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: June 11, 2015, 18:30-21:15

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. Attached to this exam is a collection of formulas that have been discussed during the lectures. You are allowed to use them.

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 on the other hand, with the restriction that the grade for the written exam must be at least 5.5.

3. 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	3	3	3	3	3	3	3	3	**24**
B	6	6	6	6					**24**
C	3	3	3	5	4				**18**
D	9	9	3	3	6				**30**

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 2 calls per minute *per square kilometer* (i.e., km^2), and the call duration has a gamma distribution with mean 1 minute. 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.

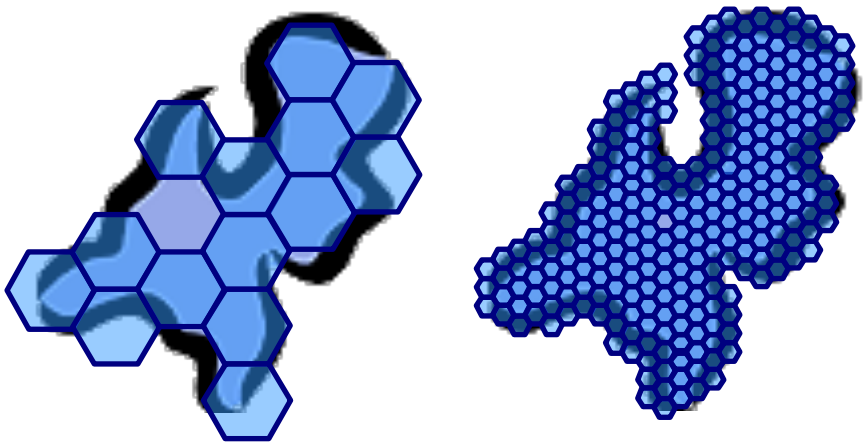


Figure 1: Illustration of GSM network dimensioning problem.

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.

A.2 The Poisson process is a natural way to model voice call arrivals. Why is that?

A.3 Give a formula for the call-blocking probability, depending on the size of the cell.

A.4 The call blocking probability is known to be insensitive with respect to the distribution of the call duration. But 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*? Be precise!

A.6 The Poisson process is known to have two important properties: (1) the *superposition* property, and (2) the *thinning-out* property. What do these properties mean?

Now suppose the service provider wants to offer a new *additional* service to its customers, video conferencing, requiring 2 parallel channels for each connection. Video conferencing calls arrive according to a Poisson process with rate 1.5 calls *per hour per* km^2, and the mean conference call duration is 15 minutes. Assume that the cell size is 2 km^2. Recall that each cell has 4 channels. Call attempts are blocked when there are not enough lines available.

A.7 Formulate a simple model description for the problem.

A.8 The Kaufman-Roberts recursion provides a powerful means to calculate the blocking probabilities for each of the call classes. Explain in words what the basic idea of the Kaufman-Robert recursion is (no formulas are required!).

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 2.

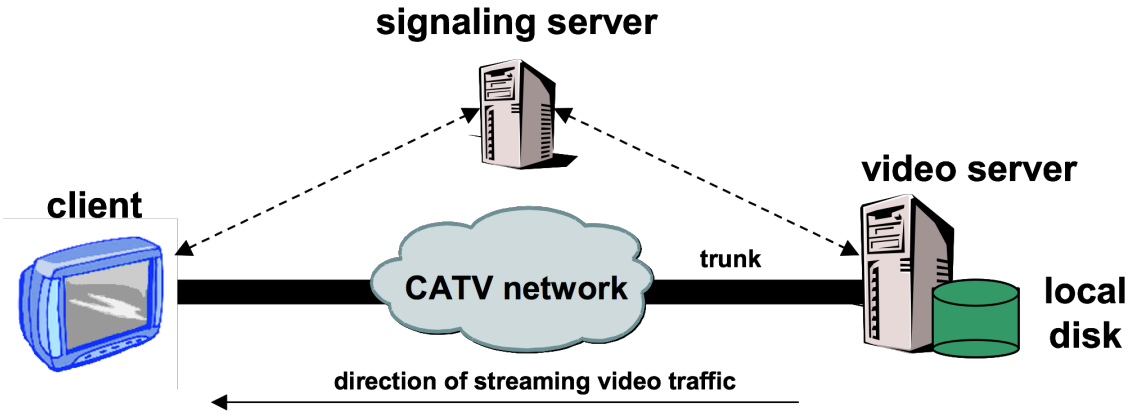


Figure 2: CableCom's Video-on-Demand service.

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 0.8 seconds. 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 20 Megabits per second for the duration of a movie. The duration of a movie has a gamma distribution with mean 1.5 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 3 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 Supposed you have a network connection with 1Mbit/s bandwidth and 100ms roundtrip latency (RTT). How big must TCP's send window be to allow the sender to send at 1Mbit/s for a long time?

C.2 How big must the receive window be to sustain the same bandwidth either when there is no packet loss or when there is a small amount of packet loss?

C.3 What is network congestion and how can an implementation of the TCP protocol detect it?

C.4 Explain how TCP implements congestion control, esp. when creating a new connection, and later in the stable state of a running connection.

C.5 What are "cumulative acknowledgements"? How does TCP use these in order to avoid the situation where transmission is stalled until a timeout occurs?

QUESTION D: Performance of HTTP

D.1 Supposed you have a network connection with 100ms roundtrip latency (RTT). If we ignore the size of all packets being transferred, except for the content of the HTML file being requested, how long does is take to fetch a HTML file of 1KB using HTTP when the available network bandwidth is either 1Mbit/s or 10Mbit/s ?

D.2 Nowadays, HTTP requests do have non-negligible size, mostly due to cookies that have to be transferred. As a reaction to this, the TCP specification has been adapted by increasing the congestion window from 1 to (finally) 10 packets. Assume RTT=100ms, assume bandwidth is 10Mbits/s, and the size of the HTTP request is 10KB (including cookies). What is the time it takes to fetch a HTML file of 1KB using HTTP, when TCP starts with a congestion window of either 1 packet or 10 packets? For simplicity, assume that each packet can hold 1KB of data.

D.3 HTTPS (compared to HTTP) adds encryption of the transmitted data by using TLS. If we ignore all optimisations of TLS, how does adding TLS affect the time to fetch a single file from a server, when compared to plain HTTP?

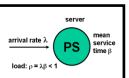
D.4 What is the reason for using symmetric encryption of the actual data that is exchanged whereas the initial key negotiation is based on the much more secure public key encryption?

D.5 Consider a web page that consists of tens of separate files on the server (HTML, CSS, images). Explain how HTTP/2 improves the page load time, compared to HTTP1.0. Why is HTTP 1.1's connection keepalive not good enough for the same purpose?

Rules of Thumb Queueing

Performance of Networked Systems

Processor Sharing model:

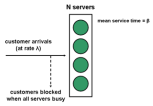


Expected time spent = $\frac{\beta}{1-\rho}$ and slow-down factor $\frac{1}{1-\rho}$

Insensitivity: result holds for any service-time distribution

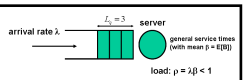
Erlang blocking formula:

Blocking probability = $\frac{\frac{(\lambda\beta)^N}{N!}}{1+\frac{(\lambda\beta)^1}{1!}+\frac{(\lambda\beta)^2}{2!}+\ldots+\frac{(\lambda\beta)^N}{N!}}$



Insensitivity: result holds for any service-time distribution

M/G/1 Queueing model



Average waiting time: $E[W]=\frac{\rho}{2(1-\rho)}\frac{Var[B]+\beta^2}{\beta}$

Average time spent: $E[T]=E[W]+\beta$

If service times fixed: $Var[B]=0$, if exponential: $E[T]=\frac{\beta}{1-\rho}$

Insensitivity property does NOT hold

Rules of Thumb Probability

Performance of Networked Systems

Exponential: $f_X(x)=\lambda e^{-\lambda x}(x>0),\ E[X]=1/\lambda, Var[X]=1/\lambda^2$

Uniform: $f_X(x)=\frac{1}{b-a}(a<x<b),\ E[X]=\frac{a+b}{2}/2, Var[X]=\frac{(b-a)^2}{6}$

Erlang-k: $f_X(x)=\mu^k\frac{x^{k-1}}{(k-1)!}e^{-\mu x}(x>0),\ E[X]=\frac{k}{\lambda}, Var[X]=\frac{k}{\lambda^2}$

Hyper-exponential (with 2 phases):

$f_X(x)=p\mu_1 e^{-\mu_1 x}+(1-p)\mu_2 e^{-\mu_2 x}(x>0),\quad E[X]=\frac{p}{\mu_1}+\frac{(1-p)}{\mu_2}$

Geometric: $\Pr\{N=k\}=(1-p)p^k(k=0,1,\ldots),\ E[N]=\frac{p}{1-p}, Var[N]=\frac{p}{(1-p)^2}$

Geometric: $\Pr\{N=k\}=(1-p)p^{k-1}(k=1,2,\ldots),\ E[N]=\frac{1}{1-p}, Var[N]=\frac{p}{(1-p)^2}$

Poisson: $\Pr\{N=k\}=e^{-\lambda}\frac{\lambda^k}{k!}(k=0,1,\ldots),\ E[N]=Var[N]=\lambda$

Binomial: $\Pr\{N=k\}=\binom{n}{k}p^k(1-p)^{n-k}(k=0,1,\ldots,n),\ E[N]=np,\ Var[N]=np(1-p)$