Canadian Bioinformatics Workshops

www.bioinformatics.ca
@Bioinfodotca
@bioinfojobs
Facebook.com/bioinformaticsdotca
Bioinformatics-ca.github.io
Learning Objectives of Module

Participants will gain practical experience and skills to be able to:

- Understand the current status of cloud solutions and providers available
- Understand what Docker is, its benefits and the differences between a Docker container and a virtual machine
- Docker’s importance in bioinformatics
- Overview of Dockstore, a bioinformatics focused Docker registry developed by OICR and its current state
Reproducibility Evolution

Months → Days → Minutes → Seconds

Cloud Technology Providers

- Public
  - Amazon Web Services
  - Google
  - Azure

- Private
  - OpenStack
Cloud computing with Openstack

• Openstack is a free and open-source software platform for cloud computing
• Initially developed by NASA and Rackspace in 2010 and rapidly adopted by most hardware and software companies
• Written in Python, it has the same “building blocks” as AWS, Google Cloud Compute and Microsoft Azure, without the cost and lack of access to the source code.

Openstack

• Release cycle of 6 months

• Initially focused on core services (compute, networking, block storage, identity, image repository, object storage and dashboard).

• Currently, there are more than 19 projects providing additional services (https://www.openstack.org/software/project-navigator/)
  – Telemetry
  – Orchestration
  – Database
  – Bare-metal provisioning, etc.
Openstack – a complex distributed system

Openstack dashboard

Source: http://docs.openstack.org/ops-guide/architecture.html

Module 3  bioinformatics.ca
Openstack Neutron

Network Topology

Restrict the canvas by scrolling up/down with your mouse/trackpad or the topology. Pan around the canvas by clicking and dragging the space behind the topology.

Toggle Colors  Toggle Network Collapse

Launch Instance  Create Network  Create

Openstack – large users

at&t  verizon  Deutsche Telekom  comcast

Walmart  Best Buy  ebay  PayPal  SONY

Volkswagen  BMW
**Openstack used heavily in research**

- Most academic and research institutions have adopted Openstack in order to replace or complement their traditional HPC clusters.
- Benefits include flexibility, self-service, and root access provided to its users.

**Cloud extended functionality**

- Software defined networking (SDN) – allows users to create virtual networks that replicate their physical environments, including IP addressing scheme.
- Block storage – create raw block storage on demand (like an external hard drive) and attach it to an instance. This feature can satisfy additional storage needs, and persist critical data in case the instance is unavailable or terminated.
- Object storage – save files in an HTTP-accessible object store (like Google drive or Dropbox).
**Cloud-init**

A package pre-installed on most recent cloud Linux distributions that can:

- grow the root file system to match the VM’s flavor
- set an SSH key or a root password
- make available a file provided by the user
- run a script provided by the user
- do a callback to an URL when a VM is finished booting or execute commands at boot

More info and examples are available at:

---

**Docker**

- An open-source project that automates the deployment of Linux applications inside software containers

- Containers wrap a piece of software in a complete filesystem that contains everything needed to run:
  - code, runtime, system tools, binaries

- Guarantees that the software will always run the same, regardless of its environment
Docker

- Docker containers allow a user to run an application that requires the latest version of a library, although the system has an older library that cannot be updated.
- Docker containers are much smaller than full VMs, making it easier to share.
- More containers can run on a physical server (than VMs)
- Faster start-up times than VMs
- Closer to bare-metal performance
- Because of security concerns though, most environments run container(s) on top of VMs
- Using Docker containers from untrusted sources is a major security risk
Docker

Containers vs. VMs

Containers are isolated, but share OS and, where appropriate, bins/libraries.

Result is significantly faster deployment, much less overhead, easier migration, faster restart.

Docker

Why are Docker containers lightweight?

Original App
No OS to take up space, resources, or require restart

Copy of App
No OS. Can share bins/libs

Modified App
Copy on write capabilities allow us to only save the diffs between container A and container A'.
Dockstore

Dockstore is a project actively developed with new features and containers added monthly.

Dockstore is an OICR-developed Docker registry for bioinformatics (https://dockstore.org)

It currently contains 27 bioinformatics tools packaged in Docker containers, and a number of workflows are being added.

It allows researchers to collaborate and share privately their workflows packaged in containers.
Dockstore and Dockerhub

- Dockerhub: the official registry for Docker containers
- Dockstore: built specifically for bioinformatics and allows you to describe your containers with the Common Workflow Language (CWL), an emerging standard used by the GA4GH Containers and Workflow working group.