Computers & Operations Research 37 (2010) 1174–1180

Contents lists available at ScienceDirect

Computers & Operations Research

journal homepage: www.elsevier.com/locate/caor

On tandem blocking queues with a common retrial queue

K. Avrachenkov [a,*], U. Yechiali [b]

[a] INRIA Sophia Antipolis, France
[b] Tel Aviv University, Israel

ARTICLE INFO

Available online 17 October 2009

Keywords:
Retrial networks
Mean value analysis
Fixed point approach

ABSTRACT

We consider systems of tandem blocking queues having a common retrial queue. The model represents dynamics of short TCP transfers in the Internet. Analytical results are available only for a specific example with two queues in tandem. We propose approximation procedures involving simple analytic expressions, based on mean value analysis (MVA) and on fixed point approach (FPA). The mean sojourn time of a job in the system and the mean number of visits to the orbit queue are estimated by the MVA which needs as an input the fractions of blocked jobs in the primary queues. The fractions of blocked jobs are estimated by FPA. Using a benchmark example of the system with two primary queues, we conclude that the approximation works well in the light traffic regime. We note that our approach becomes exact if the blocking probabilities are fixed. Finally, we consider two optimization problems regarding minimizing mean total sojourn time of a job in the system: (i) finding the best order of queues and (ii) allocating a given capacity among the primary queues.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

Majority of TCP transfers in the Internet are small in volume, consisting of only few packets [6]. The TCP congestion control mechanism does not have a chance to influence the dynamics of the traffic originated from short TCP transfers. Many short TCP transfers fit in the minimal size congestion window and hence the rate of the TCP transfer cannot be controlled by means of congestion window. We argue that for such type of TCP traffic, a network of blocking queues with retrials is an appropriate model. Then, an additional motivation for the study of retrial networks with blocking finite buffer capacity queues is the drop tail queue management policy employed in the Internet routers. A router using drop tail policy drops packets from the end of the queue when the queue size increases beyond some value. The dropped packets are then retransmitted by the sender.

Explicit analytic results were derived in [4] for a system comprised of a single $M/M/1/1$ primary (blocking) queue and an associated $M/M/1/\infty$ retrial (orbit) queue from which blocked jobs from the primary queue retry to be processed. Further explicit results were obtained for a system with *two* $M/M/1/1$ queues in tandem and a *common* associated $M/M/1/\infty$ orbit queue. The case with two queues in tandem turned out to be involved enough to predict that exact analytic solutions for $r>2$ tandem queues with blocking and common associated retrial queue will be very difficult to achieve, and even if achieved, the expressions for the various performance measures will be extremely complicated and hence with no significant insight. Therefore, in this work, we propose approximation procedure consisting of two parts. In one we use mean value analysis (MVA) to derive simple analytic expressions for the mean number of visits to the orbit queue and the mean sojourn time of a job in the system. The obtained expressions use as parameters the fractions of blocked jobs. Thus, in the other part of our approximation procedure we estimate the fraction of blocked jobs with the help of a fixed point approach (FPA). By comparing the approximation results with the exact results for the case of $r=2$ queues, we show that the proposed approximation is good when the system load is light.

Specifically, in the mean value analysis, assuming a *fixed* probability p_j of blocking in queue j, we calculate the probability generating function (PGF) and mean of N_j, the number of times an arbitrary job visits the orbit queue before passing queue j $(1\le j\le r)$ for the *first time*, where N_r specifies the *total* number of times an arbitrary job visits the retrial queue before leaving the system. We then derive the Laplace–Stieltjes transform (LST) and calculate the mean of Y_j, the total sojourn time of an arbitrary job in the system until it passes queue j for the *first time*. Similarly to N_r, Y_r specifies the *total* sojourn time of a job in the system. In the fixed point approach we assume that the input flows are Poissonian and we use Erlang's loss formula for the $M/M/c/c$ queue.

Having these results we consider two optimization problems:

(i) Finding the best order of arranging the queues so as to minimize the mean total sojourn time of a customer in the

* Corresponding author. Tel.: +33 4 92387751.
E-mail addresses: k.avrachenkov@sophia.inria.fr (K. Avrachenkov), uriy@post.tau.ac.il (U. Yechiali).

0305-0548/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cor.2009.10.004

system, when the orbit queue is either an $M/M/1/\infty$ system or an $M/M/\infty/\infty$ system. We show that the optimal order is to arrange the queues in an *increasing* order of the index $(1-p_j)E[B_j]/p_j$, where B_j is the processing time of a job in primary queue j.

(ii) Given a fixed total capacity C to all r queues, how this amount of resource should be allocated to the various queues so as to minimize the total sojourn time of a job through the system.

In comparison with the single node retrial queues [1,3,7,8], the networks of queues with retrials receive significantly less attention. In [2] the authors prove the non-existence of product-form solutions for certain queueing networks with retrials. Jackson-type systems with r tandem non-blocking $M/M/1/\infty$ queues and with feedback to (i) the first queue, and (ii) to a common $M/M/1/\infty$ retrial queue, where feedback from each queue j to the retrial queue is applied only *after* a job passes queue j, have been analysed in [5]. The following related model was also studied in [9]: a single job is made up of r independent tasks, all of which must be successfully performed for the job to be completed. Upon failure at any stage, the job has to be started all over again.

2. The model

Consider a system with r *blocking* primary queues in tandem, and a common associated retrial (orbit) queue to which all blocked jobs from the various primary queues are dispatched. Each blocked job, after spending a sojourn time in the orbit queue, tries to be admitted to the first queue and then continue traversing successfully through all r queues, until finally leaving the system. Thus, a job may traverse $m<r$ queues only to be blocked in the $(m+1)$-th queue, and then, after spending time in the orbit queue, start all over again from the first queue. A schematic presentation of the system is depicted in Fig. 1.

Assume that the outside arrival rate of new jobs to the system is λ jobs per unit time. Assume for a while that the blocking probabilities P_j $(j=1,2,\ldots,r)$ in the various primary queues are *fixed*. That is, $P_j=p_j$. (Further assumptions will be introduced for the various scenarios treated in the ensuing sections.)

We first calculate the probability generating function (PGF) and mean of the number of times a job visits the orbit queue until leaving the system. We then derive the Laplace–Stieltjes transform (LST) and mean of the time it takes to achieve that.

3. Number of visits at the orbit queue

Let N_j be the number of times a job visits the orbit queue until it passes successfully queue j *for the first time*. For $j\geq 1$ we have $(N_0=0)$

$$N_j=\begin{cases} N_{j-1} & \text{w.p. } 1-p_j,\\ N_{j-1}+1+N_j' & \text{w.p. } p_j,\end{cases}$$

where N_j' is an independent replica of N_j. We thus have that $N_j^*(z)$, the PGF of N_j, is given by

$$N_j^*(z)=E[z^{N_j}]=\frac{(1-p_j)N_{j-1}^*(z)}{1-zp_jN_{j-1}^*(z)},$$

and

$$E[N_j]=\frac{E[N_{j-1}]+p_j}{1-p_j}.$$

Iterating with $N_1^*(z)=(1-p_1)/(1-zp_1)$ and with $E[N_1]=p_1/(1-p_1)$ we get that

$$N_j^*(z)=\frac{\prod_{i=1}^{j}(1-p_i)}{1-z(1-\prod_{i=1}^{j}(1-p_i))},$$

and

$$E[N_j]=\frac{1-\prod_{i=1}^{j}(1-p_i)}{\prod_{i=1}^{j}(1-p_i)}=\sum_{m=1}^{j}\frac{p_m}{\prod_{i=m}^{j}(1-p_i)}.$$

It follows that N_j has a geometric distribution (shifted to 0) with "success" probability $1-\prod_{i=1}^{j}(1-p_i)$.

Clearly, as mentioned, N_r is the total number of times a job visits the orbit queue until it successfully leaves the system. It follows that with *fixed* blocking probabilities, the total number of times a job visits the orbit queue, until successfully passing queue j, is *independent* of the *order* of any set of j primary queues, for every $1\leq j\leq r$. Indeed, a job passes queue j if and only if it is *not* blocked in any of the first j queues, which occurs with probability $\prod_{i=1}^{j}(1-p_i)$. This explains why N_j is independent of the order of those queues.

Remark 1. For the calculation of $N_j^*(z)$ and $E[N_j]$ when the blocking probabilities are fixed, the primary queues can be of any blocking type and they need not be all the same.

4. Sojourn time of a job in the system

Let the service time of a job in queue j be a random variable, B_j $(j=1,2,\ldots,r)$, having a general probability distribution function. The sojourn time of a job in queue j is denoted by W_j.

Assume further that each time a job visits the orbit queue it resides there for a random time, W_0. Naturally, this random time depends on the assumptions on the type of queue the orbit queue is (e.g. $G/G/1/\infty$, $M/G/1/\infty$, or $M/G/\infty/\infty$, etc.). Thus, if for example the orbit queue is an $\cdot/G/\infty/\infty$, where the service time is B_0, then $W_0=B_0$.

Let Y_j be the length of time until a job *first* passes successfully primary queue j.

Then, similarly to the derivation of N_j, we can write $(Y_0=0)$

$$Y_j=\begin{cases} Y_{j-1}+W_j & \text{w.p. } 1-p_j,\\ Y_{j-1}+W_0+Y_j' & \text{w.p. } p_j,\end{cases}$$

where Y_j' is an independent replica of Y_j.

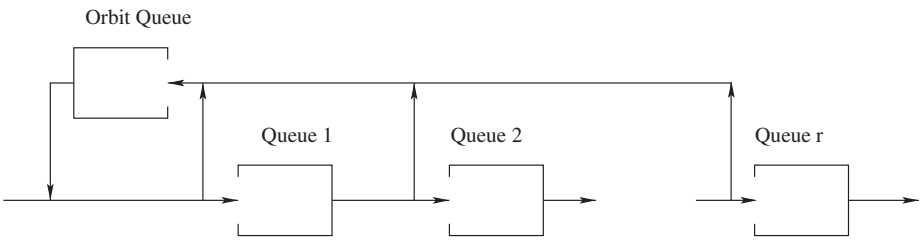

Fig. 1. Scheme of the system.

Thus, the LST of Y_j, $Y_j^*(s)=E[\exp(-sY_j)]$, is given by

$$Y_j^*(s)=\frac{(1-p_j)Y_{j-1}^*(s)W_j^*(s)}{1-p_jY_{j-1}^*(s)W_0^*(s)},$$

and its mean by

$$E[Y_j]=\frac{E[Y_{j-1}]}{1-p_j}+\frac{p_j}{1-p_j}E[W_0]+E[W_j].$$

Iterating with $E[Y_1]=[p_1/(1-p_1)]E[W_0]+E[W_1]$, we obtain

$$\begin{aligned}E[Y_j]&=\sum_{m=1}^{j}\frac{E[W_m]}{\prod_{i=m+1}^{j}(1-p_i)}+E[W_0]\sum_{m=1}^{j}\frac{p_m}{\prod_{i=m}^{j}(1-p_i)}\\&=\sum_{m=1}^{j}\frac{E[W_m]}{\prod_{i=m+1}^{j}(1-p_i)}+E[N_j]E[W_0].\end{aligned}\tag{1}$$

Now, the mean sojourn time of a job in the system is given by $E[Y_r]$.

5. Minimizing the mean sojourn time (when blocking probabilities are fixed)

Our objective now is to arrange the queues so that $E[Y_r]$, the mean total sojourn time of a job in the whole retrial network, is minimized. Since $E[N_r]$ is independent of the order of the queues, it suffices (see (1)) to find the order of queues that minimizes

$$\sum_{m=1}^{r}\frac{E[W_m]}{\prod_{i=m+1}^{r}(1-p_i)}.$$

Let $\pi_0=(1,2,\ldots,j-1,j,j+1,j+2,\ldots,r)$ be the order (policy) that arranges the queues according to some initial order $(1,2,\ldots,r)$. Let $\pi_1=(1,2,\ldots,j-1,j+1,j,j+2,\ldots,r)$ be the policy in which the order of queues j and $j+1$ is interchanged with respect to π_0. Set

$$\alpha_m=\frac{E[W_m]}{\prod_{i=m+1}^{r}(1-p_i)}.$$

Then, under π_0, we have

$$E[Y_r|\pi_0]=\sum_{m=1}^{j-1}\alpha_m+\frac{E[W_j]}{\prod_{i=j+1}^{r}(1-p_i)}+\frac{E[W_{j+1}]}{\prod_{i=j+2}^{r}(1-p_i)}+\sum_{m=j+2}^{r}\alpha_m,$$

while, under π_1, we have

$$\begin{aligned}E[Y_r|\pi_1]=&\sum_{m=1}^{j-1}\alpha_m+\frac{E[W_{j+1}]}{(1-p_j)\prod_{i=j+2}^{r}(1-p_i)}\\&+\frac{E[W_j]}{\prod_{i=j+2}^{r}(1-p_i)}+\sum_{m=j+2}^{r}\alpha_m.\end{aligned}$$

Thus, after multiplying throughout by $\prod_{i=j+2}^{r}(1-p_i)$, it follows that $E[Y_r|\pi_0]\le E[Y_r|\pi_1]$ if and only if

$$\frac{E[W_j]}{1-p_{j+1}}+E[W_{j+1}]\le\frac{E[W_{j+1}]}{1-p_j}+E[W_j].$$

That is, π_0 is better than π_1 if and only if

$$\frac{1-p_j}{p_j}E[W_j]\le\frac{1-p_{j+1}}{p_{j+1}}E[W_{j+1}].\tag{2}$$

By repeating queue interchanges we conclude that $E[Y_r]$ is *minimized* if and only if the queues are arranged in an *increasing* order of the index

$$\frac{1-p_j}{p_j}E[W_j].$$

That is, if p_j is large, then the mean number of attempts until first passing queue j, namely $p_j/(1-p_j)$, is also large, and hence it is better to place queue j at the beginning of the network of tandem queues. Similarly, small $E[W_j]$ has the same effect.

Remark 2. If each of the primary queues is a $\cdot/G/\infty/\infty$ queue with B_j being the service time of a job, and $(1-p_j)$ being the admission probability, independent of the state of the system, then $W_j=B_j$ for every $1\le j\le r$ and the optimizing index is

$$\frac{1-p_j}{p_j}E[B_j].$$

6. Fixed point approach

Let λ be the external arrival rate to primary queue 1. We first calculate the overall input rate to each primary queue, as well as to the orbit queue. Let Λ_j denote the overall input rate (= mean number of arrivals per unit of time) at the gate of primary queue j. If the blocking probability at queue j is P_j (P_j can be interpreted as the long time average fraction of jobs sent from queue j to the orbit queue), the arrival rate to queue r must be $\Lambda_r=\lambda/(1-P_r)$, since $\Lambda_r(1-P_r)=\lambda$ jobs enter and leave the stationary system per unit of time. The blocked rate Λ_rP_r is directed to the orbit queue. Similarly, $\Lambda_{r-1}=\Lambda_r/(1-P_{r-1})$ and $\Lambda_j=\Lambda_{j+1}/(1-P_j)$ for $1\le j\le r-1$. This implies that $\Lambda_j=\lambda/\prod_{i=j}^{r}(1-P_i)$. Thus, the overall rate of blocked jobs arriving at and leaving the orbit queue is

$$\Lambda_0=\sum_{j=1}^{r}\Lambda_jP_j=\lambda\sum_{j=1}^{r}\frac{P_j}{\prod_{i=j}^{r}(1-P_i)}=\lambda E[N_r].\tag{3}$$

Indeed, since $E[N_r]$ is the mean number of times a job visits the orbit queue, the output rate of that queue is $\Lambda_0=\lambda E[N_r]$. Now, clearly,

$$\Lambda_1=\lambda+\Lambda_0=\lambda\left(1+\sum_{j=1}^{r}\frac{P_j}{\prod_{i=j}^{r}(1-P_i)}\right)=\frac{\lambda}{\prod_{i=1}^{r}(1-P_i)}.\tag{4}$$

Suppose now that each primary queue j is a $\cdot/G/K_j/K_j$ queue. Assume further that the arrival rate to each queue is approximately Poisson, implying that each primary queue is an $M/G/K_j/K_j$ queue with arrival rate Λ_j. Then, the blocking probability P_j of queue j can be approximated by the Erlang loss formula, namely,

$$\tilde{P}_j=\frac{\rho_j^{K_j}/K_j!}{\sum_{i=0}^{K_j}\rho_j^i/i!},\quad j=1,2,\ldots,r,\tag{5}$$

where the approximated offered load at queue j is calculated as

$$\rho_j=\Lambda_jE[B_j]=\frac{\lambda E[B_j]}{\prod_{i=j}^{r}(1-\tilde{P}_i)}=\frac{\Lambda_{j+1}E[B_j]}{1-\tilde{P}_j}.$$

Thus, for queue r,

$$\rho_r=\Lambda_rE[B_r]=\frac{\lambda E[B_r]}{1-\tilde{P}_r}=\lambda E[B_r]\frac{\sum_{i=0}^{K_r}\rho_r^i/i!}{\sum_{i=0}^{K_r-1}\rho_r^i/i!}=\lambda E[B_r]\left(1+\frac{\rho_r^{K_r}/K_r!}{\sum_{i=0}^{K_r-1}\rho_r^i/i!}\right).$$

The above equation determines the value of ρ_r, from which $\tilde{P}_r$ is readily calculated. Now, we can write

$$\Lambda_{r-1}=\frac{\Lambda_r}{1-\tilde{P}_{r-1}},$$

and

$$\begin{aligned}\rho_{r-1}&=\Lambda_{r-1}E[B_{r-1}]\\&=\frac{\lambda E[B_{r-1}]}{(1-\tilde{P}_r)(1-\tilde{P}_{r-1})}=\frac{\lambda E[B_{r-1}]}{(1-\tilde{P}_r)}\left(1+\frac{\rho_{r-1}^{K_{r-1}}/K_{r-1}!}{\sum_{i=0}^{K_{r-1}-1}\rho_{r-1}^i/i!}\right).\end{aligned}$$

Then, going down from $r-1$ to 1, all ρ_j can be calculated along with all $\tilde{P}_j$.

To check the validity of this fixed point approach we will compare, for each j, the above probability $\tilde{P}_j$ with the *fraction* of times P_j a job is blocked at queue j.

7. Calculating the load-dependent blocking probabilities for a network with $M/G_j/1/1$ primary queues

Suppose (see Section 6) that each queue is an $M/G/1/1$ type queue. That is, we make the approximation that the arrival flow to queue j, at a rate of $\Lambda_j = \lambda/(\prod_{m=j}^{r}(1-P_m))$ is Poissonian. This assumption implies that the mean interarrival time to queue j is $1/\Lambda_j = (\prod_{m=j}^{r}(1-P_m))/\lambda$. Hence, the long run average blocking probability in queue j (being an $M/G/1/1$ queue, or using Erlang's loss formula with $K_j = 1$) is

$$\tilde{P}_j = \frac{b_j}{(\prod_{m=j}^{r}(1-\tilde{P}_m))/\lambda + b_j} = \frac{\lambda b_j}{\prod_{m=j}^{r}(1-\tilde{P}_m) + \lambda b_j} = \frac{\gamma_j}{\prod_{m=j}^{r}(1-\tilde{P}_m) + \gamma_j}, \tag{6}$$

where $b_j := E[B_j]$ and $\gamma_j := \lambda b_j$ for $j = 1, 2, \ldots, r$. Under π_0 we have

$$\tilde{P}_r = \frac{\gamma_r}{(1-\tilde{P}_r) + \gamma_r}. \tag{7}$$

Eq. (7) is a quadratic equation in $\tilde{P}_r$ and its solution is $\tilde{P}_r = \gamma_r$ (the solution $\tilde{P}_r = 1$ is not of interest). Indeed, since every job enters queue r once and only once, the load on this queue is $\gamma_r = \lambda b_r$ and this is the fraction of time queue r is busy and hence, it is also its blocking probability. It follows that $\Lambda_r = \lambda/(1-\tilde{P}_r) = \lambda/(1-\gamma_r)$. Now, for queue $j = r-1$, the inter-arrival time is $1/\Lambda_{r-1} = [\prod_{m=r-1}^{r}(1-\tilde{P}_m)]/\lambda$. This implies, using (7), that

$$\tilde{P}_{r-1} = \frac{\gamma_{r-1}}{(1-\tilde{P}_{r-1})(1-\tilde{P}_r) + \gamma_{r-1}} = \frac{\gamma_{r-1}}{(1-\tilde{P}_{r-1})(1-\gamma_r) + \gamma_{r-1}}. \tag{8}$$

The solution of the quadratic equation (8) is $\tilde{P}_{r-1} = \gamma_{r-1}/(1-\gamma_r)$. We therefore claim.

Lemma 1. *The blocking probabilities are given by*

$$\tilde{P}_j = \frac{\gamma_j}{1-\sigma_{j+1}}, \quad j = r, r-1, \ldots, 2, 1, \tag{9}$$

where $\sigma_j = \sum_{m=j}^{r}\gamma_m$ ($\sigma_{r+1} = 0$).

Proof. The lemma has been shown to be true for $j = r$ and $r-1$. We assume that it holds for all $j = r, r-1, \ldots, k+1$ and prove its validity for $j = k$. We first claim that $\prod_{m=k+1}^{r}(1-\tilde{P}_m) = 1-\sigma_{k+1}$. This follows by substituting from (9) the values of $\tilde{P}_j$, $j = r, r-1, \ldots, k+1$. Thus,

$$\tilde{P}_k = \frac{\gamma_k}{\prod_{m=k}^{r}(1-\tilde{P}_m) + \gamma_k} = \frac{\gamma_k}{(1-\tilde{P}_k)(1-\sigma_{k+1}) + \gamma_k}. \tag{10}$$

Again, the solution of (10) is $\tilde{P}_k = \gamma_k/(1-\sigma_{k+1})$, which completes the proof by induction. □

We note that from (9) it follows that $\tilde{P}_k < 1$ if and only if $\sigma_k = \sum_{j=1}^{k}\lambda b_j < 1$. Indeed, it has been shown in [4] that for a retrial tandem network with two $M/M/1/1$ primary queues, where $\mu = \mu_1 = 1/b_1 = \mu_2 = 1/b_2$, a necessary condition for stability is $\mu > 2\lambda$. That is $1 > 2\lambda/\mu = \lambda b_1 + \lambda b_2 = \sigma_2$. Moreover, when the retrial queue is a $\cdot/M/1/\infty$ queue with mean service time $b_0 = 1/\mu_0$, it has been shown in [4] that when $\mu_0 \to \infty$, a necessary and sufficient condition for stability becomes again $\sigma_2 < 1$.

8. Capacity allocation

Assume that the total capacity budgeted to the primary nodes of the tandem network is μ, that is, $\sum_{j=1}^{r}\mu_j = \mu$. We would like to distribute the total capacity in some optimal way among the primary queues. We consider separately two case.

8.1. Blocking probabilities are fixed

If $P_j = p_j$, independent of the queue load, then the optimization problem is (when $E[W_m] = b_m = 1/\mu_m$)

$$\min\left\{E[Y_r] = \sum_{m=1}^{r}\frac{1/\mu_m}{\prod_{i=m+1}^{r}(1-p_i)} + E[N_r]E[W_0]\right\}$$
$$\text{subject to} \quad \sum_{m=1}^{r}\mu_m = \mu, \quad \mu_m > 0, \quad m = 1, 2, \ldots, r. \tag{11}$$

With $E[N_r]$ independent of the μ_j's, by using Lagrange multipliers and differentiation one gets that the optimal values of $\mu_j's$ satisfy

$$\mu_{j+1}^{*2} = (1-p_{j+1})\mu_j^{*2} = \prod_{i=2}^{j+1}(1-p_i)\mu_1^{*2}, \quad 1 \le j \le r-1.$$

Thus, we have

$$\mu_1^* = \left(1 + \sum_{m=2}^{r}\sqrt{\prod_{i=2}^{m}(1-p_i)}\right)^{-1}\mu, \tag{12}$$

and

$$\mu_j^* = \left(\sqrt{\prod_{i=2}^{j}(1-p_i)}\right)\mu_1^*, \quad 2 \le j \le r. \tag{13}$$

That is, in the optimal capacity allocation, the first queue gets the largest capacity and then each following queue j gets a smaller capacity, reduced by a factor of $\sqrt{1-p_j}$.

8.2. Blocking probabilities estimated by $P_j = \gamma_j/(1-\sigma_{j+1})$

In the case when the blocking probabilities are estimated by $P_j = \gamma_j/(1-\sigma_{j+1})$, then $E[N_r]$ (Section 3) does play a role. We use $1-P_j = (1-\sigma_j)/(1-\sigma_{j+1})$ and $\prod_{i=m}^{r}(1-P_i) = 1-\sigma_m$. Thus, the optimization problem becomes:

$$\min\left\{E[Y_r] = \sum_{m=1}^{r}\frac{1/\mu_m}{(1-\sigma_{m+1})} + E[W_0]\sum_{m=1}^{r}\frac{\gamma_m/(1-\sigma_{m+1})}{(1-\sigma_m)}\right\}$$
$$\text{subject to} \quad \sum_{m=1}^{r}\mu_m = \mu, \quad \mu_m > 0, \quad m = 1, 2, \ldots, r. \tag{14}$$

Recall that $\sigma_m = \sum_{i=m}^{r}\lambda b_i = \sum_{i=m}^{r}\lambda/\mu_i$ and $\gamma_m = \lambda/\mu_m$. Using Lagrange multipliers for problem (14) does not yield a "nice" solution, but it can readily be solved numerically by standard procedures.

As we have noted above, the term with $E[N_r]$ cannot be neglected in this case. However, when $E[N_r]$ is small (e.g., when the retrial queue is $\cdot/M/\infty/\infty$ queue and μ_0 is large), we can apply the results of Section 5. In particular, in Section 5 it was shown that $E[Y_r]$ is minimized if the index

$$\frac{1-\tilde{P}_j}{\tilde{P}_j}b_j = \frac{1-\sigma_j}{\gamma_j}b_j = \frac{1}{\lambda}(1-\sigma_j)$$

is increasing. However, $(1-\sigma_j)$ is increasing for *any* order of the queues. That is, *all* orders give the same mean total sojourn time. This result seems at first to be somewhat surprising. However, numerical calculations performed in [4] for an analytic, non-approximating, solution of a network of two ($r = 2$) $M/M/1/1$ type queues (with common $M/M/1/\infty$ retrial queue) showed that L_{system}, the mean overall number of jobs in the system is *symmetric* with respect to the mean service rates μ_1 and μ_2 for a given value

of $\mu_1+\mu_2$. That is, any order of the two queues will result in the same value of L_{system}.

9. Numerical results

Here we perform numerical comparison of proposed approximations versus Monte Carlo simulations and exact results available for a particular case.

Specifically, in [4] we *explicitly* solved the model with two ($r=2$) $M/M/1/1$ tandem queues and an $M/M/1/\infty$ orbit queue. We shall refer to the results of [4] as the *exact model*. Let us recall some results from [4].

The mean total sojourn time of a job in the system T_{system} is, using Little's law,

$$T_{system}=\frac{1}{\lambda}L_{system},$$

where L_{system} denotes the average number of jobs in the system, given by (see Eq. (31) in [4])

$$L_{system}=L_{orbit}+P_{10}(\cdot)+P_{01}(\cdot)+2P_{11}(\cdot),$$

where $P_{ij}(\cdot)$ is the probability of i jobs in queue 1 and j jobs in queue 2 $(i,j=0,1)$. The probabilities $P_{10}(\cdot)$, $P_{01}(\cdot)$ and $P_{11}(\cdot)$, representing the fraction of time the system is in state (1,0), (0,1) or (1,1), respectively, were found to be (see Proposition 3 in [4])

$$P_{10}(\cdot)=\frac{\lambda}{\mu_1},$$

$$P_{01}(\cdot)=\frac{\lambda(\mu_1\mu_2(\mu_1+\mu_2+\mu_0)-\lambda(\mu_0(\mu_1+\mu_2)-\mu_1\mu_2)-\lambda^2(\mu_1+\mu_2))}{\mu_1\mu_2^2(2\lambda+\mu_1+\mu_2+\mu_0)},$$

$$P_{11}(\cdot)=\frac{\lambda}{\mu_2}-P_{01}(\cdot),$$

while L_{orbit} was shown to be

$$L_{orbit}=L_{00}+L_{10}+L_{01}+L_{11},$$

with L_{00}, L_{10}, L_{01} and L_{11} being calculated from the set of linear equations (26)–(29) in [4]:

$$\begin{aligned}
&(\lambda+\mu_0)L_{00}-\mu_2L_{01}=0,\\
&(\lambda+\mu_0)L_{00}-\mu_1L_{10}+\mu_2L_{11}=-\lambda P_{00}(\cdot)-(\lambda-\mu_1)P_{10}(\cdot)-\mu_2P_{11}(\cdot),\\
&-\mu_1L_{10}+(\lambda+\mu_2+\mu_0)L_{01}-\mu_1L_{11}=\mu_1P_{11}(\cdot),\\
&\mu_0L_{00}-\lambda L_{10}+\mu_0L_{01}-(\lambda+\mu_1)L_{11}=\lambda P_{10}(\cdot)+(\lambda+\mu_1)P_{11}(\cdot).
\end{aligned}$$

Let us compare T_{system} and $E[Y_2]$, where

$$E[Y_2]=E[W_2]+\frac{E[W_1]}{1-\tilde{P}_2}+E[N_2]E[W_0], \tag{15}$$

with

$$E[W_j]=E[B_j],\quad j=1,2,\quad \text{and}\quad E[N_2]=\frac{1}{1-\tilde{P}_2}\left(\frac{\tilde{P}_1}{1-\tilde{P}_1}+\tilde{P}_2\right).$$

To estimate $E[W_0]$ we assume the orbit queue to be of an $M/M/1/\infty$ type with arrival rate Λ_0 and mean service time $E[B_0]=1/\mu_0$. Thus, $E[W_0]$ is given by

$$E[W_0]=\frac{1}{\mu_0-\Lambda_0}=\frac{1}{\mu_0-\lambda E[N_2]}.$$

For the 2-queue in tandem and $M/M/1/\infty$ orbit queue from [4] we can calculate the exact long time average fraction of jobs blocked at each primary queue. Namely, the blocking rate at the gate of the first primary queue is

$$\Lambda_1P_1=\lambda(P_{10}(\cdot)+P_{11}(\cdot))+\mu_0((P_{10}(\cdot)-P_{10}(0))+(P_{11}(\cdot)-P_{11}(0))),$$

where $P_{ij}(n)$ is the probability of i jobs in queue 1, j jobs in queue 2 and n jobs in the orbit queue,

$$\Lambda_1=\lambda+\Lambda_0,$$

and

$$\Lambda_0=\mu_0(1-(P_{00}(0)+P_{10}(0)+P_{01}(0)+P_{11}(0))),$$

is the rate of jobs coming out of the orbit queue, while $P_{00}(0)$, $P_{10}(0)$, $P_{01}(0)$ and $P_{11}(0)$ are given in Proposition 3 of [4]. Thus, we have

$$P_1=\frac{\lambda(P_{10}(\cdot)+P_{11}(\cdot))+\mu_0((P_{10}(\cdot)-P_{10}(0))+(P_{11}(\cdot)-P_{11}(0)))}{\lambda+\mu_0(1-(P_{00}(0)+P_{10}(0)+P_{01}(0)+P_{11}(0)))}. \tag{16}$$

The rate Λ_2 is given by

$$\Lambda_2=\mu_1(P_{10}(\cdot)+P_{11}(\cdot)),$$

and the rate of blocking at the gate of the second primary queue is

$$\Lambda_2P_2=\mu_1P_{11}(\cdot).$$

Thus, we can write

$$P_2=\frac{P_{11}(\cdot)}{P_{10}(\cdot)+P_{11}(\cdot)}=\frac{\lambda/\mu_2-P_{01}(\cdot)}{\lambda/\mu_1+\lambda/\mu_2-P_{01}(\cdot)}. \tag{17}$$

Specifically,

$$\frac{1}{P_2}=1+\frac{P_{10}(\cdot)}{P_{11}(\cdot)}=\frac{\mu_2^2(2\lambda+\mu_1+\mu_2+\mu_0)}{\lambda[(\lambda+\mu_0)(\mu_1+\mu_2)+\mu_1\mu_2]}.$$

We refer to Eq. (15) together with Eqs. (16) and (17) as the *mean value approach with exact fractions of blocked jobs*. On the other hand, using Lemma 1, we can approximate the fractions of blocked jobs by

$$\tilde{P}_1=\frac{\lambda}{\mu_1}\Big/\left(1-\frac{\lambda}{\mu_2}\right),\quad \tilde{P}_2=\frac{\lambda}{\mu_2}.$$

We shall refer to Eq. (15) with the above approximations in place of P_1 and P_2 as the *fixed point approach*.

We note that the fractions P_1 and P_2 have not been calculated in [4]. We have indicated there that the comparison of the exact model with the fixed point approximation is the topic of the ensuing research.

We have also performed Monte Carlo simulations.

First we plot the expected total sojourn time of a job in the system obtained by four approaches: the exact model, the mean value approach with exact fractions of blocked jobs, the fixed point approach and Monte Carlo simulations. Similarly to the scenario considered in [4], we vary μ_1 keeping the sum $\mu_1+\mu_2$ constant. One can see in Fig. 2 that the mean value approach with the exact fractions of blocked jobs gives more precise results than

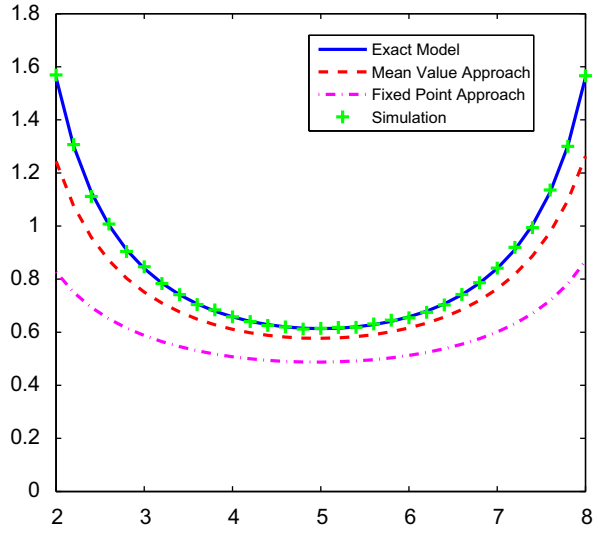

Fig. 2. Expected sojourn time as a function of μ_1, given $\mu_1+\mu_2=10$, $\lambda=1$ and $\mu_0=20$.

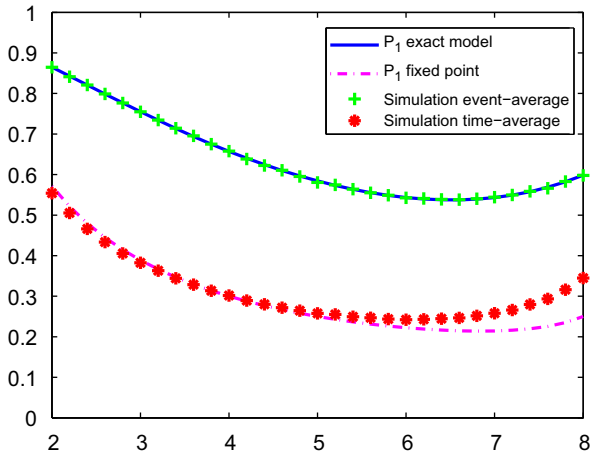

Fig. 3. Fraction of blocked jobs at the first primary queue as function of μ_1, given $\mu_1+\mu_2=10$, $\lambda=1$ and $\mu_0=20$.

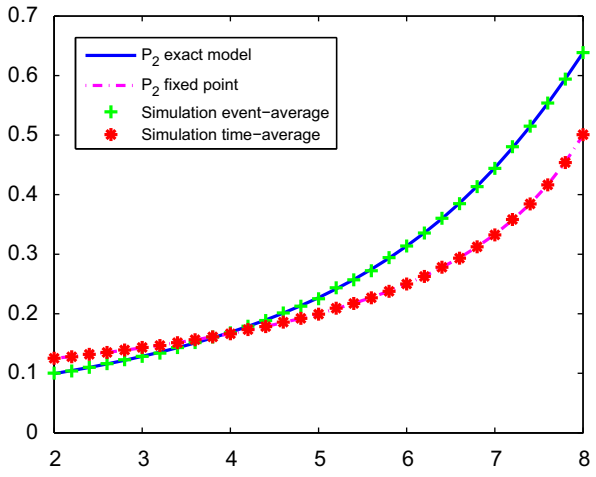

Fig. 4. Fraction of blocked jobs at the second primary queue as function of μ_1, given $\mu_1+\mu_2=10$, $\lambda=1$ and $\mu_0=20$.

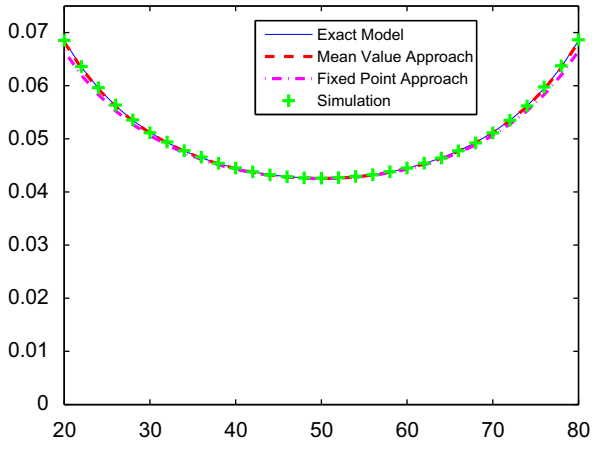

Fig. 5. Expected sojourn time as a function of μ_1, given $\mu_1+\mu_2=100$, $\lambda=1$ and $\mu_0=20$.

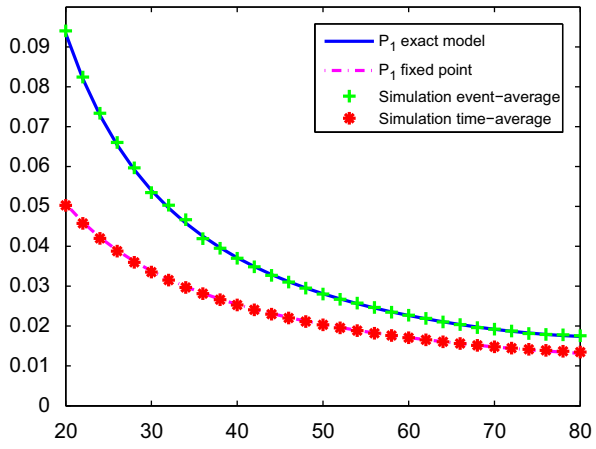

Fig. 6. Fraction of blocked jobs at the first primary queue as a function of μ_1, given $\mu_1+\mu_2=100$, $\lambda=1$ and $\mu_0=20$.

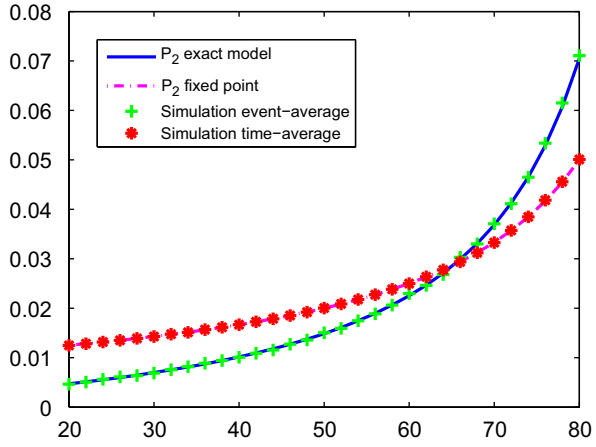

Fig. 7. Fraction of blocked jobs at the second primary queue as a function of μ_1, given $\mu_1+\mu_2=100$, $\lambda=1$ and $\mu_0=20$.

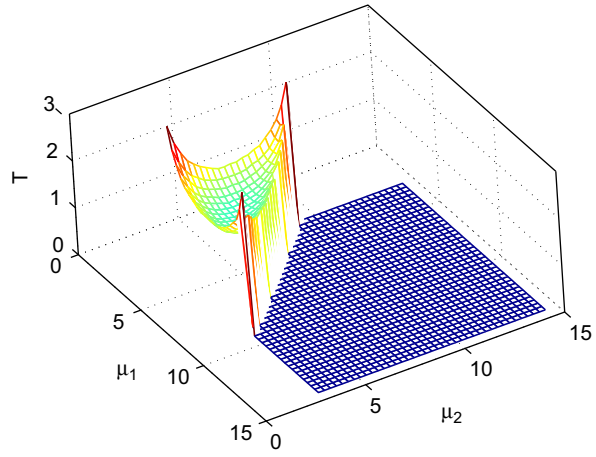

Fig. 8. Expected sojourn time as a function of μ_1 and μ_2, given $\mu_1+\mu_2+\mu_3=15$, $\lambda=1$ and $\mu_0=20$.

the fixed point approach. In Figs. 3 and 4 one can see that there is a gap between the exact values of the fractions of blocked jobs and their approximations obtained via the fixed point approach. In fact, the probabilities obtained by FPA approximate well the time-average probabilities of full queues but not the event-average fractions of blocked jobs. Nevertheless, the behaviour of the fractions of blocked jobs is captured qualitatively well by the fixed point approach. In particular, we can see that the value of the fraction of the jobs blocked in the first primary queue is not monotone with respect to the capacity of the first primary queue.

As confirmed by Figs. 5–7, the fixed point approach approximates better the system performance as both capacities

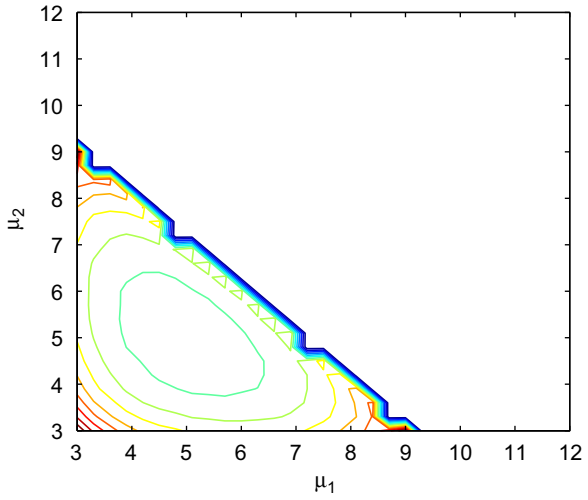

Fig. 9. Expected sojourn time as a function of μ_1 and μ_2 (the same value levels), given $\mu_1+\mu_2+\mu_3=15$, $\lambda=1$ and $\mu_0=20$.

of the primary queues increase or equivalently the traffic load decreases. We observe from Figs. 2 and 5 that if one uses exact fractions of blocked jobs, the mean-value analysis produces quite accurate results.

From Figs. 2 and 5 it appears that the expected total sojourn time of a job in the system is minimized when $\mu_1=\mu_2$. We have also performed Monte Carlo simulations for the model with three M/M/1/1 tandem queues ($r=3$). We have varied μ_1 and μ_2, keeping $\mu_1+\mu_2+\mu_3$ constant (see Figs. 8 and 9). In the case of three tandem queues it appears that the minimum of the expected total sojourn time of a job in the system is achieved at the point $\mu_1=\mu_2=\mu_3$. This is our conjecture that we plan to study in the future.

10. Conclusion

We have analysed networks of tandem blocking queues having a common retrial queue, for which explicit analytic results are not available. We have proposed approximation procedures involving simple analytic expressions, based on mean value analysis and on fixed point approach. The mean sojourn time of a job in the system and the mean number of visits to the orbit queue are estimated by the MVA which needs as an input the fractions of blocked jobs in the primary queues. The fractions of blocked jobs are estimated by FPA. Using a benchmark example of the system with two primary queues, we conclude that the approximation works well in the light traffic regime. We have formulated a number of optimization problems such as capacity allocation problem. We note that our approach becomes exact if the blocking probabilities are fixed.

Acknowledgements

We would like to thank Alain Jean-Marie for his very helpful advices about Monte Carlo simulations. We also would like to thank anonymous reviewers and the editor Antonis Economou for their suggestions. We also acknowledge the funding from Euro-NF Network of Excellence.

References

[1] Artalejo JR. Accessible bibliography on retrial queues. Mathematical and Computer Modelling 1999;30:223–33.
[2] Artalejo JR, Economou A. On the non-existence of product-form solutions for queueing networks with retrials. Electronic Modeling 2005;27:13–9.
[3] Artalejo JR, Gómez-Corral A. Retrial queueing systems: a computational approach. Berlin: Springer; 2008.
[4] Avrachenkov K, Yechiali U. Retrial networks with finite buffers and their application to Internet data traffic. Probability in the Engineering and Informational Sciences 2008;22:519–36.
[5] Brandon J, Yechiali U. A tandem Jackson network with feedback to the first node. Queueing Systems 1991;9:337–52.
[6] Collange D, Costeux J-L. Passive estimation of quality of experience. Journal of Universal Computer Science 2008;14:625–41.
[7] Falin GI. A survey of retrial queues. Queueing Systems 1990;7:127–67.
[8] Falin GI, Templeton JGC. Retrial queues. Boca Raton: CRC Press; 1997.
[9] Yechiali U. Sequencing an *N*-stage process with feedback. Probability in the Engineering and Informational Sciences 1988;2:263–5.