Delay in a tandem queueing model with mobile queues : an analytical approximation

Delay in a Tandem Queueing Model with Mobile Queues: An Analytical Approximation

Ahmad Al Hanbali[a]*, Roland de Haan[b], Richard J. Boucherie[b], and Jan-Kees van Ommeren[b]

[a] EURANDOM, QPA group, The Netherlands.
[b] Dep. of Applied Mathematics, SOR group, University of Twente, The Netherlands.

alhanbali@eurandom.tue.nl
haanr,r.j.boucherie,J.C.W.vanOmmeren@ewi.utwente.nl

September 7, 2009

Abstract

In this paper, we analyze the end-to-end delay performance of a tandem queueing system with mobile queues. Due to state-space explosion there is no hope for a numerical exact analysis for the joint-queue length distribution. For this reason, we present an analytical approximation that is based on queue length analysis. Through extensive numerical validation, we find that the queue length approximation exhibits excellent performance for light and moderate traffic load.

Keywords: Tandem queueing model; Mobile queues; Autonomous server; Performance analysis; Delay analysis; Stability; Ad hoc networks.

AMS Classification: 60K25; 68M20.

1 Introduction

In this paper, we analyze the end-to-end delay performance of a customer in a tandem queueing system with mobile queues. This is a typical scenario in the context of mobile ad hoc networks where the wireless devices move autonomously [1, 2].

The network model of our interest is reminiscent of a multi-queue tandem model with multiple alternating servers which move among the queues autonomously. In the literature, it is usually assumed that the server can be controlled. Tandem models with controlled

*The major part of this work was carried out when Ahmad Al Hanbali was at the SOR group at the University of Twente.

servers have been analyzed under various servicing strategies in the special case of a single server (see, e.g., [3]). In a two-queue setting, [4] analyzes the model via boundary value techniques. Unfortunately, the analysis along these lines for more than two queues appears intractable. Time-limited service models with server control have also been studied in the context of polling systems (see, e.g., [5, 6, 7]), where the server moves to another queue when it becomes empty. In the mobility-driven model of our interest, the server is autonomous and there is no possibility to control its movement.

As a primary step towards understanding the impact of mobility on the end-to-end delay, we studied in [8] a model comprising a fixed source and destination queue, and a single mobile queue operating as a relaying device. Modeling this network as a tandem of queues, we performed an exact analysis for the joint queue length by extending the techniques developed in [6] and [9]. Due to the state-space explosion, the computation time of the joint queue length probabilities may grow large for certain model parameters. Therefore, as a complementary tool, we presented an analytical approximation for the case that the service requirements at each queue are exponential. In this paper, we are interested in the model comprising *multiple* mobile queues. Unfortunately, the exact analysis carried out in [8] is numerically intractable in this model due to the increase in the number of queues. For this reason, we will focus on the approximation. As a generalization, we will allow the distribution of the service requirements at the different queues to be general.

Our main interest is in the end-to-end delay in the network described above. The main complexity in our model is the correlation between the queue lengths at different queues. A numerically efficient approximation will be presented. The main idea is to relate the sojourn time at a mobile queue and its queue length process at specific embedded epochs. The queue length process at these embedded epochs is then analyzed in isolation as a discrete-time queue with geometric batch arrivals. The key element is to approximate the batch arrival process with correlated batch sizes with a batch process of independent batch size. This approximation is referred to as *queue length approximation*.

Note that the arrivals to a queue are the departures of the upstream queue in the tandem. Therefore, to derive the queue length of queue i, it is required to first analyze queue $i-1$, and so on. Thus, our approximation is based on an iterative scheme that derives the delay at queue one first, then at queue two, and so on. A similar iterative scheme was used recently in [10] for the analysis of multi-server tandem queues with finite buffers and blocking.

The rest of the paper is organized as follows. Section 2 presents our model. The stability condition of the system is derived in Section 3. In Section 4, we present exact results for the sojourn time in the source queue. Section 5 proposes and analyzes the approximation for the sojourn time in the mobile queue via queue lengths. In Section 6, we numerically validate the accuracy of the approximations and present additional results which give insight in the delay of the network. Section 7 concludes the paper. Proofs of our results are given in Section 8.

2 Model

We consider a tandem model consisting of N first-in-first-out (FIFO) systems with unlimited queue, Q_i, $i = 1, \ldots, N$, in which customers arrive to Q_1 and subsequently require service at Q_2, Q_3,..., and Q_{N-1} before reaching their destination at Q_N. The special feature of the model is that Q_i, $i = 2, \ldots, N-1$, alternates between positions L_{i-1} and L_i such that Q_{i-1}'s server is available for service (i.e. customers at Q_{i-1} are served) only when both Q_{i-1} and Q_i are at L_{i-1} and Q_i's server is available for service when both Q_i and Q_{i+1} are at L_i. The servers of Q_{i-1} and Q_i are two different servers that cannot be serving at the same time. Q_1 and Q_N are fixed and they remain at location L_1 and L_{N-1} respectively. Q_N is a sink and will not be included in our analysis.

Customers arrive to Q_1 according to a Poisson process with arrival rate λ. The service requirement B_i at Q_i has general distribution $B_i(t)$ with mean b_i. We assume that the service requirements are independent and identically distributed (iid) random variables (rvs).

The queues Q_i, $i = 2, \ldots, N-1$, move autonomously. Q_i remains at location L_{i-1} (resp. L_i) a random duration $X_{i,n}^{i-1}$ (resp. $X_{i,n}^{i}$) before it migrates to L_i (resp. L_{i-1}) during its n-th visit, see Figure 1. After the n-th visit to L_{i-1}, Q_i incurs a switch-over time $C_{i,n}^{+}$ from L_{i-1} to L_i, and similarly a switch-over time $C_{i,n}^{-}$ after the n-th visit to L_i. The location of Q_i is driven by an underlying continuous-time, discrete-state, process $\{L_i(t) : t \geq 0\}$ with state-space $\{-2, -1, 0, 1\}$. More precisely, $L_i(t) = 1$ ($L_i(t) = 0$) when Q_i is at L_{i-1} (resp. L_i) at time t, and $L_i(t) = -1$ ($L_i(t) = -2$) when Q_i switches from L_{i-1} to L_i (L_i to L_{i-1}). Without loss of generality, let $L_i(0) = 1$. We assume $\{X_{i,n}^{i-1}, X_{i,n}^{i}, C_{i,n}^{+}, C_{i,n}^{-}\}$ are iid and mutually independent, and also independent of the inter-arrival times and service requirements. We further assume that $X_{i,n}^{i-1}$ ($X_{i,n}^{i}$) is an iid sequence of exponentially distributed rvs with rate α_i^1 (α_i^0).

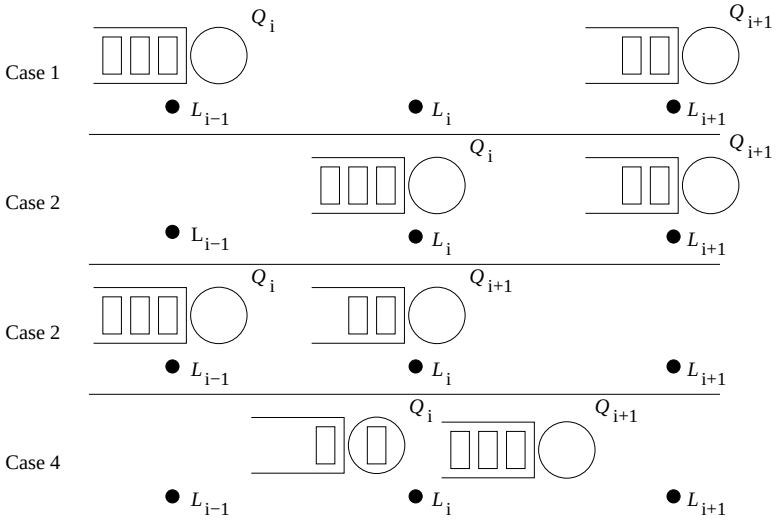


Figure 1: Possible locations of Q_i and Q_{i+1}.

We will refer to the time period during which the server is available for service at Q_i as

Q_i *service period*. Due to the tandem structure, the Q_i *service period* represents the Q_{i+1} *arrival period*, i.e., the period of time during which Q_{i+1} receives customers that completed their service at Q_i. The Q_i service period occurs when the process $\big(L_i(t), L_{i+1}(t)\big)$ is in state $(0,1)$, see Case 4 in Figure 1. The duration of the n-th Q_i service period, denoted by $Y_{i,n}$, is the minimum of the exponentially distributed rvs $X^i_{i,m}$ and $X^i_{i+1,l}$, for some m and l. That is, $Y_{i,n}$ is an iid sequence of exponentially distributed rvs with rate $\xi_i := \alpha^0_i + \alpha^1_{i+1}$. Let Y_i denote the generic rv of $Y_{i,n}$. During a Q_i service period, the server alternates between service and idle states depending on whether or not customers are present at Q_i. When the server is serving a customer at the end of a server visit to Q_i, service will be preempted. At the beginning of the next visit to Q_i, the service time will be re-sampled according to $B_i(\cdot)$. This discipline is commonly referred to as *preemptive-repeat-random*.

In the following, given a continuous rv X, $X(t)$ will denote its distribution function, $\tilde{X}(s)$ its Laplace-Stieltjes Transform (LST) and x its expectation. Similarly, given a discrete rv Y, $Y(n)$ will denote its distribution function, $\hat{Y}(z)$ its probability generating function (p.g.f.) and y its expectation.

The preemptive-repeat-random discipline induces that the amount of work generated by a customer to Q_i, referred to as generalized work, can be written as

$$B^g_i = B^*_i + \sum_{l=1}^{L} Y^*_{i,l}, \tag{1}$$

where B^*_i is the conditional Q_i service time given that it is smaller than Q_i service period, L is the total number of interruptions during the customer's service, and $Y^*_{i,l}$ is the conditional Q_i service period given that it is smaller than Q_i service time. Since a Q_i service period is exponentially distributed, it is easily seen that the distribution of L is geometric with parameter $\mathbb{P}[B_i > Y_i] = 1 - \tilde{B}_i(\xi_i)$. Conditioning on L, the LST of generalized work is

$$\tilde{B}^g_i(s) = \frac{(s+\xi_i)\tilde{B}_i(s+\xi_i)}{s+\xi_i\tilde{B}_i(s+\xi_i)}, \quad Re(s) \geq 0, \tag{2}$$

where $Re(s)$ denotes the real part of s. In particular, its expectation reads

$$\mathbb{E}\big[B^g_i\big] = \frac{1-\tilde{B}_i(\xi_i)}{\xi_i\tilde{B}_i(\xi_i)}. \tag{3}$$

Let $N_i(t)$ denote the number of customers in Q_i, $i = 1, \ldots, N$, at time t. Assume $N_i(0) = 0$, $i = 1, \ldots, N$. Let D_i denote the sojourn time of an arbitrary customer in Q_i, $i = 1, \ldots, N-1$. In the following, we will study $\tilde{D}_i(s)$, the LST of the sojourn time in Q_i, $i = 1, \ldots, N-1$.

3 Stability

Stability is considered on a per-queue basis as service capacity cannot be exchanged between the queues. The system is stable if and only if all the queues in the system are stable.

For an individual queue to be stable, we must have that the average work per unit time brought by a customer to the queue, $\lambda\mathbb{E}[B_i^g]$, is strictly smaller than the average fraction of time the queue server is available for service. In the following, we will derive the average fraction of time the Q_i server is available for service which corresponds to the *probability* that the Q_i server is available. At time t, the Q_i server is available when both Q_i and Q_{i+1} are at location L_i, i.e., when $\big(L_i(t), L_{i+1}(t)\big) = (0,1)$. By a renewal reward argument, we have that

$$\mathbb{P}\big(L_i(t) = k\big) = \frac{\alpha_i^{1-k}}{\alpha_i^1 + \alpha_i^0 + \alpha_i^1\alpha_i^0(c_i^+ + c_i^-)}, \quad k = 0, 1, \tag{4}$$

and since the mobility processes of Q_i and Q_{i+1} are independent, we obtain that

$$\begin{aligned}\mathbb{P}\big(\big(L_i(t), L_{i+1}(t)\big) = (0,1)\big) &= \mathbb{P}\big(L_i(t) = 0\big)\mathbb{P}\big(L_{i+1}(t) = 1\big) \\ &= \prod_{l=i}^{i+1} \frac{\alpha_l^{i+1-l}}{\alpha_l^1 + \alpha_l^0 + \alpha_l^1\alpha_l^0(c_l^{l-1,l} + c_l^{l,l-1})}.\end{aligned} \tag{5}$$

Note that Q_1, the source node, remains always at location L_1, i.e., $L_1(t) = 0$, $t \geq 0$. This can be included in (5) by letting $\alpha_1^1 \to \infty$ and $\alpha_1^0 = 0$. Moreover, since Q_N remains at location L_N, we let $\alpha_N^1 = 0$ and $\alpha_N^0 \to \infty$, so that (5) is valid for $i = 1, \ldots, N$.

Stability condition: Q_i is stable iff

$$\begin{aligned}\rho_i &:= \frac{\lambda\mathbb{E}\big[B_i^g\big]}{\mathbb{P}\big(\big(L_i(t), L_{i+1}(t)\big) = (0,1)\big)} \\ &= \lambda\frac{1 - \tilde{B}_i(\xi_i)}{\xi_i\tilde{B}_i(\xi_i)} \prod_{l=i}^{i+1} \frac{\alpha_l^1 + \alpha_l^0 + \alpha_l^1\alpha_l^0(c_l^{l-1,l} + c_l^{l,l-1})}{\alpha_l^{i+1-l}} < 1,\end{aligned} \tag{6}$$

where ρ_i is referred to as the generalized load at Q_i.

Notice that under stability the arrival rates to Q_{i+1} and Q_i are equal.

4 Exact analysis of queue one

The server visit process is autonomous and the service is according to the preemptive-repeat-random discipline. It is then easily seen that Q_1 in isolation is an M/G/1 queue with on-off server with arrival rate λ, mean service time b_1, exponential on-period X_2^1 with rate α_2^1, and off-period R^{off} equal to the switch-over times plus the Q_2 sojourn time at L_2, i.e., $R^{off} = C_2^{1,2} + C_2^{2,1} + X_2^2$. By a renewal reward argument, P_{on}, the probability that the server is on, satisfies $P_{on} = \mathbb{P}\big(L_2(t) = 1\big)$ which is given in (4), and $P_{off} := 1 - P_{on}$.

The M/G/1 queue with on-off server has extensively been studied in the literature (see, e.g., [11, 12]). Let us state here only the results that are relevant for our analysis. The LST of the sojourn time of a customer is denoted by $\tilde{D}_1(s)$ and follows from a decomposition argument [12]

$$\tilde{D}_1(s) = \tilde{W}_1(s)\tilde{B}_1^{eff}(s)\ , \tag{7}$$

where $\tilde{W}_1(s)$ and $\tilde{B}_1^{eff}(s)$ denote the LST of the waiting time of a customer (until it is taken into service for the first time) and the effective service time (including possible service interruptions), respectively. Note that B_1^{eff} includes the service interruption time and is therefore not equal to B_1^g. The LSTs $\tilde{W}_1(s)$ and $\tilde{B}_1^{eff}(s)$ are given by [8]

$$\tilde{W}_1(s) = \tilde{W}_{M/G/1}(s)(P_{on} + P_{off}\tilde{R}_e^{off}(s)\big), \tag{8}$$

$$\tilde{B}_1^{eff}(s) = \frac{(\alpha_2^1+s)(\alpha_2^0+s)\cdot\tilde{B}_1(\alpha_2^1+s)}{(\alpha_2^1+s)(\alpha_2^0+s) - \alpha_2^1\alpha_2^0(1-\tilde{B}_1(\alpha_2^1+s))\tilde{C}_2^{1,2}(s)\tilde{C}_2^{2,1}(s)}, \tag{9}$$

where $Re(s) \geq 0$, $\tilde{R}_e^{off}(s)$ denotes the LST of the residual time of an off-period, and $\tilde{W}_{M/G/1}(s)$ is the LST of the waiting time in the corresponding M/G/1 queue with service time with LST $\tilde{B}_1^{eff}(s)$.

It follows that $\hat{N}_1(z)$, the p.g.f. of the Q_1 queue length, can be expressed as function of $\tilde{D}_1(s)$, using the so-called functional form of Little's law, as follows (see, e.g., [13])

$$\hat{N}_1(z) = \tilde{D}_1\big(\lambda(1-z)\big), \quad |z| \leq 1. \tag{10}$$

Let us denote by $\hat{N}_1^v(z)$ the p.g.f. of Q_1 queue length at the start time of its service period. It can then be shown by using Eq. (10), the PASTA property and conditioning on the position of the server, that

$$\hat{N}_1^v(z) = \tilde{W}_{M/G/1}\big(\lambda(1-z)\big)\cdot\tilde{B}_1^{eff}\big(\lambda(1-z)\big)\cdot\tilde{R}^{off}\big(\lambda(1-z)\big)\ . \tag{11}$$

Moreover, let $K_{2,n}$ denote the total number of arrivals to Q_2 during its n-th arrival period. Since in our tandem model two successive queues cannot be on service at the same time, the results derived for the p.g.f. of the joint queue length in a time-limited polling model in [14] can be used to find that

$$\hat{K}_{2,n}(z) = \frac{1}{1-z\tilde{B}_1(\alpha_2^1)}\left[1-\tilde{B}_1(\alpha_2^1) + \frac{\alpha_2^1\tilde{B}_1(\alpha_2^1)(1-z)}{\alpha_2^1+\lambda\big(1-\mu(\alpha_2^1,z)\big)}\hat{N}_1^v\big(\mu(\alpha_2^1,z)\big)\right], \tag{12}$$

where $\mu(\alpha_2^1,z)$ is the smallest root of $x = z\tilde{B}_1\big(\alpha_2^1+\lambda(1-x)\big)$ with $|\mu(\alpha_2^1,z)|<1$. $\hat{K}_{2,n}(z)$ will be required later in the approximative analysis for Q_2. In the following, we will study each mobile queue in isolation.

5 Sojourn time approximation via queue length

In this section, we present an approximation for the LST of the sojourn time of a customer in Q_i, denoted by $\tilde{D}_i(s)$, $i = 2,\ldots,N-1$, via queue lengths. We refer to this approximation as the queue length approximation. We consider the queue length process of Q_i when $(L_i(t), L_{i+1}(t)) = (0,1)$, i.e., during a Q_i service period. It turns out that this queue length process corresponds to the waiting time in a Geo/G/1 discrete-time queue with geometric inter-arrival time distribution and general service requirement distribution. The delay D_i follows from adding the total time a customer spends in service to the latter waiting time.

5.1 Queue length of Q_i

Consider the queue length process of Q_i only during its service periods. This is done by removing the time intervals where the Q_i server is not present, i.e., during Q_i off-periods. This new process can be seen as the queue length in a batch arrival queue with inter-arrival times distributed as Q_i service period. Let y_n, $n = 0, 1, \ldots$, denote the ending times of Q_i service periods. Let $N^e_{i,n}$ denote the queue length of Q_i at epoch y_n. Assume that the queue length is left-continuous, i.e., arriving batches are not counted as being in the system until (just) after they arrive.

Let M_n denote the total number of Q_i arrival periods that occur between the n-th and $(n+1)$-st Q_i service period. Note that due to the tandem structure in our model it is clear that the Q_i arrival period represents the Q_{i-1} service period. Let $K^m_{i,n}$ denote the total number of arrivals to Q_i during the m-th Q_i arrival period for $m = 1, \ldots, M_n$. Thus, between the end of the n-th and $(n+1)$-st service period $\sum_{m=1}^{M_n} K^m_{i,n}$ customers arrive to Q_i. So that, in our interpretation of the batch arrival queue with off-service periods removed, at time y_n a batch of size $\sum_{m=1}^{M_n} K^m_{i,n}$ arrives to the queue. Note that it is possible that $M_n = 0$, in this case the batch size is simply equal to zero. Let $E_{i,n+1}$ denote the number of customers that complete their service in Q_i during the $(n+1)$-st service period in the case where at the beginning of this period the Q_i queue length is infinite. A sample path of the evolution of $N_i(t)$ as a function of t is depicted in Figure 2. It is then easily seen that during the n-th cycle, $[y_n, y_{n+1})$, $N^e_{i,n+1}$ can be written as function of $N^e_{i,n}$ as follows

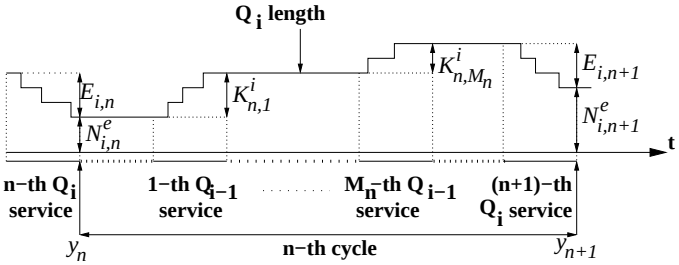


Figure 2: Sample path of the Q_i queue length.

$$N^e_{i,n+1} = \Big(N^e_{i,n} + \sum_{m=1}^{M_n} K^m_{i,n} - E_{i,n+1}\Big)^+, \quad n \geq 0. \tag{13}$$

where $(\cdot)^+ := \max(0, \cdot)$. Recall that $Y_{i,n+1}$ denotes the duration of the $(n+1)$-st Q_i service period that is an exponentially distributed rv with rate $\xi_i = \alpha^0_i + \alpha^1_{i+1}$. Recall that the customers service requirements are iid rvs with general distribution. It follows that $E_{i,n}$, $n = 0, 1, \ldots$, are iid rvs which are geometrically distributed with parameter $\nu_i := \mathbb{P}(B_i < Y_{i,n+1}) = \tilde{B}_i(\xi_i)$. Thus, the probability of the event $\{E_{i,n+1} = l\}$ reads

$$\mathbb{P}(E_{i,n+1} = l) = (1 - \nu_i)\nu_i^l, \quad l = 0, 1, \ldots \ . \tag{14}$$

Note that $E_{i,n+1}$ is independent of $K^m_{i,n}$ and M_n, $K^m_{i,n}$ depends on the index m, however, it is independent of the rv M_n, and that $K^m_{i,n}$ depends on the queue length process of Q_{i-1} during the time interval between the n-th and $(n+1)$-st Q_i service periods. Moreover,

the queue length of Q_{i-1} does not form a Markov chain. For this reason, the rvs $K_{i,n}^m$, $n = 0, 1, \ldots$, $m = 1, \ldots, M_n$, are not independent rvs. In addition, $E_{i,n}$ and $K_{i,n}^m$ are not independent. For the sake of model tractability, we make the following approximating assumption:

Assumption A: $K_{i,n}^m$, $n = 0, 1, \ldots$, $m = 1, \ldots, M_n$ *are iid and also independent of* $\{E_{i,l} : l = 0, 1, \ldots, n\}$.

By **Assumption A**, Eq. (13) represents the waiting time of an arrival in a *discrete-time* single-server queue with inter-arrival times $E_{i,n+1}$ and service requirements $F_{i,n} := \sum_{m=1}^{M_n} K_{i,n}^m$. The main advantage in this model is that the distribution of $E_{i,n+1}$ is geometric. It is known that $\hat{N}_i^e(z)$, the steady-state p.g.f. of $N_{i,n}^e$, is given by (see [15, Corollary 4.3] with U and B equal in distribution and $\gamma = p = \nu_i$)

$$\hat{N}_i^e(z) = \frac{(1 - \nu_i - \nu_i \mathbb{E}[F_i])(z-1)}{z - 1 + \nu_i(1 - z\hat{F}_i(z))}, \quad |z| \leq 1, \tag{15}$$

where $\hat{F}_i(z)$ is the steady-state p.g.f. of $F_{i,n}$. Since $K_{i,n}^m$ is independent of M_n, the p.g.f. $\hat{F}_i(z)$ can be written as follows

$$\hat{F}_i(z) \;=\; \mathbb{E}\Big[\mathbb{E}\big[z^{K_{i,n}}\big]^{M_n}\Big] = \hat{M}_n\big(\hat{K}_{i,n}(z)\big). \tag{16}$$

We emphasize that $\hat{K}_{i,n}(z)$ follows from the analysis of Q_{i-1}. For this reason, to complete the analysis of Q_i, in Section 5.4 we will derive $\hat{K}_{i+1,n}(z)$.

To derive the LST of the sojourn time at Q_i we need $\hat{N}_i^c(z)$, the p.g.f. of Q_i queue length seen by an arbitrary customer, and $\hat{M}_n(z)$, which will be determined in Section 5.2.

Lemma 1. *The p.g.f. of the queue length of Q_i seen by an arbitrary arriving customer is given by*

$$\hat{N}_i^c(z) = \hat{N}_i^e(z) \frac{z\big(1 - \hat{F}_i(z)\big)}{(1-z)\mathbb{E}[F_i]}. \tag{17}$$

Proof. Let $\hat{N}_i^j(z)$ denote the p.g.f. of Q_i queue length seen by the j-th customer within a batch upon arrival including himself. Since the size of the batches is independent of the queue length of Q_i present upon arrival, $\hat{N}_i^j(z)$ reads,

$$\hat{N}_i^j(z) = z\hat{N}_i^{j-1}(z), \quad j = 1, 2, \ldots, \tag{18}$$

with $\hat{N}_i^0(z) = \hat{N}_i^e(z)$. The probability, $\mathbb{P}(J = j)$, that a customer is the j-th customer within the batch is equal to the fraction of customers who are j-th arrival in their own batch, which gives

$$\mathbb{P}(J = j) = \frac{\mathbb{P}(F_i \geq j)}{\mathbb{E}[F_i]}. \tag{19}$$

Removing the condition on the customer position in a batch in (18) and using (19) gives the desired result. □

As can be seen $\hat{N}_i^c(z)$ is function of $\hat{F}_i(z)$ and eventually of $\hat{M}_n(z)$. Now, we derive $\hat{M}_n(z)$.

5.2 P.g.f. of M_n

The rv M_n only depends on the mobility process of Q_{i-1}, Q_i, and Q_{i+1} and can be fully represented as the number of visits to a state in a Markov chain. For clarity of presentation, we will restrict ourselves to the simple case where the switch-over times are equal to zero. We emphasize that similar analysis can be done when the switch-over time distribution is phase-type for which the cardinality of the state space is enlarged, see the discussion in Section 7.

The p.g.f. of M_n can be written as follows:

$$\hat{M}_n(z) \quad = \quad \mathbb{P}(M_n = 0) + \big(1 - \mathbb{P}(M_n = 0)\big)\hat{M}_n^+(z), \tag{20}$$

where M_n^+ is M_n given that it is strictly positive. In the following lemmas, we will first derive $\hat{M}_n^+(z)$ and next $\mathbb{P}(M_n = 0)$.

Lemma 2. *The p.g.f. of M_n^+ is*

$$\hat{M}_n^+(z) \quad = \quad -bz(\mathbf{A} + z\mathbf{B})^{-1}u, \quad |z| \leq 1, \tag{21}$$

where

$$\mathbf{A} = \begin{pmatrix} A_{11} & 0 & 0 & \alpha_{i-1}^0 & 0 & 0 \\ \alpha_i^1 & A_{22} & \alpha_{i+1}^0 & 0 & \alpha_{i-1}^0 & 0 \\ 0 & \alpha_{i+1}^1 & A_{33} & 0 & 0 & \alpha_{i-1}^0 \\ \alpha_{i-1}^1 & 0 & 0 & A_{44} & \alpha_i^0 & 0 \\ 0 & 0 & 0 & \alpha_i^1 & A_{55} & \alpha_{i+1}^0 \\ 0 & 0 & 0 & 0 & \alpha_{i+1}^1 & A_{66} \end{pmatrix},$$

$$\mathbf{B} = \begin{pmatrix} 0 & \alpha_i^0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \alpha_{i-1}^1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \alpha_{i-1}^1 & 0 & 0 & 0 \end{pmatrix}, \quad u = \begin{pmatrix} \alpha_{i+1}^0 \\ 0 \\ \alpha_i^1 \\ \alpha_{i+1}^0 \\ 0 \\ \alpha_i^1 \end{pmatrix},$$

and where the diagonal entries of $\mathbf{A}$ are such that $(\mathbf{A} + \mathbf{B})e + u = \mathbf{0}$. The vector b is the row vector of order six and of non-zero entries

$$b(2) = h(1), \quad b(3) = 1 - h(1),$$

$$h \quad = \quad -\frac{(\alpha_{i-1}^1\alpha_{i+1}^1, 0, \alpha_{i-1}^0\alpha_{i+1}^1, 0, 0, \alpha_{i-1}^0\alpha_i^0)}{(\alpha_{i-1}^0 + \alpha_{i-1}^1)(\alpha_i^0 + \alpha_{i+1}^1)} \cdot \mathbf{H}^{-1} \cdot \mathbf{V},$$

$$\mathbf{H} = \begin{pmatrix} H_{11} & \alpha_{i+1}^0 & \alpha_{i-1}^0 & 0 & 0 & 0 \\ \alpha_{i+1}^1 & H_{22} & 0 & \alpha_{i-1}^0 & 0 & 0 \\ \alpha_{i-1}^1 & 0 & H_{33} & \alpha_{i+1}^0 & \alpha_i^0 & 0 \\ 0 & \alpha_{i-1}^1 & \alpha_{i+1}^1 & H_{44} & 0 & \alpha_i^0 \\ 0 & 0 & \alpha_i^1 & 0 & H_{55} & \alpha_{i+1}^0 \\ 0 & 0 & 0 & \alpha_i^1 & \alpha_{i+1}^1 & H_{66} \end{pmatrix}, \mathbf{V} = \begin{pmatrix} \alpha_i^0 & 0 \\ 0 & \alpha_i^0 \\ 0 & 0 \\ 0 & 0 \\ \alpha_{i-1}^1 & 0 \\ 0 & \alpha_{i-1}^1 \end{pmatrix}.$$

The diagonal entries of $\mathbf{H}$ are such that $\mathbf{H}e + \mathbf{V} \times (1,1)^T = \mathbf{0}$, where e is the column vector of order six and with all entries equal to 1, and x^T is the transpose of vector x.

Proof: See Appendix 8.1.

Lemma 3. *The probability that no Q_i arrival period occurs during the n-th cycle reads*

$$\mathbb{P}(M_n = 0) \quad = \quad -g \cdot \mathbf{F}^{-1} \cdot w, \tag{22}$$

where

$$g \quad = \quad \frac{1}{(\alpha_{i-1}^0 + \alpha_{i-1}^1)(\alpha_i^0 + \alpha_{i+1}^1)}(\alpha_{i-1}^1\alpha_{i+1}^1, \alpha_{i-1}^0\alpha_{i+1}^1, 0, \alpha_{i-1}^0\alpha_i^0),$$

$$\mathbf{F} \quad = \quad \begin{pmatrix} F_{11} & \alpha_{i-1}^0 & 0 & 0 \\ \alpha_{i-1}^1 & F_{22} & \alpha_i^0 & 0 \\ 0 & \alpha_i^1 & F_{33} & \alpha_{i+1}^0 \\ 0 & 0 & \alpha_{i+1}^1 & F_{44} \end{pmatrix}, \; w = \begin{pmatrix} \alpha_{i+1}^0 \\ \alpha_{i+1}^0 \\ 0 \\ \alpha_i^1 \end{pmatrix}.$$

The diagonal entries of $\mathbf{F}$ *are such that*

$$\mathbf{F}e + w + (\alpha_i^0, 0, \alpha_{i-1}^1, \alpha_{i-1}^1)^T = \mathbf{0}.$$

Proof: See Appendix 8.2.

5.3 Sojourn time in Q_i

Recall that D_i, the sojourn time in Q_i, consists of two parts: the time required to serve N_i^c customers, and the time a customer is in Q_i but Q_i is not served. Let B_i^{eff} denote the effective service time at Q_i, $i = 2, \ldots, N-1$, that starts when a customer receives the service for the first time and ends when the customer departs from Q_i. Clearly, B_i^{eff} includes the time when the Q_i service is interrupted. Let L denote the total number of interruptions during the service of a customer. It is easily seen that B_i^{eff} can be written as

$$B_i^{eff} = B_i^* + \sum_{l=1}^{L} \Big(Y_{i,l}^* + \Xi_{i,l} \Big), \tag{23}$$

where B_i^* is the conditional B_i given that it is smaller than Y_i, the exponential rv with rate $\xi_i = \alpha_i^0 + \alpha_{i+1}^1$, $Y_{i,l}^*$ is the conditional Y_i given that it is smaller than B_i, and $\Xi_{i,l}$ is the duration of the service interruption in Q_i. Let $\tilde{\Xi}_i(s)$ denote the steady-state LST of $\Xi_{i,l}$. Since we are considering the preemptive-repeat discipline, the distribution of L is geometric with parameter $\mathbb{P}[B_i > Y_i] = 1 - \tilde{B}_i(\xi_i)$. Conditioning on L, we find the LST of B_i^{eff} that reads

$$\tilde{B}_i^{eff}(s) = \frac{(\xi_i + s) \cdot \tilde{B}_i(\xi_i + s)}{(\xi_i + s) - \xi_i(1 - \tilde{B}_i(\xi_i + s))\tilde{\Xi}_i(s)}, \quad i = 2, \ldots, N-1, \quad Re(s) \geq 0. \tag{24}$$

An arriving customer to Q_i joins the queue when Q_i is not served. Therefore, the customer has to wait for the server to return to Q_i in order that her service starts. This occurs when Q_i and Q_{i+1} are both at location L_i. Let Ξ_i^* denote the first time after t that the server returns to Q_i given that an arrival joins Q_i at t. In the following lemmas, we give the LST of Ξ_i and Ξ_i^*.

Lemma 4. *The LST of* Ξ_i *is*

$$\tilde{\Xi}_i(s) = y(s\mathbf{I} - \mathbf{A} - \mathbf{B})^{-1}u, , \quad Re(s) \geq 0, \tag{25}$$

where

$$y = \frac{1}{(\alpha_{i-1}^0 + \alpha_{i-1}^1)(\alpha_i^0 + \alpha_{i+1}^1)}(\alpha_{i-1}^1\alpha_{i+1}^0, 0, \alpha_{i-1}^1\alpha_i^0, \alpha_{i-1}^0\alpha_{i+1}^1, 0, \alpha_{i-1}^0\alpha_i^0), \tag{26}$$

$\mathbf{A}$, $\mathbf{B}$, *and* u *are given in Lemma 2.*

Proof: See Appendix 8.3.

Lemma 5. *The LST of* Ξ_i^* *is*

$$\tilde{\Xi}_i^*(s) = y^*(s\mathbf{I} - \mathbf{A} - \mathbf{B})^{-1}u, \quad Re(s) \geq 0, \tag{27}$$

where

$$y^* = \frac{1}{\alpha_{i+1}^0 + \alpha_{i+1}^1}(0, \alpha_{i+1}^1, \alpha_{i+1}^0, 0, 0, 0). \tag{28}$$

Proof: See Appendix 8.4.

We are now ready to formulate our main result for the sojourn time in Q_i, the queue length approximation.

Theorem 1. (Sojourn time via queue length)
Under **Assumption A**, *the sojourn time in* Q_i *is*

$$D_i = \Xi_i^* + \sum_{i=1}^{N_i^c} B_i^{eff}. \tag{29}$$

The LST of D_i *reads*

$$\tilde{D}_i(s) = \tilde{\Xi}_i^*(s)\hat{N}_i^c\big(\tilde{B}_i^{eff}(s)\big), \quad Re(s) \geq 0. \tag{30}$$

Proof: Eq. (29) is due to the fact that the queue length of Q_i seen by an arriving customer is N_i^c (including himself) and the customer in service has to wait for Ξ_i^* before that the service restarts in Q_i.

Since N_i^c depends on the history of the Markov chain $\big(L_{i-1}(t), L_i(t), L_{i+1}(t)\big)$ and B_i^{eff} depends on the future of $\big(L_{i-1}(t), L_i(t), L_{i+1}(t)\big)$ the rvs N_i^c and B_i^{eff} are independent. Moreover, B_i^{eff} is independent of Ξ_i^*, e.g., see (23), and N_i^c is independent Ξ_i^*. All these independencies together readily give (30). □

Remark 1. *For the exponential service times, we note that in [8] we proposed a different approximation of the sojourn time at* Q_i *via the workload analysis in the queue. We emphasize that the queue length approximation proposed in this paper is much easier to derive and to extend to the general service times distribution.*

5.4 P.g.f. of $K_{i+1,n}$

In our tandem model, we note that the arrivals to a queue are the departures of the upstream queue. Therefore, to derive the queue length of Q_i, it is required to first analyze Q_{i-1}, and so on. For this reason, we emphasize that in our iterative scheme the p.g.f. $\hat{K}_{i,n}(z)$ should be computed in the analysis of Q_{i-1}. Therefore, to close the iteration loop of Q_i, we will derive in the following $\hat{K}_{i+1,n}(z)$. The rv $K_{i+1,n}$ represents the total number of arrivals to Q_{i+1} during a Q_i service period. Let N_i^v denote the queue length of Q_i just after the beginning of a Q_i service period. Therefore, N_i^v is the sum of N_i^e, the queue length of Q_i seen by arriving batch, and F_i, the batch size. Note that during a Q_i server period, there are no arrivals to Q_i and the distribution of the duration of that server period is Y_i, an exponential rv with rate $\xi_i = \alpha_i^0 + \alpha_{i+1}^1$. Consequently, using (12) with $\lambda \to 0$ and replacing α_2^1 by ξ_i and $\tilde{B}_1(s)$ by $\tilde{B}_i(s)$ gives

$$\hat{K}_{i+1,n}(z) \;=\; \frac{1}{1-\tilde{B}_i(\xi_i)z}\Big[\tilde{B}_i(\xi_i)(1-z)\hat{N}_i^v\Big(\tilde{B}_i(\xi_i)z\Big) + 1 - \tilde{B}_i(\xi_i)\Big], \tag{31}$$

where $\hat{N}_i^v(z) := \hat{N}_i^e(z)\hat{F}_i(z)$, which are given in (15) and (16).

6 Numerical results

We consider a tandem network of N queues including the source and the destination queue. The mean service time at Q_i is equal to $b_i = b$ for $i = 1, \ldots, N$. Recall that Q_i remains at locations L_{i-1} and L_i an exponentially distributed period of time with rate α_i^1 and α_i^0. We will consider that case where $\alpha_i^0 = \alpha_i^1 = \alpha_i$. The queues Q_1 and Q_N remain always at locations L_1 and L_N, respectively. We assume that the switch-over times incurred when queues alternate between locations are equal to zero. Our objectives are to validate the approximations and to give insights into the sojourn time behavior as a function of the system parameters.

To validate our approximation we will compare its results with those of the simulation. The simulation of the above tandem model scenario was implemented in the C++ programming language. To generate the random variables we used the pseudo-random generator package of C++. We note that a simulation result consists of an average over multiple runs with different seeds. The number of runs considered is high enough in order to guarantee a small 95% confidence interval.

Let $\mathbb{E}[D_i^{ql}]$ denote the mean sojourn time in Q_i using the queue length approximation given in (30). Let $\mathbb{E}[D_i^{sim}]$ denote the mean sojourn time in Q_i using simulation for the tandem network. Moreover, let us refer to the relative difference between the mean sojourn time at Q_i using the approximation and simulation as follows.

$$\tau_i^{ql}(\%) := 100 \times \Big|1 - \frac{\mathbb{E}[D_i^{ql}]}{\mathbb{E}[D_i^{sim}]}\Big|.$$

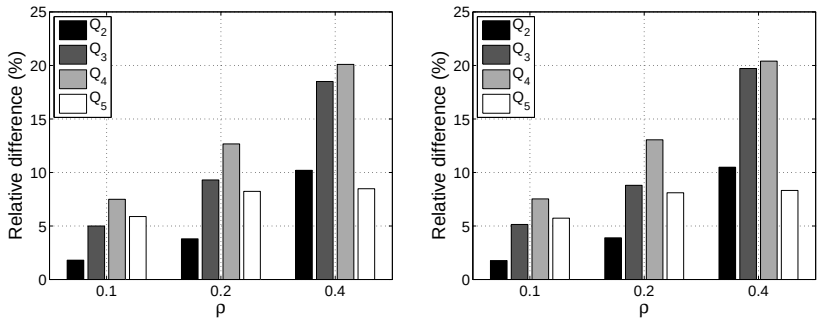
In the sequel, we will consider four different service time distributions: the deterministic distribution, the Erlang-2 distribution, the exponential distribution, and the two-phase

hyper-exponential distribution. The two-phase hyper-exponential distribution is uniquely determined by its mean value b, and by the mean m_1 and probability p_1 of the first phase.

6.1 Accuracy of queue length approximation vs. load

In this section, we study the accuracy of the mean sojourn time in Q_i using the queue length approximation by comparing it to the simulation results as function of the queue load ρ_i. This will be done for both the symmetric case when $\alpha_i = \alpha$ for $i = 2, \ldots, N-1$, and asymmetric case when $\alpha_i \neq \alpha_j$ for some i and j.

Symmetric case: we consider a tandem network of six queues, i.e., $N = 6$, with mean service time $b = 1$ and $\alpha_i = 0.05$, $i = 2, \ldots, 5$. Note that in the case of exponential services the load at the queues satisfies $\rho_2 = \rho_3 = \rho_4 = 2\rho_5 = \rho$. However, in the case of deterministic or hyper-exponential the load at the queues satisfies $\rho_2 = \rho_3 = \rho_4 = \rho \approx 2\rho_5$, see Eq. (6). Figures 3 and 4 show the relative difference as function of ρ for exponential, deterministic, and hyper-exponential service distribution. Observe that τ_i^{ql} is smaller than 20% for $\rho \leq 0.4$ and for all service distributions. For this reason, the queue length approximation is accurate in the cases of light and moderate load at Q_i. Moreover, we note that the accuracy of the approximation is almost the same for the considered service distributions.



(a) (b)

Figure 3: Relative difference between the queue length approximation and the simulation as function of ρ for $\alpha = 0.05$ and $b = 1$ with: (a) exponential service requirement, (b) deterministic service requirement.

Asymmetric case: we consider a tandem network with $N = 7$ queues including source and destination. Our objective is to show that the approximated and simulated mean sojourn time follow the same pattern for $i = 2, \ldots, 6$. We consider two different settings for $\{\alpha_2, \ldots, \alpha_5\}$: the allocation set $A = \{0.05, 0.025, 0.1, 0.0375, 0.05\}$ and the set $B = \{0.05, 0.1, 0.15, 0.2, 0.05\}$. Figure 5 displays the mean sojourn time at Q_i as function of α_i for exponential service requirement. Observe that the approximation predicts the behavior of the simulation very well. Moreover, the queues with the highest and lowest mean sojourn time are the same in the simulation and approximation. These observations also hold for the hyper-exponential service distribution as shown in Figure 6.

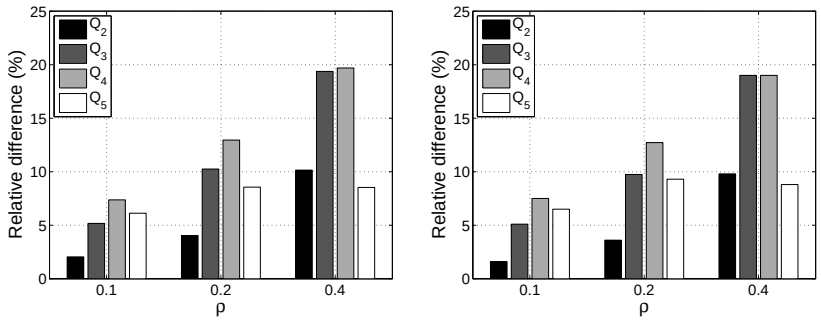


Figure 4: Relative difference between the approximation and the simulation as function of ρ for $\alpha = 0.05$ and $b = 1$ with: (a) hyper-exponential service time with $p_1 = 0.6$, $m_1 = 0.1$, and SCV= 3.43, (b) hyper-exponential service time with $p_1 = 0.8$, $m_1 = 0.1$, and SCV= 7.48.

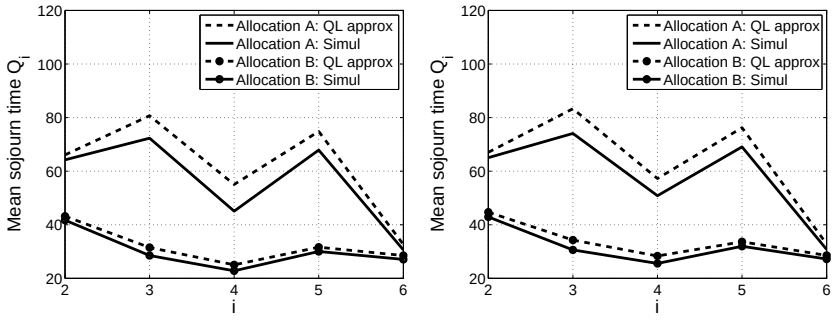


Figure 5: Mean sojourn time at Q_i using queue length approximation and simulation for $\lambda = 0.05$ and $b = 1$ with: (a) exponential service time, (b) deterministic service time.

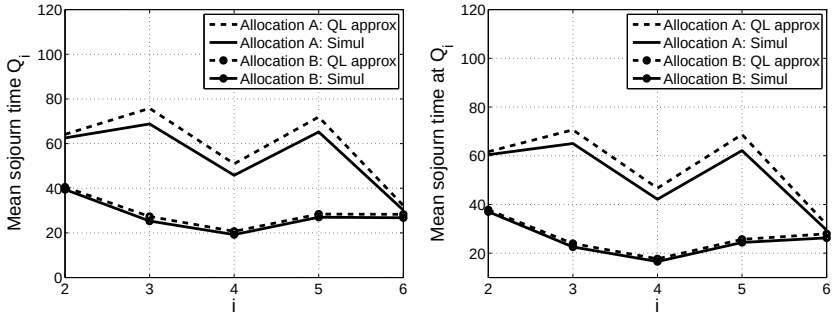
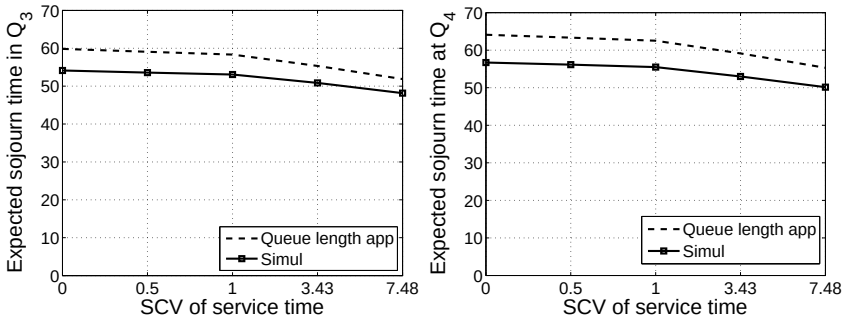


Figure 6: Mean sojourn time at Q_i using queue length approximation and simulation for $\lambda = 0.05$ and $b = 1$ with: (a) hyper-exponential service time with $p_1 = 0.6$ and $m_1 = 0.1$, and SCV= 3.43, (b) hyper-exponential service time with $p_1 = 0.8$, $m_1 = 0.1$, and SCV= 7.48.

6.2 Mean approximate sojourn time vs. service times distribution

Let us check the behavior of the queue-length approximation as function of the service times distribution. We consider the symmetric scenario of a tandem network of six queues, i.e., $N = 6$, with mean service requirement $b = 1$ and $\alpha_i = 0.05$, $i = 2, \ldots, 5$. Figure 7 displays the expected sojourn time at Q_3 and Q_4 using the approximation as function of the square coefficient of variation (SCV) of the service times. For $\lambda = 0.05$ and $\lambda = 0.1$ respectively, observe that the accuracy of the queue length approximation is almost insensitive of the SCV of the service times. Furthermore, for all parameter values considered the approximated mean delay in Q_i gives an upper bound of the simulated mean delay in Q_i. This observation is in support of the result in [16] which proves that in the correlated M/G/1 queue a positive correlation between the service time and the last inter-arrival time reduces the mean sojourn time. We should emphasize that in our model $K_{i,n}^m$ and the last inter-arrival time are positively correlated, i.e., an increase of the last inter-arrival time induces stochastically an increase of $K_{i,n}^m$.



(a) (b)

Figure 7: Expected sojourn time at Q_3 and Q_4 as function of the SCV of the service times for $\alpha = 0.05$ and $\lambda = 0.05$.

6.3 Squared coefficient of variation of sojourn time in Q_i

Next, we compare the squared coefficient of variation σ_i of the sojourn time at Q_i, $\sigma_i := \mathbb{V}\mathrm{ar}[D_i]/\mathbb{E}[D_i]^2$, following from the queue length approximations with the simulation denoted as σ_i^{ql} and σ_i^{sim}, respectively. Tables 1, 2, and 3 show σ_i^{ql} and σ_i^{sim}, and also the second moments $\mathbb{E}\big[(D_i^{ql})^2\big]$ and $\mathbb{E}\big[(D_i^{sim})^2\big]$ for the exponential, the deterministic, and the hyper-exponential service times distribution. Observe that the squared coefficient of variation of the approximations are accurate.

6.4 Impact of α_i on mean sojourn time

Our objective is to show the impact of α_i on the mean sojourn time at Q_i as function of the service time distribution. We consider the symmetric case where $\alpha_i = \alpha$. Table 4 shows the mean sojourn time at Q_i as function of α in the case of exponential service times. Note that for $\lambda = 0.075$ and $b = 1$, the load ρ_i, $i = 2, \ldots, 5$, is equal to 0.3 and

	Q_2	Q_3	Q_4	Q_5	Q_6
α_i	0.05	0.1	0.15	0.2	0.05
σ_i^{ql}	0.785	0.811	0.830	1.084	0.791
σ_i^{sim}	0.778	0.782	0.795	1.099	0.801
$\mathbb{E}\left[(D_i^{ql})^2\right]$	3328.0	1796.4	1147.9	2080.9	1453.0
$\mathbb{E}\left[(D_i^{sim})^2\right]$	3094.0	1449.8	933.9	1891.0	1318.8
α_i	0.05	0.025	0.1	0.0375	0.05
σ_i^{ql}	0.998	0.761	0.981	0.815	0.761
σ_i^{sim}	1.022	0.741	0.985	0.795	0.756
$\mathbb{E}\left[(D_i^{ql})^2\right]$	8726.9	11466.6	6000.3	10163.8	1875.6
$\mathbb{E}\left[(D_i^{sim})^2\right]$	8338.7	9098.3	4780.2	8266.8	1643.7

Table 1: Coefficient of variation and second moment of the sojourn time at Q_i using queue length and workload approximation and simulation for: $\lambda = 0.05$, exponential service with $b = 1$, for the α_i allocation set A (Top) and B (bottom).

	Q_2	Q_3	Q_4	Q_5	Q_6
α_i	0.05	0.1	0.15	0.2	0.05
σ_i^{ql}	0.775	0.799	0.817	1.052	0.780
σ_i^{sim}	0.767	0.763	0.769	1.060	0.791
$\mathbb{E}\left[(D_i^{ql})^2\right]$	3542.7	2112.5	1465.4	2322.6	1449.5
$\mathbb{E}\left[(D_i^{sim})^2\right]$	3254.1	1650.2	1155.2	2102.5	1324.8
α_i	0.05	0.025	0.1	0.0375	0.05
σ_i^{ql}	0.99	0.755	0.968	0.809	0.748
σ_i^{sim}	1.012	0.734	0.964	0.787	0.742
$\mathbb{E}\left[(D_i^{ql})^2\right]$	8941.5	12182.8	6441.7	10513.9	1882.0
$\mathbb{E}\left[(D_i^{sim})^2\right]$	8511.1	9513.6	5075.6	8650.5	1650.1

Table 2: Coefficient of variation and second moment of the sojourn time at Q_i using queue length and workload approximation and simulation for: $\lambda = 0.05$, deterministic service with $b = 1$, for the α_i allocation set A (Top) and B (bottom).

	Q_2	Q_3	Q_4	Q_5	Q_6
α_i	0.05	0.1	0.15	0.2	0.05
σ_i^{ql}	0.805	0.833	0.854	1.142	0.816
σ_i^{sim}	0.897	0.903	0.912	1.081	0.907
$\mathbb{E}\left[(D_i^{ql})^2\right]$	2958.7	1368.5	793.9	1729.5	1456.0
$\mathbb{E}\left[(D_i^{sim})^2\right]$	2805.4	1162.1	675.9	1582.4	1306.1
α_i	0.05	0.025	0.1	0.0375	0.05
σ_i^{ql}	1.015	0.773	1.010	0.826	0.789
σ_i^{sim}	1.018	0.871	1.011	0.899	0.885
$\mathbb{E}\left[(D_i^{ql})^2\right]$	8288.3	10180.8	5214.7	9449.5	1857.9
$\mathbb{E}\left[(D_i^{sim})^2\right]$	7956.4	8321.6	4241.5	7687.5	1620.7

Table 3: Coefficient of variation and second moment of the sojourn time at Q_i using queue length and workload approximation and simulation for: $\lambda = 0.05$, hyper-exponential service with $b = 1$, $p_1 = 0.6$, and $m_1 = 0.1$, for the α_i allocation set A (Top) and B (bottom).

$\rho_6 = 0.15$. Observe that the mean sojourn time decreases at Q_i with α. Moreover, the mean sojourn time at Q_i, $i = 2, \ldots, 5$, converges to the mean sojourn time in an M/M/1 queue with load 0.3 and arrival rate $\lambda = 0.075$ that is equal to 5.71. A similar result holds for Q_6 which gives that its limiting mean sojourn time is equal to 2.38. Table 5 displays the mean sojourn time at Q_i as function of α in the case of deterministic service. The mean sojourn time of the deterministic service as function of α has an optimum value for α. Additional experiments show that this optimum is around 0.4. The hyper-exponential service gives similar results as the case of exponential service. That is, the mean sojourn time in Q_i, $i = 2, \ldots, 5$, is decreasing with α and it converges to a limit value, which is approximately equal to the mean sojourn time in an M/M/1 queue with arrival rate λ and load ρ_i. For the deterministic service, we note that the optimal value of α is sensitive to the value of λ and b in such a way that the higher the load at the queue the smaller the optimal value of α.

α	ρ_2	Q_2	Q_3	Q_4	Q_5	Q_6
0.05	0.3	60.55	71.64	76.54	78.14	35.62
0.1	0.3	33.23	38.04	41.43	42.42	19.01
0.2	0.3	19.55	22.25	23.36	23.75	10.75
0.4	0.3	12.69	13.96	14.39	14.52	6.34
0.8	0.3	9.24	9.82	9.96	10.00	4.29
50	0.3	5.77	5.78	5.78	5.79	2.38

Table 4: Mean sojourn time at Q_i using the queue length approximation as function of α for: $\lambda = 0.075$, exponential service with $b = 1$.

α	ρ_2	Q_2	Q_3	Q_4	Q_5	Q_6
0.05	0.315	62.49	74.37	23.52	20.95	22.99
0.1	0.33	35.77	42.21	44.94	45.91	19.34
0.2	0.37	23.52	27.10	28.31	28.67	11.03
0.4	0.46	20.98	23.30	23.79	23.90	7.24
0.8	0.74	49.57	53.01	53.37	53.58	6.26

Table 5: Mean sojourn time at Q_i using the queue length approximation as function of α for: $\lambda = 0.075$, deterministic service with $b = 1$.

α	ρ_2	Q_2	Q_3	Q_4	Q_5	Q_6
0.05	0.27	56.83	66.39	70.96	72.58	35.00
0.1	0.24	29.29	33.58	35.87	36.89	18.31
0.2	0.21	15.33	17.05	17.97	18.39	9.64
0.4	0.16	8.19	8.82	9.15	9.31	5.2
0.8	0.124	4.5	4.73	4.84	4.89	2.88
50	0.05	0.74	0.74	0.75	0.75	0.38

Table 6: Mean sojourn time at Q_i using the queue length approximation as function of α for: $\lambda = 0.075$, hyper-exponential service with $b = 1$, $p_1 = 0.6$, and $m_1 = 0.1$.

6.5 Large tandem model

In Table 7, we show the mean sojourn time at Q_i and the relative difference τ_i between approximation and simulation as function of N for $\lambda = 0.1$, exponential service with $b = 0.5$, and $\alpha = 0.05$. We note that up to a certain threshold the relative difference is increasing and it starts to decrease slightly after that value. Remark that the sojourn time approximation at Q_{N-1} is more accurate since it is next to the destination queue.

7 Conclusion and possible generalization

In this paper, we have addressed the performance of a tandem queueing system with mobile queues. We have proposed an analytical approximation for the LST of the delay in Q_i. The approximation is called queue length approximation. Through extensive numerical validation we have shown that the queue length approximation gives nice results for light and moderate load in the case of general service times distribution.

For the sake of clarity, we restricted ourselves in Section 5.2 to the case where the switch-over times are zero. As a generalization, we consider here the case of non-zero switch-over times. In particular, we assume that when Q_i migrates from location L_i to L_{i-1} it requires an exponentially distributed switch-over time with mean c_i^-. Similarly, when Q_i migrates from location L_{i-1} to L_i it requires an exponentially distributed switch-over time with mean c_i^+. The state space of the Markov chain $\big(L_{i-1}(t), L_i(t), L_{i+1}(t)\big)$ is equal to $\Omega = \{-2, -1, 0, 1\}^3$. Following the footprints of Section 5.2, one can easily show that the $\hat{M}_n(z)$ has exactly the same form as depicted in Lemmas 2 and 3. The matrices $\mathbf{A}$ and

	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8
$N=5$ $\mathbb{E}[D_i^{ql}]$	48.94	55.94	29.51				
$N=5$ τ_i^{ql}	3.6	10.3	7.5				
$N=6$ $\mathbb{E}[D_i^{ql}]$	48.94	55.94	60.32	30.53			
$N=6$ τ_i^{ql}	3.6	10.3	13.3	8.6			
$N=7$ $\mathbb{E}[D_i^{ql}]$	48.94	55.94	60.32	62.63	31.02		
$N=7$ τ_i^{ql}	3.6	10.3	13.3	14.5	8.6		
$N=8$ $\mathbb{E}[D_i^{ql}]$	48.94	55.94	60.32	62.63	63.73	31.24	
$N=8$ τ_i^{ql}	3.6	10.3	13.3	14.6	14.2	8	
$N=9$ $\mathbb{E}[D_i^{ql}]$	48.94	55.94	60.32	62.63	63.73	64.24	31.34
$N=9$ τ_i^{ql}	3.6	10.3	13.3	14.6	14.2	13.6	7.2

Table 7: Mean sojourn time at Q_i and relative difference of queue length approximation as function of the tandem network size N for: $\lambda = 0.1$, exponential service with $b = 0.5$, and $\alpha = 0.05$.

B in this case have a much larger dimension. More precisely, **A** (resp. **B** and **H**) is a 60-by-60 matrix.

In this paper we restricted ourselves to the Tandem model case. The case of a general network of queues with fork and join traffic and with mobile queues remains an open problem to be addressed in the future. Moreover, we considered the case where there is a single mobile node moving between two consecutive locations. The scenario of multiple mobile nodes moving between two consecutive locations is important to address some applications such as in vehicular networks where for example the mobile nodes represent the busses moving between to stations.

Acknowledgment

In the Netherlands, the 3 universities of technology have formed the 3TU.Federation. This article is the result of joint research in the 3TU.Centre of Competence NIRICT (Netherlands Institute for Research on ICT). The authors would thank De Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) for their financial support.

8 Appendix: Proofs

In this section, we will use the theory of the finite-state continuous-time absorbing Markov chains to compute $\hat{M}_n(z)$. Lemma 6 summarizes some known results, e.g. see [17], of this theory that will be used afterwards.

Lemma 6. *Consider a finite-state, continuous-time, Markov chain* $\{MC(t), t \geq 0\}$*, with*

state space $\zeta = \{1, \cdots, m+n\}$ *and with infinitesimal generator matrix,* $\mathbf{G}$*, of the form*

$$\mathbf{G} = \left(\begin{array}{c|c} \mathbf{U} & \mathbf{V} \\ \hline \mathbf{0}_m & \mathbf{0}_n \end{array} \right),$$

where $\mathbf{U}$ *is an* m*-by-*m *matrix,* $\mathbf{V}$ *is an* m*-by-*n *matrix,* $\mathbf{0}_m$ *is an* n*-by-*m *matrix with entries equal to* 0*,* $\mathbf{0}_n$ *is an* n*-by-*n *matrix with entries equal to* 0*. The states* $\{m+1, \cdots, m+n\}$ *are absorbing. Then,*

(a) the states $\{1, \cdots, m\}$ *are all transient if and only if* $\mathbf{U}$ *is a non-singular matrix.*

(b) the probability distribution, $F(.)$*, of the time until absorption in one of the absorbing states* $\{m+1, \cdots, m+n\}$*, given that* $MC(0) = i$*,* $i = 1, \ldots, m$*, reads*

$$F(t) = 1 - \alpha_i exp(\mathbf{U}t)e, \quad t \geq 0, \tag{32}$$

where α_i *is the* m*-dimensional row vector with entries equal to* 0 *except the* i*-th one that is equal to* 1*,* e *is the* m*-dimensional column vector with entries all equal to* 1*, and where*

$$exp(\mathbf{U}t) := \sum_{i=0}^{\infty} \frac{(\mathbf{U}t)^i}{i!},$$

with $(\mathbf{U}t)^0 = \mathbf{I}_m$ *the* m*-by-*m *identity matrix. Similarly, the Laplace-Stieltjes Transform,* $\tilde{F}(s)$*, of the time until absorption in one of the states* $\{m+1, \cdots, m+n\}$*, given that* $MC(0) = i$*,* $i = 1, \ldots, m$*, reads*

$$\tilde{F}(s) = \alpha_i (s\mathbf{I}_m - \mathbf{U})^{-1} \mathbf{V} e_n, \quad Re\ s \geq 0, \tag{33}$$

where e_n *is the* n*-dimensional column vector with entries are all equal to* 1*.*

(c) given that $MC(0) = i$*, the expected amount of time spent in the transient state* j *is equal to the (i,j)-entry of* $-\mathbf{U}^{-1}$*,* $i, j = 1, \ldots, m$*.*

(d) given that $MC(0) = i$*, the probability that absorption occurs in state* j *is equal to the (i,j)-entry of* $-\mathbf{U}^{-1}\mathbf{V}$*,* $i = 1, \ldots, m$ *and* $j = m+1, \ldots, m+n$*.*

8.1 Proof of Lemma 2

Recall that M_n^+ is equal to M_n given that $M_n > 0$, where M_n is the total number of Q_i arrival periods during the time interval that separates two consecutive Q_i service periods.

Let $W(t) = \big(L_{i-1}(t), L_i(t), L_{i+1}(t), M(t)\big)$ denote the continuous-time Mar-kov chain with discrete state-space $\{0,1\}^3 \times \{1, 2, \ldots\}$, where $M(t)$ is the number of Q_i arrival periods until time t given that it is strictly positive. Assume that $\big(L_{i-1}(0), L_i(0), L_{i+1}(0)\big)$ is in steady-state and that a Q_i arrival period has just started at 0, i.e., time 0 is the first time that $\big(L_{i-1}(0), L_i(0), L_{i+1}(0)\big) = (0, 1, \cdot)$ with $\big(L_{i-1}(0-), L_i(0-)\big) \neq (0, 1)$. Moreover, we set $M(0) = 1$ and $M(0-) = 0$, and make the states $(\cdot, 0, 1, \cdot)$ of $W(t)$ to be absorbing.

Merging these absorbing states into one state, referred to as a, will not impact the dynamics of $W(t)$ before absorption. Since $\big(L_{i-1}(t), L_i(t), L_{i+1}(t)\big)$ is an irreducible Markov chain, the probability of transition to state a is equal to one and thus the time until absorption, T_a, is a proper rv. We will refer to the previous absorbing chain as **AMC**. Now writing M_n^+ in terms of $M(t)$ gives that $M_n^+ = M(T_a)$. The probability distribution $\mathbb{P}(M_n^+ = m)$ is the probability that the transition to a occurs from one of the states $\{(i,j,k,m) : i,j,k = 0,1 \text{ and } (j,k) \neq (0,1)\}$.

We derive now $\hat{M}_n^+(z)$, the p.g.f. of M_n^+. Let us define a level $l(m)$, $m = 1, 2, \ldots$, to be the transient states of **AMC** with $M(t) = m$ and ordered as follows

$$l(m) := \big\{(0,0,0,m), (0,1,0,m), (0,1,1,m), (1,0,0,m), (1,1,0,m), (1,1,1,m)\big\},$$

Observe that there are in total six states in $l(m)$. We order the infinite number of **AMC** states as follows: $l(1)$, $l(2)$, ..., and finally the absorbing state a. It is easily seen that the generator matrix $\mathbf{P}$ of **AMC** can be written as

$$\mathbf{P} = \begin{pmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{0}^T & 0 \end{pmatrix},$$

where $\mathbf{Q}$ represents the generator matrix of transitions between the transient states of **AMC**, $\mathbf{R}$ represents the rate vector of transitions from the transient states to the absorbing state a, $\mathbf{0}^T$ is the row vector with all entries equal to zero. Let u denote a column vector that designates the transition rate vector from $l(m)$ states to the state a. Therefore, $u = (\alpha_{i+1}^0, 0, \alpha_i^1, \alpha_{i+1}^0, 0, \alpha_i^1)^T$. Since u is independent of m, the vector $\mathbf{R}^T = (u^T, u^T, \ldots)$. Note that on leaving $l(m)$ the **AMC** either jumps to $l(m+1)$ or to a. For this reason, $\mathbf{Q}$ is an infinite upper-bidiagonal block matrix of the following form

$$\mathbf{Q} = \begin{pmatrix} \mathbf{A} & \mathbf{B} & \mathbf{0} & \cdots & \cdots \\ \mathbf{0} & \mathbf{A} & \mathbf{B} & \mathbf{0} & \cdots \\ \vdots\ddots & \ddots & \ddots & \ddots & \end{pmatrix}, \tag{34}$$

where, $\mathbf{A}$ is a 6-by-6 matrix that represents the transition rates between the states of $l(m)$, $m = 1, 2, \ldots$, which reads

$$\mathbf{A} = \begin{array}{c|cccccc} & 0,0,0,m & 0,1,0,m & 0,1,1,m & 1,0,0,m & 1,1,0,m & 1,1,1,m \\ \hline 0,0,0,m & A_{11} & 0 & 0 & \alpha_{i-1}^0 & 0 & 0 \\ 0,1,0,m & \alpha_i^1 & A_{22} & \alpha_{i+1}^0 & 0 & \alpha_{i-1}^0 & 0 \\ 0,1,1,m & 0 & \alpha_{i+1}^1 & A_{33} & 0 & 0 & \alpha_{i-1}^0 \\ 1,0,0,m & \alpha_{i-1}^1 & 0 & 0 & A_{44} & \alpha_i^0 & 0 \\ 1,1,0,m & 0 & 0 & 0 & \alpha_i^1 & A_{55} & \alpha_{i+1}^0 \\ 1,1,1,m & 0 & 0 & 0 & 0 & \alpha_{i+1}^1 & A_{66} \end{array}$$

$\mathbf{B}$ is a 6-by-6 matrix that represents the transition rates from the states of $l(m)$ to $l(m+1)$,

$m = 1, 2, \ldots$, which reads

$$\mathbf{B} = \begin{array}{c|cccccc} & 0,0,0,n & 0,1,0,n & 0,1,1,n & 1,0,0,n & 1,1,0,n & 1,1,1,n \\ \hline 0,0,0,m & 0 & \alpha_i^0 & 0 & 0 & 0 & 0 \\ 0,1,0,m & 0 & 0 & 0 & 0 & 0 & 0 \\ 0,1,1,m & 0 & 0 & 0 & 0 & 0 & 0 \\ 1,0,0,m & 0 & 0 & 0 & 0 & 0 & 0 \\ 1,1,0,m & 0 & \alpha_{i-1}^1 & 0 & 0 & 0 & 0 \\ 1,1,1,m & 0 & 0 & \alpha_{i-1}^1 & 0 & 0 & 0 \end{array}$$

where $n = m + 1$. The diagonal entries of $\mathbf{A}$ are such that $(\mathbf{A} + \mathbf{B})e + u = \mathbf{0}$, where e is the column vector of order six and with all entries equal to 1.

Next, we will derive $\mathbb{P}(M_n^+ = m)$ as function of the blocks of the inverse of $\mathbf{Q}$. Since $\mathbf{Q}$ is an upper-bidiagonal block matrix, it is easily verified that $\mathbf{Q}^{-1}$ is an upper-triangular block matrix of blocks $\mathbf{U}_{l,m} = \big(-\mathbf{A}^{-1}\mathbf{B}\big)^{m-l}\mathbf{A}^{-1}$ for $l \geq 1$ and $m \geq l$. Note that the matrix $\mathbf{A}$ is invertible since it is a generator matrix of a transient chain. Moreover, $-\mathbf{A}^{-1}\mathbf{B}$ is a sub-stochastic probability matrix whose entries give the probability of jumping to level $l(m+1)$ given that the **AMC** starts in $l(m)$. For this reason, $(-\mathbf{A}^{-1}\mathbf{B})^m \to \mathbf{0}$ as $m \to \infty$.

From the theory of absorbing Markov chains, given that **AMC** starts in $l(1)$ with probability distribution vector b, the probability that the absorption occurs from one of the states of level $l(m)$ is given by (see Lemma 6.(d))

$$\mathbb{P}(M_n^+ = m) = -b\mathbf{U}_{1,m}u = -b\big(-\mathbf{A}^{-1}\mathbf{B}\big)^{m-1}\mathbf{A}^{-1}u. \tag{35}$$

The p.g.f. of M_n^+ then reads

$$\begin{aligned} \hat{M}_n^+(z) &= -bz\sum_{m\geq 0}\big(-z\mathbf{A}^{-1}\mathbf{B}\big)^m\mathbf{A}^{-1}u, \\ &= -bz(\mathbf{A}+z\mathbf{B})^{-1}u, \quad |z| \leq 1, \end{aligned} \tag{36}$$

To complete the proof of Lemma 2 it remains to find b. We assumed that at time 0 the Q_i arrival period has just started. This means that time 0 is the first time after $s(<0)$ that $\big(L_{i-1}(0), L_i(0),\ L_{i+1}(0)\big) = (0,1,\cdot)$ and $\big(L_{i-1}(s), L_i(s), L_{i+1}(s)\big) \neq (0,1,\cdot)$. More specifically, given that $\big(L_{i-1}(s), L_i(s),\ L_{i+1}(s)\big)$ starts in $\{(0,0,1),(1,0,1)\}$ with steady-state distribution, the process $\big(L_{i-1}(t), L_i(t),\ L_{i+1}(t)\big)$, $s < t \leq 0$, either jumps first into $\{(0,1,0),(0,1,1)\}$, or first into $\{(0,0,0),\ (1,0,0),(1,1,0),(1,1,1)\}$ and later on into $\{(0,1,0),(0,1,1)\}$, see Figure 8. Given that $\big(L_{i-1}(s), L_i(s),\ L_{i+1}(s)\big) = (\cdot,0,1)$ with steady-state distribution, the former event occurs with a probability vector that is equal to the probability of transition to $\{(0,1,0),(0,1,1)\}$, considered as absorbing set, that reads (see Lemma 6.(d))

$$\begin{aligned} f &= -\frac{(\alpha_{i-1}^1, \alpha_{i-1}^0)}{\alpha_{i-1}^0 + \alpha_{i-1}^1}\begin{pmatrix} -\alpha_{i-1}^0 - \alpha_i^0 - \alpha_{i+1}^1 & \alpha_{i-1}^0 \\ \alpha_{i-1}^1 & -\alpha_{i-1}^1 - \alpha_i^0 - \alpha_{i+1}^1 \end{pmatrix}^{-1} \\ &\quad \times \begin{pmatrix} 0 & \alpha_i^0 \\ 0 & 0 \end{pmatrix} \\ &= \frac{(0, \alpha_{i-1}^1\alpha_i^0)}{(\alpha_{i-1}^0 + \alpha_{i-1}^1)(\alpha_i^0 + \alpha_{i+1}^1)}. \end{aligned} \tag{37}$$

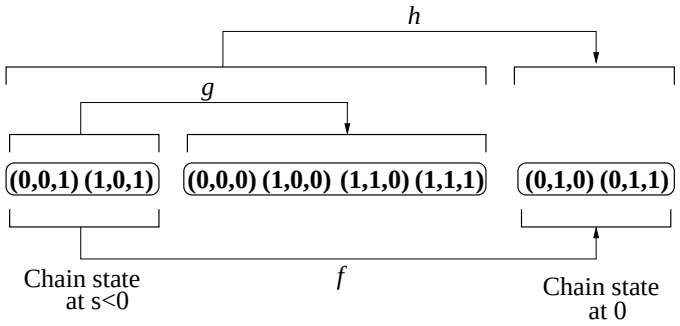


Figure 8: Initial probability distribution of **AMC** that is equal to $f + h$.

Given that $\big(L_{i-1}(s), L_i(s),\, L_{i+1}(s)\big) = (\cdot, 0, 1)$, the latter event is composed of two consecutive steps: the first one occurs when the process $\big(L_{i-1}(t), L_i(t),\, L_{i+1}(t)\big)$ jumps first into $\{(0,0,0), (1,0,0), (1,1,0), (1,1,1)\}$ and the second one occurs when it jumps into $\{(0,1,0), (0,1,1)\}$, see Figure 8. The probability vector of the first step is equal to g, see Eq. (43). For the second step, given that the process $\big(L_{i-1}(t), L_i(t), L_{i+1}(t)\big)$ starts in $\{(0,0,0),\, (1,0,0), (1,1,0), (1,1,\, 1)\}$ with probability g, it is possible that the process visits $\{(0,0,1), (1,0,1)\}$ several times before it first jumps into $\{(0,1,0), (0,1,\, 1)\}$. This occurs with probability (see Lemma 6.(d))

$$h \;=\; -\frac{(\alpha^1_{i-1}\alpha^1_{i+1}, 0, \alpha^0_{i-1}\alpha^1_{i+1}, 0, 0, \alpha^0_{i-1}\alpha^0_i)}{(\alpha^0_{i-1} + \alpha^1_{i-1})(\alpha^0_i + \alpha^1_{i+1})} \cdot \mathbf{H}^{-1} \cdot \mathbf{V}, \tag{38}$$

where,

$$\mathbf{H} = \begin{pmatrix} H_{11} & \alpha^0_{i+1} & \alpha^0_{i-1} & 0 & 0 & 0 \\ \alpha^1_{i+1} & H_{22} & 0 & \alpha^0_{i-1} & 0 & 0 \\ \alpha^1_{i-1} & 0 & H_{33} & \alpha^0_{i+1} & \alpha^0_i & 0 \\ 0 & \alpha^1_{i-1} & \alpha^1_{i+1} & H_{44} & 0 & \alpha^0_i \\ 0 & 0 & \alpha^1_i & 0 & H_{55} & \alpha^0_{i+1} \\ 0 & 0 & 0 & \alpha^1_i & \alpha^1_{i+1} & H_{66} \end{pmatrix}, \mathbf{V} = \begin{pmatrix} \alpha^0_i & 0 \\ 0 & \alpha^0_i \\ 0 & 0 \\ 0 & 0 \\ \alpha^1_{i-1} & 0 \\ 0 & \alpha^1_{i-1} \end{pmatrix}, \tag{39}$$

and the diagonal entries of $\mathbf{H}$ are such that $\mathbf{H}e + \mathbf{V}(1,1)^T = \mathbf{0}$. Finally, $f + h$ gives the probability distribution of $\{(0,1,0), (0,1,1)\}$ at time 0. Therefore, the non-zero entries of b read

$$b(0,1,0) = (f+h)(1), \quad b(0,1,1) = (f+h)(2) = 1 - (f+h)(1), \tag{40}$$

which completes the proof.

8.2 Proof of Lemma 3

The probability $\mathbb{P}(M_n = 0)$ is the probability that no Q_i arrival period occurs during the n-th cycle. This happens when no Q_{i-1} service period occurs between the n-th and $(n+1)$-st Q_i service periods. In terms of the Markov chain $(L_{i-1}(t), L_i(t), L_{i+1}(t))$, the probability of the latter event reduces to the probability that the chain first visits the set $\{(0,0,1), (1,0,1)\}$ and later on $\{(0,1,0), (0,\, 1,1)\}$, given the initial conditions that

$$(L_{i-1}(0-), L_i(0-), L_{i+1}(0-)) \in \{(0,0,1), (1,0,1)\}, \tag{41}$$

$$(L_{i-1}(0), L_i(0), L_{i+1}(0)) \in \{(0,0,0), (1,0,0), (1,1,0), (1,1,1)\}. \tag{42}$$

Making the sets $\{(0,0,1), (1,0,1)\}$ and $\{(0,1,0),\ (0,1,1)\}$ absorbing, $\mathbb{P}(M_n = 0)$ is the probability of absorption in $\{(0,0,1), (1,0,1)\}$ given the conditions in (41) and (42). First, let us derive the initial probability vector of the absorbing Markov chain. This initial probability vector is equal to the probability that the process jumps into $\{(0,0,0),$ $(1,0,0), (1,1,0), (1,1,1)\}$, given that at initial time this process starts with the steady state distribution in $\{(0,0,1),\ (1,0,1)\}$, which can be written as (see Lemma 6.(d))

$$\begin{aligned} g &= -\frac{(\alpha^1_{i-1}, \alpha^0_{i-1})}{\alpha^0_{i-1} + \alpha^1_{i-1}} \begin{pmatrix} -\alpha^0_{i-1} - \alpha^0_i - \alpha^1_{i+1} & \alpha^0_{i-1} \\ \alpha^1_{i-1} & -\alpha^1_{i-1} - \alpha^0_i - \alpha^1_{i+1} \end{pmatrix}^{-1} \\ &\quad \times \begin{pmatrix} \alpha^1_{i+1} & 0 & 0 & 0 \\ 0 & \alpha^1_{i+1} & 0 & \alpha^0_i \end{pmatrix} \\ &= \frac{(\alpha^1_{i-1}\alpha^1_{i+1}, \alpha^0_{i-1}\alpha^1_{i+1}, 0, \alpha^0_{i-1}\alpha^0_i)}{(\alpha^0_{i-1} + \alpha^1_{i-1})(\alpha^0_i + \alpha^1_{i+1})}. \end{aligned} \tag{43}$$

It then follows from absorbing Markov chain analysis that (see Lemma 6.(d))

$$\mathbb{P}(M_n = 0) \ = \ -g \cdot \mathbf{F}^{-1} \cdot w, \tag{44}$$

where,

$$\mathbf{F} \ = \ \begin{pmatrix} F_{11} & \alpha^0_{i-1} & 0 & 0 \\ \alpha^1_{i-1} & F_{22} & \alpha^0_i & 0 \\ 0 & \alpha^1_i & F_{33} & \alpha^0_{i+1} \\ 0 & 0 & \alpha^1_{i+1} & F_{44} \end{pmatrix}, \ w = \begin{pmatrix} \alpha^0_{i+1} \\ \alpha^0_{i+1} \\ 0 \\ \alpha^1_i \end{pmatrix},$$

and where the diagonal entries of $\mathbf{F}$ are such that

$$\mathbf{F}e + w + (\alpha^0_i, 0, \alpha^1_{i-1}, \alpha^1_{i-1})^T = \mathbf{0},$$

which completes the proof.

8.3 Proof of Lemma 4

Ξ_i is the duration of service interruption in Q_i. Therefore, in terms of the Markov chain $\big(L_{i-1}(t), L_i(t), L_{i+1}(t)\big)$, Ξ_i is the return time of the Markov chain $\big(L_{i-1}(t), L_i(t), L_{i+1}(t)\big)$ to the set $\{(0,0,1), (1,0,1)\}$, given that the chain has just left this set at initial time. Let y denote the row vector that represents the probability distribution of the states $\{(0,0,0), (0,1,0), (0,1,1), (1,0,0),\ (1,1,0), (1,1,1)\}$ at the initial time. Hence, y can be written as (see Lemma 6.(d))

$$\begin{aligned} y &= -\frac{(\alpha^1_{i-1}, \alpha^0_{i-1})}{\alpha^0_{i-1} + \alpha^1_{i-1}} \begin{pmatrix} -\alpha^0_{i-1} - \alpha^0_i - \alpha^1_{i+1} & \alpha^0_{i-1} \\ \alpha^1_{i-1} & -\alpha^1_{i-1} - \alpha^0_i - \alpha^1_{i+1} \end{pmatrix}^{-1} \\ &\quad \times \begin{pmatrix} \alpha^1_{i+1} & 0 & \alpha^0_i & 0 & 0 & 0 \\ 0 & 0 & 0 & \alpha^1_{i+1} & 0 & \alpha^0_i \end{pmatrix} \\ &= \frac{(\alpha^1_{i-1}\alpha^0_{i+1}, 0, \alpha^1_{i-1}\alpha^0_i, \alpha^0_{i-1}\alpha^1_{i+1}, 0, \alpha^0_{i-1}\alpha^0_i)}{(\alpha^0_{i-1} + \alpha^1_{i-1})(\alpha^0_i + \alpha^1_{i+1})}. \end{aligned} \tag{45}$$

Considering the set $\{(0,0,1),(1,0,1)\}$ as an absorbing set, $\tilde{\Xi}_i(s)$ becomes the LST of the time to absorption of $\big(L_{i-1}(t), L_i(t), L_{i+1}(t)\big)$ with generator matrix between transient states **A**+**B**, given in Section 5.2, u transition rate column vector from transient states to the absorbing set, and with initial probability distribution y. Lemma 6.(b) gives the desired result of $\tilde{\Xi}_i(s)$.

8.4 Proof of Lemma 5

Ξ_i^* is the first time after t that the server returns to Q_i given that an arrival to Q_i occurs at t. Therefore, in terms of the Markov chain $\big(L_{i-1}(t), L_i(t), L_{i+1}(t)\big)$ the duration of Ξ_i^* is equal to the first passage time of the Markov chain $\big(L_{i-1}(t),\ L_i(t), L_{i+1}(t)\big)$ to the set $\{(0,0,1),(1,0,1)\}$ given that the chain starts in $\{(0,1,0),\ (0,1,1)\}$ at initial time. By analogy with the derivation of $\tilde{\Xi}_i(s)$, assuming that $\{(0,0,1),(1,0,1)\}$ is an absorbing set, $\tilde{\Xi}_i^*(s)$ becomes the LST of the time to absorption of $\big(L_{i-1}(t), L_i(t), L_{i+1}(t)\big)$ with $\mathbf{A}+\mathbf{B}$, the transient states generator, u transition rate column vector from transient states to the absorbing set, and with initial probability distribution y^*. That is, the probability distribution that the chain starts in $\{(0,1,0),(0,1,1)\}$ is given by

$$y^* = \frac{(0, \alpha_{i+1}^1, \alpha_{i+1}^0, 0, 0, 0)}{\alpha_{i+1}^0 + \alpha_{i+1}^1}. \tag{46}$$

Lemma 6.(b) gives the desired result of $\tilde{\Xi}_i^*(s)$.

References

[1] (Delay Tolerant Networking Research Group) Web site: http://www.dtnrg.org.

[2] Grossglauser, M., Tse, D.: Mobility increases the capacity of ad hoc wireless networks. IEEE Transactions on Networking **10** (2002) 477–486

[3] Wang, C., Wolff, R.: Work-conserving tandem queues. Queueing Syst. **49** (2005) 283–296

[4] Coffman, E.G., Fayolle, G., Mitrani, I.: Two queues with alternating service periods. In: Performance '87: Proc. of the 12th IFIP WG 7.3 International Symposium on Computer Performance Modelling, Measurement and Evaluation. (1988) 227–239

[5] Frigui, I., Alfa, A.: Analysis of a time-limited polling system. Computer Communications **21(6)** (1998) 558–571

[6] Leung, K.: Cyclic-service systems with non-preemptive time-limited service. IEEE Transactions on Communications **42** (1994) 2521–2524

[7] Yechiali, U., Eliazar, I.: Polling under the randomly-timed gated regime. Stochastic Models **14** (1998) 79–93

[8] Al Hanbali, A., de Haan, R., Boucherie, R.J., van Ommeren, J.K.: A tandem queueing model for delay analysis in disconnected ad hoc networks. In: Proc. of ASMTA, LCNS 5055, Nicosia, Cyprus (2008) 189–205

[9] de Haan, R., Boucherie, R.J., van Ommeren, J.K.: A polling model with an autonomous server. Research Memorandum 1845, University of Twente (2007)

[10] Van Vuuren, M., Adan, I., Resing-Sassen, S.: Performance analysis of multi-server tandem queues with finite buffers and blocking. OR Spectrum **27** (2005) 315–338

[11] Doshi, B.: Queueing systems with vacations - a survey. Queueing Systems **1** (1986) 29–66

[12] Katayama, T.: Waiting time analysis for a queueing system with time-limited service and exponential timer. Naval Research Logistics **48** (2001) 638–651

[13] Zazanis, M.: A Palm calculus approach to functional versions of Little's law. Stochastic Processes and their Applications **74** (1998) 195–201

[14] de Haan, R., Al Hanbali, A., Boucherie, R.J., van Ommeren, J.K.: Analysis of polling systems under exponential time-limited service disiplines. Research Memorandum 1894, University of Twente (2008)

[15] van Ommeren, J.K.: The discrete-time single-server queue. Queueing Systems **8** (1991) 279–294

[16] Borst, S., Boxma, O., Combé, M.: Collection of customers: a correlated M/G/1 queue. Performance Evaluation **20** (1992) 47–59

[17] Neuts, M.: Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach. Johns Hopkins University Press (1981)