Cloud Computing with Amazon Web Services

Amazon Web Services provides scalable computational power in the form of virtual computer "instances" which are then loaded with a OS image. Several options for increased processing power or memory exist, however sequence assembly projects are typically limited by memory constraints as the graph generated by the reads must be held in active memory so the algorithm can plot a path through reads.

This protocol assumes an Amazon AWS account exists and a computer capable of running Mac or Linux (preferably) terminal is available.

1. Log in to your account. The AWS console will display.
2. Click the "Launch Instance" button and select the Classic Launch Wizard
3. You will be prompted to choose an AMI. This is a machine image which loads an operating system into your instance. Most 64Bit Linux images from the community AMIs will work, and you can purchase and modify one. Some programs will not run on certain Linux distributions.
4. Select the number of instances to launch and the size. Notice the number of CPU cores and amount of memory in the click-down menu. Some programs are multi-threaded and extra cores can speed up processing. Memory constraints are often a problem with assemblers. If a process kills or the instance crashes, run "TOP" alongside your program and see if the memory usage hits 100%. If so, launch a larger instance.
5. Click through details without changing anything
6. (**First time only**) You must create a key pair which functions as a password to access your instances. You can reuse key files when you launch more instances, but you must create a new one or transfer an existing key to log in with other computers. To log into an instance created at work from your home office you must transfer the key file.
7. (**First time only**) Create a new security group. In the click-down menu select SSH. Name it something descriptive and select "Add Rule". This protocol allows you to log into your instance from a terminal.
8. Select the SSH protocol and continue.
9. Click launch. After about a minute the instance will be up and running. You can monitor the status from the console.

To log in with terminal, you must highlight the instance and find the Public DNS. Copy that address to clipboard. It will look like this:

    ec2-153-129-111-124.compute-1.amazonaws.com

From terminal enter the following:

    ssh -i ~/Desktop/key.pem root@ec2-153-129-111-124.compute-1.amazonaws.com

The ssh command is followed by the location of your key file then root@<public DNS>

The terminal is now connected to an instance which you control remotely. You can now upload data files, install programs, and run scripts on a virtual machine. Remember to "Terminate Instance" when you're done using it since Amazon charges hourly.

Resources

SEQanswers Wiki and Forum

The wiki describes the function of many useful tools and links to the files. There are also great how-to articles which describe some basic workflows and provide reviews for various software packages. The forum is a great resource for advice, help, and scripts from some very talented members.

BioPython Tutorial and Cookbook

Python is a very useful language for parsing sequence data. BioPython has many scripts useful for bioinformatics as built-in functions allowing for easy sequence conversion, generating histograms and statistics, etc. It also has an installation guide.

Python Documentation - docs.python.org/index.html

Contains tutorials and a database of every aspect of Python and sample usages. A bit technical for a quick-reference, but very complete.

LinuxCommand - linuxcommand.org/writing_shell_scripts.php

Beginners guide to shell scripts. How to create workflows and pipelines that will automate otherwise time-consuming tasks and using powerful tools to process large quantities of data quickly.

Software Carpentry – software-carpentry.org

Beginners guides to various programming languages and shell scripts.

Stack Overflow – stackoverflow.com

A question and answer site for any programming language. Posting a question will usually result in a quick reply or link. Very useful for cleaning up scripts and problem solving. Many beginner questions have already been asked and answered with useful snippets of code or scripts.

SeqHack – seqhack.blogspot.com

Scripts related to my bioinformatic projects will be archived at this website.