

**Assignment 1** In this problem we assume that a single disk inside a RAID 5 array has died. In order to recompute the missing data, the data stored on all the remaining disks has to be read. The task is to consider the number of expected read errors during the reconstruction of the data.

When using consumer hard disks in RAID 5 configuration, compute the expected number of URE errors during RAID 5 array rebuild. Use a Bernoulli process model the URE errors with the typical consumer URE rate of 1 bit error per  $10^{15}$  bits read.

Assume the RAID 5 arrays are full of data and consist of the following amounts of storage space (without the parity disk):

- a) 10 TB (1p)
- b) 20 TB (1p)
- c) 40 TB (1p)

**Assignment 2** One of the difficulties in scaling out distributed systems is that of worst case latency. Imagine for example a Web search application, where the search index is distributed over  $N$  servers, and in order to give the final search result to the client one query is sent to each one of the  $N$  servers and summarized together in a centralized fashion before giving the client back any results. Assume that the response time of each of the servers is a Bernoulli process with response time of 1 ms for 99% of the queries, and the response time of 50 ms for 1% of the queries (due to e.g. hard disk seek latencies, Java Garbage collection, or other bookkeeping processes running on the same server). Assume for simplicity there are no other delays.

- a) What is the expected latency for a client query in the case  $N = 10$ ? (1p)
- b) What is the expected latency for a client query in the case  $N = 100$ ? (1p)
- c) What is the expected latency for a client query in the case  $N = 1000$ ? (1p)

**Assignment 3** Briefly (using at most half a page of text maximum) describe what is the statement called as the FLP Theorem and what implications does it have for design of distributed systems both on the theoretical as well as on the practical level. (4p)

**Assignment 4** Bloom filters are a probabilistic data structure for storing sets of items. Consider the case of a Bloom filter with 4 megabytes of memory, where we would like to insert at most  $n = 9000000$  items. What is the approximate optimal number of hash functions  $k$  to minimize the number of false positives? What is the false positive probability with using that  $k$  (rounded to the nearest integer) after having inserted  $n$  unique items? (4p)

$$\left(1 - e^{-\frac{kn}{m}}\right)^k$$
$$k =$$

**Note! Non-programmable calculators are allowed in this exam.**  
**More assignments on the other side of the paper.**

---

The name of the course, the course code, the date, your name, your student id, and your signature must appear on every sheet of your answers.

Course feedback questionnaire is now open. Thank you for feedback!

**Assignment 5** Please briefly (using maximum of three sentences for each case) define what are the following concepts as used the the course lectures:

- a) Consistent hashing (1p)
- b) RAID 0 (1p)
- c) RAID 10 (1p)
- d) BASE (1p)
- e) Scaling up (1p)
- f) Paxos algorithm (1p)

**Assignment 6** Briefly (using at most half a page of text maximum) describe the MapReduce programming paradigm and its relation to fault tolerance and functional programming. (4p)

---

The name of the course, the course code, the date, your name, your student id, and your signature must appear on every sheet of your answers.

Course feedback questionnaire is now open. Thank you for feedback!