Aalto University, Department of Computer Science KH

CSE-E5430 Scalable Cloud Computing

Examination, December 18, 2015

Assignment 1 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 5 ms for 99% of the queries, and the response time of 120 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 2 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 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:

a) 12 TB (1p)

b) 24 TB (1p)

c) 48 TB (1p)

Assignment 3 Briefly (using at most half a page of text maximum) describe the CAP Theorem and the implications it has for the design of distributed systems. What are the main use cases of the CA, AP, and CP systems? Please explain through the use of examples how these kinds of systems combined? (4p)

Assignment 4 Bloom filters are a probabilistic data structure for storing sets of items. Consider the case of a Bloom filter with 8 megabytes of memory, where we would like to insert at most $n = 10000000$ 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)

Note! 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. All calculators are allowed in this exam.

Aalto University, Department of Computer Science KH

CSE-E5430 Scalable Cloud Computing

Examination, December 18, 2015

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

a) Asynchronous consensus problem (1p)

b) Bernoulli process (1p)

c) RAID 10 (1p)

d) Lambda architecture (1p)

e) Distributed Coordination System (1p)

f) RAID write hole problem (1p)

Assignment 6 Briefly (using at most half a page of text maximum) describe the Apache Spark programming paradigm and compare its benefits and drawbacks compared to the MapReduce programming paradigm. (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. All calculators are allowed in this exam.