AALBORG UNIVERSITY

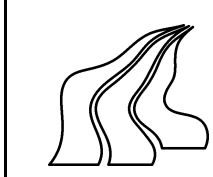
A processor sharing model for wireless data communication

by

Martin Bøgsted Hansen

March 2004 R-2004-06

Department of Mathematical Sciences
Aalborg University
Fredrik Bajers Vej 7 G ▪ DK - 9220 Aalborg Øst ▪ Denmark
Phone: +45 96 35 80 80 ▪ Telefax: +45 98 15 81 29
URL: www.math.auc.dk/research/reports/reports.htm



ISSN 1399–2503 ▪ On-line version ISSN 1601–7811

A processor sharing model for wireless data communication

Martin Bøgsted Hansen[1]
Department of Mathematical Sciences
Aalborg University

Abstract

The use of loss models has a long history in the analysis of telecommunication networks, see e.g. the seminal book by Ross (1995). Typically, one model jobs that arrive at a loss station, with c servers and request access to k servers according to a Poisson process with rate λ and each job occupies these servers for an exponentially distributed holding time with mean $1/(k\mu)$. However, in lack of requested resources some Time Division Multiple Access (TDMA) implementations for mobile data communication like High Speed Circuit Switched Data (HSCSD) and General Packet Radio Service (GPRS) allow already established resources for data connections to be downgraded to allow a new connection to be established. As noted by Litjens and Boucherie (2002) this resembles classical processor sharing models, and in this spirit we formulate a variant of the processor sharing model with a limited and unevenly distributed number of allocated resources. The model is illustrated on a typical HSCSD setup. Performance characteristics, such as blocking probabilities, utilization, average allocated bandwidth, sojourn- and response times are studied. The maximum likelihood principle is suggested for inferential purposes.

Keywords: Processor sharing models, performance characteristics, statistical inference.

1 Introduction

Many wireless communication systems, like GSM, shares the radio spectrum resource by performing frequency division multiple access (FDMA) and time division multiple access (TDMA). FDMA divides the spectrum into a number of carrier frequencies. A certain number of frequencies are allocated to the

[1]Postal address: Fredrik Bajers Vej 7G, 9220 Aalborg East, Denmark; Email: mbh@math.aau.dk; Phone: +45 9635 8801; Fax: +45 9815 8129.

base station of a call. Each of these frequencies are further divided into time slots. When a mobile station wants to download, say, from the Internet it syncronises its time with the base station and negotiate what frequency and time slots that can be used for the data transmission. As an example consider a TDMA frame consisting of $c = 7$ slots and a mobile station which requests 4 time slots for downlink. In the present paper we assume that the base station can pack resources such that the time slots are consecutive. If we assume the transition rate for one time slot is x kbit/s the maximal bandwidth for the communication described above is $4x$ kbit/s. Both HSCSD and GPRS can dynamically downgrade the number of allocated time slots to allow new jobs to enter the system or to give a more fair share of resources upon a service completion. A direct physical realization can be examined by the so-called Round-Robin model. Each job which arrives at the system enter an ordered queue of length c, if there are less than c jobs at service, otherwise it is lost. When the job arrives at the service points it is allocated a multiple of, the length of the time slot q, service time. If the job completes within this time it simply leaves. If it requires more service it is immediately returned to the queue. When q is small compared to the service requests one can approximate the system by a processor sharing system, where each job in the system is allocated a service rate which is a multiple of μ, see e.g. Schassberger (1984).

In the present paper we will treat 2 different policies for allocating the resources. Assume that c resources are available and each of n jobs request simultaneously k resources. We then allocate $\min(k, [c/n])$ resources to each job, in an egalitarian way. However, when $k > [c/n]$ the question is what we shall do with the c mod n remaining resources. In the first in first allocate policy (FIFA) we allocate one extra residual resource to the c mod n jobs with longest sojourn time (time spent in the system). In the last in first allocate (LIFA) we do the opposite and allocate one extra resource to the c mod n jobs with shortest sojourn time.

The model to be considered is a variant of the model analysed in Litjens and Boucherie (2002). The present paper does not include a traffic mix of voice and data calls and a possible queueing scenarios of the data calls. Making this restriction, the model is easy to formulate and derive results for. Moreover, it turns out it is possible to derive a rather explicit expression for the likelihood function for inferential purposes. This is an important aspect, as data collection on cell level in the base station is traditionally restricted to simple overall counters. Moreover, the present paper presents an extension of the blocking probability concept and two alternative allocation policies, which could be compared to the fair channel allocation policy of

Litjens and Boucherie (2002). Consequently, the present paper could be considered as a stepstone between purely simulation oriented papers and the more analytically complex model of Litjens and Boucherie (2002).

The purpose of the paper is to provide an initial set of tools which mimics the standard tools for single slot voice and data traffic. In Section 2, we formalize a model, via a Markovian birth and death process, for the system. Thereafter we study in Section 3 in detail the performance of the system by blocking probabilities, average bandwidth allocation, utilization, sojourn time and mean response time (conditional mean sojourn time given the service request). A possible way of infering the parameters of the model is sketched in Section 4. In Section 5 we study numerically the performance of a typical HSCSD set up with 7 time slots and mobile stations requesting 4 time slots. Finally, Section 6 concludes the paper with a number of suggestions for future work.

2 Modeling issues

Jobs arrive at the station according to a Poisson process with rate λ, requiring an exponentially distributed amount of service with mean $1/\mu$ and access to service rate $k\mu$ if possible. There is room for c jobs at the system. Resources are shared in an egalitarian way. As resources are shared the rate of service received fluctuates with time and its sojourn time depends not only on the jobs in the processor at its arrival, but also on subsequent arrivals. The model is formalized in the following way.

Let $\gamma_n = (\gamma_{1,1}, \ldots, \gamma_{n,n})$ be defined for $n = 1, 2, \ldots, c$ in the following way

$$\gamma_{n,j} = \begin{cases} k & n \leq c/k \\ [n/c] + 1 & n > c/k, j \leq c \bmod n \\ [c/n] & \text{otherwise} \end{cases} \tag{1}$$

and

$$\Gamma_n = \sum_{j=1}^{n} \gamma_{n,j} = \min\{nk, c\} \tag{2}$$

$$\Gamma_{n,<j} = \sum_{i=1}^{j-1} \gamma_{n,i} \tag{3}$$

$$\Gamma_{n,>j} = \sum_{i=j+1}^{n} \gamma_{n,i}. \tag{4}$$

Let X_t denote the number of jobs at the system at time t. To model the system we let X_t denote a Markov process with state space $E = \{0, 1, \ldots, c\}$ and infinitesimal generator given by

$$q_{i,j} = \begin{cases} \Gamma_i \mu & j = i - 1 \\ \lambda & j = i + 1 \\ 1 - \lambda 1_{\{i<c\}} - \Gamma_i \mu 1_{\{i>0\}} & i = j. \end{cases} \tag{5}$$

Then the steady state probabilities $\pi_i = P(X_t = i)$ for $i = 0, 1, 2, \ldots, c$ are given by

$$\pi_i = \rho^i \prod_{n=1}^{i} \Gamma_n / \sum_{l=0}^{c} \rho^l \prod_{k=1}^{l} \Gamma_k$$

where $\rho = \lambda/\mu$.

Assume, at a given point in time there are n jobs $x_1, \ldots, x_n$ in the system and that the they are sorted increasing according to their arrival times in the system. In classical processor sharing systems with a limited number of servers, see e.g. Heyman *et al.* (2003), resources are allocated in an egalitarian way so that each job is serviced at rate Γ_n/n. However, in the present paper we allow resources to be downgraded with an integral number and to have an unevenly distributed number of resources. E.g., we say the allocation policy is FIFA if the i'th job is serviced with rate $\gamma_{n,i}$, and on the contrary if the i'th job is serviced with rate $\gamma_{n,n-i+1}$ we say the allocation policy is LIFA.

3 Performance characteristics

The mathematical model for downgradable wireless data connections as described above helped us in clarifying the basic features of the loss system. From a practical point of view interest now gather around evaluating the performance of a given system in order to assessing effects of change. The first step is to define appropriate performance characteristics, which reflects the end users quality perception.

The performance of mobile data networks can be measured using multiple different key characteristics. The focus of the present paper is to study the loss system in its own right, so we will not touch upon specific radio characteristics. Traditionally, end users perception is described by blocking probabilities, throughput and delay.

3.1 Utilization, blocking probabilities and throughput

A measure of how much the system resources are being utilized is given by the *utilization*, which measures the average number of time slots which are loaded with data services. Network with utilization close to 100% are close to saturation and user performance is likely to be very poor. Utilization is defined as

$$U = \frac{1}{c} \sum_{n=0}^{c} \Gamma_n \pi_n. \tag{6}$$

The *blocking probability* shows how close the system is to saturation. When the blocking limit is reached jobs have to be dropped. In the analysis of loss systems, blocking probabilities are an important performance indicator. In the following we shall extend that concept to multislot access systems. In steady state let B_i, where $i = 1, \ldots, k$ denote the probability for obtaining less than i servers, when requesting access to k servers. More formally,

$$B_i = \pi_c + \sum_{n=0}^{c-1} 1_{\{\gamma_{n+1,n+1} < i\}} \, \pi_n. \tag{7}$$

Obviously, $B_1 \leq B_2 \leq \cdots \leq B_k$ and B_1 is the loss or blocking probability for the whole system.

Throughput in data communication systems is measured as the amount of data delivered per unit of time. In the present system this is measured by the number of servers each user is allocated. As resources are shared the number of allocated slots fluctuates with time depending on the number of jobs at service upon arrival and subsequent arrivals. As an indication of the system throughput allocated to each job one can calculate the average number of allocated servers to each job at service. The *average allocated bandwidth* at a given point in time is defined by

$$AB = \sum_{n=0}^{c} \frac{\Gamma_n \pi_n}{n} \tag{8}$$

3.2 Delay times

In the present paper we shall refer to delay as the excess time a given service request is amenable to, due to the traffic load.

As a basis for this consider the sojourn time, i.e. th time spent in the system. In the following we will derive the sojourn time distribution and the conditional expectation of the sojourn time given the requested service time,

also termed *the mean response time*, which can been interpreted as the delay time! Consequently, if we let U denote the random amount of service a job arriving in equilibrium requests and let the W denote the sojourn time, we are interested in the distribution of W and $E[W|U]$.

The theory is mainly inspired by Asmussen (2003, pp. 111-113) and Coffman *et al.* (1970) who derived similar characteristics for the classical egalitarian M/M/1 processor sharing queue.

3.2.1 The sojourn time distribution

In classical processor sharing theory we say a job is of type n if it meets n other jobs in the system upon arrival, this occurs with probability π_n.

In order to calculate the sojourn time distribution we extend the type n job concept to a type n, j job. Assume, a newly arrived job in steady state is placed as the j'th arrived job and assume that there are n jobs upon arrival, then $H_{n,j}(y)$ denote the probability that the sojourn time strictly exceeds y. The following result yields a recursive algorithm for obtaining the steady state distribution of the sojourn time.

Theorem 1 *Let $\tilde{\pi}_n = \pi_n / \sum_{n=0}^{c-1} \pi_n$ for $n = 0, \ldots, c-1$ be the steady state probabilities for X_t restricted to $n = 0, \ldots, c-1$ and $h^{(k)} = \sum_{n=0}^{c-1} \tilde{\pi}_n h_{n,n}^{(k)}$, then the distribution of the sojourn time equals*

$$P(W > y) = \sum_{k=0}^{\infty} h^{(k)} \frac{y^k}{k!}.$$

Where the numbers $h_{n,j}^{(k)}$ for $n = 0, \ldots, c-1$ and $j = 0, \ldots, n$ are recursively defined by

$$\begin{aligned} h_{n,j}^{(k)} &= \mu(\Gamma_{n+1,<j+1} h_{n-1,j-1}^{(k-1)} + \Gamma_{n+1,>j+1} h_{n-1,j}^{(k-1)}) 1_{\{n>0\}} \\ &- (\lambda 1_{\{n<c-1\}} + \mu \Gamma_{n+1}) h_{n,j}^{(k-1)} \\ &+ \lambda 1_{\{n<c-1\}} h_{n+1,j}^{(k-1)} \qquad (9) \\ h_{n,j}^{(0)} &= H_{n,j}(0) = 0. \end{aligned}$$

Proof We use standard techniques of birth-death processes, as described by e.g. Feller (1971). For an arriving job of type n, j, where $n < c-1$, the probability of another arrival during the next infinitesimal small time interval

h, neglecting terms of the order h^2, is λh. If $n = c-1$ a new arrival is lost. The tail probability is under this condition $H_{n+1,j}(y)$, as the position of j is preserved. On the contrary if a job leaves the system in the interval, we have to split into two situations, a) the job is older than the j'th job and b) the job is younger than the j'th job. For a) the probability is $\mu h \Gamma_{n+1,<j+1}$ and the position of j is decreased by 1, whereby the tail probability becomes $H_{n-1,j-1}(y)$. In situation b) the probability is $\mu h \Gamma_{n+1,>j+1}$ and the position of j is preserved. The tail probability then becomes $H_{n-1,j}(y)$. Note that a death can only occur whenever $n > 0$. The probability of no occurrence at all is $1 - \lambda h 1_{\{n<c-1\}} - \mu h \Gamma_{n+1}$, again with the condition that no birth is allowed whenever $n < c-1$. Hence, we may write

$$\begin{aligned} H_{n,j}(y+h) &= (1 - \lambda h 1_{\{n<c-1\}} - \mu h \Gamma_{n+1}) H_{n,j}(y) \\ &+ \lambda h H_{n+1,j}(y) 1_{\{n<c-1\}} \\ &+ \mu h (\Gamma_{n+1,<j+1} H_{n-1,j-1}(y) + \Gamma_{n+1,>j+1} H_{n-1,j}(y)) 1_{\{n>0\}}. \end{aligned}$$

If we properly rearrange terms and take the limit $h \to 0$, we obtain

$$\begin{aligned} H^{(1)}_{n,j}(y) &= \mu (\Gamma_{n+1,<j+1} H_{n-1,j-1}(y) + \Gamma_{n+1,>j+1} H_{n-1,j}(y)) 1_{\{n>0\}} \\ &- (\lambda 1_{\{n<c-1\}} + \mu \Gamma_{n+1}) H_{n,j}(y) \\ &+ \lambda H_{n+1,j}(y) 1_{\{n<c-1\}}. \end{aligned} \tag{10}$$

By induction we see that $H_{n,j} \in C^\infty[0,\infty)$ and $|H^{(k)}_{n,j}(y)| \leq (2\lambda + 2c\mu)^k$. Henceforth, the following Taylor series expansion holds

$$H_{n,j}(y) = \sum_{k=0}^{\infty} H^{(k)}_{n,j}(0) \frac{y^k}{n!}.$$

Using the obvious fact that $H_{n,j}(0) = 0$ and defining $h^{(k)}_{n,j} = H^{(k)}_{n,j}(0)$, we obtain by conditioning on the type of the job and the job is not lost at the system together with an interchange of a summation order, the result follows:

$$\begin{aligned} P(W > y) &= \sum_{n=0}^{c-1} H_n(y) \tilde{\pi}_n \\ &= \sum_{n=0}^{c-l} \sum_{k=0}^{\infty} h^{(k)}_{n,j} \frac{y^k}{n} \tilde{\pi}_n \\ &= \sum_{k=0}^{\infty} h^{(k)} \frac{y^k}{k!}. \end{aligned}$$

□

3.2.2 The response time

Let $m(u) = E[W|U = u]$ be the mean response time, i.e. the time spend in the system for a particular job which is not lost, arrive in steady state and requires u amount of service upon arrival. The following result states how the response time can be recursively determined under the FIFA and LIFA policies. Let the numbers $g_{n,j}^{(k)}$ for $n = 0, \ldots, c-1$ and $j = 0, \ldots, n$ be recursively defined by

$$\begin{aligned} g_{n,j}^{(k)} &= \frac{1}{\gamma_{n+1,j+1}}(1_{\{k=1\}} \\ &+ \mu(\Gamma_{n+1,<j+1} g_{n-1,j-1}^{(k-1)} + \Gamma_{n+1,>j+1} g_{n-1,j}^{(k-1)}) 1_{\{n>0\}} \\ &- (\lambda 1_{\{n<c-1\}} + \mu(\Gamma_{n+1} - \gamma_{n+1,j+1})) g_{n,j}^{(k-1)} \\ &+ \lambda 1_{\{n<c-1\}} g_{n+1,j}^{(k-1)} \qquad (11) \\ g_{n,j}^{(0)} &= m_{n,j}(0) = 0 \end{aligned}$$

Theorem 2 *Let $\tilde{\pi}_n = \pi_n / \sum_{n=0}^{c-1} \pi_n$ for $n = 0, \ldots, c-1$, be the steady state probabilities for X_t restricted to $n = 0, \ldots, c-1$ and $g^{(k)} = \sum_0^{c-1} \tilde{\pi}_n g_{n,n}^{(k)}$, then the response time equals*

$$m(u) = \sum_{k=0}^{\infty} g^{(k)} \frac{u^k}{k!}.$$

Proof By $m_{n,j}(u)$ we understand the conditional expectation of the sojourn time for a type n job with service time u. Proceeding in the same way as in the proof of Theorem 1, we can up to order h^2 split into a death, birth or no occurrence in the infinitesimal interval. However, in the interval the job has received the amount $h\gamma_{n+1,j}$ of service and the responsetime is increased by h. Therefore we can write

$$\begin{aligned} m_{n,j}(u) &= h \\ &+ (1 - \lambda h 1_{\{n<c-1\}} - \mu h(\Gamma_{n+1} - \gamma_{n+1,j+1})) m_{n,j}(u - h\gamma_{n+1,j+1}) \\ &+ \lambda h m_{n+1,j}(u) 1_{\{n<c-1\}} \\ &+ \mu h(\Gamma_{n+1,<j+1} m_{n-1,j-1}(u) + \Gamma_{n+1,>j+1} m_{n-1,j}(u)) 1_{\{n>0\}}. \end{aligned}$$

Rearranging, taking limits as in the proof of Theorem 1 and using the fact that $m_{n,j}(0) = 0$, we arrive at the stated result. □

4 Statistical analysis using an embedded Markov chain

Extensive statistical analyses of wired data communication systems have recently caused interest in the study of heavy tailed distribution functions of service requests, see e.g. Mah (1997) and Crovella and Bestavros (1997). Similar analyses of wireless data services are still in its infacy. One of the problems faced is that data collection in the base stations is based on simple overall counters. Interest then center around what information that can be inferred from data. In the following we show how the maximum likelihood principle can be used to infer characteristics of the traffic data from a simple overall counter.

Embedded Markov chains facilitates the study of the underlying queueing systems as a Markov chain. Let Q_n be the number of jobs in the system immediately after the n'th departure in the queueing system described in Section 2. It can be shown that $\{Q_n, n = 0, 1, 2, \ldots\}$ is a Markov chain, by the same procedure as used for the M/G/1 queue in Asmussen (2003, p. 281). let p_{ij} be the transition probability $P(Q_{n+1} = j | Q_n = i)$. Consider e.g. the problem of estimating the parameter $\rho = \lambda/\mu$. Assume we observe the process until the total number of departured jobs reach a fixed value N. As one notices that the service times are state dependent one can use the results of Goyal and Harris (1972) to obtain a rather explicit expression for the likelihood function:

$$
\begin{aligned}
L(\rho) &= P(Q_0 = i_0) \prod_{k=1}^{N} P(Q_k = i_k | Q_{k-1} = i_{k-1}) \\
&= \pi_{i_0} \prod_{k=1}^{N} \int_0^\infty \exp(-\lambda t) \frac{(\lambda t)^{i_k - i_{k-1}}}{(i_k - i_{k-1})!} \Gamma_{i_k} \mu \exp(-\Gamma_{i_k} \mu t) dt \\
&= \pi_{i_0} \prod_{k=1}^{N} \frac{\Gamma_{i_k} \mu}{\lambda + \Gamma_{i_k} \mu} \left(\frac{\lambda}{\lambda + \Gamma_{i_k} \mu} \right)^{i_k - i_{k-1}} \qquad (12) \\
&= \left(\frac{\rho^{i_0} \prod_{n=1}^{i_0} \Gamma_n}{\sum_{l=0}^{c} \rho^l \prod_{k=1}^{l} \Gamma_k} \right) \prod_{k=1}^{N} \frac{\Gamma_{i_k}}{\rho + \Gamma_{i_k}} \left(\frac{1}{1 + \Gamma_{i_k} \rho^{-1}} \right)^{i_k - i_{k-1}} . \qquad (13)
\end{aligned}
$$

Which is seen to be a product of negative binomial probability functions, whence the likelihood function can be maximized in the usual manner to determine the maximum likelihood estimates. Following Billingsley (1974) one can determine the precision of the derived estimates and the Fisher's information matrix to determine simultaneous confidence regions.

5 An example

In the following we will consider a typical HSCSD set-up, where one frequency is allocated and the time frame is divided into 8 timeslots. One time slot is usually devoted to signaling and modern terminals on the market typically requests 1 to 4 timeslots for downlink. We therefore consider an example where $c = 7$ and k is varied between 1 and 4. Data transfer requests arrive according to a Poisson process with arrival rate λ and the service requests are exponentially distributed with mean $1/\mu$.

5.1 Utilization

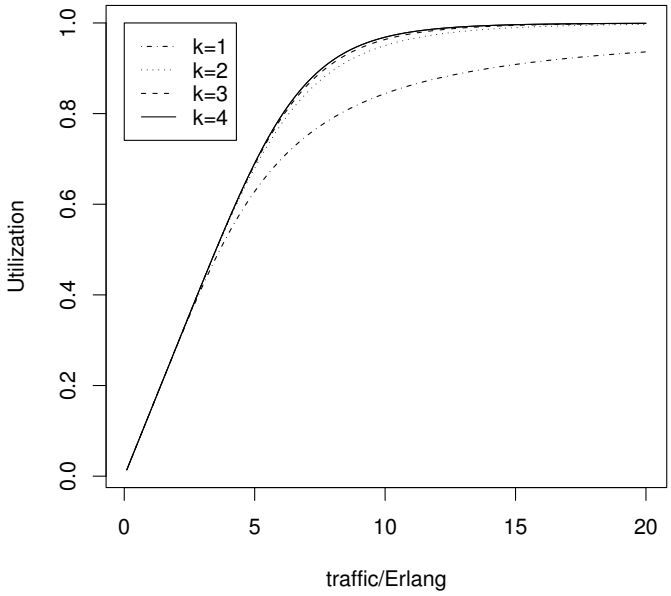


Figure 1: Utilization for request types $k = 1, 2, 3, 4$ are plotted as a function of the traffic load $\rho = \lambda/\mu$.

A useful way to measure the load is to measure the average amount of timeslots utilized by the jobs, as defined above. It is straightforward to see that for fixed traffic load $\rho = \lambda/\mu$ the utilization will increase as a function of requested number of timeslots k. In Figure 1 utilization is plotted as a function of the traffic load for various request types. In general we see the obvious

fact that increasing traffic yields an increasing utilization and increasing the request type yields a higher utilization. In general one spots a nonlinear relationship between request type and utilization.

5.2 Blocking probabilities

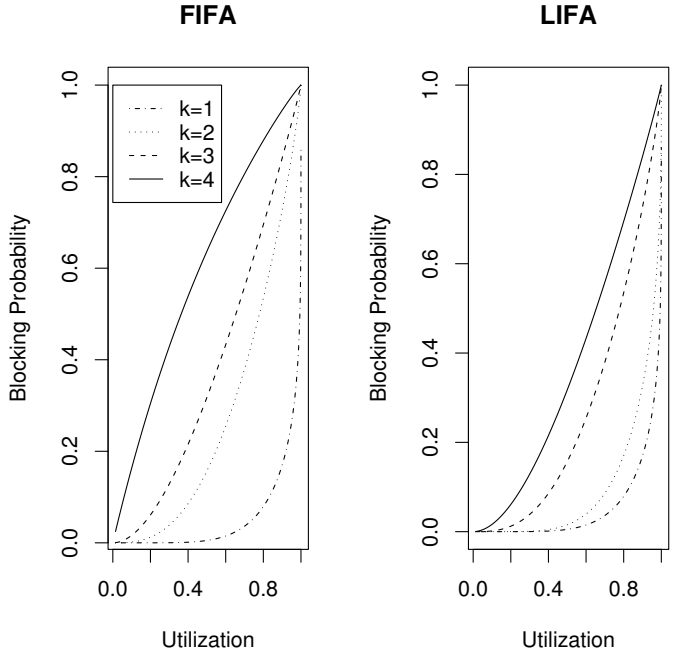


Figure 2: The blocking probabilities B_i are plotted for $k = 4$ as function of the utilization.

In order to study blocking probabilities under various system load situations, we plot them as a function of utilization. In Figure 2 they are plotted for the FIFA and LIFA allocation policies. Naturally, we see that the blocking probabilities are higher for the FIFA policy. Furthermore, we see that B_1 is first severely affected around a 50% utilization but both 2,3 and 4 slots blocking are levering of much earlier, indicating that ambitious service levels for higher bandwidth allocations can be hard to meet for HSCSD data.

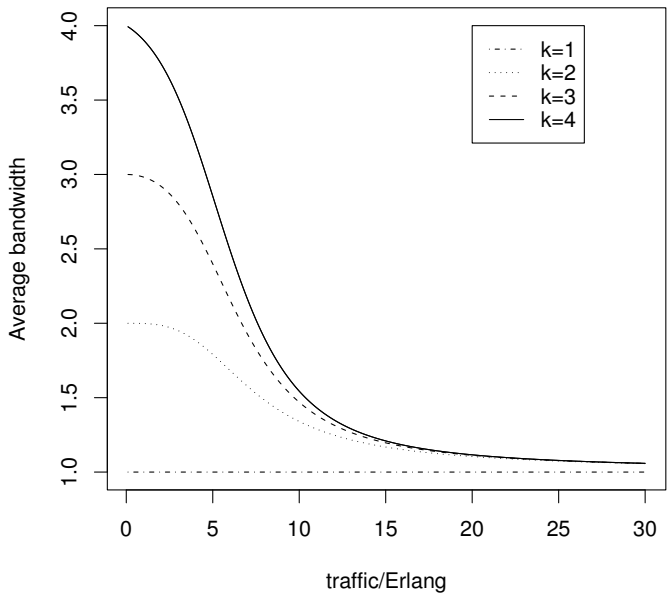


Figure 3: Average bandwidth for different request types $k = 1, 2, 3, 4$ are plotted as a function of the traffic load $\rho = \lambda/\mu$.

5.3 Average bandwidth

As the load increases in wireless applications there are two main reasons for throughput reductions - increasing interference and downgrading available timeslots. The latter effect is illustrated in Figure 3. For increasing load the average allocated bandwidth is illustrated for different request types, $k = 1, \ldots, 4$. For low traffic we see that each connection on average get the requested number of resources, but as the load increases the average number of allocated slots decrease quickly to one time slot.

5.4 Sojourn time

5.4.1 The sojourn time distribution

We have assumed that the distribution of the requested service is exponential with mean $1/\mu$. It is now of interested to see how the varying bandwidth allocation impact the actually time spent in the system, i.e. the sojourn time. This measure describes the lack of throughput the user experiences.

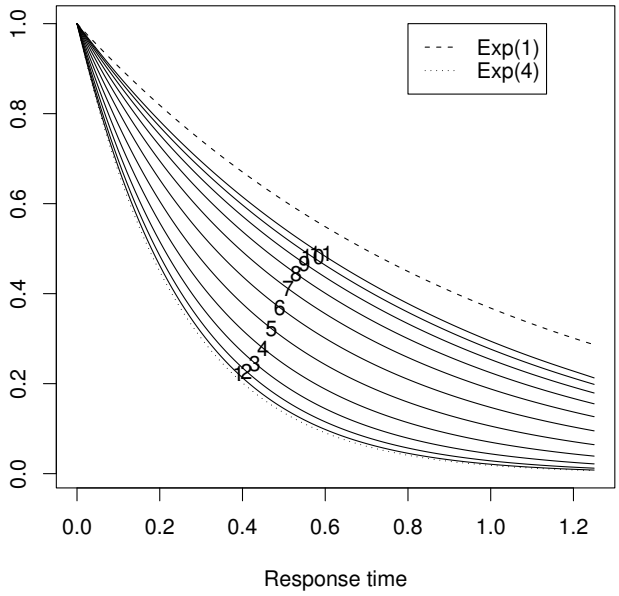


Figure 4: The tail probabilities for the sojourn time distribution is plotted for fixed $\mu = 1$ and increasing $\lambda = 1, \ldots, 11$. Tail probabilities are plotted for the exponential distribution with mean 1 and 1/4, respectively.

If there is small load on the system one would expect the sojourn time to be nearly exponentially distributed with mean $1/(k\mu)$ and on the contrary under heavy load, only one timeslot would be allocated, and consequently one would expect that the sojourn time for a non lost job is nearly exponentially distributed with mean $1/\mu$. In Figure 4 the expected minimum and maximal tail probabilities are plotted. For a constant $\mu = 1$, an increasing load is constructed by increasing the input rate λ from 1 to 11. The expected behavior is confirmed - in the light traffic case we see that the performance is nearly optimal and in the heavy traffic case the the performance i nearly minimal.

5.4.2 The response time

A more operational useful tool is to use the mean responsetime, which gives the expected responsetime for a given service request. In light traffic we would expect the jobs to be served straightaway with service rates $1/(k\mu)$

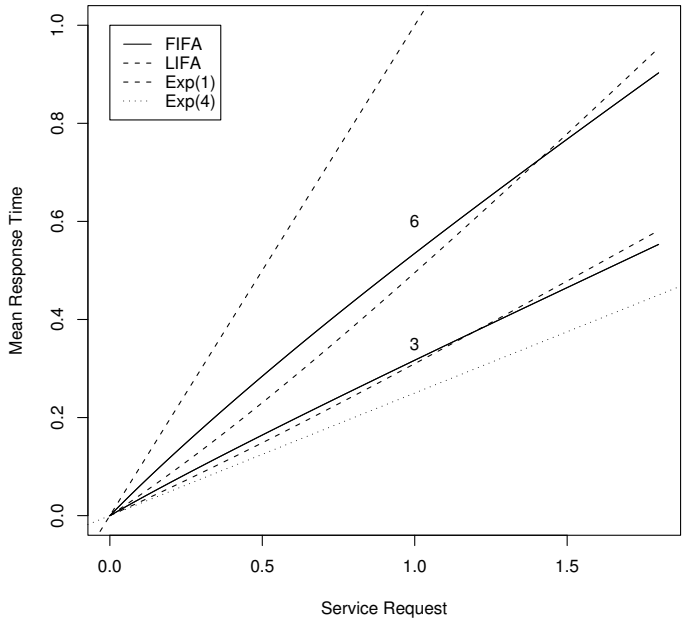


Figure 5: Mean response time plotted for a fixed value of $\mu = 1$ and $\lambda = 3, 6$. The FIFO policy is illustrated with the fulldrawn line and LIFA as the dashed line. Theoretical maximal and minimal performance are given by the dashed and dotted line, respectively.

and in heavy traffic all non-lost jobss are allocated only one time slot and we would expect the service rates to be $1/\mu$. This means we would expect the mean response time in the light and heavy traffic case to be proportional to the service requests with factor $1/k$ and 1, respectively. In Figure 5 the expected light and heavy traffic behaviour is depicted. Furthermore, as the fulldrawn line one sees the responsetime for the FIFA policy under mean service request $\mu = 1$ and input rates 3 and 6, respectively. The same is shown for the LIFA policy in red. One notices that for short service requests the LIFA policy performs better, but for larger policies the FIFA policy performs better. The intuition behind this is that for very short service requests, the LIFA policy allocates a maximal number of resources and short service requests are therefore expedited quickly. On the hand, when service requests are longer, they stay longer in the system and the LIFA policy at some point give them a lower service rate than the FIFA policy would have given them.

6 Open problems

The present paper gave a framework for analyzing downgradable processor sharing systems, without resolving to simulations. But in order to be applicable in more complex systems with composed traffic, such as the coexistence of voice calls, HSCSD as well as GPRS calls one has to extent the present theory to the multiservice situation, as done in Litjens and Boucherie (2002). Furthermore, extensions to more general queuing systems like the GI/G/1 PS springs in mind. The combination of an analytically tractable model with a theory for access control (Alanyali, 1999), could be very beneficial when one consider choosing between various policies from a revenue increasing point of view. The focus on the behaviour of the bandwidth of the single job should be combined with the packet switching properties of GPRS (Foh *et al.*, 2001) and make a more realistic calculations of end-to-end QoS with respect to throughput, jitter and delay. Lastly, other vendor specific counters could be brought into play via maximum likelihood estimation. The statistical considerations should be brougt in action on real network elements in order to derive partly observed traffic characteristics.

References

Alanyali, M. (1999) Three management policies for a resource with partition constraints. *Adv. Appl. Prob.*, **31**, 795–818.

Asmussen, S. (2003) *Applied Probability and Queues*. 2nd edn. Application of Mathematics, Springer Verlag.

Billingsley, P. (1974) *Statistical Inference for Markov Processes*. Chicago: University of Chicago Press.

Coffman, E. G., Muntz, R. R. and Trotter, H. (1970) Waiting time distribution for processor-sharing systems. *J. of ACM*, **17**, 123–130.

Crovella, M. E. and Bestavros, A. (1997) Self-similarity in World Wide Web traffic: Evidence and possible causes. *IEEE/ACM Transaction on Networking*, **5**, 835–846.

Feller, W. (1971) *An Introduction to Probability Theory and Its Applications*, vol. II. 2nd edn. New York: John Wiley and Sons.

Foh, C. H., Meini, B., Wydrowski, B. and Zukerman, M. (2001) Modelling and performance evaluation of GPRS. In *Proceedings of IEEE VTC*, vol. 16. Rhodes, Greece.

Goyal, T. and Harris, C. (1972) Maximum likelihood estimation for queues with state dependent service. *Sankhya A*, **34**, 65–80.

Heyman, D. P., Lakshman, T. V. and Neidhardt, A. L. (2003) A new method for analyzing feedback-based protocols with applications to engineering web traffic over the Internet. *Computer Communications*, **26**, 785–803.

Litjens, R. and Boucherie, R. (2002) Performance analysis of fair channel sharing policies in an integrated cellular voice/data network. *Telecommunication Systems*, **19**, 147–186.

Mah, B. A. (1997) An empirical model of HTTP network traffic. In *Proceedings of INFOCOM '97*. Kobe, Japan.

Ross, K. W. (1995) *Multiservice Loss Models for Broadband Telecommunication Networks*. Telecommunication Networks and Computer Systems, Springer Verlag.

Schassberger, R. (1984) A new approach to the M/G/1 processor-sharing queue. *Adv. Appl. Prob.*, **16**, 202–213.