This is a written exam for the course "*Performance of Networked Systems*"

Lecturers: prof.dr. R.D. van der Mei and dr. T. Kielmann
Date: March 27, 2015

Rules for the exam:

1. Allowed material: This is a closed-book exam. You are not allowed to use any kind of written material or your laptop, and electronic communication during the exam is strongly prohibited.

2. Calculation of end grade for the course: the end grade for the course is built up in two parts: homework assignments and a written exam.

   - *Homework assignments*: Both homework assignment grades count for 20% each of the final grade.
   - *Written exam*: for this written exam you get a grade between 1 and 10. This grade will count for the remaining 60% of the final grade.
   - *Final grade*: the final grade is calculated as the weighted average of the grade for the written exam and the two homework assignment grades, with the restriction that the grade for the written exam must be at least 5.5, in order to pass.

4. Credits: This written exam consists of four questions (A, B, C and D), each of which consists of a number of sub-questions. The maximum number of credits you can get is distributed as follows amongst the sub-questions:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**total**
A	4	4	4	4	4	4			**24**
B	6	6	6	6					**24**
C	6	10							**16**
D	6	6	8	6	6				**32**

Good luck!

QUESTION A: Some basics

A.1 What is a Poisson process, and why do Poisson processes provide a natural way to model randomly occurring events?

A.2 A discrete random variable N is said to have a Poisson distribution with mean λ if for k=0,1,2,…

$$\Pr\{N = k\} = \frac{e^{-\lambda}\lambda^{k}}{k!}.$$

What is the relation between a Poisson *process* and a Poisson *distribution*? Be precise.

A.3 Mean Value Analysis (MVA) is a powerful way to calculate performance metrics for closed queueing networks. Explain in words what the basic idea behind MVA is (no mathematical notation is required).

During the course, a number of mobile and wireless network technologies have been discussed, including GSM, GPRS, UMTS, and HSDPA and WLAN.

A.4 What are the main differences between GSM and GPRS from a performance point of view?

A.5 What are the main differences between UMTS and HSDPA from a performance point of view?

A.6 For the WLAN 802.11 protocol, the performance model of Bianchi was discussed extensively. Describe what the basic ideas behind the Bianchi model are. Be short, but be clear.

QUESTION B: Capacity planning for Video-on Demand for CableCom

Cable TV company CableCom plans to offer Video-on-Demand (VoD) services, allowing their subscribers to watch videos upon request *at any time*. This is fundamentally different from the traditional situation, where CableCom used to offer only standard cable TV services: for each TV channel pre-scheduled TV programs (see your TV guide) used to be simply broadcast to all customers at specific times. CableCom has installed two types of servers: a signalling server and a video server. The process of setting up a VoD session consists of two phases. First, the client send a request to the signalling server to set up a connection between the video server and the client (phase I). Once a connection between the video server and the client has been established, the video server immediately starts to send the video traffic stream to the client TV (phase II). See Figure 1.

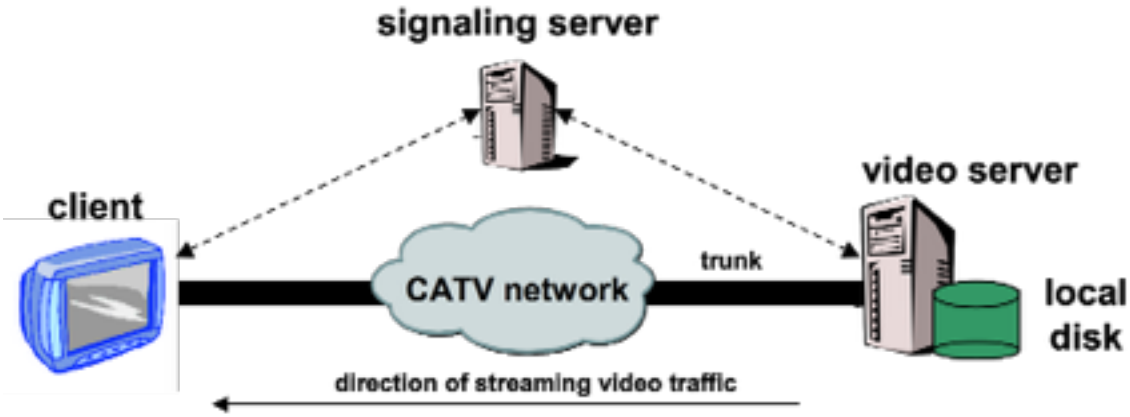

Figure 1: CableCom's Video-on-Demand service.

CableCom expects that VoD will be a commercial success, and therefore, wants to plan the capacity of its infrastructure properly and timely anticipate on performance problems when the number of users increases in the near future. In this context, CableCom wants to make sure that the signalling server is well-dimensioned so that the connection set-up phase does not take too long. Moreover, they want to make sure that the amount of network bandwidth is sufficiently large.

We make the following simplifying assumptions:

Regarding Phase I: The amount of time involved in processing a connection set-up request by the signalling server is exponentially distributed with mean 1 second. The signalling server handles connection set-up requests in the order of arrival, and can handle one request at a time (single-threaded). The network latency and bandwidth consumption involved in establishing a connection are negligible.

Regarding Phase II: Once the connection is established the video server will generate a traffic steam over the CATV network trunk at a constant rate of 5 Megabits per second for the duration of a movie. The duration of a movie has a gamma distribution with mean 2 hours. The CATV network trunk is shared by all clients of CableCom and its capacity is 1 Gigabit per second. When there is insufficient bandwidth available for running a VoD session over a newly established connection, the session is blocked and the connection is terminated. The time involved in terminating a connection is negligible.

B.1 Formulate a performance model that encompasses *both* the delay involved in setting up a connection and the blocking of VoD sessions. Define the relevant notation and the performance metrics. Motivate your assumptions and be precise!

B.2 What is the expected time it takes to set up a connection between the video server and the client?

B.3 CableCom wants to deliver good service to its VoD customers, and requires that the average duration of the connection establishment phase (including both waiting time and processing time) is less than 5 seconds. What is maximum number of connection set-up requests per minute that can be handled while meeting this constraint?

B.4 How many VoD sessions can the CATV network trunk handle simultaneously? Give an expression for the session blocking probability.

QUESTION C: TRANSPORT LAYER PERFORMANCE

C.1 Explain the term “bandwidth-latency product” and its impact on the performance of a TCP connection. How big must the send window be to allow the sender to send without interruption (in the absence of packet loss)? How big should the receive window be to maximise the achieved bandwidth?

C.2 Explain how TCP includes congestion control in its sliding-window algorithm. Explain how TCP detects network congestion and how it deals with it?

QUESTION D: PERFORMANCE OF HTTP

D.1 Explain the impact of latency on the completion time of a simple HTTP (1.0) request. Draw a sketch showing the messages that are exchanged between client (browser) and server.

D.2 In recent years, the TCP specification has been adapted such that the start size for the congestion window in slow start has been raised from 1 to 4 to 10 segments. How does HTTP performance benefit from this?

D.3 Which performance problem is created by using HTTP 1.0 or 1.1 over TCP? Which workaround strategies have to be followed by designers of web applications?

D.4 Which particular problems arise for web applications on mobile networks? What should a mobile application do to deal with these problems?

D.5 By which major change does HTTP/2 improve over HTTP 1.1? Explain how HTTP/2 is expected to have higher performance than HTTP1.1.