Alex Wu

CSE40822 Cloud Computing

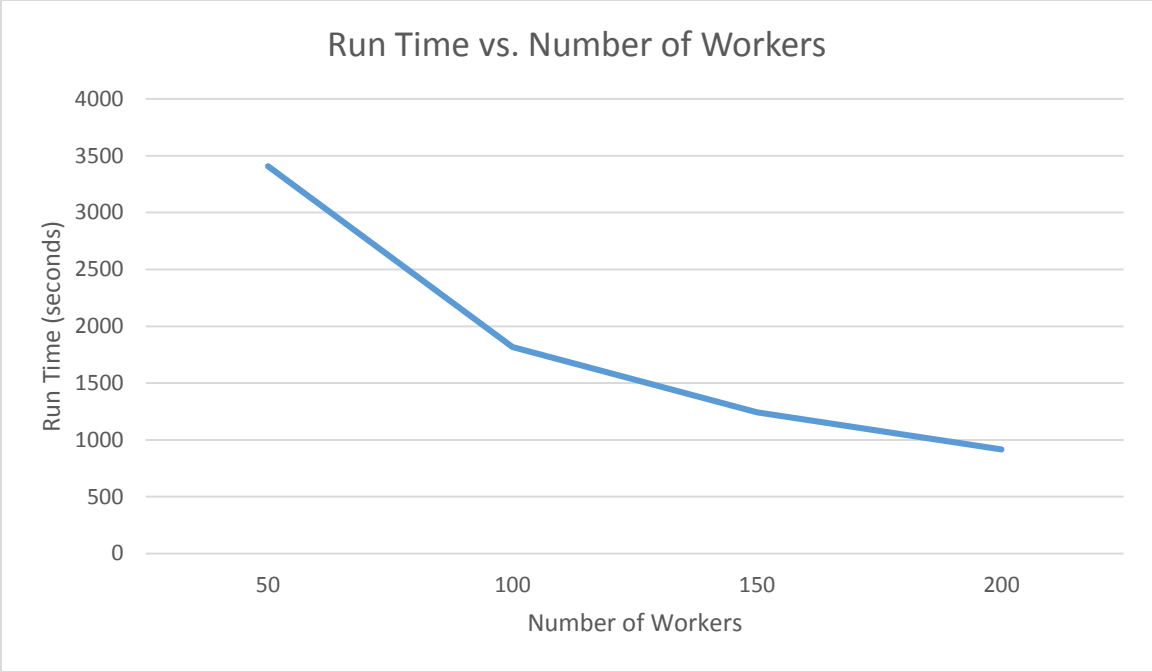
Assignment 2

1. It took 79.347 seconds to run comparisons of a snippet of agambiae.small.fasta with 7 other snippets of agambiae.small.fasta. Given this, approximately 5.293 comparisons can be made in one minute. In order to compare every sequence to every other sequence sequentially, for n sequences, it would take (n-1)*n/2 comparisons divided by 5.293 comparisons per minute for 5880.408 minutes. In order to complete this job in one hour, it would take approximately 98.007 machines or 99 machines. The snippet given only contains 250 sequences. The complete a.gambiae data contains 100,000 sequences. Using the same formula, this would take 94463442.823 minutes on one machine, which is equal to 15743907.047 hours = 655996.127 days = 1797.250 years.

   This estimate is based off of the benchmark of timing the swalign tool for a certain number of compares (in this case, 7). To calculate the comparisons per minute, I took this comparison, even though it may be off by a bit, as a basis.

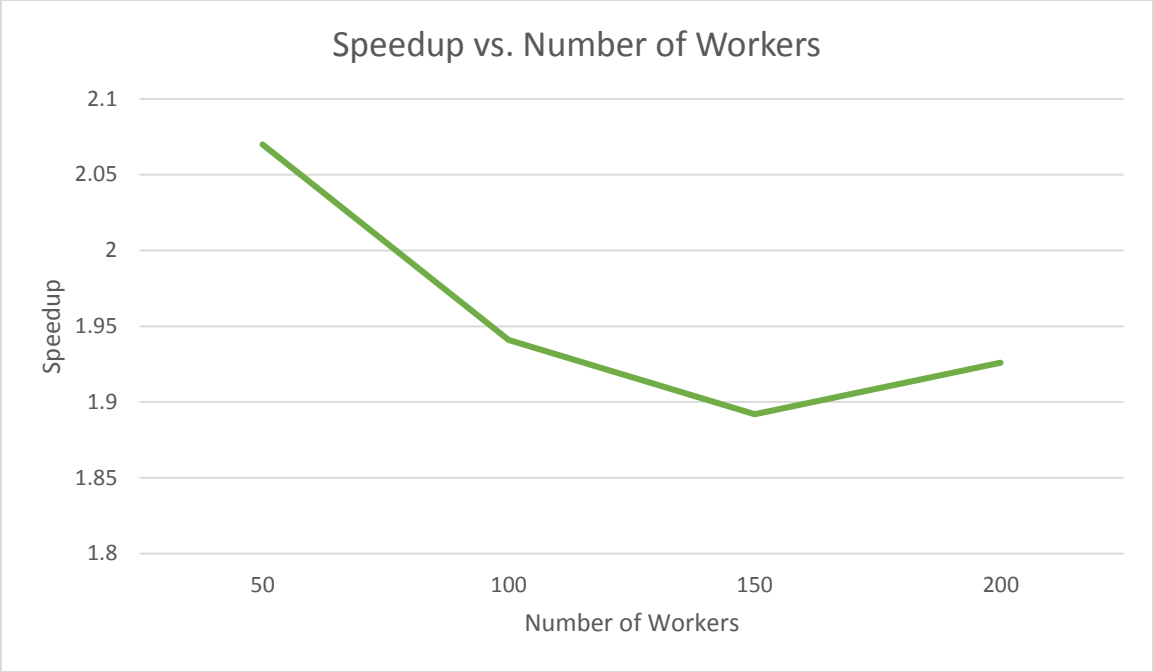
2. Top 10 Matches
   1: sequence 1102140143903 matches 1102140176186 with a score of 1539
   2: sequence 1102140177182 matches 1102140177183 with a score of 874
   3: sequence 1102140167168 matches 1102140172678 with a score of 818
   4: sequence 1101555423227 matches 1102140171813 with a score of 777
   5: sequence 1101555423815 matches 1102140164074 with a score of 767
   6: sequence 1102140171192 matches 1101671610328 with a score of 764
   7: sequence 1102140171192 matches 1102140170611 with a score of 744
   8: sequence 1102140178449 matches 1102140171192 with a score of 742
   9: sequence 1102140171192 matches 1101555423803 with a score of 741
   10: sequence 1102140178493 matches 1101671610328 with a score of 736

3. Run Time (Parallel):
   For 50 workers, the run time is: 3409.00028419 seconds
   For 100 workers: 1817.60285783 seconds
   For 150 workers: 1243.22674584 seconds
   For 200 workers: 916.066500902 seconds

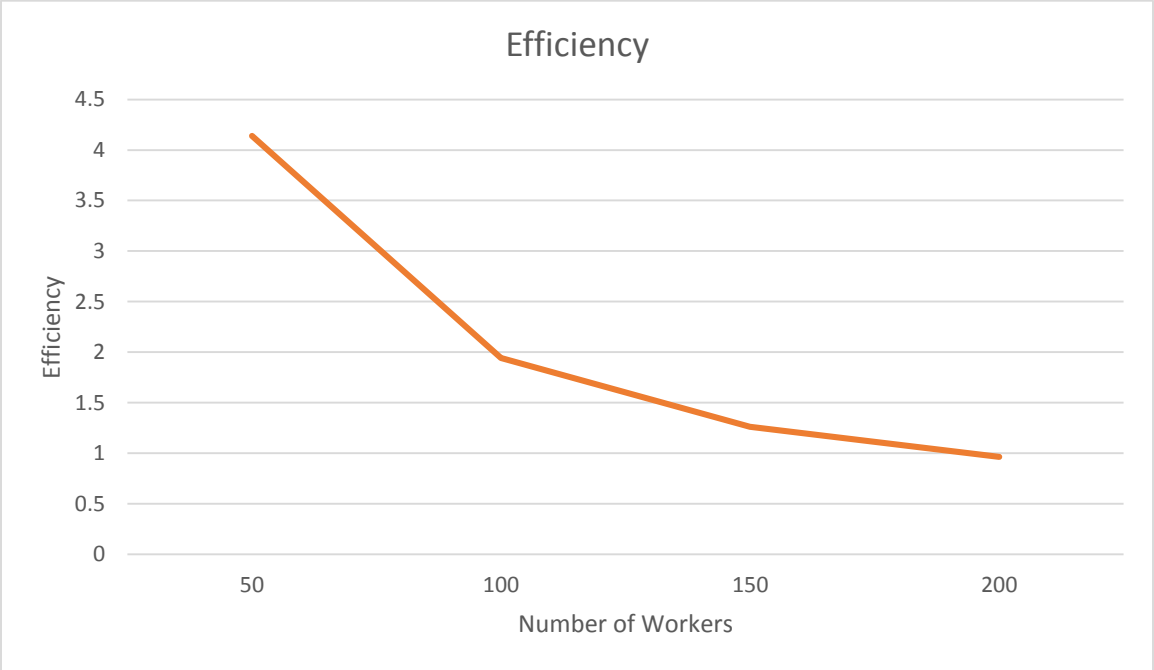


Estimated Sequential Time:
For 50 workers, the estimated sequential time is 5880.408/50=117.608 minutes.
For 100 workers, 5880.408/100=58.80408 minutes
For 150 workers, 5880.408/150=39.203 minutes
For 200 workers, 5880.408/200=29.402 minutes

Speedup:
For 50 workers, the speedup is 60*117.608/3409.00028419=2.070
For 100 workers, 60*58.80408/1817.60285783=1.941
For 150 workers, 60*39.203/1243.22674584=1.892
For 200 workers, 60*29.402/916.066500902=1.926



Efficiency
For 50 workers, the efficiency is 100*2.070/50=4.14%
For 100 workers, 100*1.941/100=1.941%
For 150 workers, 100*1.892/150=1.2613%
For 200 workers, 100*1.926/200=.963%



Efficiency and Run Time both experience exponential decay as the number of workers goes up. Like efficiency and run time, the speedup, for the most part, seems to decrease when the number of workers increases. Interestingly, 200 workers seems to break this trend as there is a spike in speedup. This seems to be due to the fact that the run time for 150 workers is a bit out of the ordinary, as compared to the other worker levels. The performance, of course, did not match the estimates of just multiplying sequential time. This is to be expected as ideal parallelized time is not possible, since there are many factors (the overhead of starting a worker, for example).