Aalto University, Department of Computer Science KH

CS-E4120 Scalable Cloud Computing

Examination, February 13, 2017

Assignment 1 Amdahl's law states the maximum available speedup of a program which has proportion P of parallel code and $1-P$ of sequential code, and N processors working in parallel. Compute the speedup of code with $N = 128$.

a) In the case $P = 90\%$. (1p)

b) In the case $P = 95\%$. (1p)

c) In the case $P = 99\%$. (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 (without the parity disk):

a) 50 TB (1p)

b) 100 TB (1p)

c) 150 TB (1p)

Assignment 3 Briefly (using at most half a page of text maximum) describe the consistent hashing approach to implementing a distributed hash table (DHT), together with the potential benefits or drawbacks of the approach when adding or removing servers to a DHT. (4p)

Assignment 4 Bloom filters are a probabilistic data structure for storing sets of items. Consider the case of a Bloom filter with 10 megabytes of memory, where we would like to insert at most $n = 7000000$ 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

CS-E4120 Scalable Cloud Computing

Examination, February 13, 2017

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) Apache Spark (1p)

b) RAID 6 (1p)

c) FLP Theorem (1p)

d) Lambda architecture (1p)

e) Scaling up (1p)

f) ACID properties (1p)

Assignment 6 Briefly (using at most half a page of text maximum) describe the Hadoop Distributed Filesystem (HDFS) architecture. What are the techniques used in HDFS to implement fault tolerance, and are the techniques different or similar to existing RAID based fault tolerance techniques? Can you analyze the HDFS architecture from the CAP theorem perspective? What are the strong points and weak points of the HDFS architecture regarding scalability to different workloads as well as with regards to fault tolerance?

(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.