COMPUTAÇÃO DE ALTO DESEMPENHO

2018/2019

Test 2

Note: justify all your answers

1. [2 values] What distinguishes parallel computing on a distributed memory architecture from parallel computing on a shared memory architecture? What new concerns you need to have in mind when moving from a shared memory to a distributed memory architecture?

2. [1 value] Is it possible to have shared memory programming model on top of a distributed memory architecture? If not, justify your answer. If so, explain how it could be done.

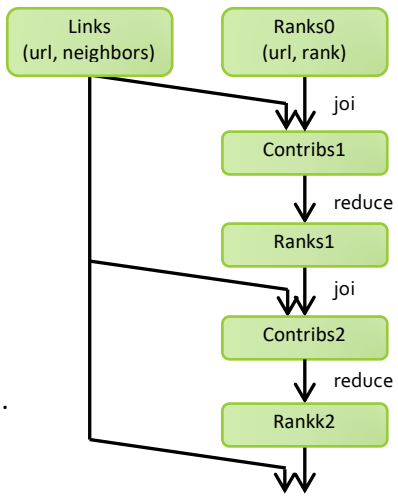
3. Consider the following sequential Java method that returns the size of the longest string in a set of strings:

```
int maxStringSize(Set<String> set) {
  int result = 0;
  for (String s : set) {
   if (s.length > result)
       result = s.length;
  }
  return result;
}
```

   a. [3 values] Present a Spark implementation of such method. You may present a solution in Java or Scala, and use RDDs or Datasets. A possible signature for the method is **int** maxStringSize(JavaRDD<String> set) Moreover, assume that Spark has already been initialized.
   b. [1.5 value] Considering that you are now deploying your solution in a cluster of four nodes. Is your solution suitable for distributed parallel execution?
   c. [1 value] Your solution makes use of a data or of a task decomposition technique?
   d. [1 value] Are these decompositions guided by the input or output parameters?
   e. [1 value] Are you certain that your code will execute on all four machines? If not, are you able to programmatically ensure (in Spark) that will effectively happen?

4. [1.5 values] Explain the Single Program Multiple Data execution model. Do Spark and/or MPI apply a Single Program Multiple Data execution model?

5. Consider the PageRank Spark implementation studied on the course and that iteratively processes a graph of linked web pages, in order to compute a ranking of such pages.



   a. [1 value] Identify the narrow and wider dependencies.
   b. [1 value] Decompose the computation into stages and identify all network communication.
   c. [1.5 values] What is the impact of tuning data locality in this example. How can you do this tuning?

6. In the context of static and dynamic mapping of tasks into processing elements.
   a. [1 value] Is mapping in Spark static or dynamic?
   b. [1 value] Explain the purpose and the mechanics of the work-stealing algorithm.
   c. [0.5 values] Why does the work-stealing algorithm uses a deque rather than a queue?