A multi-server queueing model with locking

Ivo Adan, Ton de Kok and Jacques Resing
Department of Mathematics and Computing Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

August 20, 1998

Abstract

In this paper we analyse a multi-server queueing model with locking. The model is motivated by a situation we encountered at a maintenance facility for trains. Maintenance is done at parallel tracks, where each track offers space to two trains. Trains can enter and leave the tracks from one and the same side only. This gives rise to *locking of the front train*: in order to leave the maintenance track the front train has to wait till maintenance of the back train (if there is one) has also been completed. Hence, part of the maintenance (or track) capacity is lost. The queueing model is used to investigate the loss of capacity and its effect on sojourn times. The performance of this system is also compared with other designs. A surprising result is that in *light traffic* it is better to use only half of the track capacity by allowing no more than one train at a maintenance track.

1 Introduction

In this paper we consider a multi-server queueing model, in which the servers are clustered into small groups of two servers. Within a group, one of the servers is called the front server and the other one is called the back server. A special feature of the model is that if the service of a customer at the front server has been completed and there is another customer in service at the back server, the customer can not leave the system (is locked in) until also the service of the customer at the back server has been completed. During this period the front server is blocked and can not serve a new customer. Hence, part of the capacity of this server is lost.

Performance measures of interest are the (loss of) capacity of the system, the equilibrium distribution of the number of customers in the system and the waiting time distribution. Under the assumption of Poisson arrivals and exponential service times, we are able to derive a closed form expression for these performance measures.

The model is motivated by a situation we encountered at the maintenance department of the Dutch railway company. Maintenance on trains is done on a number of separate parallel tracks. On each of these tracks there is room for two trains. Trains can leave the tracks only on the same side as they enter the tracks. Hence, a train may, after completion of its maintenance, be locked in by a train that arrived later on the same track and that has not yet completed its maintenance. Clearly the front part and the back part of a maintenance track correspond to a group of a front and a back server in the queueing model described above. This queueing model may be used to investigate the impact of locking on the performance of the maintenance facility, and to compare the performance of this system with other designs. A surprising result is that, in light traffic, allowing at most one train at a maintenance track reduces the sojourn times. So, use of only half of the track capacity may be better than use of the full capacity.

The remainder of this paper is organized as follows. In section 2 we describe the queueing model in detail. Next, in section 3 we determine the equilibrium distribution of the number of customers in the system. From that, we derive the waiting time distribution in section 4. In section 5 we compare the performance of this system with several other systems. In particular, we are interested in a comparison with the system in which only the front servers may be used and with the system in which the servers are not clustered but can work independently. The first system corresponds to the situation in which only the front part of a maintenance track may be used and in the second system there are twice as many tracks, but they now have room for only one train. In section 6 we relax the assumption of exponential service times. Finally, section 7 is devoted to conclusions and comments.

2 Model description

Customers arrive according to a Poisson process with rate λ. Service times are exponentially distributed with parameter μ. There are s groups of servers, each group consisting of two servers (see figure 1). Within a group, one of the servers is called the front server

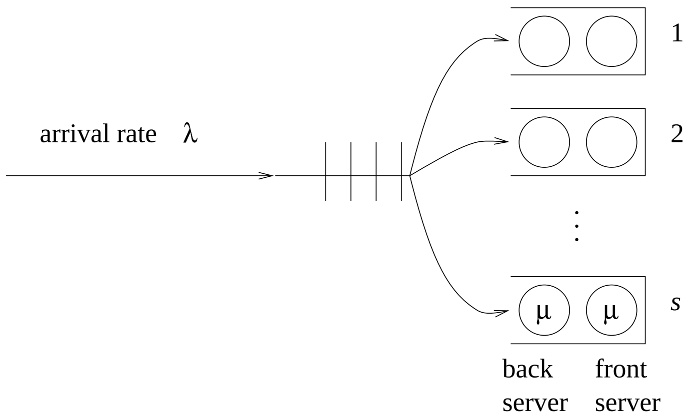

Figure 1: Queueing model for a maintenance facility for trains

and the other one is called the back server. Customers are served by front servers as long as these are available. If all front servers are occupied, new customers are served by back servers. If the service of a customer at the front server has been completed and there is another customer in service at the back server, the customer can not leave the system until also the service of the customer at the back server has been completed. During this period the front server is blocked and can not serve a new customer.

This system can be described by a continuous-time Markov process. To specify the states of the Markov process we distinguish between the situation where all the server positions are occupied and the situation where some of the server positions are not occupied. In the first situation we can describe the state of the system by the pair (n, m) where n denotes the number of customers waiting in the queue and m the number of customers that already completed their service but that are locked in by a customer at the back server. The states (n, m) will be termed the *saturated states*. So the possible saturated states are the pairs of integers (n, m) where n ranges from 0 to ∞ and m from 0 to s.

In the situation where not all server positions are occupied the description of the system is a little bit more complicated. Then we can describe the state of the system by the quadruple (k_0, k_1, k_2, k_3). Here k_0 denotes the number of groups of servers for which both server positions are empty, k_1 the number of groups of servers for which only the front server is working, k_2 the number of groups of servers for which only the back server is working (and at the front server a customer is locked in) and k_3 the number of groups of servers for which both servers are working. The states (k_0, k_1, k_2, k_3) will be termed the *unsaturated states*. So the possible unsaturated states are all quadruples of non-negative integers (k_0, k_1, k_2, k_3) which satisfy $k_0 + k_1 + k_2 + k_3 = s$ and $k_2 + k_3 < s$. Hence, the total number of unsaturated states equals $\binom{s+3}{3} - (s + 1)$.

Note that when there are always customers waiting in the queue (i.e., in heavy traffic), the front server in a group will lose half of its capacity, because the server always has to wait for a service completion of the server at the back position. So each group can serve $3\mu/2$ (instead of 2μ) customers per unit of time. Hence, the capacity of the system is equal to $3\mu s/2$, and to guarantee that the system can handle all customers, we have to require that

$$\lambda < \frac{3}{2}\mu s. \tag{1}$$

The stability condition (1) can be rigorously proved by using Neuts' mean drift condition, i.e. formula (1.7.11) in [9]. From now on we assume that condition (1) holds. Note that when the servers are not clustered, but can work independently (i.e. an $M/M/2s$ system) the capacity is $2\mu s$. Hence, the locking of customers due to clustering reduces the capacity with 25%.

We denote the equilibrium probabilities for the saturated states (n, m) and the unsaturated states (k_0, k_1, k_2, k_3) by $p(n, m)$ and $p(k_0, k_1, k_2, k_3)$, respectively. In the next section we formulate the equilibrium equations and determine a closed form expression for the equilibrium probabilities $p(n, m)$ in the saturated states.

3 The equilibrium distribution

The flow diagram for the saturated states is shown in figure 2. From this diagram we

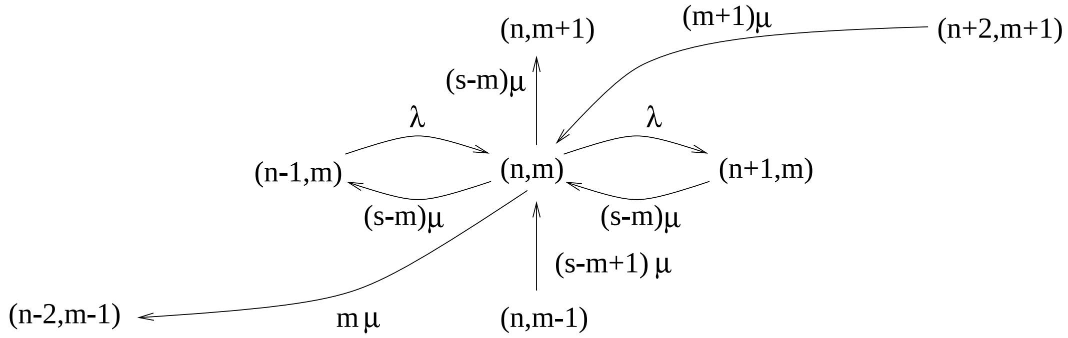

Figure 2: Flow diagram for the saturated states (n, m)

obtain by equating in each state (n, m) with $n > 0$ the flow out of and into that state the following balance equations for $p(n, m)$:

$$\begin{aligned}[\lambda + (2s - m)\mu]p(n, m) &= \lambda p(n-1, m) + (s - m + 1)\mu p(n, m-1) \\ &\quad + (m+1)\mu p(n+2, m+1) + (s-m)\mu p(n+1, m), \\ &\qquad n = 1, 2, \cdots, \; m = 0, \cdots, s, \end{aligned} \tag{2}$$

where by definition $p(n, -1) = p(n, s+1) = 0$. The balance equations for the states $(0, m)$ and the unsaturated states can be obtained similarly. We will refer to these equations as the *boundary equations*. The precise form of the boundary equations is not relevant to the analysis, and therefore, they are omitted.

The approach to solve the balance equations is inspired on earlier work [1, 2, 5, 6, 10] and proceeds as follows. We first construct a set of solutions of the balance equations (2) of the form

$$p(n, m) = y(m)x^n. \tag{3}$$

Then we use the solutions in this set to construct a linear combination which also satisfies the boundary equations and the normalization equation.

Substitution of the form (3) into (2) and dividing by the common power x^{n-1} yields

$$\begin{aligned}&(s - m + 1)\mu x y(m-1) + [(s-m)\mu x^2 - [\lambda + (2s - m)\mu]x + \lambda]y(m) \\ &\quad + (m+1)\mu x^3 y(m+1) = 0, \qquad m = 0, \cdots, s, \end{aligned} \tag{4}$$

where by convention $y(-1) = y(s+1) = 0$. For given x this is a system of linear homogeneous equations for $y(0), \ldots, y(s)$. Now we have to find the values of x for which this system has a non-null solution. Since we must be able to normalize the solution afterwards, only values of x satisfying $|x| < 1$ are useful.

Note that the linear system (4) has a non-null solution if and only if the determinant of this system is equal to zero. Hence, the desired values of x are the zeros inside the unit circle of the determinant, which is a high-degree polynomial in x. It will be difficult to prove directly that there are sufficiently many zeros inside the unit circle and to numerically determine these zeros. Therefore we employ an idea of [10, 6, 3] to transform the difference equations (4) into a single differential equation for the generating function of the sequence $y(m)$. This approach, in fact, leads to a nice *factorization* of the determinantal equation. So, let us introduce the generating function

$$Y(z) = \sum_{m=0}^{s} y(m) z^m.$$

Multiplying (4) by z^m and adding with respect to m we obtain the differential equation

$$\begin{aligned} s\mu x z Y(z) - \mu x z^2 Y'(z) + [\lambda + s\mu x^2 - (\lambda + 2s\mu)x]Y(z) \\ + \mu x(1-x)zY'(z) + \mu x^3 Y'(z) = 0, \end{aligned}$$

which may be rewritten in the form

$$\frac{Y'(z)}{Y(z)} = \frac{s\mu x z + [\lambda + s\mu x^2 - (\lambda + 2s\mu)x]}{\mu x z^2 - \mu x(1-x)z - \mu x^3} = \frac{A(x)}{z - z_1(x)} + \frac{s - A(x)}{z - z_2(x)} \tag{5}$$

where

$$z_1(x) = \frac{1 - x + \sqrt{5x^2 - 2x + 1}}{2}, \qquad z_2(x) = \frac{1 - x - \sqrt{5x^2 - 2x + 1}}{2}$$

and $A(x)$ is such that

$$2A(x) - s = \frac{s(x-3) + 2\lambda(1-x)/(\mu x)}{\sqrt{5x^2 - 2x + 1}}.$$

The general solution of the differential equation (5) is

$$Y(z) = K(z - z_1(x))^{A(x)}(z - z_2(x))^{s - A(x)},$$

with K an arbitrary constant. Now the key idea to proceed is that, since $Y(z)$ is a polynomial in z, the exponents $A(x)$ and $s - A(x)$ should be equal to a non-negative integer, i.e. $A(x) = j, j = 0, \ldots, s$. Hence, for each j we get an equation for x and it can be shown that, under the stability condition (1), this equation has exactly one root, x_j say, in the interval $(0, 1)$. This is summarized in the following lemma.

Lemma 3.1 *Provided condition (1) holds, we have for all $j = 0, \ldots, s$ that the equation*

$$2j - s = \frac{s(x-3) + 2\lambda(1-x)/(\mu x)}{\sqrt{5x^2 - 2x + 1}} \tag{6}$$

has a unique solution $x = x_j$ in the interval $(0, 1)$. Further, the solutions x_j satisfy

$$x_0 > x_1 > \cdots > x_{s-1} > x_s.$$

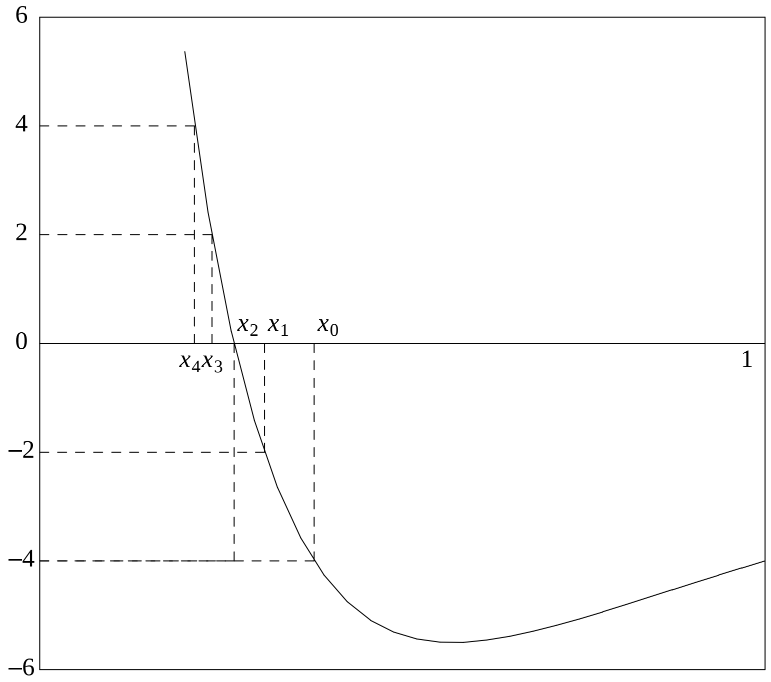

Figure 3: The right-hand side of equation (6) for $s = 4$, $\lambda = 2$ and $\mu = 1$

Proof: Let $f(x)$ denote the right-hand side of (6). Since $f'(0) = -\infty$ and $f'(1) > 0$ by virtue of (1) it holds that $f'(x)$ has at least one zero in $(0, 1)$. Straightforward algebra shows that there is exactly one zero, say $\tilde{x}$. So $f(x)$ decreases on $(0, \tilde{x})$ and increases on $(\tilde{x}, 1)$. Now the lemma follows by using that $f(0) = +\infty$ and $f(1) = -s$ (see figure 3). $\square$

Remark 3.2 *(Factorization)* From Lemma 3.1 we can obtain a nice factorization of the determinant of linear system (4). Denoting the determinant by $D(x)$, it holds that

$$D(x) = \left(\frac{1}{2}\right)^{s+1} \prod_{j=0}^{s} \left(\mu x s(x-3) + 2\lambda(1-x) - \mu x(2j-s)\sqrt{5x^2 - 2x + 1}\right).$$

The lemma provides a simple and efficient way to determine the desired values of x. For each j we determine the unique root x_j of equation (6) in $(0, 1)$ by using, e.g., bisection. For given $x = x_j$ the corresponding $y_j(m)$ follow from the recursion (4) and the initial values $y_j(s+1) = 0$ and $y_j(s) = 1$. Summarizing, we have found a set of $s+1$ basis solutions of the equations (2). The general solution of (2) can be expressed as a linear combination of these basis solutions, i.e.,

$$p(n, m) = \sum_{j=0}^{s} C_j y_j(m) x_j^n, \qquad n = 0, 1, 2, \cdots, \ m = 0, \ldots, s.$$

The remaining equations still to be satisfied are the boundary equations and the normalization equation. These equations (where one of the boundary equations may be dropped, since the balance equations are dependent) form a system of $\binom{s+3}{3}$ linear, inhomogeneous equations for, just as many, unknown coefficients C_j and unknown equilibrium probabilities $p(k_0, k_1, k_2, k_3)$. It is now a matter of routine to prove that this system indeed has a unique solution (cf. section 4 in [2]). The results are formulated in the following theorem.

Theorem 3.3 *There exist unique coefficients C_j such that the equilibrium probabilities $p(n,m)$ in the saturated states can be expressed as*

$$p(n,m) = \sum_{j=0}^{s} C_j y_j(m) x_j^n, \qquad n = 0,1,2,\cdots,\ m = 0,\cdots,s, \tag{7}$$

where x_j is the unique root of equation (6) in the interval $(0,1)$ and the corresponding $y_j(m)$ are a non-null solution of the linear homogeneous system (4) with $x = x_j$.

Remark 3.4 *(Stability condition)* If stability condition (1) does not hold, then the left-hand side of (6) is decreasing on the whole interval $(0,1)$. This implies that for $j = 0$ equation (6) has no solution in $(0,1)$, and thus we have a set of s instead of $s+1$ basis solutions. This number is not sufficient to construct a linear combination (other than the null solution) which also satisfies the boundary equations.

Remark 3.5 *(Mean values)* Based on expression (7) for the probabilities $p(n,m)$ it is easy to derive expressions for $E[L_q]$, the mean number of jobs in the queue, and $E[W]$, the mean waiting time in the queue. It follows from (7) that

$$E[L_q] = \sum_{n=0}^{\infty} \sum_{m=0}^{s} np(n,m) = \sum_{j=0}^{s} C_j \left(\sum_{m=0}^{s} y_j(m) \right) \frac{x_j}{(1-x_j)^2},$$

and by applying Little's law,

$$E[W] = \sum_{j=0}^{s} C_j \left(\sum_{m=0}^{s} y_j(m) \right) \frac{x_j}{\lambda(1-x_j)^2}. \tag{8}$$

4 Waiting time distribution

In this section we show that the probability distribution of the waiting time is a finite mixture of exponentials. The proof follows the same line of thoughts as in [4].

Theorem 4.1 *The waiting time distribution is given by*

$$\Pr[W > t] = \sum_{j=0}^{s} C_j \left(\sum_{m=0}^{s} y_j(m) \right) \frac{1}{1-x_j} e^{\lambda(1-1/x_j)t}, \qquad t \geq 0. \tag{9}$$

Proof: Define $F_{n,m}(t)$ as the probability that the waiting time of a customer is greater than t, given that the customer sees the saturated state (n,m) on arrival. Clearly, by the PASTA property and substitution of expression (7) for the probabilities $p(n,m)$ we have

$$\Pr[W>t] = \sum_{n,m} p(n,m)F_{n,m}(t) = \sum_{j=0}^{s} C_j \left(\sum_{m=0}^{s} y_j(m)\right) \sum_{n=0}^{\infty} F_{n,m}(t)x_j^n. \tag{10}$$

It is easy to check that the functions $F_{n,m}(t)$ satisfy the set of differential equations

$$\begin{aligned}\frac{d}{dt}F_{n,m}(t) + (2s-m)\mu F_{n,m}(t) &= (s-m)\mu F_{n,m+1}(t)\\ &\quad +(s-m)\mu F_{n-1,m}(t) + m\mu F_{n-2,m-1}(t),\\ &\qquad n=0,1,2,\cdots,\ m=0,\cdots,s,\end{aligned}$$

where, by convention, the function $F_{n,m}(t)=0$ if (n,m) is not feasible. Hence, the Laplace transform $F^*_{n,m}(\theta) = \int_0^\infty e^{-\theta t}F_{n,m}(t)dt$ satisfies

$$\begin{aligned}[\theta + (2s-m)\mu]F^*_{n,m}(\theta) &= 1 + (s-m)\mu F^*_{n,m+1}(\theta)\\ &\quad +(s-m)\mu F^*_{n-1,m}(\theta) + m\mu F^*_{n-2,m-1}(\theta).\end{aligned} \tag{11}$$

Now, if we introduce $G_{j,m}(\theta) = \sum_{n=0}^{\infty} F^*_{n,m}(\theta)x_j^n$ and $H_j(\theta) = \sum_{m=0}^{s} y_j(m)G_{j,m}(\theta)$, then we obtain from (11) by straightforward calculations that

$$\begin{aligned}[\theta + (2s-m)\mu]G_{j,m}(\theta) &= \frac{1}{1-x_j} + (s-m)\mu G_{j,m+1}(\theta)\\ &\quad +x_j(s-m)\mu G_{j,m}(\theta) + x_j^2 m\mu G_{j,m-1}(\theta),\end{aligned}$$

which, by adding over all m, subsequently yields

$$\begin{aligned}\theta H_j(\theta) + \sum_{m=0}^{s}(2s-m)\mu y_j(m)G_{j,m}(\theta) &= \frac{1}{1-x_j}\sum_{m=0}^{s} y_j(m) + \sum_{m=0}^{s} G_{j,m}(\theta)\cdot\\ &\quad \Big[(s-m+1)\mu y_j(m-1) + x_j(s-m)\mu y_j(m) + x_j^2(m+1)\mu y_j(m+1)\Big].\end{aligned}$$

Finally, by using equation (4) we can conclude from the relation above that

$$H_j(\theta) = \left(\sum_{m=0}^{s} y_j(m)\right)\frac{1}{1-x_j}\,\frac{1}{\theta - \lambda(1-1/x_j)}$$

and hence, by (10),

$$W^*(\theta) = \int_0^\infty e^{-\theta t}\Pr[W>t]dt = \sum_{j=0}^{s} C_j\left(\sum_{m=0}^{s} y_j(m)\right)\frac{1}{1-x_j}\,\frac{1}{\theta-\lambda(1-1/x_j)},$$

from which the theorem immediately follows. □

5 Performance evaluation

As mentioned in the introduction the possibility of locking in customers, and consequently that of blocking the servers reduces the service capacity. The loss of service capacity can be quantified by the mean number of servers that is blocked divided by $2s$, the total number of servers. Figure 4 shows the loss of capacity as a function of λ. Obviously, it increases in λ, and tends to a maximal loss of 25% of the available service capacity as λ tends to $3\mu s/2$.

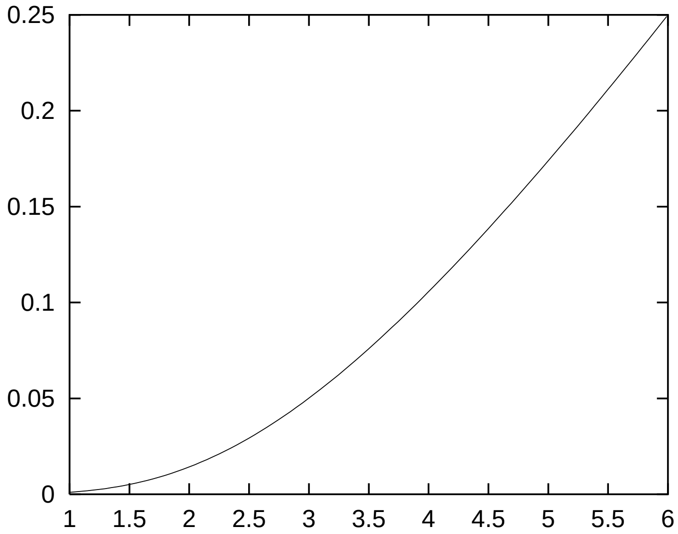

Figure 4: The loss of service capacity as a function of the arrival rate in the situation $s = 4$ and $\mu = 1$

Next we compare the performance of our train model with several related systems. First of all we compare the model with the $M/M/s$ and the $M/M/2s$ systems. The first system corresponds to the situation in which only the front part of a maintenance track may be used and in the second system there are twice as many tracks, but they now have room for only one train. The performance measure we primarily focus on is the mean sojourn time. In the train model the sojourn time S is the sum of the waiting time in the queue, the service time and the time a train is locked in. By Little's law, we have

$$E[S] = \frac{1}{\lambda} E[L],$$

where $E[L]$ is the mean number of customer in the system, which is given by

$$\begin{aligned} E[L] &= \sum_{(k_0,k_1,k_2,k_3)} (k_1 + 2k_2 + 2k_3) p(k_0, k_1, k_2, k_3) + \sum_{(n,m)} (n + 2s) p(n, m) \\ &= \sum_{(k_0,k_1,k_2,k_3)} (k_1 + 2k_2 + 2k_3) p(k_0, k_1, k_2, k_3) \\ &\quad + \sum_{j=0}^{s} C_j \left(\sum_{m=0}^{s} y_j(m) \right) \left(\frac{x_j}{(1 - x_j)^2} + \frac{2s}{1 - x_j} \right). \end{aligned}$$

Clearly, one expects that the mean sojourn time in the train model is less than the mean sojourn time for the $M/M/s$ model and greater than the one in the $M/M/2s$ model. In figures 5 and 6 we compare the mean sojourn time in the train model with 4 tracks with the mean sojourn times in the $M/M/4$ and $M/M/8$ model. The mean sojourn time is shown as a function of the arrival rate λ. The mean service time is equal to 1.

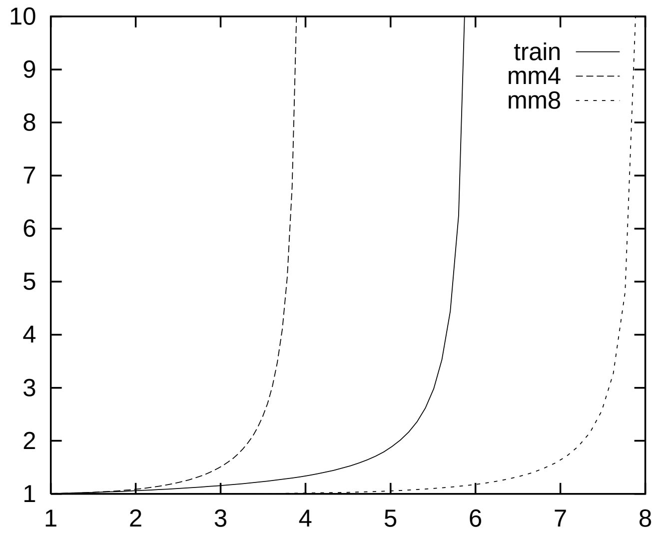

Figure 5: The mean sojourn time as a function of the arrival rate for the train model and for the $M/M/s$ and $M/M/2s$ models in the situation $s = 4$ and $\mu = 1$

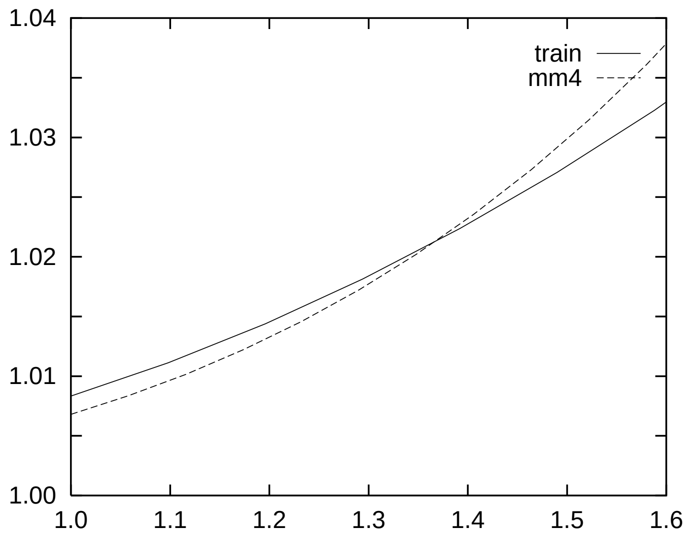

Figure 6: The mean sojourn time as a function of the arrival rate in light traffic for the train model and the $M/M/s$ model in the situation $s = 4$ and $\mu = 1$

At first glance, figure 5 shows the expected behaviour that the performance of the train model is sandwiched between those of the $M/M/4$ and $M/M/8$ models. However, figure 6 shows the surprising phenomenon that in light traffic the train model behaves even worse than the $M/M/4$ system. An explanation of this phenomenon is the following. In light traffic there is a trade-off between waiting time before service (in the queue) versus waiting time after service (locking). Now, consider the situation where a customer arrives when all s front servers are occupied. Assume furthermore that no new customers arrive until all $s+1$ customers have left (which is likely in light traffic). Then the expected remaining sojourn time of all customers in the train model is equal to $1/\mu$, except for the one that is locked in by the new customer. His expected remaining sojourn time is equal to $1/\mu+1/2\mu$. Hence the sum of the remaining sojourn times of the $s+1$ customers in the train model is equal to $(s+3/2)/\mu$. But the corresponding quantity in the $M/M/s$ system is equal to $(s+1+1/s)/\mu$. This suggests that in light traffic it is better to use only the servers in the front position, provided s is greater than 2.

Figure 5 illustrates that the mean sojourn times explode when the systems nearly operate at maximum capacity. The capacity of our train model with s tracks is equal to $3\mu s/2$, so the capacity of the $M/M/s$ and $M/M/2s$ system is less and greater, respectively. It is interesting to compare the train model with a multi-server system with the same capacity, namely the one with $3s/2$ servers, which of course only makes sense if s is even.

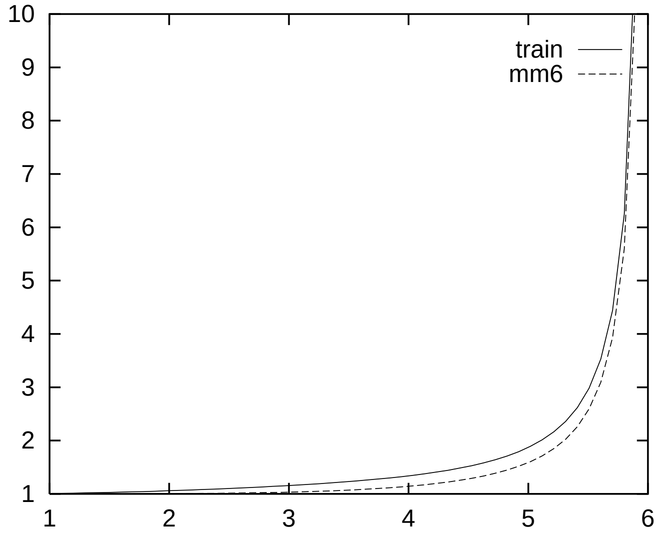

Figure 7: The mean sojourn time as a function of the arrival rate for the train model and the multi-server model with $3s/2$ servers in the situation $s=4$ and $\mu=1$

In figure 7 we see that for all values of λ the train model with 4 tracks behaves worse than the $M/M/6$ model. This may be intuitively explained by the random availability of the servers in the trainmodel (on the average there are 6 servers available) versus the deterministic availability of the servers in the $M/M/6$ model. The difference in performance of the two systems can be explicitly quantified when the workload ρ tends to 1, where ρ is defined as

$$\rho = \frac{2\lambda}{3\mu s}\,.$$

The performance of the train model for ρ close to 1 can be determined by exploiting that the first term in (7) dominates the others, since x_0 converges to 1 whereas the other roots stay away from 1. In fact, it follows from (6) that as ρ tends to 1, then

$$x_0 = 1 + \frac{13}{12}(\rho - 1) + O(\rho - 1)^2.$$

which implies that

$$E[L] = \frac{13}{12}\frac{1}{1-\rho} + O(1).$$

In the multi-server system with $3s/2$ servers the mean number of customers in the system satisfies

$$E[L(M/M/\frac{3}{2}s)] = \frac{1}{1-\rho} + O(1),$$

as ρ tends to 1. Hence, we can conclude that, when ρ is close to 1, the mean number of customers in the train model is approximately 13/12 times the mean number of customers in the multi-server model with $3s/2$ servers. By Little's law, the same holds for the mean sojourn times.

6 General service times

So far we assumed exponential service times. In this section we discuss the generalization to more general service times. First, we concentrate on the loss of capacity of the system and after that on the analysis of the equilibrium probabilities.

Consider a single group of two servers and assume that there are always customers waiting in the queue. Then the time points that a customer leaves the front server form regeneration points for that particular group of servers. Denote with X_0 the service time of a customer at the front server and with $X_1, X_2, \ldots$ successive service times of customers at the back server. The X_i's have common distribution function $F(\cdot)$ and finite mean EX. From the theory of regenerative processes it follows that the long-run fraction of time that the front server is working is equal to EX divided by the expected length of the regeneration cycle. If we define $N = \inf\{n : X_1 + \cdots + X_n \geq X_0\}$, then the expected length of the regeneration cycle is given by $E(X_1 + \cdots + X_N) = EN \cdot EX$. The last equality follows from Wald's formula and the fact that N is a stopping time for the sequence $X_1, X_2, \ldots$ Hence, the fraction of time that the front server is working equals $1/EN$. This implies that, due to locking, the capacity of the system is reduced with $(1 - 1/EN) \cdot 50\%$.

We can calculate EN by using that $EN = 1 + \int_0^\infty E[N(x)]\, dF(x)$, where $E[N(x)]$ is the expected number of renewals in $(0, x)$ of the renewal process corresponding to $F(\cdot)$. Straightforward calculations yield that $EN = 2$ for the exponential distribution, $1.5 < EN < 2$ for the Erlang distribution and $2 \leq EN \leq 4$ for the hyperexponential distribution of order 2. For distributions with heavy tails we expect that EN can be arbitrarily large, and thus the loss of capacity arbitrarily close to 50%.

When the service time distribution is taken from the general class of Coxian distributions, the queueing model can still be formulated as a Markov process. The state description is of course more complicated than in the exponential model. For Coxian service times with k stages the saturated states can be described by $(n, \vec{m})$ where

$$\vec{m} = (m(1,1), \ldots, m(k+1,1), m(1,2), \ldots, m(k+1,2), \ldots, m(1,k), \ldots, m(k+1,k)),$$

in which the variable $m(i,j)$ denotes the number of groups for which the customer at the front position is in service stage i and the one at the back position in service stage j. Stage $i = k+1$ means here that the customer at the front position completed service but is locked in by the one at the back position.

We can try to follow the same approach as in section 3 to determine the equilibrium probabilities $p(n, \vec{m})$. In this case, substitution of the form

$$p(n, \vec{m}) = y(\vec{m})x^n$$

into the balance equations leads to a set of difference equations for the sequence $y(\vec{m})$, which can be transformed into a partial differential equation for the multi-dimensional generating function of the sequence $y(\vec{m})$. The solution of this partial differential equation, however, is not as easy as the solution of (5). In [5] it is shown that the analysis of the balance equations for the multi-server queue with Coxian interarrival and service times leads to a similar partial differential equation. There, the equation is elegantly solved by using the method of characteristics (see e.g. [7]). It is worthwhile to investigate whether or not this technique also works for the system with locking and Coxian service times.

7 Conclusions and comments

In this paper we analysed an exponential multi-server model for train maintenance. The characteristic feature of the model is that customers can be locked in after service completion. This model has been used to gain insight into the effect of this locking on the performance of the system, where we mainly focused on sojourn times and loss of capacity. The conclusions can be summarized as follows:

- The loss of maintenance capacity is increasing in the workload, and it is 25% in a heavily loaded system.
- In light traffic usage of only half of the track capacity by allowing at most one train at a track reduces the sojourn times of trains;
- In heavy traffic the mean sojourn time in a maintenance facility with s tracks for two trains is approximately 8% more than in a facility with $3s/2$ tracks for only one train;

In our model customers are served by front servers if these are available and by back servers otherwise. It turned out that, at least in light traffic, this allocation of customers to servers is not optimal. An interesting open problem is to determine the optimal customer allocation strategy.

The train model can be described by a Markov process whose state space is a semi-infinite strip. We presented a method to express the equilibrium probabilities of this process as a finite sum of terms which are geometric in the number of waiting customers. The geometric factors are the roots (inside the unit circle) of a determinantal equation. This approach is closely related to the one in [8]. By using a generating-function technique we have shown that the determinant can be factorized (see remark 3.2), which considerably simplifies the determination of the roots. It seems important to investigate when such a factorization is possible, and in particular, when it can be established by using the generating-function technique presented in this paper.

References

[1] I.J.B.F. Adan, W.A. van de waarsenburg, J. Wessels, Analyzing $E_k|E_r|c$ queues, *EJOR*, **92** (1996), pp. 112–124.

[2] I.J.B.F. Adan, J. Wessels and W.H.M. Zijm, A compensation approach for two-dimensional Markov processes, *Adv. Appl. Prob.*, **25** (1993), pp. 783–817.

[3] D. Anick, D. Mitra and M.M. Sondhi, Stochastic theory of a data-handling system with multiple sources. *BSTJ*, **61** (1982), pp. 1871–1894.

[4] D. Bertsimas, An exact FCFS waiting time analysis for a class of $G/G/s$ queueing systems. *QUESTA*, **3** (1988), pp. 305–320.

[5] D. Bertsimas, An analytic approach to a general class of $G/G/s$ queueing systems. *Opns. Res.*, **38** (1990), pp. 139–155.

[6] D. Bertsimas and X.A. Papaconstantinou, Analysis of the stationary $E_k/C_2/s$ queueing system. *EJOR*, **37** (1988), pp. 272–287.

[7] P.F. Garabedian, *Partial differential equations*, Wiley, Chichester, 1967.

[8] I. Mitrani and D. Mitra, A spectral expansion method for random walks on semi-infinite strips, in: R. Beauwens and P. de Groen (eds.), *Iterative methods in linear algebra*, North-Holland, Amsterdam (1992), 141–149.

[9] M.F. Neuts, *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*, The John Hopkins University Press, Baltimore (1981).

[10] S. Shapiro, The M-server queue with Poisson input and Gamma-distributed service of order two. *Opns. Res.*, **14** (1966), pp. 685–694