SURPI on AWS

June, 2014

Below are some notes regarding usage on SURPI on AWS.

Drive Space

• The default size of the boot drive is 8GB, which is likely too small to use for SURPI. For this reason, it is recommended that you instead use the SSD instance drives for running the pipeline (there are 4 800GB drives on an i2.4xlarge instance).

• Portions of SURPI require scratch space in order to execute properly. For this reason, it is necessary that the `temporary_files_directory` parameter is set to a path with plenty of scratch space.

• SURPI contains code for running in a cluster, however this method is experimental, undocumented, and is not yet fully implemented.

First run

Approximately 1TB of SURPI reference data is stored on an EBS volume that is automatically attached to the EC2 instance at the mount point `/reference`. Due to the lazy loading of EBS volumes, there is a 5-50% loss of IOPS the first time this storage is accessed - which lengthens the first SURPI run after starting up a new AWS instance. Pre-Warming the EBS drive will restore normal performance to the first run, though it takes approximately 12 hours to complete. The time taken is similar to the time penalty of the first run, so does not effectively remove the penalty.

Normal performance is restored to the 2nd and all subsequent SURPI runs on an instance. Below is some AWS performance data:

Here is a description of the readcounts within a particular (unreleased) testset:

	# of Reads (250bp each)
Original FASTQ	1,370,478
Preprocessed	1,271,472
Human Unmatched	1,270,501

Below are timing values using the above testset on Amazon using SURPI 1.0.13 on our released AMI, on an i2.4xlarge instance. All times are in seconds except for the total runtime.

	Standard EBS		EBS Optimized, 4000 PIOPS	
	Run1	**Run2**	**Run1**	**Run2**
Preprocessing	99	100	98	99
Human Mapping	4209	542	3635	313
SNAP to NT	46546	10406	48096	6220
RAPSearch to NR	1742	875	3540	499
Total SURPI runtime	14h 53m 27s	3h 33m 29s	15h 41m 51s	2h 12m 37s

Recall that prewarming takes approximately 12 hours. Here are some points to note:

• As expected, the first-run penalty only manifests within steps using EBS drives (SNAP & RAPSearch). Preprocessing does not use the EBS drive, and so there is no first run time penalty, nor does using an EBS optimized drive give any benefit.

• There appears to be no benefit to using EBS optimized drives for the first run after instantiation of a new instance.

• Using an EBS optimized volume with 4000 PIOPS completes the SURPI pipeline in approximately 2/3 of the time compared to using a Standard EBS volume.

• This document was prepared before AWS released SSD-backed EBS volumes. When we benchmark this storage, we will add the results above.

Steps to run SURPI on AWS

1. Start up AWS EC2 AMI (ami-d86697b0) on an i2.4xlarge instance.

2. ssh to instance, and format the instance drives:

   execute the program **/home/ubuntu/format_instance_drives.sh**

   This will format all 4 SSDs on the i2.4xlarge instance, mount them, and create several folders:

```
/ssd2/SURPI_runs    for running SURPI
/ssd4/tmp           for temporary data
```

3. Transfer your FASTQ file to `/ssd2/SURPI_runs/`

   Here is a sample command line that may be helpful if transferring from OSX or Linux. Change the bold text as necessary with your setup.

```
scp -i <keyfile> -o StrictHostKeyChecking=no <sample.fastq>
ubuntu@<hostname>:/ssd2/SURPI_runs
```

4. Move to the SURPI run directory, and create a config file

```
cd /ssd2/SURPI_runs
SURPI.sh -z sample.fastq
```

Two files should be created by the above commands:

a. `sample.config`
This file contains all configuration information for the SURPI run.

b. `go_sample`
This program is used to initiate the SURPI run.

5. Verify SURPI installation

```
SURPI.sh -f sample.config -v
```

The -v switch will verify software dependencies and reference data. All dependencies/databases should have an 'OK' displayed next to them. If any instead say 'BAD', then SURPI cannot find this dependency or database. Check the config file, and your $PATH to be sure they are configured properly.

6. Edit the config file

When running on AWS the following parameter should be adjusted within the config file:

```
temporary_files_directory="/ssd4/tmp/"
```

Adjust the other parameters in the config file as needed.

7. Start SURPI

```
./go_sample &
```