CMSE 401 - Homework 3

Jack Hamel

February 15, 2019

Dear Awesome PI,

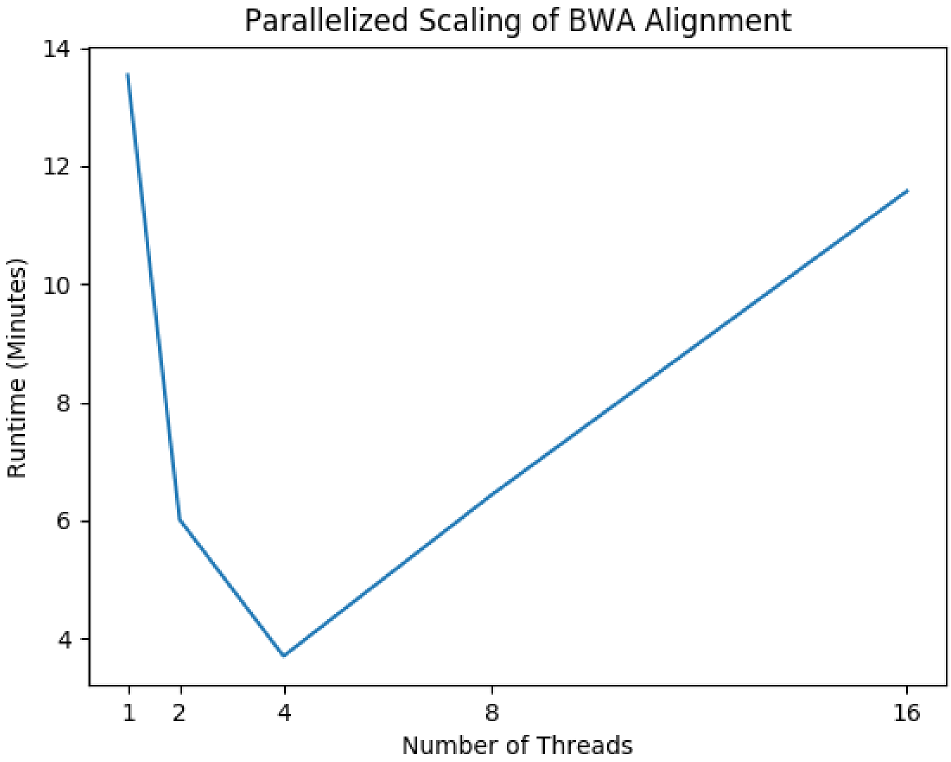
I have written a file titled README for you to reference. It is a fully executable bash script that you can submit to the HPC using `sbatch README` or run it interactively. This file will (you should be able to follow this flow by reading the comments in the file) first download the worm reference data file to your current directory, load BWA and its dependent modules, then copy over the worm datasets. The downloading and copy can be removed/changed if you are working with a different genome. Following these setup steps, the script will index the two datasets and reference file. This is done so asking to run on one node, but it may be worth investigating whether or not this step can be parallelized on multiple nodes. When indexing larger genomes, like humans, it may be nice, but it is quite fast with the worm genome. The next step is alignment. I did some testing and found that the alignment is fastest when ran with four parallel threads. Lastly, I clean up the current directory from intermediate files. This last step you may or may not want to leave.

In order to run the README you have two options: run it interactively with `./README`, or submit it to the SLURM with `sbatch README`.

Below I have included some timing studies. For the indexing, I ran each (dataset 1, dataset 2, reference) three times to get an average runtime. See these in the table below.

	Dataset 1	Dataset 2	Reference
Time (sec)	315.677	364.805	92.160

I also benchmarked the alignment performance for a variety of threads. I found that 4 threads is the optimal number to use. You can see this in the plot below.



I hope this helps.

-Jack Hamel