Flexible use of virtualization for Galaxy

Marius van den Beek
Drosophila Genetics and Epigenetics
UMR7622 – Jussieu
mvandenb@snv.jussieu.fr
drosophile.com
Who we are

• Mostly genetics and mol. biology background
• Research focused on small RNA biology (miRNA, piRNA, siRNA)
• Over the years, a number of (command-line) workflows have been developed in the lab, for our own research or for collaborators
The need for a data analysis platform ...

Lab publications


Our galaxy server – alive since December ‘11...
Our galaxy server – some stat’s...

Dell PowerEdge T610
2x X5650  @ 2.67GHz, 12 cores total
32GB Ram
2x300 System / in RAID1, managed by LVM
6x4TB Storage in RAID6, 16TB net capacity, LVM
21 users, of which 6 regular users
~ 60 custom tools

mostly dedicate to small RNAseq and RNAseq analysis

... and a second, more powerful one and a storage server are expected next week!
A flexible backup strategy

Root backups are performed with make_snapshot.sh in an incremental fashion:

```
mv folder backup.0 to backup.1,
cp –al backup.1 to backup.0
```

LVM snapshot, mounted read only

mount backup filesystem, rsync into folder backup.0

unmount & destroy snapshot, remount backup file-system read only
Virtualizing a backup ...

We can now use any of the backups to create a bootable virtual hard disk! p2v.sh does all of the following:

Create a virtual disk, format with a filesystem, mount it, copy a root file system, set new hostname, deletes old udev rules, changes /etc/fstab and install grub.

The resulting image is bootable on OSX, Windows and Linux, as long as it is the 64 bit version!

Perfect start for beginners, students, collaborators, to test the latest upgrades and especially to test new tools!
Using multiple virtual machines to test cluster setup

To setup a cluster, all we need is a shared nfs folder for user files and a job_workflow directory.

Then install the opengrid batch-queing system, inform the galaxy job runner and we’re good to go:

```bash
$ qstat -f
queuename    qtype  resv/used/tot. load_avg arch               states
main.q@galaxy-exec2  BIP   0/1/1   0.23   lx26-amd64  galaxy r// 12/02/2013 00:57:37 1
main.q@galaxy-exec3  BIP   0/1/1   -NA-   lx26-amd64  au   11/30/2013 16:22:48 1
main.q@galaxy-exec4  BIP   0/1/1   0.01   lx26-amd64  galaxy r// 12/02/2013 00:57:37 1
main.q@vboxgalaxy   BIP   0/1/1   0.45   lx26-amd64  galaxy r// 12/02/2013 00:57:37 1
main.q@virtualged   BIP   0/1/1   0.43   lx26-amd64  galaxy r// 12/02/2013 00:57:37 1

$ qstat
```

Running a bowtie Workflow on the server vs. 4 iMacs ...

```bash
gedserver   galaxy-exec2   VirtualGED   vboxgalaxy   galaxy-exec4
20:11:00,335  19:45:07   19:45:07   19:45:07   19:45:07
~7 minutes   ~11 minutes   ~11 minutes   ~11 minutes   ~11 minutes
```
GED lab

Christophe ANTONIEWSKI
Clement CARRE
Bruno DA SILVA
Hélène THOMASSIN-BOURREL
Margarita Angelova

Bioinformatics Post-doc
available!

Thanks!

mvandenb@snv.jussieu.fr