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 29, 2017, 08:45-11:30

Rules for the exam:

1. Allowed material: This is a closed-book exam. You are not allowed to use any kind of written material nor your laptop. 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, with the restriction that the grade for the written exam must be at least 5.5.

3. Points: This written exam consists of five questions (A, B, C, D, and E), each of which consists of a number of sub-questions. The maximum number of points 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	8	8	8						24
C	6	10							16
D	6	4							10
E	6	8	8						22

Good luck!

QUESTION A: On Poisson processes and performance of wireless networks

A.1 The Poisson process is a natural way to model random events. Why is that?

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 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?

A.4 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, only the basic idea).

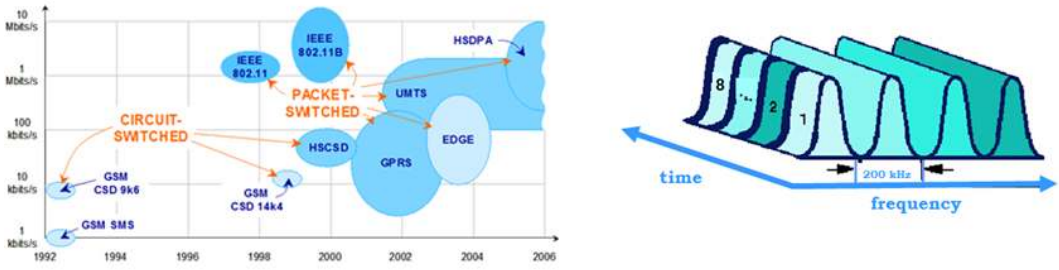


Figure 1: Evolution of mobile and wireless network technologies (left) and illustration of GSM medium access (right).

During the course, a number of mobile and wireless network technologies and their performance models have been discussed, including GSM, GPRS, UMTS, HSDPA and WLAN (see Figure 1).

A.5 The GSM protocol is based on both Time Division Medium Access (TDMA) and Frequency Division Medium Access (FDMA)? Explain what that mean? (see also Figure 1)

A.6 What are the main differences between GSM and GPRS from a performance point of view?

A.7 The HSDPA protocol is based on the concept of what is called *opportunistic scheduling*. What does that mean?

A.8 For the WLAN 802.11 protocol, the performance model of Bianchi was discussed extensively. Describe what the basic ideas of the Bianchi model are (no details are required, just give the intuition and the basic ideas). Be short and to the point.

QUESTION B: Design evaluation of a Web server system with caching

FastAccess Inc. is an Internet Service Provider (ISP) providing high-speed access to the Internet to its subscribers. In addition to providing Internet access, FastAccess also provides Web hosting services. To this end, FastAccess has invested in buying two Web servers, a load balancer and a server cache, which are connected via a high-speed LAN. Original copies of the Web pages hosted by FastAccess are stored on both Web servers, throughput referred to as WS1 and WS2. Copies of recently used Web documents may be stored on the server cache (see Figure 2 below).

A document-retrieval request initiated by a subscriber proceeds along the following steps: First, it is checked whether a copy of the requested document is stored in the server cache, and if so, the document is sent immediately from the cache to the end user within a negligible amount of time;

otherwise, the request is forwarded to the Load Balancer (LB) which, in turn, forwards the requests for the non-cached documents to one of the two Web servers. Both Web servers handle the incoming requests in the order of arrival, and can handle one request at a time.

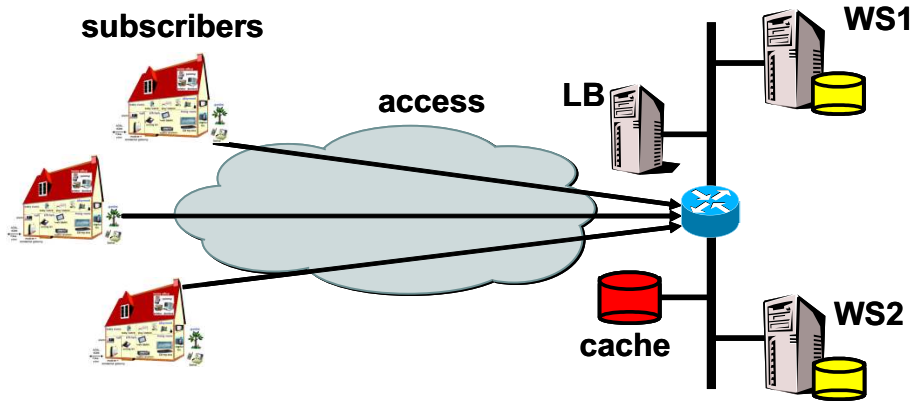


Figure 2. Illustration of FastAccess' infrastructure.

FastAccess wants to be known as a high-quality ISP by delivering Web hosting services with minimal response times for its own subscribers when retrieving documents stored at the Web servers (retrieval times for Web documents outside of the domain of FastAccess cannot be fully controlled and are not considered here). To this end, FastAccess want to evaluate the following design for the handling of non-cached requests.

The LB immediately forwards a request *randomly* to one of the two Web servers (with equal probability), both of which have implemented an infinite-sized buffer to temporarily store requests that arrive when the server is busy. Note that in this design, buffering only takes place at the Web servers, not at the LB.

Suppose FastAccess has hired you as a well-paid consultant to assess and predict quality of its infrastructure and wants to know how the response times depend on the number of customers in order to timely anticipate on system upgrades if the number of customers grows in the near future.

Note 1: We assume that both the access network and the LAN are over-dimensioned and that network latency is negligible. Instead, we focus on the effectiveness of the server cache and the performance of the Web servers.

Note 2: Do not take into account the local caching policy. Instead, assume that the cache-hit ratio is fixed (with some probability p).

B.1 Formulate a performance model for the response time and loss probabilities. Introduce notation and formulate your assumptions.

B.2 What is the overall mean response time of an arbitrary transaction? Motivate your findings.

B.3 Suppose the load balancer would distribute the request in round-robin order (i.e., 1, 2, 1, 2, 1, 2, etc) instead of randomly. Will that have an influence on the mean response time at the Web servers, and if so, will the mean response times get larger or smaller? Motivate your answer.

QUESTION C: Transport Layer

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 with TCP?

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: Web Applications and Mobile Networks

D.1 How does a wireless network pose extra performance challenges for web applications, compared to a wired network? How does a mobile connection (e.g. 4G) compare to Wifi?

D.2 How can a mobile web application save on device battery?

QUESTION E: HTTP

E.1 Google's web engineer Eric Bidelman sketches in his blog the difference between HTTP1.1 and HTTP/2 in the image shown below. Explain what is the improvement of HTTP/2 shown in the image. (How does this improvement work?)

E.2 Explain how web applications should be optimised for performance when deployed using HTTP1.1. In comparison to this, how should web applications be optimised when using HTTP/2?

E.3 In web applications, the browser can request information from the web server without reloading the current page. Explain how XmlHttpRequest (XHR) accomplishes this. How can XHR be used for getting server-initiated updates? How does long-polling improve scalability issues with the latter?

Rules of Thumb Queueing

Performance of Networked Systems

Processor Sharing model:

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

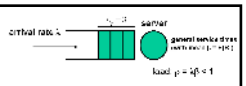
Insensitivity: result holds for any service-time distribution

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