Increasing reproducibility in training by using VM images

Ideas from the EXCELERATE workshop "Using clouds and VMs in bioinformatics training"

Eija Korpelainen, CSC – IT Center for Science, Finland

www.elixir-europe.org

Workshop Using clouds and VMs in bioinformatics training

- May 2016 at ELIXIR-FI, 33 trainers and technical experts from 13 different countries
- Talks, discussions, hands-ons
- Materials and videos available at https://github.com/ekorpela/cloud-vm-workshop
- Speakers
  - *Ola Spjuth (UPPMAX, ELIXIR-SE)*
  - *Markus van Dijk (SURFsara, ELIXIR-NL)*
  - *Andrew Lonie & Simon Gladman (EMBL-ABR)*
  - *Annette McGrath (CSIRO, Australia) & Jerico Revote (Monash eResearch Centre, Autralia)*
  - *Pedro Fernandes (ELIXIR-PT)*
  - *Stefano Nicotri (INFN, ELIXIR-IT)*
  - *Christophe Blanchet and Victoria Dominguez (ELIXIR-FR)*
  - *Diego Scardaci (EGI.eu/INFN) and Fotis Psomopoulos (Aristotle University of Thessaloniki)*
  - *Abdulrahman Azab (ELIXIR-NO)*
  - *Nicolas Delhomme (Umeå Plant Science Center, Sweden)*
  - *Kalle Happonen, Jukka Nousiainen, Olli Tourunen, Kimmo Mattila (ELIXIR-FI)*

Outline

- Introduction
- VM and Docker image
  - What are they?
  - How do they differ?
  - How to make them?
- What kind of systems are there in different countries?
- How can we make it easier for trainers to use images?

What is the problem in setting up a training environment?

- Typically a lot of software and reference data need to be installed → installation takes time, you need somebody with admin rights
- Students need to have identical installation → if they come with their own laptops, this is difficult to achieve
- Your course will be repeated in different location → the same installation hassle again!
- NGS analysis tools change rapidly → need to update the tools used in training
- Students need access to an identical environment after the course

- Analysis jobs can require a lot of CPU and memory → laptop might not suffice
- 20-30 people run the analysis job at the same time → need a lot of computing resources temporarily

What is an image and can it help?

- Ready-made package of analysis tools and their dependencies, reference data,...
- Can be installed with one click
- Provides reproducibility: allows you to create exactly the same environment again
- Runs on your computer or in the cloud (easy to scale up)
- Two types of images
  - Virtual machine image (called VM instance when it is running)
  - Docker image (called Docker container when it is running)

What is the difference between VM and Docker?

- VM has its own operating system and needs a hypervisor software
- Docker containers use the host computer's operating system

VM instance

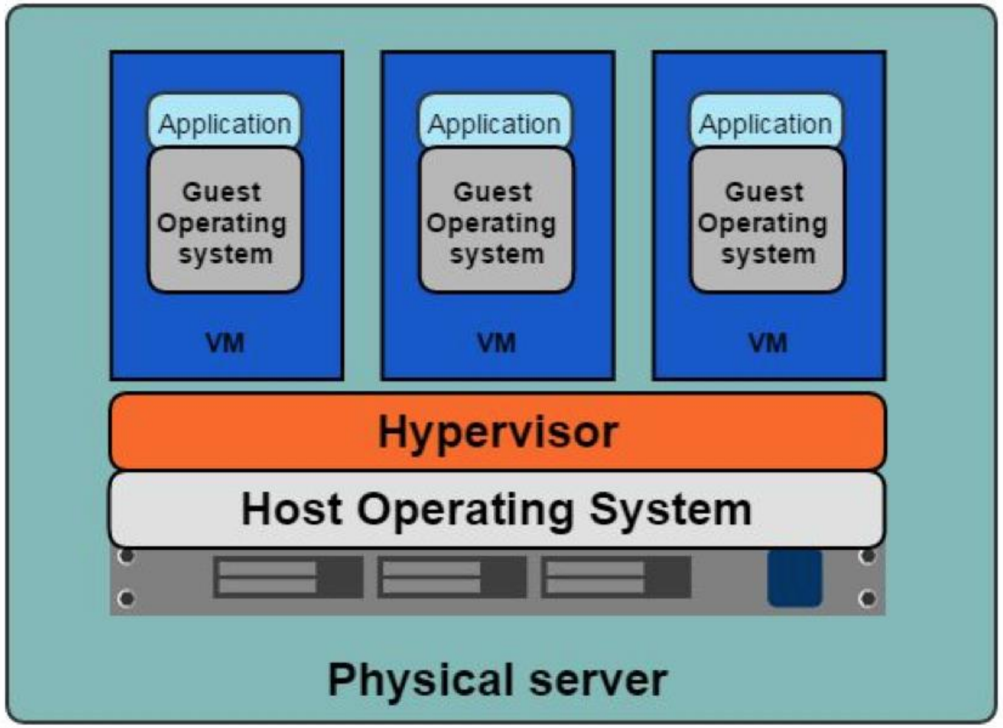
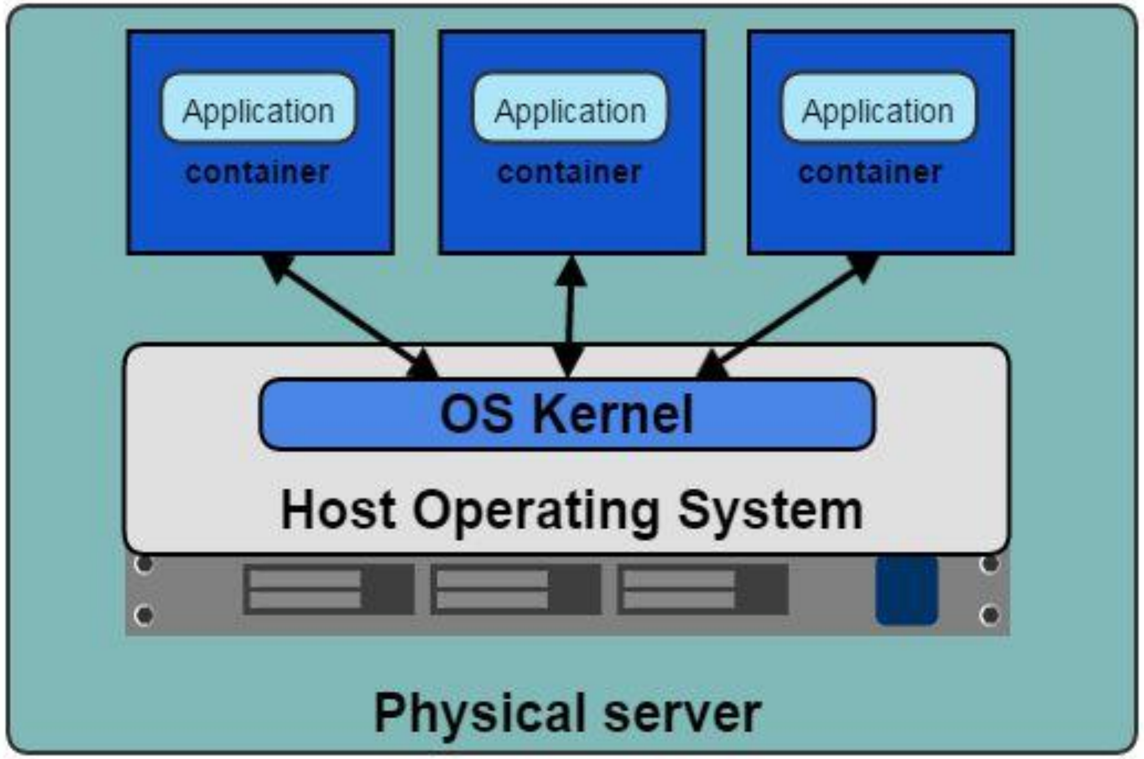


Image by Abdulrahman Azab, ELIXIR-NO

Docker container



Pros and cons (by Markus van Dijk, SURFsara)

- Speed
  - VM requires full boot
  - Docker is fast to start/stop
- Resources
  - VM needs a hypervisor (KVM, VirtualBox, VMware) and dedicated resources
  - Docker is more light-weight and has a small footprint
- Safety
  - VM is “safe” environment (if done properly)
  - Docker not yet safe enough (but you can run Docker in VM :-)
- Multi-user situation in training
  - VM can have multiple users
  - Can make multiple Docker containers

Images can be made in two different ways

- Build the image manually
  - Take a base image (e.g. Ubuntu), install the analysis tools etc, and take a snapshot
  - Pros: Easy to understand and do
  - Cons: Large image file, hard to version, different VM file format needed for different clouds,…
- Write a recipe for building the image and build it automatically
  - E.g. Ansible file for VM, Dockerfile for Docker
  - Pros: Small file, easy to version and update, easy for others to see what exactly goes to your image (admins will love you)
  - Cons: Need more expertise

Starting a VM in the cloud involves several steps, e.g.

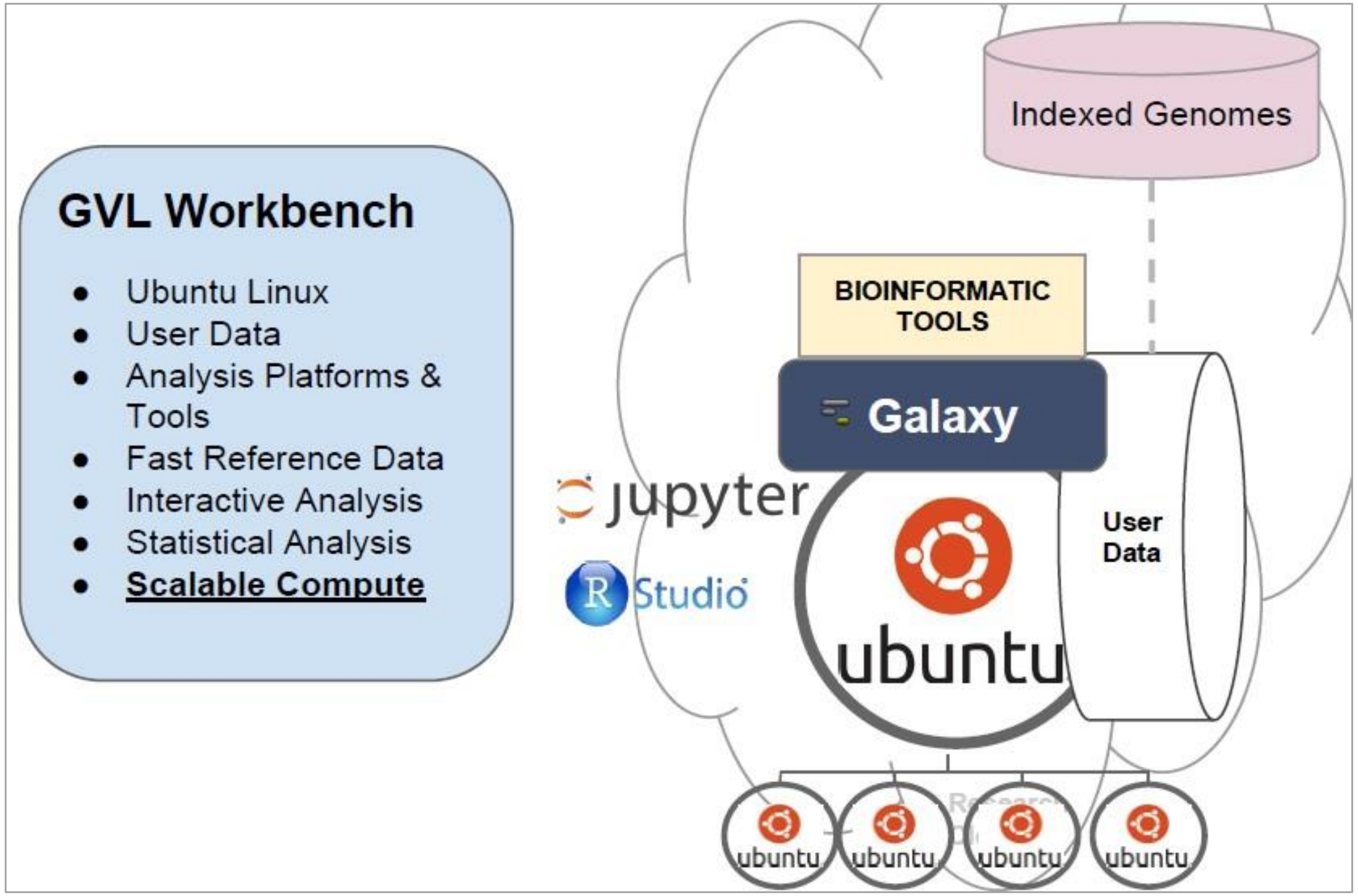
- Setup prerequisites
  - SSH key
  - Security group
- Launch a virtual machine
- Assign a floating IP to the VM
- Log into the VM
- Attach volume

What kind of systems are there in different countries?

- Cloud, planning WaaS (Workshop as a Service): NL
- Cloud, image catalogue: SE
- Cloud, ready-made images, image catalogue: EGI FedCloud
- Cloud, ready-made images, image catalogue, Galaxy-T training environment: IT
- Cloud, ready-made images, image catalogue, precalculated reference data: AU (GVL)
- Cloud, ready-made images, image catalogue, precalculated reference data, BioShaDock Dockerhub: FR
- Cloud, ready-made images, matching modular training materials: AU (BTP)
- Cloud, Blueprints software to provide easily Rstudio and Jupyter notebooks for training,: Chipster VM with a lot of tools and precalculated reference data: FI

Genomics virtual laboratory

- **Genomics Virtual Laboratory: A Practical Bioinformatics Workbench for the Cloud** Afgan et al. PLoS One 2015



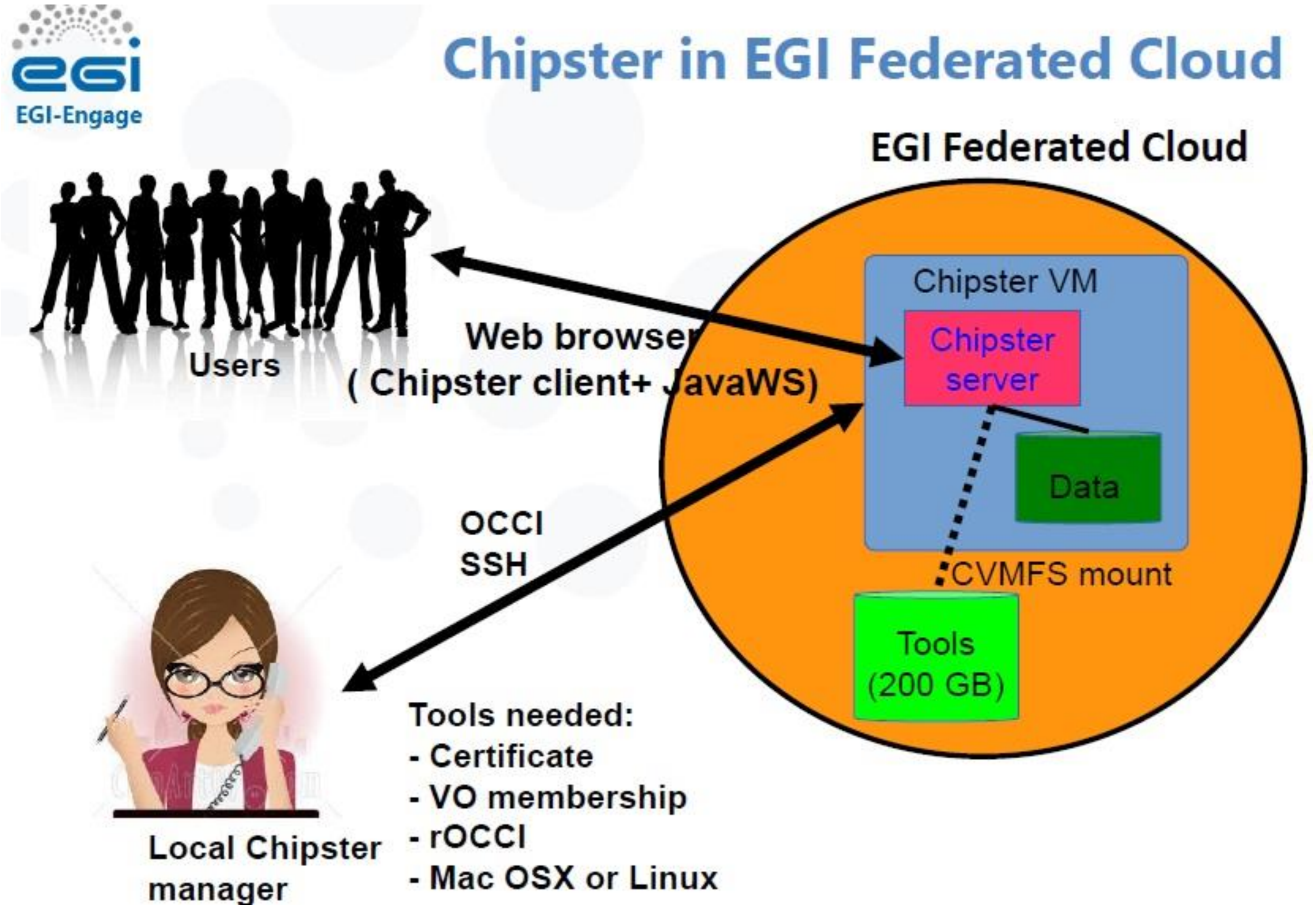
Bioinformatics training platform, BTP

- Automatically made images, with matching modular and versioned training materials (inc datasets)

- **Development of a cloud-based Bioinformatics Training Platform** Revote et al Briefings in Bioinformatics 2016
- **Towards an open, collaborative, reusable framework for sharing hands-on bioinformatics training workshops** Watson-Haigh et al Briefings in Bioinformatics 2016

EGI FedCloud: Free computing resources for training (and beyond)

- AppDb for ready-made images (Galaxy, Chipster, Jupyter notebook, etc)
- https://wiki.egi.eu/wiki/Training_infrastructure



Chipster in EGI Federated Cloud

Tools needed:
- Certificate
- VO membership
- rOCCI
- Mac OSX or Linux

How can we make it easier for trainers to use images?

- Provide ready-made images (or recipes)
- Enable trainers to make images
  - Training, instructions, help
- Make running (several) VMs / Docker containers easier
  - Simpler GUIs
  - Training, instructions, help

Summary

- Images (VM or Docker) make it easier to provide a reproducible training environment
- You can store and share images or their recipes
- Making images and running them in the cloud requires technical skills
- ...but in many countries this has been made easier
  - Ready-made images
  - Easy GUIs to deploy them in the cloud
  - Training and help for trainers
- Ongoing discussion between trainers and technical specialists is important