CS 6240 Parallel Data Processing in Mapreduce

Assignment 1
Date: Feb 4, 2015
Author: Yue Liu

Operating System: Mac OS X Yosemite Version 10.10.1
Processor: 2.4 GHz Intel Core i5
Memory: 4 GB 1067 MHz DDR3
Disk : SATA HDD(Since the disk of my machine is SATA HDD(writing speed = 25M/s), the performance is around 1/30 of those with solid-state disk)

(Q1) What is the performance difference between v1, v2 and v3 where you run v2 and v3 both in a single Hadoop process and "pseudo" distributed processes.

The result of my v1, v2, v3 are 1180s, 386s, 478s, respectively. Obviously v2 is quicker than v3, v3 is quicker than v1. In brief, the rank of speed in my machine is v2 > v3 >v1.

Version	Time(s)
V1	1180
V2	386
V3	478

(Q2) Comparing v(2|3) with v4, what is the largest value of N that does not affect performance?

I modified v3 to v4(computing Fibbonacci number with memoization) and computed N = 10, 50, 100, 150, 500, 750, 1000. The results are 480s, 440s, 446s, 447s, 442s, 437s, 456s, respectively. Obviously the performance of map function does not impact the overall compute time.

N	Time(s)
10	480
50	440
100	446
150	447
500	442
750	437
1000	456

(Q3) How many instances of the reducer are running?

Only one instance of the reducer is running since it is standalone.