**Memory Utilization Exceeds Allocated**

**Last week’s task:**

1. **Downloading input data for alignment tool**
2. **Run BWA-mem but found unexpected Slurm behavior**

**Next week’s task:**

1. **Setting cgroup correctly so that it can limit resource usage**
2. **Collecting BWA-mem systems metrics**
3. Overview

Last week's objective was to collect system metrics from the alignment tool. The collection of metrics starts with increasing the size of the common container and downloading hundreds of GB of data pairs of fastq files simultaneously. After that, BWA tried to submit it to Slurm. However, a problem was found in Slurm behavior caused by incorrect cgroup configuration, where the value of memory usage by a job exceeds the amount allocated.

1. Downloading data

First, I downloaded hundreds of GB of data from GSE183947. The average time it takes to download 1 pair of fastq files is 50 minutes. Meanwhile, we want to use 30 pairs of fastq files. It takes a lot of space and time to obtain those 30 files. Therefore, successively to overcome the problem of space and time is to enlarge the common container by adding some NFS servers and performing several download processes simultaneously in 1 instance.

The addition of NFS servers is done by making all compute nodes an NFS server. Thus, where there was only 1 NFS server before, which was in the master nodes, there are now N + 1 NFS servers and all nodes will be clients to all NFS servers where N is the number of slave nodes. Then to cut the duration of the data download process, each node will run more than 1 download process simultaneously. In this case, the architecture used is 1 master node and 5 slave nodes. For each node it will download 6 pairs of fastq files simultaneously.

1. Running alignment tool

In the previous week, there was a failure to run BWA mem as indicated by the error "paired names have different names". However, this error can be resolved by re-downloading the pairs fastq files which caused the error. To test whether a problem has been resolved, BWA mem is run using one of the randomly selected pairs. The results show that BWA mem can be executed with these files without experiencing errors.

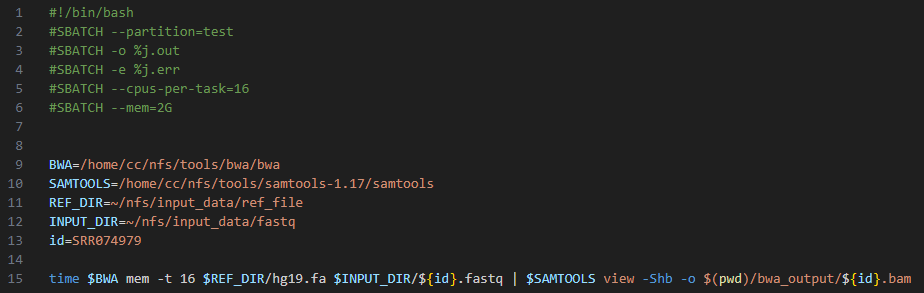


Figure 1. Script to run BWA mem for specific input data

After the alignment tool was run by submitting the script in Figure 1 using the sbatch command, things were found that were beyond our expectations. Memory utilization exceeds the allocated amount as shown in Figure 2. Memory allocation is 2GB, but memory usage can reach 7GB. Even though what we expect is if the maximum memory usage is as big as the allocated memory.



Figure 2. Memory utilization exceeds allocated memory. The memory usage expectation is below 2GB, but the job used 7GB.

1. Configurable allocation behavior

After further investigation, it turns out that the memory allocation parameter in sbatch by default has a different function than what we expected. This parameter serves to determine the minimum amount of memory needed for a job. For example, we have a node with 100GB of memory. Then, a job with a memory allocation parameter of 50GB is sent to that node. Then, in the process of executing the job, we send another job with a 51GB memory allocation. So, the job has to wait for the previous job to finish because the minimum amount of memory required by the job is 51GB while the available memory in the node is 50GB. Because the memory allocation parameter functions to determine the minimum amount of memory, it means that it is very possible for a job to use more memory than has been allocated.

Although by default the memory allocation function is to determine the minimum amount of memory, this function can be changed by applying a certain plugin and configuration. The plugin referred to here is the control group. Control groups or what can be referred to as cgroups can be useful for limiting the use of resources by jobs based on the amount that has been allocated [1]. This configuration can be modified in the cgroup.conf file. However, after trying to change the configuration with the configuration in Figure 3, no changes have occurred. Memory usage can still exceed the allocation. Due to unfamiliarity with how to use this plugin, so far memory usage has not been managed to be limited based on allocation.

So for the next task I will repeat the collection of system metrics of BWA. But first, I will find out how cgroup works so that it can limit memory usage.

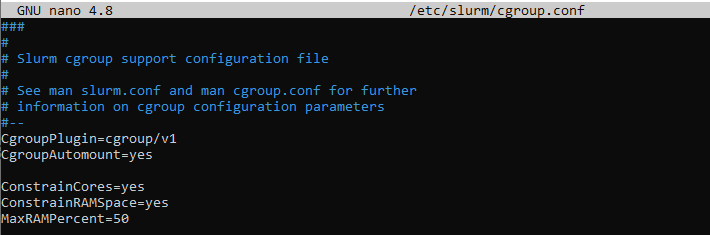


Figure 3. Cgroup configuration for limiting memory usage of a job

**Reference**

[1] Control Group in Slurm, <https://slurm.schedmd.com/cgroup.conf.html>