CPS216 Data-Intensive Computing Systems, Fall 2011
Exercise 1

Question 1 This question is based on the following SQL query over table R(A,B):

```
Select    A, MAX(B)
From      R
Where     B >= 1000 and B < 2000
Group By  A
```

Figure 1 describes the contents of the records in table R(A,B). There are 10000 records in R, with 2500 unique values of A and 5000 unique values of B.

(a) Explain how MapReduce can be used **most efficiently** to process this query. That is, explain what the Map phase will do and what the Reduce phase will do. Keep your answer brief and to the point.

(b) Suppose the records in table R are stored on 10 nodes, M1-M10, as shown in Figure 2. For example, all records with "A $\geq$ 1 and A $\leq$ 250" are stored on node M1, all records with "A $\geq$ 251 and A $\leq$ 500" are stored on node M2, and so on. The MapReduce computation is done with 10 Mappers and 2 Reducers. The Mappers in the will run on the 10 nodes M1-M10. The Reducers will run on two separate nodes R1-R2 such that R1 will process all records with "A $\geq$ 1 and A $\leq$ 1250", a and R2 will process all records with "A $\geq$ 1251 and A $\leq$ 2500". In this scenario, how many records will be shuffled from a Mapper node to a Reducer node in the most efficient MapReduce execution?

Records in R

A	B
1	1
1	1
1	2
1	2
2	3
2	3
2	4
2	4
	⋮
2499	4997
2499	4997
2499	4998
2499	4998
2500	4999
2500	4999
2500	5000
2500	5000

R has a total of T(R)=10000 records

Algebraic representation of records in R

A	B
x	2x−1
x	2x−1
x	2x
x	2x

For x in 1,2,3,...,2499,2500

Figure 1: Figure showing the contents of records in R

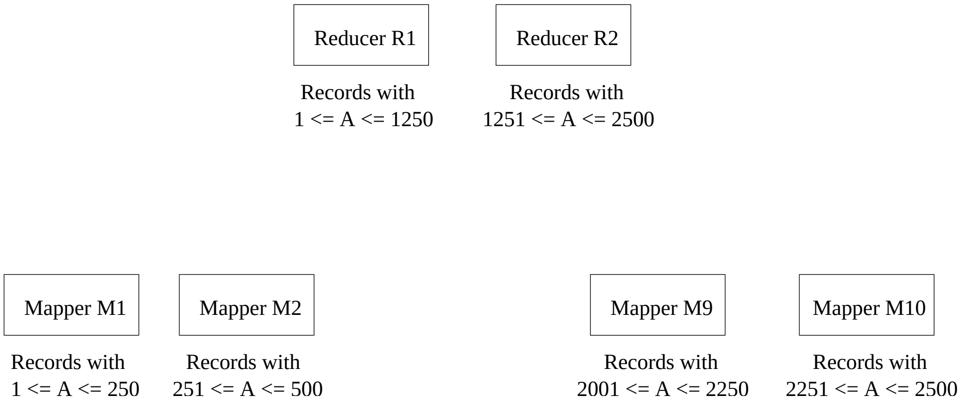

Figure 2: Figure showing the 10 mappers M1-M10 and 2 reducers R1-R2