DSC 102
Systems for Scalable Analytics

Arun Kumar

Exercise 2

Time tip: Roughly 45sec to 1min per 1pt
Exercise

Q1) [3 x 2pts] Suppose an SQL query takes 20min to run on a single worker node and x min when run on 5 worker nodes. What is the speedup for the given value of x? Is the speedup linear, sublinear, or superlinear?
A. x = 7min
B. x = 4min
C. x = 3min
Suppose an ML training procedure takes 40 min to run on a single worker node. We then triple the dataset size, say, to help improve accuracy and use 3 worker nodes. It now takes $x$ min. What is the scaleup for the given value of $x$? Is the scaleup linear, sublinear, or superlinear?

A. $x = 50$ min
B. $x = 40$ min
C. $x = 35$ min
Q3) Consider the following task graph with the task lengths shown. You are given 3 workers to execute this graph in a task-parallel manner like discussed in class.

A. [2pts] What is the lowest possible completion time of this workload?

B. [4pts] What is the highest possible speedup of this workload on 3 workers vs 1 worker?

C. [4pts] What is the total idle time across all workers in a schedule that yields the highest speedup?
Exercise

Q4) [4pts] Suppose you are given a workload with $n$ tasks, each of length $k$ units. You are allowed to use task parallelism as discussed in class. What is the lowest possible completion time of this workload? What should the task graph look like and what is the number of workers needed to achieve that completion time?