories

Code Repositories

- Publicly available, generally commercial solutions:
  - GitHub
  - BitBucket
  - Sourceforge

Data Repositories

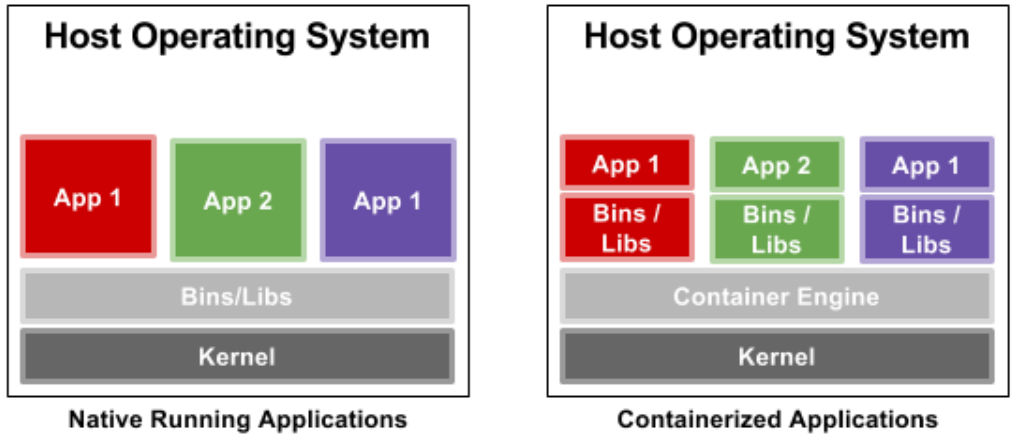
- Generally domain specific and administered:
  - Biology: GenBank, NCBI Sequence Read Archive, etc.
  - Chemistry: chEMBL, caNanoLab, STRENDA, etc.
  - Earth, Environmental, and Space: NASA Goddard Earth Sciences Data and Information Services Center, PANGAEA
  - Astronomy: SIMBAD, UK Solar System Data Centre
  - Ocean Sciences: SEANOE, Australian Ocean Data Network

…

Containerization

Linux Containers

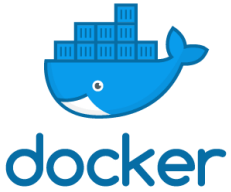
- Standard units of software that package up code and all dependencies so an application can run quickly and reliably from one computing environment to another
- Containers provide operating-system level virtualization through a virtual environment that has its own process and network space, as opposed to a full-fledged virtual machine



Containerization

- Applications distributed as containers have the same installation overhead as a normal instance of the application, but with a slight additional overhead requiring knowledge of the containerization framework.
- Once the container is generated, the full burden of application installation and configuration is abstracted from the end users of application.
- Containers are easily migrated from system to system, and are easy to distribute and provide as digital artifacts

- Docker is the clear leader in containerization frameworks for cloud computing ecosystems



- Some concern about the security of Docker containers in non-cloud environments.
- Singularity is emerging as the containerization framework of choice in HPC environments.

Singularity

- Developed in 2015 as an open-source project by researchers at Lawrence Berkeley National Laboratory led by Gregory Kurtzer
- Deployed at Ohio State University, Michigan State University, Texas Advanced Computing Center, San Diego Supercomputer Center, Oak Ridge National Laboratory, etc.

What makes Singularity different from other containerization solutions:

- Reproducible software stacks – container image allows easy verification
- Mobility of compute – containers can be transferred with standard tools
- Compatibility with complicated architectures – compatible with existing infrastructure and legacy systems
- Security model – security model of untrusted users running untrusted containers

Singularity

- Enables users to have full control of their environment
- Can package entire scientific workflows, software, libraries, even data
- Empowers users, no need to ask your IT system admin to install anything
- Compatible with Docker
- Secure!

Container Registries

- DockerHub (https://hub.docker.org)
  - Online repository of Docker container images, can be built on demand by various triggers
  - ~2,257,118 available container images
- SingularityHub (https://singularity-hub.org)
  - Developed and maintained by Stanford Research Computing and Stanford Medicine
  - ~4,096 containers across ~2,167 collections