1   Cluster Design

Read the paper on the **Google Cluster Architecture** by Barroso, Dean, and Holzle. A link to the paper is available in section C5 link L2 of the Web-page or the corresponding Subject. Try to answer after reading the paper the following questions whose answers are in the paper or may need to be further thought upon. Your answers must be well-documented based on textbook, notes, or paper-based (C5/L2) material and quantitavely-justified.

Questions Q1-Q7 prepares you to collect data to answer the remaining questions.

Q1 (10 points). Citing the 1998 (not 2003) paper, what are the Google (1998) statistics for the following questions. (Round numbers up to the next multiple of 10 for (1.2) through (1.4), and to the next integer for 1.5.)

1.1 Number of documents indexed,

1.2 Size of uncompressed and compressed text,

1.3 Index size (includes lexicon, anchor-related data),

1.4 Doc-index size, and

1.5 Links size used for PageRank computations.

Q2 (6 points). From the 2003 paper, what are the answers to the following questions. (Assume a power consumption of 200W per machine, and use the fastest machine configuration mentioned in the paper/Subject 5; for rack configuration use page 14, NOT page 15 information.)

2.1 Maximum number of machines per side of a rack (rack-side) and per rack,

2.2 Footprint of a rack, and power consumption (per rack rather than individual machine),

2.3 Switch connections per rack side (or rack) assuming one machine one network card.

Q3 (4 points). Related to Question 2, answer the following questions (Use the most powerful configuration quoted in the paper as instructed also in 1.2.)

3.1 How much main memory RAM per rack (in GBytes).

3.2 Disk space (in GBytes or TBytes) per rack.

Q4 (4 points). Answer the following questions. (These questions is for a single cluster of roughly 2000 machines.)

4.1 Are raw-page data stored redundantly and if so how many times.

4.2 Is the index stored redundantly and if so how many times?

Q5 (2 points). What percentage of the machines become query machines (use definition before text of Q8)? Justify your answer and make sure it makes sense given the load of a cluster.

Q6 (8 points). Based on the answers of Q1-Q5 (round up to closest integer)

6.1 What is the average document size.

6.2 What is the average index-size contribution per document.

6.3 What is the average DocIndex contribution per document.

6.4 Average links contribution per document.

Q7 (8 points). Based on Q6

7.1 What is the average size per document not only of the raw document but also of doc-index data structures such as thos mentioned in 6.1-6.4.

7.2 What is the average size per document of all index-related data structures as mentioned in 6.1-6.4

7.3 What is the document / index-ratio as can be derived from 7.1 and 7.2 (no rounding)?

7.4 If one maintains 2 copies of the raw documents, and 4 copies of the index, what is the modified answer?

You are given 2000 machines (PCs). Using the data available in the paper answer the following questions in a way that is fully compatible with the paper. Read all the questions first. You may assume that you will use only three types of servers: index, doc, and query servers (that will become GWS, cache, ad, spell check ones).

Q8 (20 points). How would you organize the 2000 machines in racks etc? (Note keep a backup of roughly 4% of resources. Also assume that switches are mostly 100Mbit with 2 additional Gigabit uplinks each. Few Gigabit switches are also available.)

8.1 How many racks are you going to use and how many rack-sides are there?

8.2 How many network switches?

8.3 Other resources? Explain.

Q9 (20 points). What is the collective memory (RAM), disk space (TBytes) of these resources? And also per rack (GBytes)!

Q10 (20 points). Using data from other papers you have already read for this course,

- How many documents can your cluster design support? (Be consistent with Q7 on doc/index size.)

- How many of these machines will be file servers, how many index servers, and how many query machines? (Be consistent with Q7 on doc/index size.)

- How many queries can the structure support per second?

Q11 (23 points). Does Google store compressed or uncompressed raw data at the time of writing of this (more recent) 2003 paper? Quote the paper. How about index data?