MapReduce: Simplified Data Processing on Large Clusters

NOTE: Your slides/presentation need to cover the assigned sections and 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. For a slide that covers a given question, please print the question on the slide. However, you don’t have to answer the question using a long paragraph of text on the slide. Instead, use bullet points, graph, animation, or oral explanation to answer the question. In your Q&A report, use text to more thoroughly answer the questions.

You only need to cover Section 1 -- Section 3.

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)