

A Tandem Queueing Model for Delay Analysis in Disconnected Ad Hoc Networks

Ahmad Al Hanbali, Roland de Haan, Richard J. Boucherie,
and Jan-Kees van Ommeren

Stochastic Operations Research, University of Twente, The Netherlands **
{hanbali,haanr,r.j.boucherie,j.c.w.vanommeren}@ewi.utwente.nl

Abstract. Ad hoc network routing protocols may fail to operate in the absence of an end-to-end connection from source to destination. This deficiency can be resolved by so-called delay-tolerant networking which exploits the mobility of the nodes by letting them operate as relays according to the store-carry-and-forward paradigm.

In this work, we analyze the delay performance of a small mobile ad hoc network by considering a tandem queueing system. We present an exact packet-level analysis by applying ideas from the polling literature. Due to the state-space expansion, this analysis cannot efficiently be applied for all model parameter settings. For this reason, an analytical approximation is constructed and its excellent performance has extensively been validated. Numerical results on the mean end-to-end delay show that the switch-over time distribution impacts this metric only through its first two moments. Finally, we study delay optimization under power control.

Keywords: Tandem queueing model; Ad hoc networks; Delay-tolerant networking; Autonomous server; Performance analysis.

1 Introduction

End-to-end connectivity is not a natural property of ad hoc networks. For instance, nodes may vary their transmission power, nodes may move, nodes may enter the sleep mode, or nodes may suffer from hardware failures. As a result, the network structure changes dynamically and this may lead to undesired situations of nodes becoming disconnected from parts of the network. The traditional store-and-forward routing protocols cannot be employed in highly disconnected ad hoc networks. A solution for this problem is to exploit the mobility of nodes present in the network. Such an approach has been proposed in the pioneering paper of Grossglauser and Tse [1] as an alternative to the store-and-forward paradigm and it is now known as the store-*carry*-and-forward paradigm in the context of delay-tolerant networking (DTN) [2]. In DTN, the incurred delay to send data between nodes can be very large and unpredictable due to the disconnection problem. Applications of such can be found in, e.g., disaster relief

** 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)

networks, rural networking, environmental monitoring networks, vehicular networks, and interplanetary networks.

An important aspect of DTN is the so-called contact opportunity between nodes. Two nodes are said to be in contact if they are within transmission range of one another and thus packet exchange between them is possible. The duration of a contact impacts the performance under such a networking approach. Another key factor for the performance is the inter-contact time, which is defined as the time duration between two consecutive contacts of node pairs. The inter-contact time mainly depends on the mobility of the nodes.

We will analyze the performance of DTN by taking into account, unlike [3–5], that the transmission of packets may fail due to the short contact time and a retransmission is required. Also, we assume that a source node generates a stream of packet arrivals instead of only one packet, as in [5–7]. In addition, we are interested in a more practical case of small, finite-size networks, rather than in asymptotic cases (see, e.g., [1, 7]). As a primary step towards understanding such networks, we study a model comprising a fixed source and destination node, and a single mobile node operating as a relaying device. Although it is a small model, it contains the main characteristics of a DTN and it is already non-trivial from an analytical perspective. Moreover, we emphasize that the analysis carried out in this paper can readily be extended to model, e.g., multiple relay nodes with single-copy packet approach [8], h -hop ($h \geq 2$) relay routing schemes, or mobile source and destination nodes.

The network model of our interest is reminiscent of a two-queue tandem model with a single alternating server. Such a tandem model has been analyzed under various servicing strategies (see, e.g., [9]). Typically, these strategies are based on the assumption that the server can be controlled. However, in the mobility-driven model of our interest, the server is autonomous and there is no possibility to control its movement. The research efforts on models with time-limited service periods are also closely related to our work. In a two-queue setting, [10] analyzes the model via boundary value techniques. Unfortunately, the analysis along these lines for more than two queues appears intractable. Time-limited service models have also been studied in the context of polling systems (see, e.g., [11, 12]). However, also in these models, there exists a notion of server control, since it is assumed that whenever a queue becomes empty the server moves to another queue.

Our main interest is in the end-to-end delay in the network described above. We study this network at the packet level by considering the two-queue tandem model as a particular kind of polling system with customer routing. Specifically, it is a time-limited polling system extended with the feature that the server remains at a queue even if it becomes empty. We perform an exact analysis for this system by extending the techniques developed in [12] and [13]. Due to the state-space expansion, the computation time of the joint queue-length probabilities may grow large for certain model parameters. Therefore, as a complementary tool, we present an analytical approximation for the case that the service requirements at each queue are exponential. The queue-length process at the second

queue is then analyzed in isolation as a workload process with Poisson batch arrivals. The key element is to approximate the batch size distribution as closely as possible. Numerical experiments show the excellent performance of the approximation for a broad range of parameter settings. These experiments further show that the mean sojourn time is insensitive to third and higher moments of the switch-over times. Finally, several guidelines are given for delay optimization by power control.

The rest of the paper is organized as follows. Section 2 gives the model description, discusses the stability, and presents exact results for the sojourn time in the source node and the mobile node. Section 3 proposes and analyzes an approximation for the sojourn time in the mobile queue. In Section 4, we numerically validate the accuracy of the approximation and present additional results which give insight in the delay of the network. Section 5 concludes the paper.

2 Model and exact results

2.1 Model

We consider a tandem model consisting of 3 first-in-first-out (FIFO) systems with unlimited queue, Q_i , $i = 1, 2, 3$, in which customers arrive to Q_1 and subsequently require service at Q_2 before reaching their destination at Q_3 . The special feature of the model is that Q_2 alternates between positions L_1 and L_2 such that customers at Q_1 are served only when Q_2 is at L_1 and customers at Q_2 are served only when Q_2 is at L_2 . In addition, Q_2 incurs a switching time from L_i to L_j ($i \neq j$, $i, j \in \{1, 2\}$) during which the server at neither Q_1 nor Q_2 is available. Q_3 is a sink and will not be included in our analysis.

Given a random variable (rv) X , $X(t)$ will denote its distribution function, $\tilde{X}(s)$ its Laplace-Stieltjes Transform (LST). Customers arrive to Q_1 according to a Poisson process with arrival rate λ . The service requirement S_i at Q_i has general distribution $S_i(\cdot)$ and mean $1/\beta_i$. We assume that the service requirements are independent and identically distributed (iid) rvs.

Movement of Q_2 is autonomous. Q_2 remains at location L_1 (resp. L_2) a (random) time of duration $X_n^{L_1}$ (resp. $X_n^{L_2}$) before it migrates to L_2 (resp. L_1) during its n -th visit. After the n -th visit to L_1 , Q_2 incurs a switch-over time $C_n^{1,2}$ from L_1 to L_2 , and similarly a switch-over time $C_n^{2,1}$ after the n -th visit to L_2 . We assume that $C_n^{1,2}$ ($C_n^{2,1}$) is an iid sequence with general distribution $C^{1,2}(\cdot)$ ($C^{2,1}(\cdot)$) and mean $c^{1,2}$ ($c^{2,1}$). Thus, the location of Q_2 is driven by an underlying continuous-time, discrete-state, process $\{L(t) : t \geq 0\}$ with state space $\{-2, -1, 0, 1\}$. More precisely, $L(t) = 1$ ($L(t) = 0$) when Q_2 is at L_1 (resp. L_2) at time t , and $L(t) = -1$ ($L(t) = -2$) when Q_2 switches from L_1 to L_2 (L_2 to L_1). Without loss of generality, let $L(0) = 1$. We further assume that $X_n^{L_1}$ ($X_n^{L_2}$) is an iid sequence of common exponential distribution with rate α_1 (α_2). Furthermore, we assume $\{X_n^{L_1}, X_n^{L_2}, C_n^{1,2}, C_n^{2,1}\}$ are iid and mutually independent, and also independent of the inter-arrival times and service requirements.

During the availability of the server at Q_1 and Q_2 , the server alternates between service and idle periods depending on whether customers are present. When the server is active at the end of a visit of Q_2 to L_1 or L_2 , service will be preempted. At the beginning of the next visit of Q_2 , the service time will be re-sampled according to $S_i(\cdot)$. This discipline is commonly referred to as *preemptive-repeat-random*. Let $N_i(t)$ denote the number of customers in Q_i , $i = 1, 2$, at time t . Assume $N_i(0) = 0$, $i = 1, 2$.

Further, we note that in the analysis we will use visit (time) rather than contact (time) to refer to (the duration of) a contact opportunity as to be in line with the common practice in the polling literature. Also, it is worth pointing out that the term customer throughout this paper will designate packet.

Our objective is to analyze the sojourn time of a customer in the tandem system and at the individual queues Q_1 and Q_2 . First, we will state the stability conditions for the tandem system. Second, we discuss several results for the sojourn time and queue length at Q_1 which will be required in the analysis later. Next, we determine the joint queue-length probabilities for the tandem model at specific instants. These probabilities can be related to the time-equilibrium probabilities. Finally, applying Little's law, the mean sojourn time at Q_2 is obtained.

2.2 Stability condition

The tandem model is stable if each customer in the system can be served in a finite period of time. 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.

Let a cycle define the time that separates two consecutive server visits to a queue. Due to the independence assumptions on our rvs, cycle lengths are iid, with generic rv $C := X^{L_1} + X^{L_2} + C^{1,2} + C^{2,1}$. For an individual queue to be stable, we must have that on average the number of customer arrivals per cycle is smaller than the number of customers that can be served at most per cycle. This latter number for Q_i will be denoted by N_{\max}^i , $i = 1, 2$, and is geometrically distributed (due to the exponential visit times and preemptive-repeat-random discipline), i.e., $\mathbb{P}(N_{\max}^i = k) = p_i(1 - p_i)^k$, $k = 0, 1, 2, \dots$, where $p_i = \mathbb{P}(\text{service is preempted at } Q_i) = 1 - \tilde{S}_i(\alpha_i)$, $i = 1, 2$. Thus, the stability condition for Q_i , $i = 1, 2$, reads

$$\rho_i := \frac{\mathbb{E}[\text{arrivals per cycle to } Q_i]}{\mathbb{E}[N_{\max}^i]} = \lambda \mathbb{E}[C] \cdot \frac{1 - \tilde{S}_i(\alpha_i)}{\tilde{S}_i(\alpha_i)} < 1, \quad (1)$$

where ρ_i is referred to as generalized load at Q_i and $\mathbb{E}[C] = 1/\alpha_1 + 1/\alpha_2 + c^{1,2} + c^{2,1}$, the mean cycle time. Notice that under stability the arrival rate to Q_2 and Q_1 are equal.

2.3 Queue one

Recall that the server visit process is autonomous and that 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 $1/\beta_1$, exponential on-period X^{L_1} with rate α_1 , and off-period R^{off} equal to the switch-over times plus the server visit time to Q_2 at L_2 , i.e., $R^{off} = C^{1,2} + C^{2,1} + X^{L_2}$. By a renewal reward argument P_{on} , the probability that the server is on, satisfies $P_{on} = (\alpha_1 \mathbb{E}[C])^{-1}$ and $P_{off} := 1 - P_{on}$.

The M/G/1 queue with on-off server has been extensively studied in the literature (see, e.g., [14, 15]). 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 [15]

$$\tilde{D}_1(s) = \tilde{W}_1(s) \tilde{S}^{eff}(s), \quad (2)$$

where $\tilde{W}_1(s)$ and $\tilde{S}^{eff}(s)$ 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. These latter LSTs are given by [15]

$$\tilde{W}_1(s) = \tilde{W}_{M/G/1}(s)(P_{on} + P_{off} \tilde{R}_e^{off}(s)), \quad (3)$$

$$\tilde{S}^{eff}(s) = \frac{(\alpha_1 + s)(\alpha_2 + s) \cdot \tilde{S}_1(\alpha_1 + s)}{(\alpha_1 + s)(\alpha_2 + s) - \alpha_1 \alpha_2 (1 - \tilde{S}_1(\alpha_1 + s)) \cdot \tilde{C}^{1,2}(s) \tilde{C}^{2,1}(s)}, \quad (4)$$

where $Re(s) > 0$, $\tilde{R}_e^{off}(s)$ 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 $\tilde{S}^{eff}(s)$.

It follows that the probability generating function (p.g.f.) of N_1 , the number of customers at Q_1 , which we denote by $F_1(\cdot)$, can be expressed as function of $\tilde{D}_1(\cdot)$ using the so-called functional form of Little’s law (see [16] for a general proof for FIFO queues with non-anticipating arrivals) as follows

$$F_1(z) = \tilde{D}_1(\lambda(1 - z)), \quad |z| \leq 1. \quad (5)$$

Let us denote by $F^{\{-2,1\}}(\cdot)$ the p.g.f. of the number of customers at the end of an off-period, i.e., at the transition of $L(t)$ from -2 to 1 . It can then be shown in a couple of steps by using Eq. (5), the PASTA property and conditioning on the position of the server, that

$$F^{\{-2,1\}}(z) = \tilde{W}_{M/G/1}(\lambda(1 - z)) \cdot \tilde{S}^{eff}(\lambda(1 - z)) \cdot \tilde{R}^{off}(\lambda(1 - z)). \quad (6)$$

This function will be required in Section 3 in the approximative analysis for Q_2 . For more details on its derivation, we refer to [17].

2.4 Queues in tandem

Joint queue-length probabilities at the end of a server visit: In this section, we will determine the queue-length distribution at the end of a server visit

at each queue of the tandem model. The analysis builds on the work of Eisenberg [18] and involves setting up an iterative scheme. This iterative approach was introduced by Leung [19] for the study of a probabilistically-limited polling model. Later, this model was extended in [12] to a time-limited polling model and in [13] for a time-limited model in which the server remains at a queue even if it becomes empty. A key role in the iterative scheme is played by the (auxiliary) p.g.f.'s $\phi_k^i(\mathbf{z})$ and $\phi_k^{s,i}(\mathbf{z})$ for $\mathbf{z} := (z_1, z_2)$, which will be explained below. In the final step of the iteration scheme $\gamma^i(\mathbf{z})$, the p.g.f. of the queue-length distribution at the end of a server visit to Q_i , is obtained as a function of $\phi_k^{s,i}(\mathbf{z})$.

Let us consider a service visit to a tagged queue i . We will recursively relate the number of customers present at the end of this visit to the number present at the beginning. To this end, we mark three types of events that may occur during a server visit viz., a customer arrival to an empty Q_i , a service completion at Q_i , and the departure of a server from Q_i . Further, we define for $k \geq 0$, $N_k^i := (N_{k,1}^i, N_{k,2}^i)$, where $N_{k,j}^i$, $j = 1, 2$, denotes the number of customers at Q_j just after the epoch of the k -th marked event at Q_i . We let N_0^i refer to the number of customers present at the arrival of the server to Q_i . Finally, we denote by the rv $\kappa_i \geq 1$ the number of marked events that occurs during a visit time of Q_i . It is immediately clear that the sequence $\{N_k^i\}_{k=0}^\infty$ is a Markov chain. We may then define for $k \geq 1$

$$\phi_k^i(\mathbf{z}) := \mathbb{E}[\mathbf{z}^{N_k^i} \mathbf{1}_{\{\kappa_i > k\}}], \quad (7)$$

where $\mathbf{1}_{\{A\}}$ is the indicator function of event A ($\mathbf{1}_{\{A\}} = 1$, if A is true, and 0 otherwise). That is, $\phi_k^i(\mathbf{z})$ is the joint p.g.f. of the number of customers at all queues at the k -th marked event epoch at Q_i and marked event k is not the final marked event during the visit (i.e., marked event $k+1$ will occur). Similarly, we define for $k \geq 1$

$$\phi_k^{s,i}(\mathbf{z}) := \mathbb{E}[\mathbf{z}^{N_k^i} \mathbf{1}_{\{\kappa_i = k\}}]. \quad (8)$$

That is, $\phi_k^{s,i}(\mathbf{z})$ is the joint p.g.f. of the number of customers at all queues at the k -th marked event epoch at Q_i and k is the final marked event (i.e., marked event k is a server departure event, and marked event $k+1$ will not occur). Let $N(T)$ be the number of arrivals during a random period T , I_1 be the (exponential) inter-arrival time of customers at Q_1 , and $C^{i,j}(\mathbf{z})$ be the p.g.f. of the number of arrivals during a switch-over time from Q_i to Q_j . Then, by analogy with the results of De Haan et al. [13] for a time-limited polling system, $\phi_k^i(\mathbf{z})$ and $\phi_k^{s,i}(\mathbf{z})$, $i = 1, 2$, $k = 1, 2, \dots$, are recursively given by

$$\phi_k^1(\mathbf{z}) = \phi_{k-1}^1(\mathbf{z})|_{z_1=0} \cdot \mathbb{E}[\mathbf{z}^{N(I_1)} \mathbf{1}_{\{X^{L_1} > I_1\}}] \cdot z_1 \quad (9)$$

$$+ (\phi_{k-1}^1(\mathbf{z}) - \phi_{k-1}^1(\mathbf{z})|_{z_1=0}) \cdot \mathbb{E}[\mathbf{z}^{N(S_1)} \mathbf{1}_{\{X^{L_1} > S_1\}}] \cdot z_2/z_1, \\ \phi_k^2(\mathbf{z}) = (\phi_{k-1}^2(\mathbf{z}) - \phi_{k-1}^2(\mathbf{z})|_{z_2=0}) \cdot \mathbb{E}[\mathbf{z}^{N(S_2)} \mathbf{1}_{\{X^{L_2} > S_2\}}] / z_2, \quad (10)$$

and

$$\begin{aligned} \phi_k^{s,1}(\mathbf{z}) &= \phi_{k-1}^1(\mathbf{z})|_{z_1=0} \cdot \mathbb{E}[\mathbf{z}^{N(X^{L_1})} \mathbf{1}_{\{X^{L_1} < I_1\}}] \\ &\quad + (\phi_{k-1}^1(\mathbf{z}) - \phi_{k-1}^1(\mathbf{z})|_{z_1=0}) \cdot \mathbb{E}[\mathbf{z}^{N(X^{L_1})} \mathbf{1}_{\{X^{L_1} < S_1\}}], \end{aligned} \quad (11)$$

$$\begin{aligned} \phi_k^{s,2}(\mathbf{z}) &= \phi_{k-1}^2(\mathbf{z})|_{z_2=0} \cdot \mathbb{E}[\mathbf{z}^{N(X^{L_2})}] \\ &\quad + (\phi_{k-1}^2(\mathbf{z}) - \phi_{k-1}^2(\mathbf{z})|_{z_2=0}) \cdot \mathbb{E}[\mathbf{z}^{N(X^{L_2})} \mathbf{1}_{\{X^{L_2} < S_2\}}], \end{aligned} \quad (12)$$

where

$$\begin{aligned} \phi_0^i(\mathbf{z}) &= \gamma^{3-i}(\mathbf{z}) C^{3-i,i}(\mathbf{z}), \\ \mathbb{E}[\mathbf{z}^{N(I_1)} \mathbf{1}_{\{X^{L_1} > I_1\}}] &= \lambda / (\lambda + \alpha_1), \\ \mathbb{E}[\mathbf{z}^{N(S_i)} \mathbf{1}_{\{X^{L_i} > S_i\}}] &= \tilde{S}_i(\alpha_i + \lambda(1 - z_1)), \\ \mathbb{E}[\mathbf{z}^{N(X^{L_1})} \mathbf{1}_{\{X^{L_1} < I_1\}}] &= \alpha_1 / (\lambda + \alpha_1), \\ \mathbb{E}[\mathbf{z}^{N(X^{L_i})} \mathbf{1}_{\{X^{L_i} < S_i\}}] &= \alpha_i \cdot \frac{1 - \tilde{S}_i(\alpha_i + \lambda(1 - z_1))}{\alpha_i + \lambda(1 - z_1)}, \\ \mathbb{E}[\mathbf{z}^{N(X^{L_2})}] &= \alpha_2 / (\alpha_2 + \lambda(1 - z_1)). \end{aligned}$$

Equation (9) can be explained by the following observations. First, the time between the $(k-1)$ -th and the k -th marked event epoch (and thus also the number of arriving customers) depends on whether at least one customer was present at the $(k-1)$ -th marked event epoch. This explains why the equation consists of two parts. Second, the number of customers at all queues at a marked event epoch is equal to the number present at the previous marked event epoch adjusted for the arrivals and departures in the meantime. Equation (10) consists only of one part due to the fact that once Q_2 is empty, it will remain empty during the rest of that visit. Along the same lines as Eqs. (9) and (10), Eqs. (11) and (12) are derived. Finally, we note that $\phi_0^i(\mathbf{1}) = 1$ for $\mathbf{1}=(1,1)$, while $\phi_k^i(\mathbf{1}) \leq 1$, for all $k = 1, 2, \dots$, since the k -th marked event might not occur at all during a visit to Q_i .

Notice that the final marked event is always a departure of the server from Q_i . Therefore, we can write for the number of customers at the queues at the end of a server visit to Q_i

$$\gamma^i(\mathbf{z}) = \mathbb{E}[\mathbf{z}^{N^{\kappa_i}}] = \lim_{K \rightarrow \infty} \sum_{k=1}^K \mathbb{E}[\mathbf{z}^{N_k^i} \mathbf{1}_{\{\kappa_i=k\}}] = \lim_{K \rightarrow \infty} \sum_{k=1}^K \phi_k^{s,i}(\mathbf{z}). \quad (13)$$

For computational convenience, we will numerically invert the joint p.g.f. of the number of customers at the queues using the Discrete Fourier Transform. To apply this technique, we only need the values of the joint p.g.f. to be known at a finite number of points (z_1, z_2) , with $z_i = \omega_i^{k_i}$, where $\omega_i = \exp(-2\pi I / J_i)$. Here I is the imaginary unit and J_i refers to the number of discrete points used for Q_i in the transformation process. Hence, we replace z_i , $i = 1, 2$, in the expressions above by $\omega_i^{k_i}$, so that all expressions become functions of $\mathbf{k} = (k_1, k_2)$. Also, let $\tilde{\gamma}^i(\mathbf{k})$ refer to $\gamma^i(\mathbf{z})$, with $z_i = \omega_i^{k_i}$. We start the iteration process with an

empty system and then compute $\tilde{\gamma}^1(\mathbf{k})$ using Eq. (13) up to a finite K such that $1 - \tilde{\gamma}^i(\mathbf{0}) < \delta$, for some $\delta > 0$. Next, we compute $\tilde{\gamma}^2(\mathbf{k})$ in the same way and then we continue with $\tilde{\gamma}^1(\mathbf{k})$ again, and so on. The iteration process is stopped when $\forall \mathbf{k} : |\text{Re}(\tilde{\gamma}^i(\mathbf{k})) - \text{Re}(\tilde{\gamma}^{i-1}(\mathbf{k}))| < \epsilon$, $i = 1, 2$, for some $\epsilon > 0$, where $\tilde{\gamma}^i(\mathbf{k})$ refers to the previously obtained value for $\tilde{\gamma}^{i-1}(\mathbf{k})$. The standard values for the convergence parameters that have been used are $\epsilon = 10^{-6}$ and $\delta = 10^{-9}$. Finally, via the inverse transform, the joint queue-length probabilities at visit completion instants can be found. We note that for these probabilities to be exact, we need at least that $J_i \rightarrow \infty$, $i = 1, 2$. However, the strength of the approach is that the probabilities are already close to the exact probabilities for small values of J_i . It should be noted that when the system load increases, the values J_i must typically be increased to guarantee the accurate computation of the probabilities. Thus, this iterative approach appears mainly applicable to systems with a light to moderate load.

Mean sojourn time at queue two: The sojourn time is related to the *time-equilibrium* queue-length probabilities. These probabilities can be obtained from the queue-length probabilities at the end of a server visit due to the exponential visit times. This is done by conditioning on the position of the server. Notice that the server is either at some queue or switching from one queue to another. Let us consider the queue-length process and condition on the server being at some tagged queue Q_i . This induced process consists of a series of exponentially distributed periods with positive jumps between each two periods. Due to the PASTA property, the system state just before these (Poisson) jump epochs equals exactly the time-average system state. Moreover, the system state at these epochs is exactly the state as observed by the server when it departs from Q_i . Thus, we have that a departing server observes the system in steady-state conditioned on the queue it departs from. Let us further denote the p.g.f. of the number of customers present at a random instant during a switch-over time from Q_{3-i} to Q_i by $C_R^i(\mathbf{z})$. It can readily be found that

$$C_R^i(\mathbf{z}) = \gamma^i(\mathbf{z}) \cdot \frac{1 - \tilde{C}^{3-i,i}(\lambda(1 - z_1))}{c^{3-i,i} \cdot \lambda(1 - z_1)}. \quad (14)$$

Hence, by conditioning on the position of the server, we may write for $P(\mathbf{z}) := \mathbb{E}[z_1^{N_1} z_2^{N_2}]$, the joint p.g.f. of the time-equilibrium queue lengths,

$$P(\mathbf{z}) = \frac{1}{\mathbb{E}[C]} \sum_{i=1}^2 \left(\gamma^i(\mathbf{z}) \cdot \frac{1}{\alpha_i} + C_R^i(\mathbf{z}) \cdot c^{3-i,i} \right). \quad (15)$$

The mean sojourn time at Q_2 then follows from the mean number of customer at Q_2 , $\mathbb{E}[N_2]$, and applying Little's law

$$\mathbb{E}[D_2] = \frac{\mathbb{E}[N_2]}{\lambda} = \frac{1}{\lambda} \cdot \frac{d}{dz_2} P(1, z_2) \Big|_{z_2=1}. \quad (16)$$

Remark 1. We note that using the distributional form of Little's law also higher moments can be obtained for the total sojourn time in the tandem system. However, this form cannot be applied to the individual sojourn time at Q_2 , since the arrival process to Q_2 does not satisfy the non-anticipating property [16].

3 Approximation

In this section, we present an approximation for $\tilde{D}_2(s)$, the LST of the sojourn time of a customer in the mobile queue Q_2 , so that we may also deal with the situations in which the exact approach is no longer computationally feasible. We consider the workload process in Q_2 when $L(t) = 0$, i.e. Q_2 is served. This will be done under the additional assumption that the service times are exponentially distributed at both queues with rate β_1 and β_2 respectively. It turns out that this process corresponds to the workload process in an $M/M/1$ queue with batch arrivals. The sojourn time of a customer in Q_2 then equals the sum of

- the sojourn time of a customer in the batch arrival queue,
- the time a customer is at Q_2 but $L(t) \neq 0$, i.e. Q_2 is not served.

We emphasize that in this case both the preemptive-resume and preemptive-repeat-random disciplines are stochastically identical. For the sake of simplicity, in the following we will consider the preemptive-resume discipline.

3.1 The workload in queue two

To study the workload process at Q_2 , we split the time into disjoint intervals which begin at the time instants that the $L(t)$ -process jumps from state -2 to 1 (i.e., the start of an on-period at Q_1). Denote the starting points of these intervals by $\{Z_n, n = 1, 2, \dots\}$, with, by convention, $Z_1 = 0$. Let the n -th cycle of $L(t)$ denote the time interval $[Z_n, Z_{n+1}[$, with duration $X_n^{L_1} + C_n^{1,2} + X_n^{L_2} + C_n^{2,1}$. Let $V(t)$ denote the workload (i.e., virtual waiting time) of Q_2 present at time t . Without loss of generality, we assume that $V(t)$ is left-continuous, i.e., arrivals are not counted as being in the system until just after they arrive. A sample path of the evolution of $V(t)$ as function of $L(t)$ is shown in Figure 1.

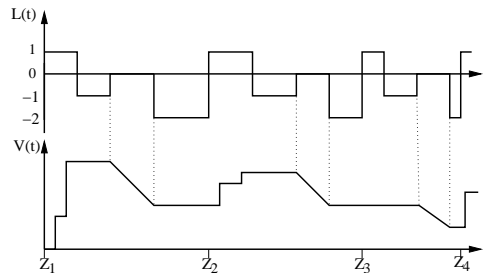


Fig. 1. Evolution of $L(t)$ and workload $V(t)$ of Q_2 .

Let W_n^B denote the workload present in Q_2 at time Z_n . Based on the evolution of $L(t)$, it is easily seen that

$$W_{n+1}^B = \left(W_n^B + \sum_{i=1}^{K_n} S_{2,i} - X_n^{L_2} \right)^+, \quad n \geq 0, \quad (17)$$

where $(\cdot)^+ = \max(\cdot, 0)$, K_n is the total number of arrivals to Q_2 (or departures from Q_1) during $X_n^{L_1}$ and $S_{2,i}$ is the service requirement of a customer in Q_2 . Note that $S_{2,i}$ is independent of $X_n^{L_2}$ and that K_n depends on $N_1(Z_n)$, the number of customers at Q_1 at time Z_n . Therefore, $K_n, n = 1, 2, \dots$, are correlated. For the sake of model tractability, we need the following

Assumption: $K_n, n = 1, 2, \dots$, are iid and independent of $\{X_m^{L_2} : m \leq n\}$.

By this assumption, Eq. (17) represents the workload seen by the first customer of a batch in a queue with Poisson batch arrivals with rate α_2 , independent batch size K_n , and exponential service requirement with rate β_2 .

Let $\tilde{G}(s) := E \left[e^{-s \sum_{i=1}^{K_n} S_{2,i}} \right]$, i.e., $\tilde{G}(s)$ denotes the LST of the service requirement of a batch. It is well known that the batch arrival queue is stable when $-\alpha_2 \tilde{G}'(0) = \alpha_2 \mathbb{E}[K_n] / \beta_2 < 1$. It can readily be verified that the latter condition is equivalent to the condition in (1) for Q_2 . Furthermore, the LST of the steady-state distribution of W_n^B can be written as

$$\tilde{W}^B(s) = \left(1 + \alpha_2 \tilde{G}'(0) \right) \frac{s}{s - \alpha_2 + \alpha_2 \tilde{G}(s)}. \quad (18)$$

By conditioning on K_n , we find that

$$\tilde{G}(s) = \mathbb{E} \left[\left(\frac{\beta_2}{\beta_2 + s} \right)^{K_n} \right]. \quad (19)$$

Finally, let $\tilde{V}^j(s)$ denote the LST of the workload seen by the j -th customer within a batch upon arrival (including the work brought in by itself). Since the service requirement of customers is independent of the workload present in the queue upon arrival and its distribution is exponential with rate β_2 , $\tilde{V}^j(s)$ reads

$$\tilde{V}^j(s) = \tilde{V}^{j-1}(s) \frac{\beta_2}{\beta_2 + s}, \quad j = 1, 2, \dots, \quad (20)$$

with $\tilde{V}^0(s) = \tilde{W}^B(s)$. Moreover, since K_n are iid, $\mathbb{P}(J = j)$, the probability 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 $\mathbb{P}(J = j) = \mathbb{P}(K_n \geq j) / \mathbb{E}[K_n]$.

Removing the condition on the customer position in a batch, the LST of the sojourn time of an arbitrary customer in the batch arrival queue is given by

$$\tilde{V}(s) = \beta_2 \tilde{W}^B(s) \frac{1 - \tilde{G}(s)}{s \mathbb{E}[K_n]}. \quad (21)$$

$$(22)$$

It remains to compute $\mathbb{E}[z^{K_n}]$ in order to find $\tilde{G}(s)$, and eventually $\tilde{W}^B(s)$. This will be done next using the matrix-geometric approach.

3.2 The p.g.f. of the batch size distribution

As remarked in the previous section, K_n is the total number of departures from Q_1 during the n -th cycle and depends on the queue length of Q_1 at time Z_n . To compute the p.g.f. of K_n , we first assume that Q_1 has a limited queue of $M - 1$ customers. This queue is denoted by Q_1^M . Later, we will let M tend to infinity to get the final results.

As we need the arriving batch size distribution in steady state, we assume that Q_1^M is in steady state at time Z_n . The probability that there are i customers in Q_1^M at Z_n is denoted by $b_M(i)$. Under the assumption that the unlimited Q_1 is stable, $\lim_{M \rightarrow \infty} b_M(i) = b(i)$ with $\sum_{i \geq 0} b(i)z^i = F^{\{-2,1\}}(z)$ (see Eq. (6)). Let $b_M = (b_M(0), \dots, b_M(M-1))$ denote the steady-state distribution of the finite capacity Q_1^M .

Let $(N_1(t), D(t))$ denote the two dimensional, continuous-time process with discrete state-space $\{0, 1, \dots, M-1\} \times \{0, 1, \dots\} \cup \{(M, 0)\}$, where $N_1(t)$ represents the number of customers in Q_1 at time t , and $D(t)$ the number of departures from Q_1 until t . State $(M, 0)$ is absorbing. We refer to this absorbing Markov chain by **AMC**. The absorption of **AMC** occurs when the server leaves the queue which happens with rate α_1 . By setting the probability that the initial state of **AMC** at $t = 0$ is $(i, 0)$ to $b_M(i)$, the probability that the absorption of **AMC** occurs from one of the states $\{(i, k) : i = 0, 1, \dots, M-1\}$ equals the steady-state batch size distribution $\mathbb{P}(K_n = k)$. The non-zero transition rates of **AMC** can be written as

$$\begin{aligned} q((i, j), (i+1, j)) &= \lambda, & 0 \leq i \leq M-2, & \quad j \geq 0, \\ q((i, j), (i-1, j+1)) &= \beta_1, & 1 \leq i \leq M-1, & \quad j \geq 0, \\ q((i, j), (M, 0)) &= \alpha_1, & 0 \leq i \leq M-1, & \quad j \geq 0. \end{aligned} \quad (23)$$

Let us order the infinite number of **AMC** states as: $(0, 0), \dots, (M-1, 0), (0, 1), \dots, (M-1, 1), \dots$, and finally $(M, 0)$. It is easily seen that the transition matrix **P** of **AMC** can be written as

$$\mathbf{P} = \left(\begin{array}{c|c} \mathbf{Q} & \mathbf{R} \\ \hline \mathbf{0} & 0 \end{array} \right),$$

where **Q** is an upper-bidiagonal block matrix of infinite dimension, **0** is the row vector with all zero entries and $\mathbf{R} = (\alpha_1, \dots, \alpha_1)^T$. The blocks of **Q**'s diagonal are all equal to **A**, a M -by- M bidiagonal matrix with diagonal $(-\lambda - \alpha_1, -\lambda - \alpha_1 - \beta_1, \dots, -\lambda - \alpha_1 - \beta_1, -\alpha_1 - \beta_1)$ and upper-diagonal $(\lambda, \dots, \lambda)$. The blocks of **Q**'s upper-diagonal are all equal to **B**, a M -by- M lower-diagonal matrix with lower-diagonal $(\beta_1, \dots, \beta_1)$.

Next, we will derive $\mathbb{P}(K_n = k)$ as function of the inverse of **Q**. Since **Q** is an upper-bidiagonal block matrix, \mathbf{Q}^{-1} is an upper-triangular block matrix. It is easy to verify that the entries of \mathbf{Q}^{-1} are equal to $\mathbf{U}_{m,l} = (-\mathbf{A}^{-1}\mathbf{B})^{l-m} \mathbf{A}^{-1}$ for $m \geq 0$ and $l \geq m$. Note that the matrix **A** is invertible since it is upper-bidiagonal with strictly negative diagonal entries.

From the theory of absorbing Markov chains, given that the initial state vector of **AMC** is b_M , the probability that the absorption occurs at one of the

states $\{(i, k) : i = 0, 1, \dots, M-1\}$ is given by (see, e.g., [20]) $P(K_n = k) = -\alpha_1 b_M (\mathbf{U}_{0,k}) e = -\alpha_1 b_M (-\mathbf{A}^{-1}\mathbf{B})^k \mathbf{A}^{-1}e$, where e denote the M -dimensional column vector with all entries equal to one. After some algebra, we find

$$E_M[z^{K_n}] = -\alpha_1 b_M (\mathbf{A} + z\mathbf{B})^{-1}e, \quad |z| \leq 1. \quad (24)$$

Therefore, it remains to find $(\mathbf{A} + z\mathbf{B})^{-1}$.

Now, define $\mathbf{Q}(z) := (\mathbf{A} + z\mathbf{B})$, let $u^T = (1, 0, \dots, 0)$ and let $v^T = (0, \dots, 0, 1)$. Observe that $\mathbf{Q}(z) = \mathbf{T}(z) + \beta_1 u u^T + \lambda v v^T$, where $\mathbf{T}(z)$ is a M -by- M tridiagonal Toeplitz matrix with diagonal entries equal to $(-\lambda - \beta_1 - \alpha_1)$, upper-diagonal entries equal to λ , and lower-diagonal entries $z\beta_1$. Let t_{ij}^* denote the (i, j) -entry of $\mathbf{T}^{-1}(z)$. By applying the Sherman-Morrison formula [21, pp. 76] we find that the (i, j) -entry of $\mathbf{Q}^{-1}(z)$ gives for $i, j = 1, \dots, M$,

$$q_{ij}^* = m_{ij} - \lambda \frac{m_{iM} m_{Mj}}{1 + \lambda m_{MM}}, \quad \text{where } m_{ij} = t_{ij}^* - \beta_1 \frac{t_{i1}^* t_{1j}^*}{1 + \beta_1 t_{11}^*}. \quad (25)$$

The inverse of a tridiagonal Toeplitz matrix has been computed in closed-form (see [22, Sec. 3.1]). Following that same approach, we obtain t_{ij}^* as function of r_1 and r_2 , the distinct roots of $\lambda r^2 - (\lambda + \beta_1 + \alpha_1)r + \beta_1 z$ with $|r_1| < |r_2|$. Inserting the values of t_{ij}^* into (24) and letting $M \rightarrow \infty$ yield that (for a detailed derivation see [17, Sec. 3.2])

$$\mathbb{E}[z^{K_n}] = \frac{\alpha_1}{\lambda(1-r_1)(r_2-1)} \left[1 + \beta_1 \frac{1-z}{\lambda r_2 - \beta_1} F^{\{-2,1\}}(r_1) \right], \quad (26)$$

where $F^{\{-2,1\}}(\cdot)$ is given in (6). Inserting $z = \beta_2/(\beta_2 + s)$ into (26) gives the closed form of $G(s)$, the LST of the service requirement of a total batch (see (19)), which in turn gives the closed form of $\tilde{W}^B(s)$.

3.3 Sojourn time in queue two

Let H_0 denote the sojourn time of a customer in the batch arrival queue. Moreover, let $\{H_t : t \geq 0\}$ denote the remaining sojourn time of a customer in Q_2 if the server would be continuously working at Q_2 from time t onwards. In other words, H_t decreases at rate 1 when $L(t) = 0$ and H_t is constant when $L(t) \in \{-2, -1, 1\}$ at time t . Let Y denote the number of service interruptions during the sojourn time of a customer.

The visit periods have an exponential length with rate α_2 . Now, given that $H_0 = v$, the number of interruptions has a Poisson distribution with

$$\mathbb{E}[z^Y | H_0 = v] = e^{-\alpha_2 v(1-z)}. \quad (27)$$

The duration of these interruptions are independent and are given by $\Xi = C^{2,1} + X^{L_1} + C^{1,2}$. Furthermore, Ξ^* , the time it takes before H_t actually starts decreasing after time 0, satisfies $\Xi^* = X_e^{L_1} + C^{1,2}$, where $X_e^{L_1}$ is the residual time of X^{L_1} . Note that $X_e^{L_1}$ and X^{L_1} are identically distributed with common

exponential distribution. It is easily seen that $D_2 = \Xi^* + H_0 + \sum_{i=1}^Y \Xi_i$. By conditioning on H_0 and Y , we find for the LST of D_2 ,

$$\tilde{D}_2(s) = \mathbb{E}[e^{-s\Xi^*}] \mathbb{E}[e^{-s(\sum_{i=1}^Y \Xi_i + H_0)}] = \mathbb{E}[e^{-s\Xi^*}] \mathbb{E}[e^{-sH_0} e^{-\alpha_2 H_0 (1 - \tilde{\Xi}(s))}].$$

where $\tilde{\Xi}(s) = \frac{\alpha_1}{\alpha_1 + s} \tilde{C}^{1,2}(s) \tilde{C}^{2,1}(s)$. Since H_0 equals the sojourn time in the batch arrival queue, we find (see, Eq. (22))

$$\tilde{D}_2(s) = \frac{\alpha_1 \tilde{C}^{1,2}(s)}{\alpha_1 + s} \times \tilde{W}^B(\Delta(s)) \times \frac{\beta_2}{\mathbb{E}[K_n]} \times \frac{1 - \tilde{G}(\Delta(s))}{\Delta(s)}, \quad (28)$$

where $\Delta(s) := s + \alpha_2(1 - \tilde{\Xi}(s))$.

4 Numerical evaluation

The evaluation of the model will be done in three parts. First, we will extensively validate the accuracy of the approximation. Second, we consider the impact of the switch-over time distribution on the mean end-to-end delay in the network. Finally, we study the optimization of the end-to-end delay in the network by adjusting the visit time parameters via power control. Throughout this section, the distribution of the switch-over times of Q_2 , $C^{1,2}$ and $C^{2,1}$, are assumed identically distributed according to an exponential distribution with mean $c^{1,2} = c^{2,1}$.

4.1 Model validation

We validate the approximate model developed in Section 3.3 for the mean sojourn time at Q_2 by comparison with the results for the exact model of Section 2.4. Note that the exact expression of the LST of the sojourn time at Q_1 was derived in Section 2.3. The approximation assumption is K_n , $n = 1, 2, \dots$, are iid and also independent of $\{X_m^{L_2} : m \leq n\}$.

Let us introduce some notation. Let $\mathbb{E}[D_2^{app}]$ (resp. $\mathbb{E}[D_2^{exa}]$) denote the mean sojourn time in Q_2 using the approximate (resp. exact) model given in Sect. 3.3 (resp. in Sect. 2.4). Let R_2 denote the absolute relative difference between the approximate and exact mean sojourn time in Q_2 , i.e., $R_2 := |1 - \mathbb{E}[D_2^{app}]/\mathbb{E}[D_2^{exa}]|$. Further, we note that the load at Q_1 and Q_2 can be rewritten as (see (1))

$$\rho_i = \frac{\lambda}{\beta_i} \left(\frac{\alpha_1 + \alpha_2}{\alpha_{3-i}} + 2\alpha_i c^{1,2} \right), \quad i = 1, 2.$$

Figure 2(a) displays R_2 as a function of λ for different values of $c^{1,2}$ with $\beta_1 = \beta_2 = 1$ and $\alpha_1 = \alpha_2 = 0.1$. Thus, in this scenario the load at Q_1 and Q_2 are equal ($\rho_1 = \rho_2$). Observe that R_2 increases with λ and that the approximate model is accurate for $\rho_1 = \rho_2 < 0.5$. This is because the probability that Q_1 is empty upon the departure of the server from Q_1 decreases with λ . For this reason, the auto-correlation of K_n increases with λ . Moreover, Figure 2(a) shows that R_2 decreases with $c^{1,2}$ for $\rho_1 = \rho_2$ (e.g., for $\rho_1 = 0.5$, $R_2 = 15\%$ when $c^{1,2} = 1$ and

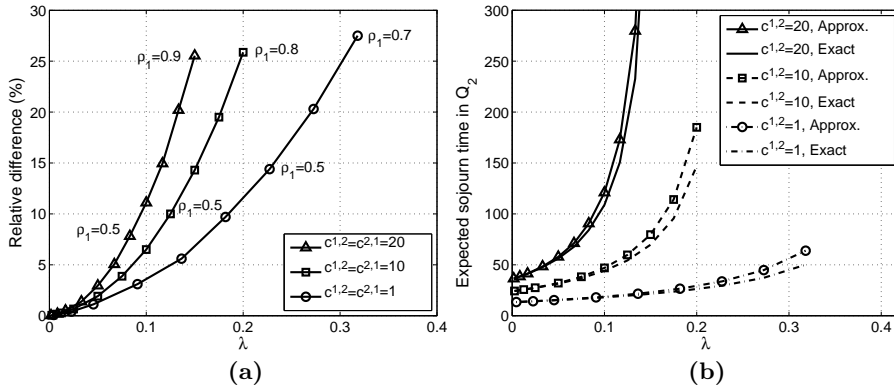


Fig. 2. For $\beta_1 = \beta_2 = 1$, $\alpha_1 = \alpha_2 = 0.1$: (a) Relative difference R_2 as a function of λ . (b) Mean sojourn time in Q_2 as a function of λ .

$R_2 = 8\%$ when $c^{1,2} = 20$). This is because in the case where $\rho_1 = \rho_2$, λ decreases with $c^{1,2}$.

Figure 2(b) shows the mean sojourn time in Q_2 using the approximate and exact models. Observe that the approximation gives an upper bound for $\mathbb{E}[D_2^{exa}]$. This observation is in support of the result in [23] which proves that in the correlated $M/G/1$ a positive correlation between the service requirement and the last inter-arrival reduces the mean sojourn time. We should emphasize that also in our model K_n and the last inter-arrival are positively correlated, i.e., an increase of the last inter-arrival time induces stochastically an increase of K_n . Finally, we found that for a given ρ_2 , the approximate model is more accurate for higher values of ρ_1 (e.g., when $\rho_2 = 0.6$, $R_2 = 9.6\%$ for $\rho_1 = 0.25$, while $R_2 = 4.7\%$ for $\rho_1 = 0.75$).

We conclude that the approximate model has the following properties:

- It is accurate for *low and moderate* load at Q_1 and Q_2 ;
- It gives an upper bound for the sojourn time at Q_2 ;
- It is accurate for *high* load at Q_1 and *moderate* load at Q_2 .

4.2 Impact of the switch-over times distribution on sojourn time

We note that in the analysis the distribution of the switch-over time was assumed to be arbitrary. This section studies the impact of the distribution of the switch-over times on the end-to-end sojourn time of a customer in the network. This will be done by considering the following two different distributions of the switch-over times in such a way that they share the same first two moments: two-phase hyper-exponential and two-phase Coxian distribution.

We evaluated the mean sojourn time as a function of $SCOV_s := Var(C^{1,2}) / (c^{1,2})^2$, the squared coefficient of variation of the switch-over times. Through an extensive numerical evaluation, we observed that the mean sojourn time is equal for the two different distributions. This suggests that the mean end-to-end sojourn time depends on the distribution of the switch-over times only through

their first two moments. Additional experiments for Weibull distributed switch-over times were performed and were in support of this conjecture. The latter experiments were done by simulation since the LST of the Weibull distribution is not known in closed form. We note that this conjecture is well known in the context of single-server queue with vacations and several polling models, but to the best of the authors' knowledge it is novel in the context of tandem models with autonomous servers.

4.3 Insight on the optimal end-to-end sojourn time

In this section, we study the evolution of α_2^{opt} , the optimal value of α_2 , that yields the minimum value of the end-to-end sojourn time under the constraints of zero switch-over time, i.e., $C^{1,2} = 0$, and constant cycle length, i.e., $\mathbb{E}[C] = 1/\alpha_1 + 1/\alpha_2$ is constant. Note that under these constraints α_1 decreases when α_2 increases. Since the mean sojourn time in Q_1 decreases with α_2 and mean sojourn time at Q_2 increases with α_2 , α_2^{opt} exists and it is unique. The adjustment of α_1 and α_2 can be done in practice by controlling the transmission power of the stations.

The optimal value α_2^{opt} can be computed by applying the numerical optimization package of MAPLE to the approximate mean sojourn time in (28). However, in practice one might prefer a more simple and intuitive rule that provides a value for α_2 which yields a close to optimal mean sojourn time. Therefore, we will discuss two alternative, heuristic optimization approaches.

The first heuristic selects the values of α_1 and α_2 such that the load is balanced at both queues, i.e., $\rho_1 = \rho_2$. This gives:

$$\alpha_i = (\beta_1 + \beta_2)/(\beta_{3-i} \cdot \mathbb{E}[C]), \quad i = 1, 2. \quad (29)$$

The second heuristic chooses α_1 and α_2 based on the analysis of a tandem model of two M/M/1 queues with shared service capacity. This means that the servers at both queues are always present, but serving at rate ν at Q_1 and at rate $1 - \nu$ at Q_2 . Then, the optimal ν , say ν^* , is the one that minimizes the end-to-end sojourn in such a tandem model, which we denote by $\mathbb{E}[D]^{M/M/1}$ and equals simply

$$\mathbb{E}[D]^{M/M/1} = 1/(\beta_1\nu - \lambda) + 1/(\beta_2(1 - \nu) - \lambda). \quad (30)$$

We choose the ratio α_1/α_2 equal to $(1 - \nu^*)/\nu^*$, such that the fraction of time that the server is at Q_1 in our model equals the optimal rate ν^* in the M/M/1 tandem model.

In Table 1, we present the results of this comparison. Here, α_2^{opt} , α_2^{LB} and $\alpha_2^{M/M/1}$ refer to the choice of α_2 in the optimal case, in the load balancing heuristic, and in the M/M/1 tandem heuristic, respectively. Further, we present the relative differences in mean sojourn time using the two heuristics (denoted by ϵ^{LB} and $\epsilon^{M/M/1}$) with respect to the optimal mean sojourn time, $\mathbb{E}[D]^{opt}$. Table 1 displays the performance of those heuristics when β_1 is increased while

β_2 , λ and $\mathbb{E}[C]$ are kept constant. We note that for the symmetric case, $\beta_1 = \beta_2$, the heuristics would also give the optimal solution $\alpha_1 = \alpha_2$. The performance using load balancing worsens rapidly when β_1 is increased. Also the M/M/1 tandem heuristic deviates from the optimum, but the relative differences remain small. We note that other values of $\mathbb{E}[C]$ were considered for which similar results were found.

We can conclude that balancing the load is not a good solution for end-to-end sojourn time optimization unless $\beta_1 \approx \beta_2$. However, using an optimization heuristic based on a simple tandem model of two M/M/1 queues will give nearly optimal results for the mean sojourn time.

β_1	1.1	2	3	6	11	16
α_2^{opt}	0.194	0.174	0.166	0.156	0.150	0.148
α_2^{LB}	0.190	0.150	0.133	0.117	0.109	0.106
$\alpha_2^{M/M/1}$	0.195	0.167	0.154	0.138	0.128	0.123
$\mathbb{E}[D]^{opt}$	14.47	12.82	12.14	11.42	11.08	10.94
ϵ^{LB} (%)	<0.1	3.9	8.6	17.2	23.6	26.3
$\epsilon^{M/M/1}$ (%)	<0.1	0.4	0.9	2.3	3.8	4.7

Table 1. Comparison of α_2 and $\mathbb{E}[D]$ for different optimization approaches for $\beta_2 = 1$, $\lambda = 0.1$, and $\mathbb{E}[C] = 10$.

5 Conclusions

This study is part of a research effort towards developing analytical models for quantifying the end-to-end delay in a delay-tolerant network. We consider here a network consisting of a fixed source node, a fixed destination node, and a mobile relay node. A closed-form expression has been derived for the delay at the packet's source node. Next, an iterative approach has been developed for the joint queue-length distribution of the source and the relay node. In addition, a simple approximate model has been proposed for the delay analysis at the relay node. The approximate model has extensively been validated and shows excellent results. Numerical results on the mean end-to-end delay show that the switch-over time distribution impacts this metric only through its first two moments. Moreover, load balancing is not an effective tool for delay optimization, while the M/M/1 tandem heuristic is near optimal.

In future work, we will study scenarios where multiple relay nodes coexist in the network. We anticipate that the exact and approximate models can be extended at least to cover the situations for which only a single copy of a packet exists at a time.

References

1. Grossglauser, M., Tse, D.: Mobility increases the capacity of ad hoc wireless networks. *ACM/IEEE Transactions on Networking* **10** (2002) 477–486
2. Delay Tolerant Networking Research Group, Web site: <http://www.dtnrg.org>.

3. Chaintreau, A., Hui, P., Crowcroft, J., Diot, C., Gass, R., Scott, J.: Impact of human mobility on the design of opportunistic forwarding algorithm. In: Proc. of IEEE INFOCOM, Barcelona, Spain (2006)
4. Groenevelt, R., Nain, P., Koole, G.: The message delay in mobile ad hoc networks. *Performance Evaluation* **62** (2005) 210–228
5. Small, T., Haas, Z.J.: Resource and performance tradeoffs in delay-tolerant wireless networks. In: Proc. ACM SIGCOM Workshop on Delay-Tolerant Networks, Philadelphia, PA, USA (2005)
6. Ibrahim, M., Al Hanbali, A., Nain, P.: Delay and resource analysis in manets in presence of throwboxes. *Performance Evaluation* **64** (2007) 933–947
7. Zhang, E., Neglia, G., Kurose, J., Towsley, D.: Performance modeling of epidemic routing. *Computer Networks* **51** (2007) 2867–2891
8. Spyropoulos, T., Psounis, K., Raghavendra, C.S.: Efficient routing in intermittently connected mobile networks: The single-copy case. To appear in *ACM/IEEE Transactions on Networking* (available online) (2007)
9. Wang, C., Wolff, R.: Work-conserving tandem queues. *Queueing Systems* **49** (2005) 283–296
10. Coffman, E.G., Fayolle, G., Mitrani, I.: Two queues with alternating service periods. In: Performance '87: Proc. of the 12th IFIP WG 7.3 Intl. Symposium on Computer Performance Modelling, Measurement and Evaluation. (1988) 227–239
11. Frigui, I., Alfa, A.: Analysis of a time-limited polling system. *Computer Communications* **21(6)** (1998) 558–571
12. Leung, K.: Cyclic-service systems with non-preemptive time-limited service. *IEEE Transactions on Communications* **42** (1994) 2521–2524
13. de Haan, R., Boucherie, R.J., van Ommeren, J.C.W.: A polling model with an autonomous server. Research Memorandum 1845, University of Twente (2007)
14. Doshi, B.: Queueing systems with vacations - a survey. *Queueing Systems* **1** (1986) 29–66
15. Katayama, T.: Waiting time analysis for a queueing system with time-limited service and exponential timer. *Naval Research Logistics* **48** (2001) 638–651
16. Zazanis, M.: A Palm calculus approach to functional versions of Little's law. *Stochastic Processes and their Applications* **74** (1998) 195–201(7)
17. Al Hanbali, A., de Haan, R., Boucherie, R.J., van Ommeren, J.C.W.: A tandem queueing model for delay analysis in disconnected ad hoc networks. Research Memorandum 1861, University of Twente (2007)
18. Eisenberg, M.: Queues with periodic service and changeover times. *Operation Research* **20** (1972) 440–451
19. Leung, K.: Cyclic-service systems with probabilistically-limited service. *IEEE Journal on Selected Areas in Communications* **9** (1991) 185–193
20. Gaver, D.P., Jacobs, P.A., Latouche, G.: Finite birth-and-death models in randomly changing environments. *Advances in Applied Probability* **16** (1984) 715–731
21. Press, W., Flannery, B., Teukolsky, S., Vetterling, W.: *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press (1992)
22. Dow, M.: Explicit inverses of Toeplitz and associated matrices. *ANZIAM J.* **44** (2003) E185–E215
23. Borst, S., Boxma, O., Combé, M.: Collection of customers: a correlated M/G/1 queue. *Performance Evaluation* **20** (1992) 47–59