Faculty of Mathematical Sciences

University of Twente
The Netherlands

P.O. Box 217
7500 AE Enschede
The Netherlands
Phone: +31-53-4893400
Fax: +31-53-4893114
Email: memo@math.utwente.nl
www.math.utwente.nl/publications

Memorandum No. 1619

Transient handover blocking probabilities in road covering cellular mobile networks

R.J. Boucherie and J. van der Wal[1]

March 2002

ISSN 0169-2690

[1]Universiteit van Amsterdam, Department of Operations Research, Roetersstraat 11, NL-1018 WB Amsterdam and Eindhoven University of Technology, Faculty of Mathematics and Computer Science, P.O. Box 513, NL-5600 MB Eindhoven

Transient handover blocking probabilities in road covering cellular mobile networks

Richard J. Boucherie*
University of Twente, Faculty of Mathematical Sciences,
P.O. Box 217, NL-7500 AE Enschede

Jan van der Wal
Universiteit van Amsterdam, department of Operations Research,
Roetersstraat 11, NL-1018 WB Amsterdam
and
Eindhoven University of Technology, Faculty of Mathematics and Computer Science,
P.O. Box 513, NL-5600 MB Eindhoven

22nd March 2002

Abstract

This paper investigates handover and fresh call blocking probabilities for subscribers moving along a road in a traffic jam passing through consecutive cells of a wireless network. It is observed and theoretically motivated that the handover blocking probabilities show a sharp peak in the initial part of a traffic jam roughly at the moment when the traffic jam starts covering a new cell. The theoretical motivation relates handover blocking probabilities to blocking probabilities in the M/D/C/C queue with time-varying arrival rates. We provide a numerically efficient recursion for these blocking probabilities.

Keywords: blocking probabilities; cellular networks; tandem of Erlang loss queues; time-dependent behaviour.
AMS Subject Classification. Primary: 90B18, 90B22; secondary: 60K25

*The research of R.J. Boucherie is partly supported by the Technology Foundation STW, applied science division of NWO and the technology programme of the Ministry of Economic Affairs, The Netherlands.

1 Introduction

Current trends in designing and building wireless communications networks include linear wireless networks dedicated to cover calls generated by subscribers traveling along a highway [2]. Such networks can be seen as one-dimensional networks with base stations placed at equal distance, and with equal capacity. Moreover, as the linear network is dedicated to covering traffic along the highway only, the network will consist of small cells with typical length 1000–2000 meters. Obviously, this road covering network must be capable of handling peaks in the load such as occurring during rush hours when a traffic jam (hot spot) is moving along the road. During rush hours, especially in a traffic jam, traffic will move along the road at considerably reduced speed, in the order of 20–40 km/hour. At this speed, it will take a subscriber roughly 3 minutes to travel along a cell, and the increase in traffic density that marks the beginning of a hot spot will be relatively sharp compared to the length of the cell. In view of the above, below we will focus our attention on a linear network (or tandem network) with equal cells along which a traffic jam with a density of almost block-shape travels at constant speed.

In a hot spot, a large number of subscribers might almost simultaneously require a handover between adjacent cells. Moreover, due to the nature of a traffic jam, a subscriber located in a group of cars is very likely to stay within that same group of cars for a long period of time while traveling along multiple cells, so that also its next handover attempt will occur at the same instant as those attempts of the other members of the group. This may lead to a serious degradation of the Quality of Service (QoS), expressed in terms of fresh call and handover blocking probabilities with typical threshold values of 1% and 0.5%, respectively, see [7]. In fact, if the traffic is dense, it may become likely for a call to terminate due to dropping, that is, for some groups of cars the probability that a call is prematurely interrupted due to handover blocking exceeds well beyond the threshold values, and may approach the probability that the call completes due to the end of the conversation. Below we will focus our attention on handover blocking probabilities for subscribers at different locations in a hot spot. As is often the case, these handover blocking probabilities cannot be obtained in closed form from a mathematical model taking into account both mobility and telecommunications aspects. Therefore, we have performed a series of dynamic simulations to investigate these blocking probabilities. *It is observed that a subscriber traveling at roughly the cell-length behind the start of a traffic jam suffers from a considerable increase in handover blocking probability at each handover attempt.*

To explain this observation, notice that the load of the telecommunications network is proportional to the density of subscribers. As a consequence, the number of handovers per time unit is proportional to this density, too. Therefore, when a traffic jam first enters an empty cell, all handovers will be successful. When the traffic jam slowly migrates into the cell, more and more handovers will be blocked. Only when the first part of the traffic jam migrates into the next cell, a substantial share of capacity is released, so subsequent handover attempts will become more successful. From a mathematical perspective, a discretized version of this intuitive justification can be captured in a recursion that indeed closely describes the qualitative behaviour of handover blocking in a

traffic jam. Reducing the discretization step, handover blocking probabilities for a traffic jam moving at constant speed turn out to coincide with fresh call blocking probabilities in an Erlang loss queue with deterministic service requirements, a model that provides an adequate quantitative description of handover blocking probabilities in our wireless network of interest. Unfortunately, analytical results for the resulting Erlang loss queue with deterministic service requirements and time-varying arrival rate are not available. Therefore, we have also performed a series of dynamic simulations to quantify blocking probabilities as function of time. In addition, we have compared these exact results with existing approximations such as the Pointwise Stationary Approximation (PSA), and the Modified Offered Load approximation (MOL), see [3], that uses the functional form of the equilibrium distribution, but with the time-dependent offered load obtained from analysis of a network with unlimited capacity (hence no blocking occurs). These commonly used approximations turn out to be fairly inaccurate for the model under consideration.

Although our study focuses on handover blocking probabilities for hot spots traveling along homogeneous linear networks, our results also describe handover blocking probabilities for homogeneous traffic entering a cell with reduced capacity, e.g., due to frequency failure, or due to the cell being larger. Also in this case, the load arriving to this deviating cell suffers a shock comparable with that of a hot spot, and the handover blocking probabilities for subscribers entering the deviating cell will show a sharp peak that may be well beyond the design standards of the wireless network. Our results clearly indicate that *the best design for road covering cellular networks is that with identical cells*, e.g., equal length cells with equal capacity.

Here is the organisation of this note. Section 2 contains our model, and the simulation study exhibiting the observed sharp peak in handover blocking probabilities. A mathematical simplification, based on a discretized version of the traveling hot spot, that provides intuition for the observed phenomenon is the topic of Section 3. This section also provides the results for the approximation of handover blocking via the Erlang loss queue with time-varying arrival rates. Section 4 provides a discussion of (the impact of) our results and completes this note.

2 Model and main observation

In this section we describe the wireless network that has been simulated and the results obtained from that. The simulation results ask for an explanation, which will be the topic of the rest of the paper.

The wireless network that we consider is linear and covered by equal cells of length L with capacity C per cell, say. To avoid boundary effects, the network is assumed to cover the entire real line $\mathbb{R}$, and cells are assumed to be non-overlapping. Along the network a hot spot of increased subscriber density is traveling at constant speed v. To isolate the effect of handover blocking for the traveling hot spot, the density of subscribers is assumed zero before the hot spot. Moreover, as the hot spot has to model a traffic jam, the density of subscribers inside the hot spot is assumed to increase sharply at the beginning of the hot spot, and then to remain fixed until the end of the hot spot, see Figure 1 for an

illustration of our model, and see [5] for a detailed description of fluid models for road traffic, and [10] for a model incorporating the teletraffic load into this fluid model.

We assume that calls are generated uniformly over the hot spot, proportional to the density of subscribers, an assumption that is justified since the number of subscribers will considerably exceed the number of subscribers making a call. Furthermore, assume that the lengths of all calls generated have generic distribution S, with mean τ. Apart from τ and S, also the speed v of the traffic jam and the length L of the cells are important. It is convenient to express the call duration in cell length units. So a call duration of τ means that, given the speed v (not specified) and the cell length L, a car travels $\tau/(L/v)$ cells during the call. By the nature of our problem, for a traffic jam traveling along a highway, the mean call length will be in the order of the time a subscriber requires to travel along a cell, i.e., τ is of the same order as (L/v). Typical values of τ we consider are $0.2L/v$, L/v and $2L/v$. E.g., for a speed of $v = 30$ km/h, and cell length $L = 2.5$ km, a call of 5 minutes travels exactly one cell length, so $\tau = L/v$.

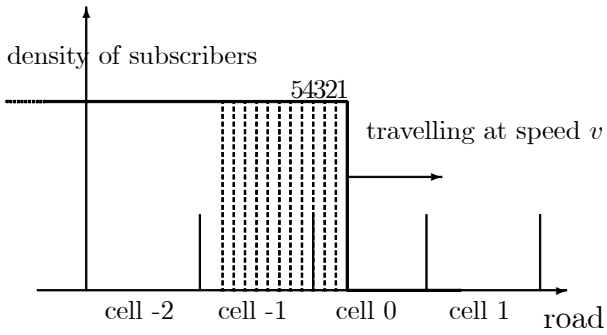


Figure 1: Road divided into cells of equal length; block shaped traffic jam with speed v.

Below, for deterministic and exponentially distributed call lengths, graphs for fresh call and handover blocking probabilities are presented for a block shaped traffic jam, modeled as a constant fresh call arrival rate λ over the cells, see Figure 1. For all experiments, the capacity of each cell is $C = 21$, and where relevant, we have taken $v = 30$ km/h, and $L = 2.5$ km. Blocking probabilities are depicted with respect to the position of the calls in the traffic jam. To this end, the traffic jam is divided into segments, labelled 1, 2, 3, $\cdots$, starting from the head of the traffic jam, and such that 10 segments cover a cell, i.e. a segment has length $0.1L$, see Figure 1. The graphs show the blocking probabilities per segment for various values of the fresh call arrival rate λ, where λ is normalised to correspond to the fresh call arrival rate in a cell that is completely covered by the traffic jam: in the example of Figure 1, where a segment has length $0.1L$, $\lambda = 0.25$ corresponds to an arrival rate of 0.025 calls per second per segment, resulting in an arrival rate of 0.25 calls per second in cell -1 that is covered by 10 segments, and an arrival rate 0.075 calls per second in cell 0 since cell 0 is covered by the first 3 segments of the traffic jam, only.

For our results, we assume that the traffic jam has been travelling along the road for a substantial amount of time, such that blocking probabilities become stationary with respect to the location in the traffic jam. (It turns out, however, that stationarity is already obtained after 2 – 3 cells.) Thus, the handover blocking probabilities depicted in Figures 2 – 5 (indicated with $\bullet$) should be interpreted as the handover blocking probabili-

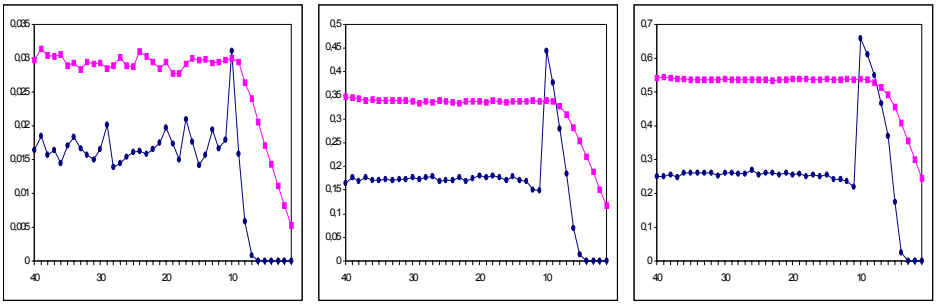
ties experienced by subscribers in a segment each time the segment enters a new cell. The fresh call blocking probability (indicated with ■) should be interpreted as the probability that a fresh call of a subscriber in a segment is blocked. The results of Figures 2 – 5 are obtained from a terminating discrete event simulation, where results are obtained with at least 95% confidence and 10% relative precision. Confidence intervals are not reported in the graphs, the irregular pattern of e.g. the left most graphs reveal part of the uncertainty in our simulation results.

The first series of graphs, Figure 2, considers the situation of deterministic call lengths with $\tau = 0.2L/v = 60\,\text{sec}$, i.e., each call lasts exactly 20% of the cell length. Arrival rates are $\lambda = 0.25/\,\text{sec}$, $\lambda = 0.50/\,\text{sec}$, and $\lambda = 0.75/\,\text{sec}$. For a non-moving traffic jam, in a cell completely covered by the traffic jam, this would result in loads 15, 30, and 45, resulting in fresh call blocking probabilities 3.15%, 35.2%, and 55.0%, respectively (readily obtained from the Erlang loss formula). The second and third series of graphs (Figure 3 and 4, resp.) consider deterministic calls of length $\tau = L/v$ sec, and $\tau = 2L/v$ sec, resp., where the arrival rates are scaled such that the load remains unchanged. The fourth series of graphs (Figure 5) reconsiders the case of Figure 3, but with negative exponentially distributed call lengths. Notice that all graphs, irrespective of the loads, the call lengths, and the distribution of these call lengths, show qualitatively the same behaviour.

As can be observed from these graphs, fresh call blocking (indicated with ■) behaves according to intuition: in the initial part of the traffic jam fresh call blocking probabilities are lower as the corresponding segments often reside in an empty cell. From segment 10 onwards, the fresh call blocking probabilities attain values close to their equilibrium values of 3.15%, 35.2%, and 55.0%. The realised values in the graphs are slightly lower as a fraction of the load is cancelled due to handover blocking.

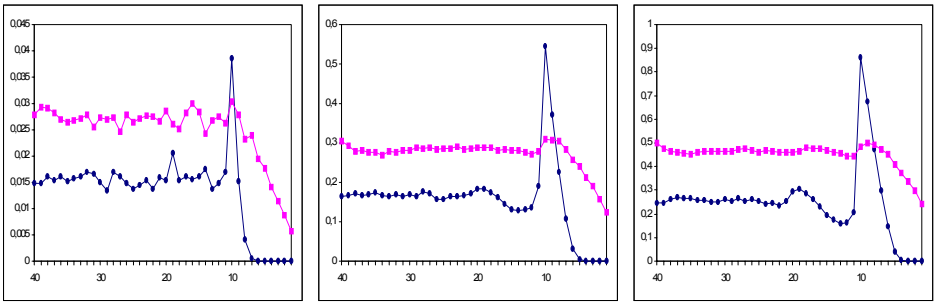
The handover blocking probabilities (indicated with •) show a remarkable peak in segment 10. At first sight, this may seem an anomaly in our simulation study. However, a more detailed investigation of the system reveals that this peak can be intuitively justified via the following argument. The load of the telecommunications network is proportional to the density of subscribers, so that the number of handovers per time unit is proportional to this density, too. When a traffic jam first enters an empty cell, all handovers will be successful. When the traffic jam slowly migrates into the cell, more and more handovers will be blocked. Only when the first part of the traffic jam migrates into the next cell, a substantial share of capacity is released, so subsequent handover attempts will become more successful. Notice that a second (smaller) peak in the handover blocking probabilities occurs at roughly $2L$ behind the start of the traffic jam for networks suffering a high load in the traffic jam. The intuition behind this peak follows the lines of that for the first peak. An additional effect occurs for $\tau = 2L/v$. Here subscribers travelling at $2L$ behind the initial part of the traffic jam will perform 2 handovers, and can therefore be blocked in both attempts. The combined effect results in a increased peak at $2L$.

From a mathematical perspective, a discretized version of this intuitive justification is captured in a recursion in Section 3.1 below.



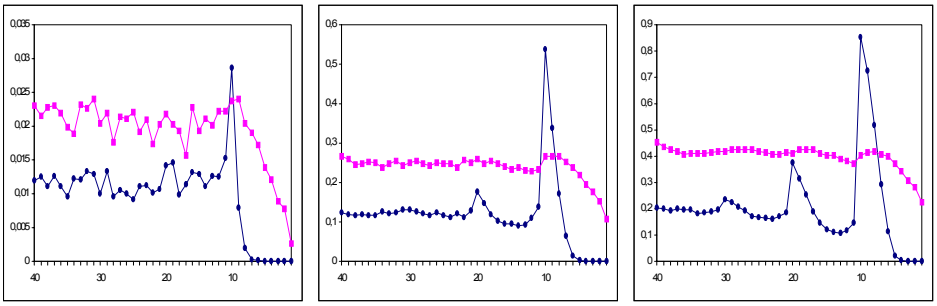
$\lambda = 0.25/\sec$ $\lambda = 0.50/\sec$ $\lambda = 0.75/\sec$

Figure 2: Deterministic call length, $\tau = 0.2L/v$.



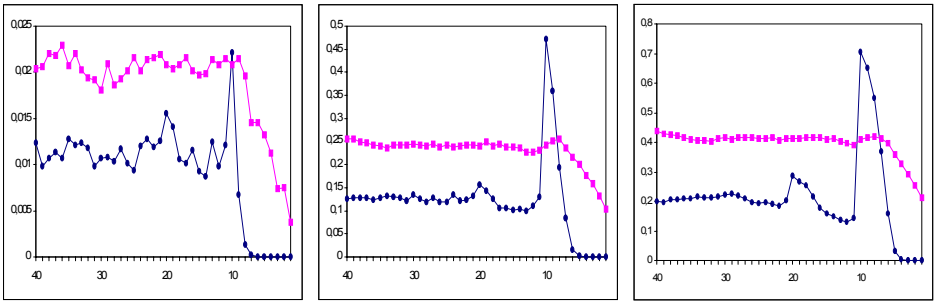
$\lambda = 0.05/\sec$ $\lambda = 0.10/\sec$ $\lambda = 0.15/\sec$

Figure 3: Deterministic call length, $\tau = L/v$.



$\lambda = 0.025/\sec$ $\lambda = 0.05/\sec$ $\lambda = 0.075/\sec$

Figure 4: Deterministic call length, $\tau = 2L/v$.



$\lambda = 0.05/\sec$ $\lambda = 0.10/\sec$ $\lambda = 0.15/\sec$

Figure 5: Exponential call length, $\tau = L/v$.

Remark 2.1 (Design of a road covering network) The peak in the handover blocking probabilities reported in Figures 2–5 is caused by a sharp increase in the teletraffic load coinciding with the initial fase of a hot spot. A similar situation occurs when homogeneous traffic travelling along a road with identical cells (same length L and capacity C) enters a cell with reduced capacity e.g. due to failure of a frequency assigned to that cell (lower C) or due to the cell being longer (larger L). In such cases, the teletraffic load of the deviating cell shows a sharp increase with resulting peak in the handover blocking probabilities. The simulation results clearly indicate that the best design for road covering cellular networks is that with identical cells.

3 Mathematical model

Below, we first present a formal discretized description of the traffic jam, in which subscribers simultaneously move along the road in discrete steps. Then, in Section 3.2, we reduce the discretization step to 0, and show that handover blocking for calls in a traffic jam can be conveniently modelled as fresh call blocking in an Erlang loss queue.

3.1 Recursion

We will consider one initially empty cell. This cell is split into d sub-cells, also called service locations, indexed j, $j = 1, \ldots, d$. These service locations share an amount of C servers that can be used at each of the service locations. We assume that customers arriving from the outside enter the cell at location 1, and move along the locations in upwards direction (location 2, etc.). Customers leave the cell upon completion at location d. A customer arriving to location 1 is accepted for service if a server is available, otherwise the customer is rejected, and cleared from the system.

The correspondence between this model and the model of Section 2 is as follows. Arrivals at location 1 are to be seen as handover attempts to cell 0. When the traffic jam moves along the road, one after the other the segments of the traffic jam enter cell 0 (see Figure 1), and calls in the segments make a handover attempt. Segments reside in cell 0 for $d = 10$ time steps until departure to cell 1. In the simple model of the present section, there are no fresh calls generated and calls are not terminated in this cell. This might seem not very realistic, but as we will argue later, a stationary model with fresh calls and call termination will have similar performance characteristics, see Remark 3.5.

Let X_n denote the number of customers arriving to service location 1 in the time interval $(n-1, n]$, $n = 1, 2, \ldots$ (we will also say at time n). We assume that the X_n are non-negative.

Remark 3.1 In our model, we include the case that X_n is non-stationary to model a non-homogeneous load of calls over the traffic jam. This seems more realistic as more fresh calls are likely to be accepted in the initial part of the traffic jam as this part more often resides in an empty cell. For short calls this effect will not be substantial. In our numerical and theoretical results we will restrict ourselves to the stationary case.

Let Y_n denote the total number of busy servers at time n, due to arrivals in $(0, n]$, and let Z_n denote the number of servers granted to the customers arriving in the time interval $(n-1, n]$. From the description of the system, we know that a customer who is accepted in location 1 will always have a server available at subsequent service locations. As a consequence, the number of servers Z_n granted at time n is released at time $d+n$. The r.v. Y_n and Z_n satisfy a number of recursions. We will distinguish the cases $n \leq d$ and $n > d$.

Case $n \leq d$. We have

$$Y_n = \min\{Y_{n-1} + X_n, C\}, \quad Z_n = Y_n - Y_{n-1}, \quad n \leq d, \tag{1}$$

where $Y_0 \equiv 0$, or equivalently, for Y_n,

$$Y_n = \min\{X_1 + \cdots + X_n, C\}, \quad n \leq d.$$

Case $n > d$. For $n > d$ the recursions are more complicated.

$$Y_n = \min\{Y_{n-1} - Z_{n-d} + X_n, C\}, \quad Z_n = Y_n - Y_{n-1} + Z_{n-d}, \quad n > d. \tag{2}$$

From the recursions (1), (2) it is clear that

$$Y_n = Z_n + Z_{n-1} + \cdots + Z_{n-d+1}, \quad n \geq 1, \tag{3}$$

and

$$Z_n = \min\{X_n, C - Z_{n-d+1} - \ldots - Z_{n-1}\}, \quad n \geq 1, \tag{4}$$

where $Z_n := 0$, for $n \leq 0$.

We are interested in f_n, the *fraction of the arriving customers that is blocked* at time n, that is defined as

$$f_n = 1 - \frac{\mathbb{E}Z_n}{\mathbb{E}X_n}, \quad n = 1, 2, \ldots \tag{5}$$

For the model of Section 2, f_n corresponds to the handover blocking probability for subscribers in the n-th segment of the traffic jam.

Example 3.2 *Let the X_n be iid Poisson distributed with mean α, so that $\sum_{k=1}^{n} X_k$ is Poisson distributed with mean αn. Then for $n \leq d$*

$$\mathbb{P}(Y_n = k) = \begin{cases} \dfrac{(\alpha n)^k}{k!} \mathrm{e}^{-\alpha n} & k < C \\ \displaystyle\sum_{j=C}^{\infty} \frac{(\alpha n)^j}{j!} \mathrm{e}^{-\alpha n} & k = C \end{cases} \tag{6}$$

so that $\mathbb{E}Z_n$ and f_n can be obtained in explicit form.

A comparison of blocking probabilities is possible using the stochastic order relation, $X \geq_{st} Y$, that implies that $\mathbb{E}X \geq \mathbb{E}Y$ (cf. [9] for a list of properties). For stationary $\{X_n\}_{n\geq 1}$, for $n \leq d$ we see from (4), that Z_n is stochastically decreasing so that f_n is increasing. For $n = d+1$, however, the situation becomes more complicated. Using (4),

$$\begin{aligned} Z_{d+1} &= \min\{X_{d+1}, C - Z_2 - \ldots - Z_d\}, \\ Z_d &= \min\{X_d, C - Z_1 - \ldots - Z_{d-1}\}, \end{aligned}$$

and $Z_i \geq_{st} Z_{i+1}$, $i = 1, \ldots, d-1$, a term by term comparison of $C - Z_2 - \ldots - Z_d$ and $C - Z_1 - \ldots - Z_{d-1}$ and invoking the stationarity of $\{X_n\}_{n\geq 1}$, implies that Z_{d+1} is stochastically larger than Z_d. Obviously, Z_{d+1} is stochastically smaller than $Z_1 = \min\{X_1, C\}$. These observations are formalized in the following lemma.

Lemma 3.3 *Assume that $\{X_n\}_{n\geq 1}$ is stationary. Then Z_n is decreasing, for $n = 1, \ldots, d$, and*

$$f_n \geq f_{n-1}, \quad n = 2, \ldots, d.$$

Furthermore, $Z_{d+1} \geq_{st} Z_d$, and $Z_1 \geq_{st} Z_{d+1}$, therefore

$$f_1 \leq f_{d+1} \leq f_d.$$

So the blocking probabilities show a peak at $n = d$.

The ordering of Z_{d+2} and Z_{d+1} depends on the distribution of X_n. To see this, consider a simple example with $C = 1$, $d = 3$, and $\mathbb{P}(X_n = 1) = p = 1 - \mathbb{P}(X_n = 0) = 1 - q$. We are now interested in a comparison of $\mathbb{P}(Z_4 = 1)$ and $\mathbb{P}(Z_5 = 1)$. To this end, one easily verifies that

$$\begin{aligned} \mathbb{P}(Z_1 &= 1) = p, \quad \mathbb{P}(Z_2 = 1) = qp, \quad \mathbb{P}(Z_3 = 1) = q^2 p, \\ \mathbb{P}(Z_4 &= 1) = p(p + q^3), \quad \mathbb{P}(Z_5 = 1) = p(q^4 + 2qp), \end{aligned}$$

and it follows that $\mathbb{P}(Z_4 = 1) \geq \mathbb{P}(Z_5 = 1)$ iff $p \geq (3 - \sqrt{5})/2 \approx 0.38$, i.e., for larger values of the arrival rates we have that $f_{d+1} \leq f_{d+2}$. Similarly, for larger values of the arrival rates we also obtain that $\mathbb{E}Z_{d+1} \geq \mathbb{E}Z_{d-1}$.

In general, it seems that for larger values of the arrival rates, i.e., for larger values of $\mathbb{E}X_n$, we will have $\mathbb{E}Z_{d+1} \geq \mathbb{E}Z_{d+2}$. The argument for the ordering of subsequent r.v's Z_{n+d+1} and Z_{n+d} strongly depends on the value of n and on the value of the arrival rates, as can also be seen from Figures 2–5. Thus, it seems that larger values of the arrival rates result in a sharper peak for the blocking probabilities.

Especially for heavily loaded systems, a second peak may occur at $n = 2d$ that has the same intuitive explanation as the peak at $n = d$. This second peak will be smaller, as it is caused by the truncated process of the first d intervals, that is, the capacity used by arrivals in interval $d + j$ is X_{d+j} truncated to Z_j. Figure 6 presents results of a numerical investigation of the blocking probabilities f_n in the setting of Example 3.1, i.e., under the assumption that X_n is Poisson distributed with mean α for varying load $\alpha = 1.5$, 2.0, 2.5, and 3.0.

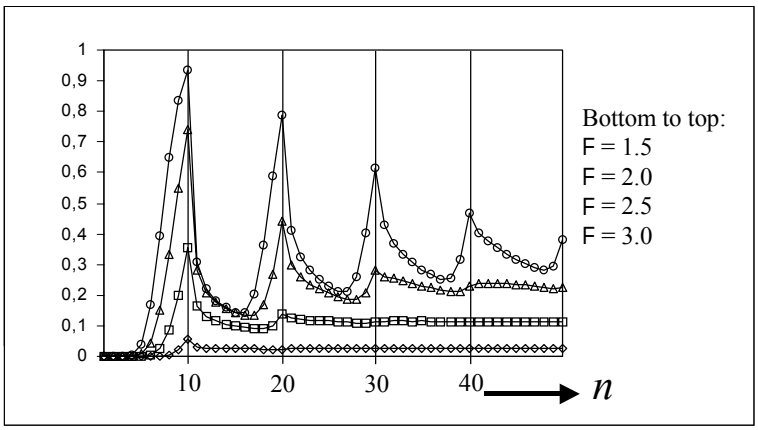


Figure 6: Blocking probabilities f_n for $C = 21$, $d = 10$, and α ranging from 1.5 to 3.0.

Remark 3.4 (Interpretation of blocking probabilities) The blocking probabilities f_n defined in (5) coincide with handover blocking probabilities for the model of Section 2. To see this, assume the traffic jam has been traveling along the road for a time-period long enough for the distribution of the number of calls in the segments to be stationary, that is each time segment n enters a cell, the distribution of the number of calls in the segment is X_n. Then, clearly, f_n coincides with handover blocking probabilities of segment n. Convergence of the distribution of the number of calls in segment n upon handover to its stationary version X_n occurs reasonably fast. In the setting of the graphs of Figure 2, with τ in the order of time a call travels through a cell, convergence was reached within 3 cells.

Remark 3.5 (Call termination and fresh call arrivals) The model of (1), (2) considers calls arriving to the system to service location 1, only. These calls correspond to handovers in the system of Section 2 (see Remark 3.4). Moreover, these calls do not terminate (except through blocking) and continue to travel along the sub-cells. In the network of Section 2, calls are terminated and fresh calls are generated throughout the traffic jam, i.e., there is an additional arrival rate of fresh calls in all segments, and in all segments calls may terminate due to call completion. We may also take these processes into account in the process of (1), (2).

To this end, consider a traffic jam that has been moving for a long time. Then, in each segment the distribution of the number of calls will be in its stationary regime. Assume that the number of subscribers substantially exceeds the number of calls in the segments, so that we may model the additional fresh call arrival process as a Poisson process at rate ν, say, and let A_i be the r.v. for the number of arrivals in segment i. In addition, in each segment a fraction p, say, of the calls will be completed. (Here, for simplicity of the argument, we assume geometrically distributed call lengths with mean $1/p$.) Clearly, if Z_n is the number of accepted calls in the interval $(n-1, n]$, then conditional on $Z_n = k$, the distribution of the number of completed calls in this segment after one time unit, B_n, is Binomial(k, p). For the number of calls in each segment of the traffic jam to have reached its stationary regime, it must be that the effects of A_i and B_n cancel.

Notice that for small blocking probabilities, Z_n closely resembles X_n, i.e., Z_n is close to

Poisson. In that case, $Z_n - B_n$ is a thinning of a Poisson process that is also Poisson, with rate $(1-p)\alpha$. So that for $\nu + (1-p)\alpha = \alpha$ we indeed obtain that $A_i \sim B_n$. In the general case, for a traffic jam that has been moving along a road until the stationary regime has been reached, we will have that the distribution of the number of calls in the segment is such that thinning with probability p, and adding a Poisson(ν) number of fresh calls will not change the distribution. Although the argument above is not rigorous, observing the graphs of Figures 2–5 and Figure 6 indeed shows that qualitatively the model of (1), (2) that ignores fresh calls provides an adequate description of handover blocking in a traffic jam.

Remark 3.6 (Layered networks) Layered networks consisting of macrocells with underlying microcells arise naturally as design option for road covering networks to increase capacity. Typically, to the network of existing macrocells with capacity C a layer of smaller microcells with capacity c is added. Layered networks of this form fit into the framework described above. To see this, consider a tandem of d service locations, indexed j, $j = 1, \ldots, d$. Let service location j be assigned c servers for private use, and let service locations j, $j = 1, \ldots, d$, share an additional amount of C servers that can be used at each of the service locations. Assume that customers arrive from the outside to location 1, and move along the locations in upwards direction. Customers leave the system upon completion at location d. A customer arriving to location j is accepted for service when a server (either from the private pool of c servers, or from the shared pool of C servers) is available at location j, otherwise the customer is rejected, and cleared from the system. Let A_n denote the number of customers arriving to service location 1 in the time interval $(n-1, n]$, $n = 1, 2, \ldots$. With $X_n := [A_n - c]^+$ the number of servers of the pool of C servers requested be these customers we arrive at the model of (1), (2), where Y_n is the total number of occupied servers from the shared pool, and blocking probabilities can be computed from (5).

3.2 Erlang loss queue with deterministic service

The recursion (1), (2) provides an algorithmic description of the load and blocking probabilities for the Erlang loss queue $M|D|C|C$ with Poisson arrivals at rate $\alpha(t)$ depending on the time t and deterministic service, that starts initially empty. To see this, let the service time D be divided into $d(\Delta)$ time steps Δ, such that $\Delta = D/d(\Delta)$. Let Y_n denote the number of calls in the queue at time $n\Delta$, $Y_0 = 0$, and let X_n denote the number of calls arriving in $((n-1)\Delta, n\Delta]$. Ignoring higher order terms, we scale upon the rate at which $\alpha(t)$ changes: we assume that X_n is Poisson distributed with rate $\alpha(n)\Delta$ for all n.

Arrivals occur throughout the interval $((n-1)\Delta, n\Delta]$, but all departures occurring in that interval are deducted at the beginning of the interval, which is commonly referred to as *'departures before arrivals'*, or late arrivals. As a result an accepted call stays in the system a little bit shorter than its service requirement D. As a consequence, for $n \leq d(\Delta)$ calls are accepted only if $Y_{n-1} + X_n \leq C$. Otherwise, only $Z_n = C - Y_{n-1}$ of these calls can be accepted. These calls will depart from the system upon completion of their service requirement D, that occurs at time step $n + d(\Delta)$. For $n \geq d(\Delta)$ first the completed calls

are removed, and all fresh calls are accepted only if $Y_{n-1} - Z_{n-d(\Delta)} + X_n \leq C$. Otherwise, only $Z_n = C - Y_{n-1} + Z_{n-d(\Delta)}$ of these calls can be accepted. The number of calls in the system at time step n then satisfies the recursion

$$\begin{aligned} Y_n &= \min\{Y_{n-1} + X_n, C\}, \quad 1 \leq n \leq d(\Delta), & (7) \\ Y_n &= \min\{Y_{n-1} - Z_{n-d(\Delta)} + X_n, C\}, \quad n > d(\Delta) & (8) \end{aligned}$$

Moreover, observe that $Z_n = Y_n - Y_{n-1} + Z_{n-d(\Delta)}$. Let $\underline{Z}_n$ be the random vector

$$\begin{aligned} \underline{Z}_n &= (Z_n, \ldots, Z_1), \quad 1 \leq n \leq d(\Delta), & (9) \\ \underline{Z}_n &= (Z_n, \ldots, Z_{n-d(\Delta)+1}), \quad n > d(\Delta) & (10) \end{aligned}$$

then $\underline{Z}_n$ records the number of calls in the queue at time step n, as well as the received amount of service of these calls (in slots). To see this, notice that for $n \leq d(\Delta)$ at time step n, the number of calls Z_n have just arrived, and therefore have received an amount of service $\theta\Delta$, with $0 \leq \theta < 1$, whereas the number of calls Z_1 arrived in time interval $(0, \Delta]$ and have therefore already received the amount of service $(n-1)\Delta + \theta\Delta$. Therefore, $\underline{Z}_n$ provides a complete description of the state of the Markov chain that records the number of calls including the remaining service requirement. For $\Delta \to 0$, thus $d(\Delta) \to \infty$, we arrive at the random vector $\underline{Z}_n$ that records the density of calls together with the remaining service requirement at time n of the continuous time $M|D|C|C$ queue.

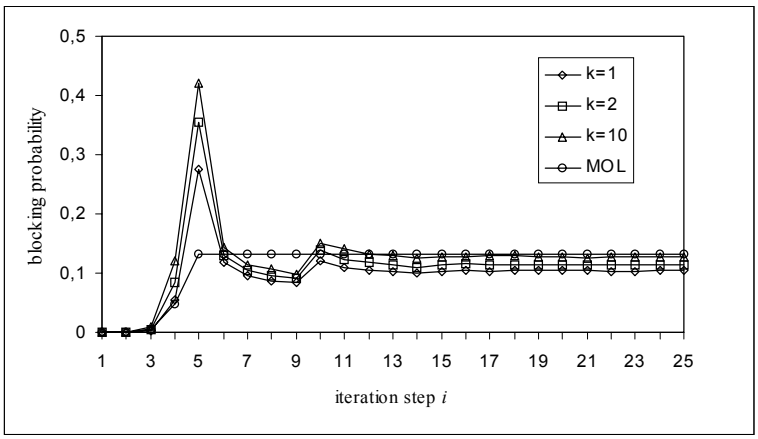


Figure 7: Convergence of blocking probabilities $f_i(k)$ for $\alpha = 2$.

The recursion (7), (8), or equivalently, the recursion (1), (2), provides a method for analysing instantaneous fresh call blocking probabilities in an $M|D|C|C$ queue with time-varying arrival rates. This recursion is particularly appealing, since blocking probabilities seem to approach those of the $M|D|C|C$ queue from below as is stated in the following conjecture, and illustrated in Figure 7. Before introducing the conjecture, let us first introduce some notation. Let $\Delta \to 0$ along the sequence $\Delta = \Delta'/2^k$, $k = 0, 1, 2, \ldots$ for $\Delta' = D/d$, where d is the initial number of time steps for $k = 0$, i.e., we will double the number of time steps in D when $k \to k+1$, and for $k = 0$ we have the situation of the beginning of the subsection. Let $X_n(k)$ denote the number of arrivals and $Z_n(k)$ be the number of accepted arrivals in $((n-1)\Delta'/2^k, n\Delta'/2^k]$, $k = 1, 2, \ldots$. Let $f_n(k) = 1 - \frac{EZ_n(k)}{EX_n(k)}$ the approximate instantaneous blocking probability that is accumulated over the interval

$((n-1)\Delta'/2^k, n\Delta'/2^k]$. The instantaneous blocking probability $f(t)$ at time t is then approximated as $f(t) \approx f_{\lceil td2^k/D \rceil}(k)$.

Lemma 3.7 ***Conjecture 3.8*** *Let $f_n(k) = 1 - \frac{EZ_n(k)}{EX_n(k)}$ be the blocking probability obtained from (7), (8) with time steps $\Delta = \Delta'/2^k$. Then $f_n(k) \leq f_{2n}(k+1)$, and $f(t) = \lim_{k\to\infty} f_{\lceil td2^k/D \rceil}(k)$.*

For the system starting in equilibrium, the ordering of blocking probabilities is straightforward, since in our recursion calls depart at the beginning of the interval, and therefore stay in the system shorter. As a consequence, the load of the queue is smaller. Clearly, the load is increasing with k. Therefore blocking probabilities increase with k, since the Erlang loss formula is a strictly increasing function of the load.

For the system with time-varying arrival rates, for $t \leq D$ the ordering of blocking probabilities can readily be demonstrated. Let $Y_n(k)$ be the number of calls in the queue at time $n\Delta'/2^k$, $k = 1, 2, \ldots$. Then

$$\begin{aligned} Y_n(k) &= \min\{Y_{n-1}(k) + X_n(k), C\}, \quad 1 \leq n \leq 2^{k-1}d, && (11)\\ Y_n(k) &= \min\{Y_{n-1}(k) - Z_{n-d}(k) + X_n(k), C\}, \quad n > 2^{k-1}d. && (12) \end{aligned}$$

Notice that, for $k = 1, 2, \ldots$,

$$\begin{aligned} X_n(k) &= X_{2n-1}(k+1) + X_{2n}(k+1), \quad n \geq 1, && (13)\\ Z_n(k) &= Z_{2n-1}(k+1) + Z_{2n}(k+1), \quad 1 \leq n \leq 2^{k-1}d, && (14)\\ Y_n(k) &= Y_{2n}(k+1) \quad 1 \leq n \leq 2^{k-1}d. && (15) \end{aligned}$$

We have assumed that $\{X_n(k)\}_n$ is iid for all k, which is possible due to the Poisson arrival process. From Lemma 3.2 we further obtain that $Z_n(k) \leq_{st} Z_{n-1}(k)$ for $1 \leq n \leq 2^{k-1}d$ and for all k. This immediately implies that

$$f_{2n}(k+1) = 1 - \frac{\mathbb{E}Z_{2n}(k+1)}{\mathbb{E}X_{2n}(k+1)} = 1 - \frac{2\mathbb{E}Z_{2n}(k+1)}{\mathbb{E}X_n(k)} \geq 1 - \frac{\mathbb{E}Z_n(k)}{\mathbb{E}X_n(k)} = f_n(k), \quad 1 \leq n \leq 2^{k-1}d, \tag{16}$$

where we have used (13) in the second equality, and (14) and $Z_{2n}(k+1) \leq_{st} Z_{2n-1}(k+1)$ in the last inequality.

The argument above cannot be continued for $t > D$, because the ordering of the $Z_{2n+j}(k+1)$ cannot be determined, see the comments following Lemma 3.3.

Below, in a series of graphs, blocking probabilities for the $M/D/C/C$ queue are depicted for varying arrival rate functions $\alpha(t)$ that naturally coincide with the behaviour of traffic jams. To this end, let α be the arrival rate of calls per stretch of road of unit length when it is fully covered by the traffic jam. The first series of graphs, Figure 8, considers a block shaped traffic jam, where the traffic mass reaches its maximum immediately at $t = 0$. Then X_i is Poisson($\alpha\Delta$) for all i, and therefore X_i is stationary. The resulting arrival rate $\alpha(t)$, for $\Delta \to 0$, to the system has a block shape:

$$\alpha(t) = \alpha, \quad t \geq 0.$$

Clearly, these results coincide with those of Section 2.

The results of Figures 9, and 10 show blocking probabilities for traffic jams with a more smooth initial phase resulting in a similar (but smaller) peak in the blocking probabilities. Figure 9 depicts blocking probabilities for a triangular initial phase of the traffic jam, such that the density of traffic increases linearly from 0 to α over the length of a cell, i.e., X_i is Poisson($i\alpha\Delta/D$) for $i = 1, \ldots, d(\Delta)$, and X_i is Poisson($\alpha\Delta$) for $i \geq d(\Delta) + 1$. The resulting arrival rate $\alpha(t)$, for $\Delta \to 0$, to the system is

$$\alpha(t) = \begin{cases} (\alpha/D)t & t \leq D, \\ \alpha & t \geq D. \end{cases} \tag{17}$$

Figure 10 depicts blocking probabilities for a traffic jam that increases linearly over two cells: the density of traffic increases linearly from 0 to α over the length two cells, i.e., X_i is Poisson($i\alpha\Delta/2D$) for $i = 1, \ldots, 2d(\Delta)$, and X_i is Poisson($\alpha\Delta$) for $i \geq 2d(\Delta) + 1$. The resulting arrival rate $\alpha(t)$, for $\Delta \to 0$, to the system is

$$\alpha(t) = \begin{cases} (\alpha/2D)t & t \leq 2D, \\ \alpha & t \geq 2D. \end{cases} \tag{18}$$

The situations of Figure 9 and 10, when translated to the setting of road covering networks, are less natural than that of Figure 8, since they smoothen the initial fase of a traffic jam over 2 – 5 km, which does not correspond to behaviour of real traffic jams (see e.g. [6]). Also in the setting of frequency defects, the situation of Figures 9 and 10 is not natural, as these correspond to gradually starting defects. From a theoretical point of view, however, these situations are of interest as they relate the sharpness of the peak to the shock in the density of subscribers. Moreover, these graphs illustrate that the peak in blocking probabilities is not due to differentiability problems of $\alpha(t)$, but is completely due to the steepness of the arrival rate curve. It is interesting to observe that the peaks in the blocking probabilities in Figures 9 and 10 occur well beyond $i = d$, corresponding to $t = D$, which is natural, since the initial phase of the traffic jam does hardly contribute to the load of the system. A precise prediction of the value for i at which the peak occurs is difficult. Observe that the peak is closer to $t = D$ for denser traffic, and that for denser traffic multiple peaks emerge. Notice that the peaks of Figures 9 and 10 are less pronounced than those of Figure 8, which is due to the more smooth increase in arrival rates.

As an aside, we have compared our results with approximate results for blocking probabilities in Erlang loss queues. In particular, we have investigated the Modified Offered Load approximation (MOL), and the Pointwise Stationary Approximation (PSA) that are commonly used in the analysis of transient blocking probabilities, see [3] for a detailed description of these approximations, and see [4] for error bounds of the MOL approximation. For our system, the approximate instantaneous blocking probabilities obtained from MOL or PSA are $\mathbb{P}(M(t) = C)$, with $M(t)$ the number of calls in the system at time t, and

$$\mathbb{P}(M(t) = k) = \frac{\rho(t)^k}{k!} \Bigg/ \sum_{j=0}^{C} \frac{\rho(t)^j}{j!}, \quad k = 0, \ldots, C, \tag{19}$$

the time-dependent distribution of an infinite server queue truncated to C, where the (offered) load $\rho(t)$ is obtained as

$$\rho(t) = \begin{cases} \int_{t-D}^{t} \alpha(s)ds & \text{for MOL} \\ \alpha(t)D & \text{for PSA} \end{cases} \tag{20}$$

see [3] for a motivation of these expressions. Computation of these (offered) loads yields, for Figure 8:

$$\rho_{MOL}(t) = \begin{cases} \int_{0}^{t} \alpha \, ds = \alpha t & \text{for } t \leq D \\ \int_{t-D}^{t} \alpha \, ds = \alpha D & \text{for } t \geq D \end{cases} \tag{21}$$

$$\rho_{PSA}(t) = \alpha D \quad t \geq 0 \tag{22}$$

for Figure 9:

$$\rho_{MOL}(t) = \begin{cases} \frac{1}{2}(\alpha/D)t^2 & t \leq D \\ \frac{1}{2}(\alpha/D)[D^2 - (t-D)^2] + \alpha(t-D) & D \leq t \leq 2D \\ \alpha D & t \geq 2D \end{cases} \tag{23}$$

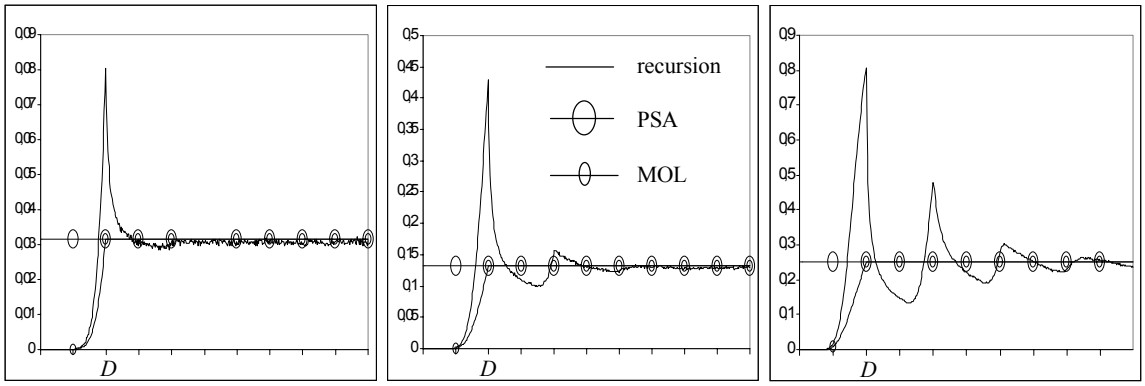
$$\rho_{PSA}(t) = \begin{cases} \alpha t & t \leq D \\ \alpha D & t \geq D \end{cases} \tag{24}$$

and for Figure 10:

$$\rho_{MOL}(t) = \begin{cases} \frac{1}{2}(\alpha/2D)t^2 & t \leq D \\ \frac{1}{2}(\alpha/2D)[t^2 - (t-D)^2] & D \leq t \leq 2D \\ \frac{1}{2}(\alpha/2D)[(2D)^2 - (t-D)^2] + \alpha(t-2D) & 2D \leq t \leq 3D \\ \alpha D & t \geq 3D \end{cases} \tag{25}$$

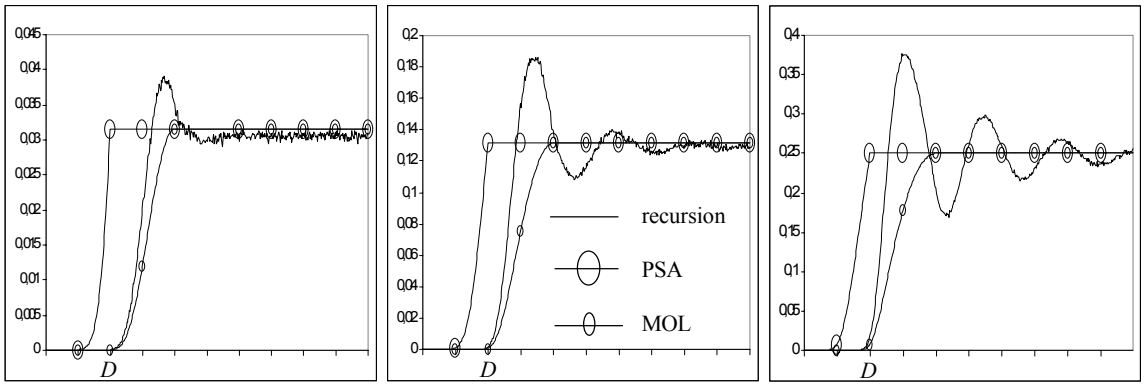
$$\rho_{PSA}(t) = \begin{cases} (\alpha/2)t & t \leq 2D \\ \alpha D & t \geq 2D \end{cases} \tag{26}$$

As can be seen from Figures 8–10, MOL outperforms PSA, but MOL completely ignores the peak in the blocking probabilities occurring at $t = D$. In fact, MOL reaches equilibrium as soon as $\alpha(t)$ becomes constant, i.e., at $t = D$ in the situation of Figure 8, where the effect of the shock for the blocking probabilities may have influence for multiple cells. Therefore, one has to be careful when dimensioning capacity of wireless networks based on these approximations.



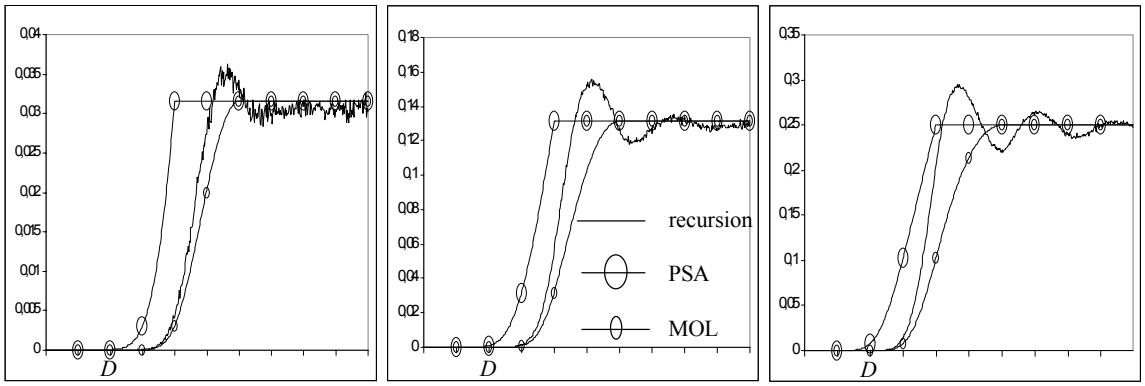
$\alpha D = 15$ $\alpha D = 20$ $\alpha D = 25$

Figure 8: block shaped traffic jam.



$\alpha D = 15$ $\alpha D = 20$ $\alpha D = 25$

Figure 9: triangular traffic jam (one cell).



$\alpha D = 15$ $\alpha D = 20$ $\alpha D = 25$

Figure 10: triangular traffic jam (two cells).

4 Concluding remarks

With the number of subscribers rapidly increasing, efficient allocation of capacity to mobile subscribers of wireless networks calls for road covering networks. Dimensioning (e.g. capacity allocation) for such networks is typically carried out assuming a (homogeneous) stationary load over the cells. For networks with load not distributed homogeneously over the cells of the network, however, the cellular nature might not be beneficial: as large bodies of mobile subscribers *(hot spots)* migrate through the cells of the network, large numbers of calls almost simultaneously transfer between cells, and have to *handover* from resources of the originating base station to resources of the destination base station. Handover blocking is considered to have a more dramatic effect on the quality of service than fresh call blocking. Hence, there is a trade-off between increasing capacity (decreasing cell sizes) resulting in lower fresh call blocking, and increasing cell sizes resulting in lower handover blocking, and adequate tools must be developed to analyse such blocking probabilities.

An often used approach towards analysis of blocking probabilities is to assume that the network operation is *as if* the network is in statistical equilibrium; tools for numerical evaluation of blocking probabilities are then provided in e.g. [1, 8]. Assuming that the network parameters are changing slowly over time, the Modified Offered Load approximation, or Pointwise Stationary Approximation, that use the functional form of the equilibrium distribution, but with the time-dependent offered load obtained from analysis of a network with unlimited capacity (hence no blocking occurs) then enable using the tools of [1, 8], and often provide an adequate approximation of transient blocking probabilities [3]. When the network parameters show a drastic deviation from equilibrium behaviour, however, such methods may not be adequate. This occurs e.g. when a traffic jam is moving along a road in a wireless network, where the number of subscribers in a traffic jam is considerably larger than that number outside the traffic jam, and blocking probabilities may show behaviour not consistent with results obtained from (approximations based on) equilibrium analysis.

This paper has provided a detailed analysis of blocking probabilities for subscribers in a traffic jam travelling along consecutive cells of a wireless network. Via a discrete event simulation, sharp peaks have been observed in the handover blocking probabilities when a traffic jam starts covering a new cell. This peak is most prominent when the traffic jam has a constant density of subscribers, a situation that can be modelled via the transient behaviour of the M/D/C/C queue, the Erlang loss queue with constant service times. A similar peak occurs when homogeneous traffic enters a cell with smaller capacity. The main conclusions from this study are that

- a subscriber travelling at roughly the cell-length behind the start of a traffic jam suffers from a considerable increase in handover blocking probability at each handover attempt,
- the best design for road covering cellular networks is that with identical cells.

Acknowledgement The authors would like to thank J. Hoeksema for the computer

programme for the graphs of Section 2.

References

[1] R.J. Boucherie, M. Mandjes and S.C. Verwijmeren, Computational methods for blocking probabilities in layered cellular mobile communications networks, Research Report, University of Twente, 2002.

[2] Ministerie van Verkeer en Waterstaat (in Dutch)
http://www.minvenw.nl/cend/dvo/persinfo/data/persberichteen/000314.5898.html

[3] W.A. Massey and W. Whitt, Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems* **13**, 183-250, 1993.

[4] W.A. Massey and W. Whitt, An analysis of the modified offered load approximation for the nonstationary Erlang loss model. *The Annals of Applied Probability* **4**, 1145-1160, 1994.

[5] G.F. Newell, *Applications of queueing theory*, Chapman and Hall, 1982.

[6] S.A. Smulders, *Control of freeway traffic flow*, CWI Tract 80, Centrum voor Wiskunde en Informatica, Amsterdam, 1996.

[7] R. Pandya, S. Tseng, and K. Basu, Some performance benchmarks for the design of wireless systems and networks, *In: Teletraffic Contributions for the Information Age, V. Ramaswami, P.E. Wirth (Editors)*, pp. 243–253, Elsevier, 1997.

[8] K.W Ross, *Multiservice loss models for broadband telecommunication networks*, Springer, 1995.

[9] S.M. Ross, *Stochastic processes*, 2nd edition, Wiley, 1996.

[10] A. Ule and R.J. Boucherie, On the distribution of customers in a wireless network driven by fluid traffic. *To appear: European Journal of Opeational Research*, 2002.