PoreCamp2016 : Using CLIMB

0 Overview

By the end of this tutorial, you will have a basic understanding of the CLIMB cloud, and be able to log into the system in preparation for the bioinformatics analysis.

1 What is CLIMB?

> *The CLIMB project (Cloud Infrastructure for Microbial Bioinformatics) is a collaboration between Warwick, Birmingham, Cardiff and Swansea Universities to develop and deploy a world leading cyber-infrastructure for microbial bioinformatics; providing **free** cloud-based compute, storage, and analysis tools for academic microbiologists in the UK.*
>
> *The project is supported by three world class Medical Research Fellows, a comprehensive training program and two newly refurbished Bioinformatics facilities in both Warwick and Swansea.*
>
> *CLIMB is funded by the UK's Medical Research Council.   (http://www.climb.ac.uk)*

It is ~£4 million of hardware, capable of supporting >1000 individual virtual servers. Microbiologists would choose to use this service instead of buying their own server, the benefits being cost and ease of analysing and sharing data.

You can run command-line or graphical software and we will be using this service to analyse our data.

2 How does CLIMB work?

The following description of CLIMB is based heavily on the launch event slides and presentations from 14 July 2016 by Tom Connor and Nick Loman:
`https://discourse.climb.ac.uk/t/climb-launch-event-slides-talks-and-notes/43`

Don't worry if you don't understand all of this section - even skimming it will help you understand what all the terms are so you know what the instructors are referring to.

2.1 Summary

- **CLIMB** is a private cloud for microbiologists in the UK.
- The cloud was set up with **OpenStack** - open source software for creating a cloud (https://www.openstack.org).
- The disk storage is a **GPFS** (General Parallel File System, http://www.ibm.com/support/knowledgecenter/SSFKCN/gpfs_welcome.html).
- A "**Group**" account is granted to the PI of a research group.
- "**Users**" are invited to the group account by the PI and allocated an individual login.

- The group account shares a set of "**Running servers**", each of which is a virtual machine (VM) with access to a certain number of cores, RAM and disk space.
- Each server can be used just like a traditional desktop server owned by a laboratory, with many users logged in and using the system at the same time.
- Servers are currently provided in two "**Flavours**". At the time of writing (Aug 2016):
  - a "**User server**" has 4 CPU cores and ~15 GB RAM; and
  - a "**Group server**" has 8 CPU cores and ~60 GB RAM.
  For comparison, a home computer/laptop might have 2-4 cores and ~8GB RAM, and an equivalent workstation with the "User" specs would be ~£1000, and with "Group" specs ~£3000.
- Each Group account has a **default allocation**, shared by the whole team:
  - 10x "instances", 20x "volumes", 10TB disk, 128x cores and 64GB RAM per instance.
- Group quotas can be increased, on request.

2.2 OpenStack terminology

- **Instances** : Each instance is a virtual server. The instance can have an external IP address, and is accessed via either ssh or through the web.
- **Volumes** : A private disk, running on a large storage system. Volumes can be treated like disk drives, and can be attached and detached from running instances.
- **Snapshots** : Effectively a digital photo of everything in a volume at the point of snapshotting. Can be used as a basis to create a new volume containing the original data.
- Projects / Tenants : All VMs are part of a project/tenant. This is a level at which quotas apply.

2.3 Components of OpenStack / CLIMB

- **Bryn** : Our interface for registering and spinning up VMs.
- **Horizon** : The OpenStack control panel (only recommended for power users).
- Keystone : The OpenStack identity service.
- Nova : The Openstack compute service.
- S3 : Amazon's storage API.
- Cinder : OpenStack block storage service.
- Glance : OpenStack image service.
- Neutron : The OpenStack Network Service.

2.4 CLIMB etiquette

While a server is running, no-one else can use those CPUs and RAM. We will leave the servers set up for this course running for the week. But in general, if you are not using a server, you should save a "snapshot", then "Terminate" the server. When you want to use it again, you can "Start" the server snapshot you saved earlier.

3 Quick Start

3.1 The PoreCamp set-up

- Each PoreCamp participant belongs to one of the 3 rotation groups called "**Rotation A/B/C**".
- The instructors have set up one CLIMB "Group" account for each rotation group, called "**PoreCamp2016 Group A/B/C**", and invited each participant to the correct group account.

- Within each group account, there should already be one "Running server" called "**PoreCamp2016 Server A/B/C**".

3.2 Find your public SSH key or create a new one

Everyone should have brought their own laptop for this course. Check if a public SSH key has already been set up. It usually resides in a file called ~/.ssh/id_rsa.pub

If you can't find it, you will need to generate a new one using `ssh-keygen`

On a Mac, the commands in your terminal window should look similar to the following. You should choose the default location (since there shouldn't be anything there already). The passphrase is optional (to skip it, just press Enter).

```
$ ssh-keygen
Generating public/private rsa key pair.
Enter file in which to save the key (/path/to/.ssh/id_rsa):
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Passphrases do not match.  Try again.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /path/to/.ssh/id_rsa.
Your public key has been saved in /path/to/.ssh/id_rsa.pub.
The key fingerprint is:
3a:b5:24:fd:d1:12:d3:fa:a3:ce:2d:9e:a2:2f:7a:21
username@computeraddress
The key's randomart image is:
+--[ RSA 2048]----+
|                 |
|            .    |
|          o .    |
|       .   =     |
|      . S + .    |
|     E .= o +    |
|      .o.. . o   |
|       o....+ .  |
|      .o.+.+*..  |
+-----------------+
$
```

For laptops running Windows, generate a key using a terminal emulator like PuTTY:
`https://docs.joyent.com/public-cloud/getting-started/ssh-keys/generating-an-ssh-key-manually/manually-generating-your-ssh-key-in-windows`

3.3 Log into CLIMB with the bryn interface

- In Chrome or Safari (avoid Firefox), open `https://bryn.climb.ac.uk`
- Enter your CLIMB `userid` and `password`

- You should now see the bryn Dashboard for "PoreCamp Group A/B/C" and a running server called "PoreCamp Server A/B/C" with an "IP address".

3.3 Register your public ssh key

- While logged in to `bryn.climb.ac.uk`, scroll down to the "Advanced interface" bar, and click on it to reveal the username (something like `bryn:NN_PoreCamp2016 Group A`) and your password.
- Click on the Horizon link (`https://birmimgham.climb.ac.uk`) and log in using this username and password.
- On the left menu bar, select Project->Access & Security.
- On the right side, select the "Key Pairs" tab.
- Press the "Import Key Pair" button, then enter a key pair name (e.g., your login name to CLIMB).
- Copy the text from the `~/.ssh/id_rsa.pub` file on your laptop (or the computer you're currently typing).
- Paste this text into the "Public Key" box in the Horizon "Import Key Pair" popup.
- Read the part of the "Description" that says you need to use "`ssh -i cloud.key username@instance_ip`" to connect to the CLIMB instance later on.
- Select "Import Key Pair" to finish registering the public key in Horizon.

3.4 Connect to your PoreCamp Server with ssh

- Log into bryn.climb.ac.uk, find the of the Running server that you'd like to connect to and note the "Name (`instance_name`) and "IP address" (`instance_ip`).
- Open a terminal window on your laptop.
- Connect to your CLIMB server instance using `ssh` with the "`-i`" flag to specify the file to use for public key authentication.
- You should see something like the following:

```
$ ssh -i ~/.ssh/id_pub.rsa ubuntu@instance_ip
Welcome to Ubuntu 16.04.1 LTS (GNU/Linux 4.4.0-31-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

  Get cloud support with Ubuntu Advantage Cloud Guest:
    http://www.ubuntu.com/business/services/cloud

19 packages can be updated.
0 updates are security updates.


*** System restart required ***
Last login: Fri Jul 22 12:54:31 2016 from 147.188.139.127
/usr/bin/xauth:  file /home/ubuntu/.Xauthority does not exist
ubuntu@instance_name:~$
```

- If you get an error message that starts with:

```
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@    WARNING: REMOTE HOST IDENTIFICATION HAS CHANGED!    @
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
```

make a backup copy of your .ssh/known_hosts file remove the line containing instance_ip from your .ssh/known_hosts file

3.6 Find data and software already on the server

```
# List the contents of the home directory
ls -l
# List the disks mounted on the server
df
# The MinION R9 data we'll be using is in /data
ls -l /data
# Third-party bioinformatics software is often installed
# by default in /usr/local/bin or /usr/bin
# To see the next page of text in "less", use the space key,
# to go back a pack type "ctrl-b", to exit to the command-line type
# the letter q
ls -l /usr/local/bin | less
ls -l /usr/bin | less
# However, if you want to check that your favourite program
# is there, try either of the following commands
which bwa
locate bwa
```

3.5 Create a directory to store your analysis files

```
# Go back to your home directory.
# If you're already there, the command won't do anything
cd
# List the contents of the home directory
ls -l
# Make a directory for all your analysis files
mkdir YOURDIRNAME
# Change the current working directory to be YOURDIRNAME
cd YOURDIRNAME
```

3.6 Run a command-line bioinformatics tool

```
# Confirm your favour program is available
which poretools
# Print a brief usage message for the program
poretools
# Print a longer help message for the program
poretools -h | less
q
# Use poretools to print the longest read in FASTQ format
```

```
poretools winner /data/basecall/pc4_deplex//pass/NB03 | less
```

3.7 Run a graphical bioinformatics tool

Please ask an instructor.

3.8 Copy a file from your laptop to the CLIMB instance

- Create a file on your laptop with `echo`:

```
$ echo "hello wonderful porecamp person" > hellome.txt
```

- Transfer a copy to `instance_ip` with `scp`. You should see something like:

```
$ scp hellome.txt ubuntu@instance_ip
hellome.txt                                        100%   14      0.0KB/s
00:00
```

3.9 Copy a file from the CLIMB instance to your laptop

- From a terminal window in your laptop, to transfer a file called "/path/to/remoteme.txt" on the CLIMB instance called `instance_ip` to your laptop:

```
cd /dir/that/you/want/file/to/be/in
scp ubuntu@instance_ip:/path/to/remoteme.txt .
```

3.8 Disconnect from your PoreCamp Server

Close the terminal or PuTTY window.

4 More information

Platforms
- OpenStack `www.openstack.org`
- GPFS `www.ibm.com/support/knowledgecenter/SSFKCN/gpfs_welcome.html`

CLIMB
- project `www.climb.ac.uk`
- community forum with tutorials `discourse.climb.ac.uk`
- bryn `bryn.climb.ac.uk`
- horizon interface `birmingham.climb.ac.uk/dashboard`
- launch event (Warwick, July 2016)
  `discourse.climb.ac.uk/t/climb-launch-event-slides-talks-and-notes/43`
- Twitter `@MRCClimb`