CS-E4640 **Autumn 2018**
Big Data Platforms
Tutorial 1

1. Amdahl's law (see Lecture 1) 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 $N = 10$, $N = 100$, and $N = 1000$.

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

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

   c) In the case $P = 99\%$.

2. Assume a 6 TB (note: hard disk TB $= 6 * 10^{12}$ bytes) hard disk that has a disk seek time of 10 ms, and a transfer rate of 150 MB/s ($150 * 10^6$ bytes/s).

   a) How much data can be read in the same time as is needed to perform one seek?

   b) What is the time needed to read all of the capacity of the hard disk sequentially?

   c) What is the time needed to read all of the capacity of the hard disk by reading data in 4 KB blocks in random order?

   d) What is the time needed to read all of the capacity of the hard disk by reading data in 64 MB blocks in random order?

3. In 2015, processors with four cores were very common. If we assume a conservative 24 month doubling rate of number of transistors in Moore's law, and also (very) hypothetically assume all the available transistors would be used to increase the number of cores a processor has (to simplify things, assume the transistor usage would be linear in the number of cores). Then, how many cores would a processor have in 2025? What about in 2035?

   Note: *Harnessing these processor core counts will be very challenging, and thus also other uses of transistors in processors are likely.*

4. Demo question: The paper: "David A. Patterson: Latency lags bandwidth. Communications of the ACM, Volume 47 Issue 10, October 2004." gives a long term observation (1982-2001) that hard disk bandwidth doubles every 2.8 years. However, in the time hard disk bandwidth has doubled, the hard disk capacity has grown 2.4 times. The

main reason for this is that when the area used to store a bit in hard disk becomes smaller, the bandwidth is increased in one physical dimension, while capacity increases in two physical dimensions, as also hard disk tracks can be packed more tightly together.

Let us consider the following **fully hypothetical scenario** based on numbers from this paper. Using the assumption these growth trends will continue from 2015 forward with the exact same bandwidth and capacity doubling rates, use an initial assumed data of 6 TB hard disk with 150 MB write speed in 2015, and 5 years planned hard disk lifetime. After how many write bandwidth doublings will writing the high end hard disk to its full capacity take more than its planned lifetime? What is the disk capacity, bandwidth, and year at that point?

Hint: *Use a small computer program to simulate the growth of both bandwidth and capacity.*

Note: *It is highly probable that flash or other non-volatile storage will replace hard disks before the scenario depicted here will happen.*