MapReduce: Simplified Data Processing on Large Clusters

NOTE: Your slides/presentation only need to cover background information necessary to answer the given questions (in a clear and well-organized manner). You are allowed to borrow contents from other resources, such as online slides, as long as you acknowledge them. The presentation should be mostly question-focused and proceed mostly in a Q&A format. Please include the questions in your slides. Don’t write detailed answers in the slides and read them to the class. Instead, use bullet points, graphs, or animations to explain your answers to the class.

In your Q&A report, use text to more thoroughly answer the questions. Include a short paragraph at the beginning of the report to summarize the paper.

(1) Compared with traditional parallel programming models, such as multithreading and MPI, what are major advantages of MapReduce?

(2) Use Figure 1 to explain a MR program’s execution.

(3) Describe how MR handles worker and master failures.

(4) The implementation of MapReduce enforces a barrier between the Map and Reduce phases, i.e., no reducers can proceed until all mappers have completed their assigned workload. For higher efficiency, is it possible for a reducer to start its execution earlier, and why? (clue: think of availability of inputs to reducers)