Parallelizing BWA Using Work Queue and Hadoop MapReduce

Christopher Ray
CSE 40822

Background: What is Ancient DNA?

- DNA recovered from biological specimens that has not been preserved specifically for later analysis
- DNA isolated from ancient specimens

Image from: http://news.psu.edu/story/141655/2010/01/19/research/mammoth-achievement-researchers-forefront-molecular-biology

Main Idea

- Plant debris that falls to lake eventually settles on lakebed, creating layers
- Ancient DNA samples found in deeper layers of lakebed should be older than those found above
- If you determine what species are present at each layer, you can trace how the plants in a surrounding area changed over time

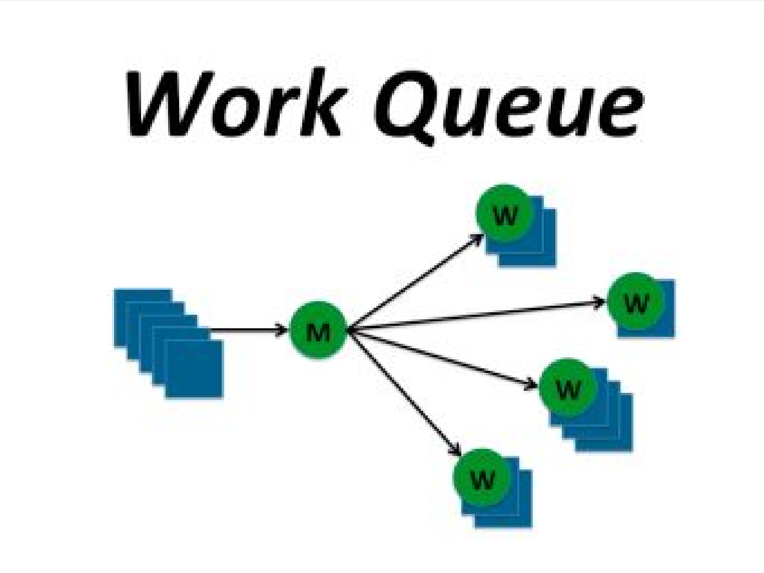
Image from: http://bactifeed.com/biological_soil_conditioner_FAQ.html

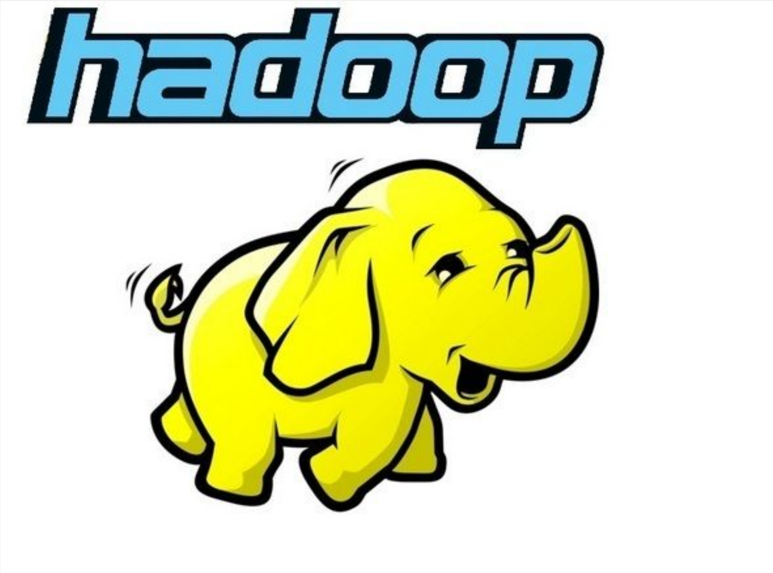
Problems

- Burrows-Wheeler Alignment Tool (BWA) is slow for large queries
  - ~23 minutes to align 8 GB fastq query
  - >1 hour to align 31 GB
- Using BWA's multithreading feature can speed up runtime, but can lead to varying results compared to running sequentially

http://bio-bwa.sourceforge.net/

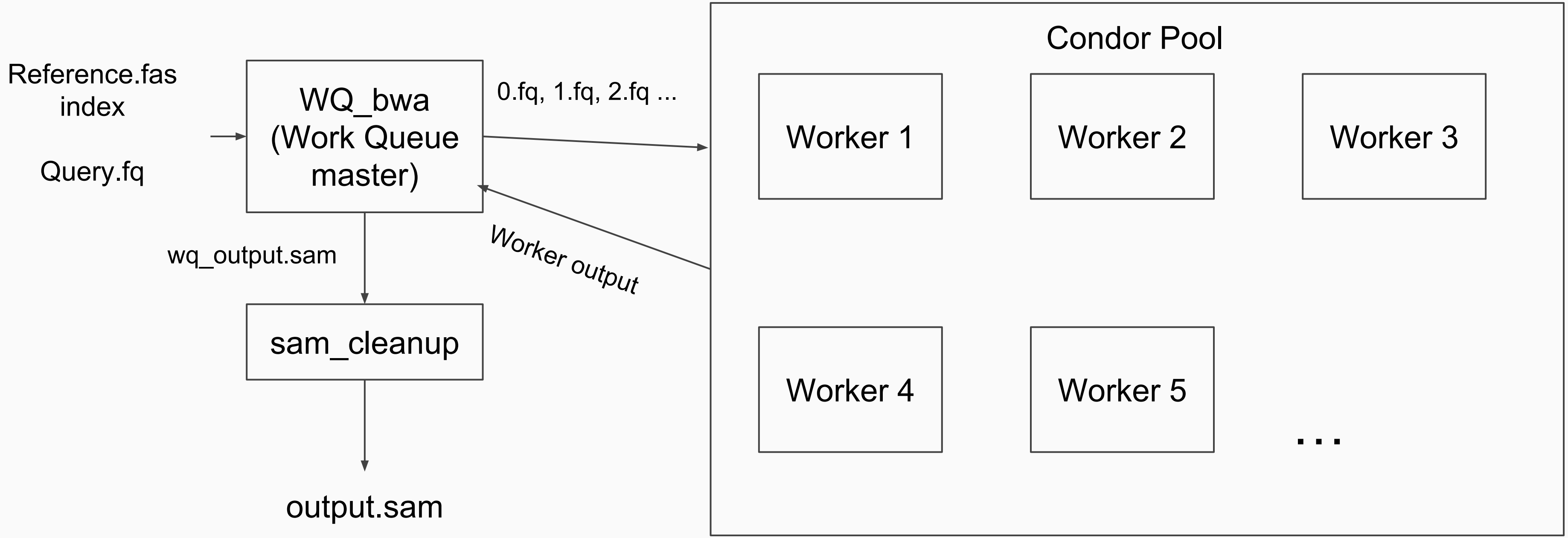
Solutions: Work Queue and Hadoop MapReduce





Images from: http://bactifeed.com/biological_soil_conditioner_FAQ.html and http://ccl.cse.nd.edu/software/workqueue/WorkQueueLogoSmall.png

Work Queue Model



Using Hadoop MapReduce

- Uploaded query fastq file to HDFS
- Ran the following command:

```
hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar -input /users/cray/poolA_dem.fq -output /users/cray/bwa_output -mapper 'bwa mem trnL_mod.fas -' -reducer /bin/cat -file /afs/crc.nd.edu/x86_64_linux/bio/BWA/0.7.12/bin/bwa -file trnL_mod.fas -file trnL_mod.fas.amb -file trnL_mod.fas.ann -file trnL_mod.fas.bwt -file trnL_mod.fas.pac -file trnL_mod.fas.sa -numReduceTasks 1
```

Results

	Runtime (seconds)	Speedup	Efficiency (%)
Sequential	1282	-	-
WQ (50 workers)	463	2.769	5.538
WQ (100 workers)	406	3.158	3.158
WQ (150 workers)	403	3.181	2.121
WQ (200 workers)	404	3.173	1.587
MapReduce	332	3.861	6.657

Challenges

- Implementing Work Queue model that continuously submitted/waited for tasks in the queue rather than submitting all tasks at once
- Configuring optimal Work Queue task size
- Configuring BWA to be usable with Hadoop MapReduce

Future Work

- Continue configuring ideal task size
- Possibly implement method for choosing different BWA algorithms within Work Queue Model and Hadoop MapReduce