This is a written exam for the course *"Performance Analysis of Communication Networks"*
Lecturers: prof.dr. R.D. van der Mei and dr. T. Kielmann
Date and location of exam: Tuesday, December 16, 2008, 8.45-11.30

Rules for the exam:

1. Allowed material: This is an open book exam. For answering the questions, you are allowed to use all kinds of written material like textbooks, printouts of the lecture slides, your own notes, etc. You are allowed to bring your laptop for looking up electronic versions of course reading material, but electronic communication during the exam is strongly prohibited.

2. Language disclaimer: You are kindly asked to answer the questions using the English language. However, if it helps clarifying your answers, you may use some Dutch here and there. Doing so, will not affect your result.

3. 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*: during the course three homework assignments have been distributed among the students and placed on the Web site. For each homework assignment each student will receive a grade between 1 and 10. The deadline for the latest homework assignment is December 24, 2008. The average of the three grades counts to 50% 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 50% of the final grade.
   - *Final grade*: the final grade is calculated as the average of the grade for the written exam on the one hand, and the average homework grade on the other hand, with the restriction that the grade for the written exam must be at least 4.0.

4. Credits: This written exam consists of three questions (A, B and C), 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**	**total**
A	2	2	2	2	2	3	3	**16**
B	5	3	5	3				**16**
C	4	4	4	4				**16**

Good luck!

QUESTION A: Dimensioning of cellular networks

A mobile operator of a cellular GSM network wants to determine how many base stations are needed to satisfy its customers' Quality of Service (QoS) demands. To this end, the operator wants to determine the maximum size of a cell for which the call-blocking probability is still below some given threshold. Voice telephone calls are generated with rate 4 calls per minute *per square kilometer* (i.e., km^2), and the call duration has a gamma distribution with mean 2 minutes. Assume that each voice call requires a single channel to the nearest base station, and that each cell can support only 4 channels in parallel.

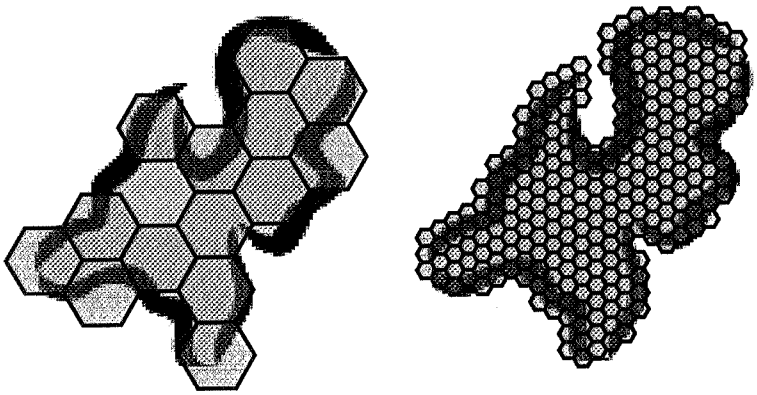

Figure 1. Illustration of GSM network dimensioning problem.

To make a proper decision on the number of base stations to be placed to offer good QoS to its customers, the operator wants to understand the impact of the cell size (in km^2) and the call-blocking probability.

A.1 Formulate a simple model description for the problem. Be precise, motivate your assumptions and clearly define any notation!

A.2 Give an expression for the call blocking probability in terms of the models parameters, and calculate the blocking probability for cell size 1 km^2.

A.3 The call blocking probability is known to be *insensitive* with respect to the distribution of the call duration. What *exactly* does that mean? Be precise!

A.4 Is the call blocking probability also insensitive with respect to the inter-arrival time distribution of the calls? If so, why, if not so, give a counter-example.

A.5 What is the relation between a Poisson *process* and a Poisson *distribution*? And what is the difference between them? Be precise!

Now suppose the service provider wants to offer a new *additional* service to its customers, video conferencing, in two flavors: (1) low-resolution video conferencing, requiring 2 parallel channels for each connection, and (2) high-resolution video conferencing, requiring 4 parallel channels for each connection. Video conferencing calls arrive according to a Poisson process with rate 4 calls *per hour per* km^2, and the mean conference call duration is 20 minutes (for both high and low resolution calls). 50% of the conference calls requires low resolution, and 50% requires high resolution. Recall that each cell has 4 channels. Call attempts are blocked when there are not enough lines available.

A.6 Determine the blocking probability for the voice telephony calls if the cell size is 1 km^2 if the new service is added.

A.7 Determine the blocking probability for the both the low-resolution video conferencing calls and the high-resolution conference calls if the cell size is 1 km^2 if the new service is added.

QUESTION B: Sliding window protocols

B.1 For a sliding-window protocol, the achievable bandwidth between a pair of sender and receiver depends on the performance characteristics of the network in between. On which characteristics precisely? (assuming the absence of transmission errors) How can the send window size be optimized for maximizing achievable bandwidth?

B.2 Somebody tells you that he managed to optimize the achievable bandwidth of a TCP connection by setting the receiver's window size to twice the size of the sender's window. Do you follow his advice?

B.3 PSockets, published in 2000, was the first software library to use multiple TCP connections between a sender-receiver pair in parallel. Compare the achievable bandwidth of

(a) one single TCP connection with a send window size W1
(b) n parallel TCP connections each with a send window size Wn, withW1 = n * Wn

In which case are the parallel TCP connections better than the single one? For answering the question it is sufficient to consider the steady state of the connection(s), thus neglecting the startup phase.

B.4 Would multiple parallel TCP connections also improve the performance of HTTP 1.0? How does HTTP 1.1 try to improve performance over HTTP1.0?

QUESTION C: Response time analysis of cash dispensers

A popular bank, FastCash, wants to offer cash dispenser services to its customers. To this end, the system is equipped with a front-end (FE) server and two database (DB) servers: an authentication DB server and a balance DB server containing information about the customers' balance. The service works as follows. The customer (1) inserts a credit card, (2) types in a 4-digit ID number, (3) types in how much money (in euros) he/she wants to have, and then presses the OK button. This cash retrieval request is then immediately forwarded to the FE server. The FE server then pre-processes the request, and forwards the request to the authentication DB server (step 1), which checks whether the 4-digit ID number is correct. The response of this request is sent back to the FE server (step 2). Subsequently, the FE server processes the response and forwards the request to the balance DB server (step 3), which checks whether the customer's balance is sufficient, and then returns a response to the FE server (step 4), which processes the response, makes a decision on whether or not the requested amount of cash is paid out or not and terminates the transaction.

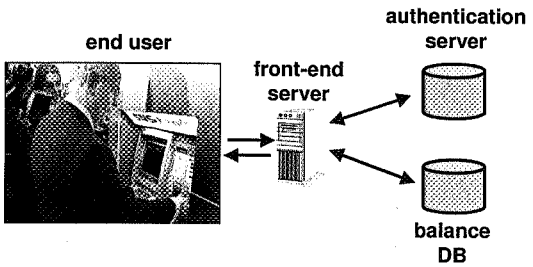

Figure 2. Processing phases of the cash dispenser service.

It is important to notice that during each customer transaction the FE server processes the request *three* times in total (namely: at the pre-processing stage, after the authentication phase and after the balance checkup phase), whereas both the authentication DB and the balance DB only process the request *once*.

Each of the three servers handles all requests *simultaneously* in a *processor-sharing* fashion; thus, for each of the servers it holds that if the are k jobs at that server, then each of these k jobs receives a fair share 1/k of the available capacity. Note that this is different from the FCFS-discipline that was assumed in the homework assignment.

The processing times of *each* request at the FE server, the authentication server and the balance DB are independent and have a gamma distribution with means β_{FE}, β_{auth} and $\beta_{balance}$, respectively. The clients' session initiation moments occur according to a Poisson process with rate λ. Our focus is on the processing times of the servers, and therefore, the network delay is assumed to be negligible. The total sojourn time of a request is this system represents the total response time of a combined cash retrieval request. FastCash now wants to be able to predict the total response time experienced by the client.

C.1 Formulate the model as an open queueing network (see Figure 2). Define the proper variables and motivate your assumptions. Be clear!

C.2 Give an expression for the *joint* probability distribution of the number of request present at each of the three queues/nodes. Be precise!

C.3 Give an expression for the expected total response time of the system. Motivate how you get this answer.

C.4 Assuming that the processing times are exponentially distributed instead of gamma-distributed. What impact would that have on the distribution of the numbers of request at each of the queues? Motivate your answer.