**Preparation of DNA-Seq Pre-alignment System Metrics Collection**

1. Overview

In this report, we will go through the preparation before collecting system metrics from DNA-Seq pre-alignment. There are several things to prepare: a script for setting up Chameleon instances automatically, a series of instructions to install pre-alignment tools, a guide to enable storage as a trackable resource.

1. Chameleon Automation

We will collect system metrics of pre-alignment jobs using several clusters with different numbers of instances. In the process of creating an instance, there are few steps that must be fulfilled, such as creating a reservation, creating an instance, allocating an ip to project and then to the instance. Although the steps are quite short, it takes a lot of time if we want to set up multiple instances (eg. 8, 16, 32). For the sake of saving time, we make a script to automate this process.

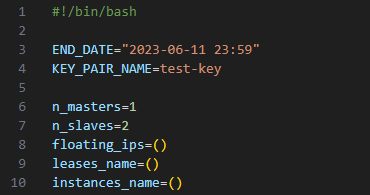


Figure 1. Initializing variables

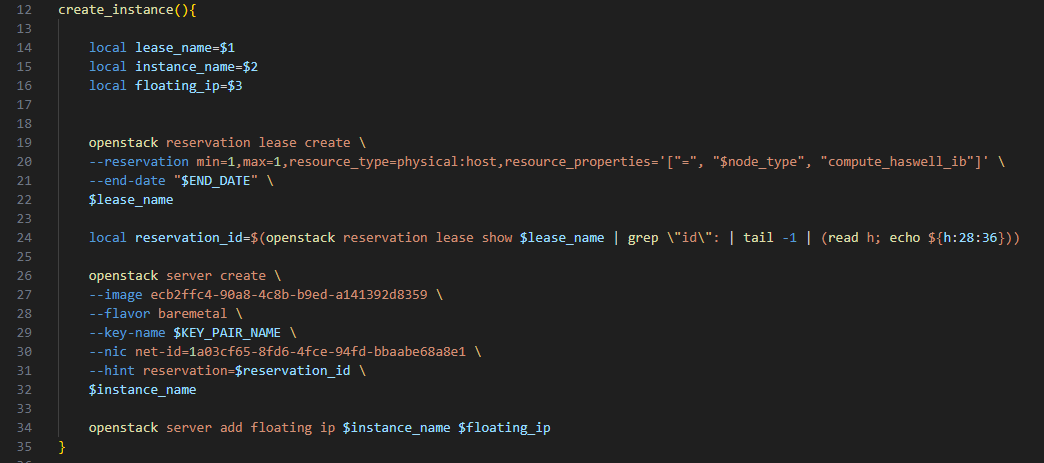


Figure 2. Defining create\_instance function

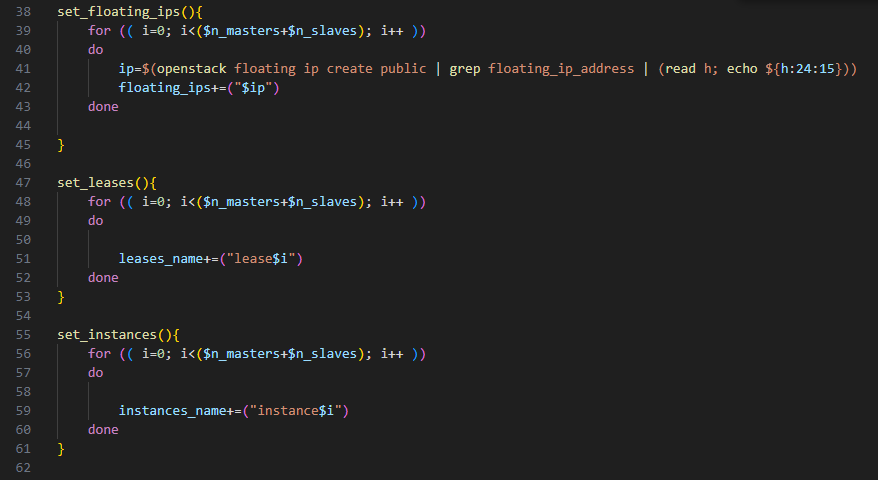


Figure 3. Defining set\_floating\_ips function

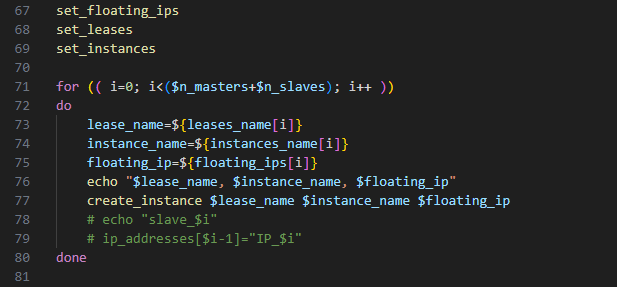


Figure 4. Function calls

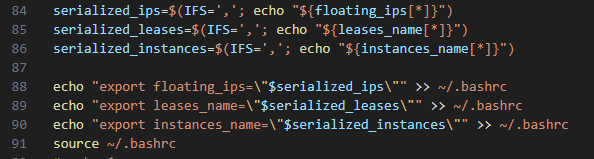


Figure 5. Storing the values inside environment variables for future use (eg. terminating instances).

1. Install pre-alignment tools

After setting up chameleon instances, we have to install the pre-alignment tools that are required by the pre-alignment script in each node. Pre-alignment script runs fastq and picard. These tools require JAVA 17. So, we will install JAVA first, then followed by installing fastq and picard with the guide below.

1. Prepare directories

$ cd ~

$ mkdir Downloads

$ mkdir bio-tools

$ cd Downloads

1. Install JAVA 17

$ wget https://download.oracle.com/java/17/archive/jdk-17.0.7\_linux-x64\_bin.tar.gz

$ tar -zxf jdk-17.0.7\_linux-x64\_bin.tar.gz --directory ~/bio-tools

$ echo "export PATH=~/bio-tools/jdk-17.0.7/bin:$PATH" >> ~/.bashrc && source ~/.bashrc

1. Install FASTQ

$ wget https://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc\_v0.12.1.zip

$ unzip fastqc\_v0.12.1.zip -d ~/bio-tools

$ echo "export PATH=~/bio-tools/FastQC:$PATH" >> ~/.bashrc && source ~/.bashrc

1. Install PICARD

# PICARD

$ cd ~/bio-tools

$ git clone https://github.com/broadinstitute/picard.git

$ cd picard/

$ ./gradlew shadowJar

$ echo "export PICARD=$(pwd)/build/libs/picard.jar" >> ~/.bashrc && source ~/.bashrc

1. Slurm Accounting Storage

To enable storage as a trackable resource, we need to define the storage system in each compute node, which is represented as a file, in the gres.conf, Slrum configuration file for GRES (Generic Resource). Based on Slurm page, we also have to specify PrologFlags, GresTypes and the number of storage in the slurm.conf To do this we can follow the instructions below:

1. Create gres.conf file by running

sudo nano /etc/slurm/gres.conf

1. Define the storage system in the gres.conf by the following line.

Name=localtmp File=/dev/sda3

1. Specify PrologFlags, GresTypes, and the number of storage in the slurm.conf by typing the following lines. (Notes: Number of storage must be defined in each compute nodes)

PrologFlags=Contain

GresTypes=localtmp

NodeName=ubuntu CPUs=48 Boards=1 SocketsPerBoard=2 CoresPerSocket=12 ThreadsPerCore=2 RealMemory=112569 TmpDisk=64000 Gres=localtmp:100G