Getting Started

August 26, 2012

Introduction

This document describes how to get started using elasticHPC library. It is divided into two sections:

- for creating a traditional EC2 cluster and running a BLAST operation.
- for creating an EMR cluster and using it run crossbow.

Perquisites

elasticHPC installed, it can be downloaded from www.elastichpc.org. A guide for retrieving AWS Security credentials is available at:
http://elastichpc.org/doc/credentials/index.html.

EC2 Cluster

The following steps describe how to run BLAST of the Windshield sequences which are already available on the AMIs against nt database which is available as a snapshot. This experiment will run on a cluster composed of 4 * m1.xlarge machines. The query will be divided using the runner tool in segments, the number of segments is to equal the number of cores available in the cluster (4 machines*4 cores/machine = 16 core).

1. Create the cluster using the following command, this step prints the domain which will be used later

```
./EHPC-Client –create -r=EU1 -ami=AMI_ID -pk=/path/to/YOUR_PRIVATE_KEY -n=4 -t=m1.xlarge
–cert=/path/to/YOUR_CERTIFICATE -kp=YOUR_KEYPAIR_NAME -sg=YOUR_SECURITY_GROUP_NAME
```

2. Attach volume snapshot containing nt database using the following command:

   ```
   ./EHPC-Client –snapshot-attach-mount –mountPoint=/blast_db/nt –snapshot-id=snap-b41bd5e2
   -d=DOMAIN_FROM_FIRST_STEP
   ```

3. Download the following python script from http://www.elastichpc.org/downloads/expcommand.py.

4. Run the following command, it will produce a command that will be used later

   ```
   python expcommand.py ALeft.fasta 16 m1.xlarge
   ```

   output command:

   ```
   /var/www/runner/py/runner.py 963271205238 ”blastn -query /home/ubuntu/cluster_experiment/ALeft.fasta##\n##8320
   -db /blast_db/nt/nt -outfmt 6 -num_threads 4 -out __o__output1__output1”
   ```

5. Run the job using the following command:

   ```
   ./EHPC-Client –run -d=DOMAIN_FROM_FIRST_STEP –command=”COMMAND_FROM_FORTH_STEP”
   -o=YOUR_NAME
   ```

6. Repeat steps 4, 5 for the following:
   - ARight.fasta
   - BLeft.fasta
   - BRight.fasta

7. Terminate the cluster using the following command:

   ```
   ./EHPC-Client –terminate -d=DOMAIN_FROM_FIRST_STEP
   ```

Notes

- AMI_ID can be retrieved from the Downloads tab in the website at http://elastichpc.org.

- Output files can be found at the following location on the main node: /var/www/runner/files/963271205238/X$_{f}asta_{o}ut$ where X is the name of the query used either ALeft, ARight, BLeft, BRight

- Notice the difference between in the commands between
  - single dash ”-”
  - double dash ”–”

EMR Cluster

The following steps describe how to run crossbow on the first using reads from the African Human Genome against build 36 of the human genome (hg18). Both the reads and genome are available on S3. The cluster will be composed of 16 c1.xlarge nodes.

1. Create the EMR cluster using the following command, this will result in a job-id and domain:

   ```
   ./EHPC-EMR –create -n=16 -t=c1.xlarge -i=s3://eg.nubios.us/crosbow.sh -r=us-east-1 -a=YOUR_ACCESS_KEY
   -p=YOUR_PRIVATE_KEY -kp=YOUR_KEYPAIR_NAME -kf=/path/to/KEYPAIR_FILE
   ```

2. Run the job using the following command:

   ```
   ./EHPC-EMR –run -d=DOMAIN_FROM_FIRST_STEP –command=”export CROSSBOW_HOME=/home/hadoop/crossbow;$
   –preprocess –input=s3://eg.nubios.us/AFGreads/reads.manifest –output=s3://eg.nubios.us/crossbow/example/hg18/output
   –reference=s3://eg.nubios.us/crossbow-refs/hg18.jar –all-haploids –tempdir=/mnt/tmp –streaming-
   jar=/home/hadoop/contrib/streaming/hadoop-streaming-0.20.205.jar –just-align” -owner=YOUR_NAME
   ```

3. Terminate the cluster using the following command:

   ```
   ./EHPC-EMR –terminate -id=JOB_ID_FROM_FIRST_STEP
   ```