Tandem queues with impatient customers for blood screening procedures

Shaul K. Bar-Lev[*], Hans Blanc[†], Onno Boxma[‡], David Perry[§]

Abstract

We study a blood testing procedure for detecting viruses like HIV, HBV and HCV. In this procedure, blood samples go through two screening steps. The first test is ELISA (antibody Enzyme Linked Immuno-Sorbent Assay). The portions of blood which are found not contaminated in this first phase are tested in groups through PCR (Polymerase Chain Reaction). The ELISA test is less sensitive than the PCR test and the PCR tests are considerably more expensive. We model the two test phases of blood samples as services in two queues in series; service in the second queue is in batches, as PCR tests are done in groups. The fact that blood can only be used for transfusions until a certain expiration date leads, in the tandem queue, to the feature of customer impatience. Since the first queue basically is an infinite server queue, we mainly focus on the second queue, which in its most general form is an S-server $M/G^{[k,K]}/S + G$ queue, with batches of sizes which are bounded by k and K.

Our objective is to maximize the expected profit of the system, which is composed of the amount earned for items which pass the test (and before their patience runs out), minus costs. This is done by an appropriate choice of the decision variables, namely, the batch sizes and the number of servers at the second service station. As will be seen, even the simplest version of the batch queue, the $M/M^{[k,K]}/1 + M$ queue, already gives rise to serious analytical complications for any batch size larger than 1. These complications are discussed in detail. In view of the fact that we aim to solve realistic optimization problems for blood screening procedures, these analytical complications force us to take recourse to either a numerical approach or approximations. We present a numerical solution for the queue length distribution in the $M/M^{[k,K]}/S+M$ queue and then formulate and solve several optimization problems. The power-series algorithm, which is a numerical-analytic method, is also discussed.

1 Introduction

Basic group testing models deal with the classification of the items of some population into two categories 'good' and 'defective'. It is assumed that the items are *group testable*, i.e., for any subset of the population it is possible to carry out a simultaneous test (group test) with two

[*]Department of Statistics, University of Haifa, Haifa 31905 Israel (barlev@haifa.ac.il)

[†]Tilburg University, Dept. Econometrics & Operations Research, P.O. Box 90153, 5000 LE Tilburg, The Netherlands (blanc@uvt.nl)

[‡]EURANDOM and Department of Mathematics and Computer Science, Eindhoven University of Technology, HG 9.14, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (boxma@win.tue.nl)

[§]Department of Statistics, University of Haifa, Haifa 31905 Israel (dperry@haifa.ac.il)

possible outcomes: 'success' (or clean, or negative), indicating that all items in the subset are good, and 'failure' (or contaminated, or positive), indicating that at least one of the items in the subset is defective, without knowing which or how many are defective. A contaminated group can be subject to further screening or be scrapped. Employing suitably designed procedures of this kind leads to a significant reduction of the number of required tests and thus of screening cost, under controlled probabilities of misclassifications.

A group testing procedure is therefore a cost-efficient technique. It has been applied in various areas, first of all for blood testing to detect various viruses and for DNA screening, but also in quality control for industrial production systems (e.g. Li [24], Bar-Lev, Boneh and Perry [5]), drug discovery (Xie et al. [39], Zhu et al. [40]) and communication networks (Wolf [38]). A key reference is the monograph by Du and Hwang [17]. Applications to HIV screening are given, among others, by Gastwirth and Johnson [19], Litvak, Tu and Pagano [25], Tu, Litvak and Pagano [34] and Wein and Zenios [37]. Uhl et al. [36] study the use of pooling in an application to genetics. In Bar-Lev et al. [6, 7, 8, 9, 10], a more detailed discussion of the literature and a classification of group testing models according to various dichotomies are given.

In this paper we focus on blood testing. We wish to analyze and optimize the performance of blood screening procedures. We analyze the delays experienced in the blood screening process, and the process of outdating, and subsequently we use the results of this analysis to minimize the costs associated with the test procedures and with the loss of blood samples due to their exceeding the expiration date.

We divide the remainder of this introductory section into three parts. In Subsection 1.1 we describe blood screening procedures and related relevant features; Subsection 1.2 presents a first global description of a tandem queueing model which can be used for the performance analysis of such blood screening procedures; and Subsection 1.3 presents aim and overview of the paper.

1.1 Blood screening procedures and related features

Blood banks worldwide aim toward the supply of uncontaminated blood. Each blood donation goes through multiple testing for the presence of various pathogens which are able to cause transfusion-transmitted diseases. In most countries screening all blood donations for hepatitis B (HBV), hepatitis C (HCV) and human immunodeficiency virus (HIV) is mandatory. The cost of this screening is rising in developed countries and is a major economic burden in developing countries. Currently, blood banks in the USA and several Western European countries have adopted pooling methods for the performance of Polymerase Chain Reaction (PCR) while screening blood donations for HIV, HBV and HCV. This is done in addition to the individual antibody Enzyme Linked Immuno-Sorbent Assay (ELISA) testing, in order to increase earlier detection of these agents and decrease morbidity (see Schottstedt et al. [29], Stramer et al. [32] and Hourfar et al. [21]).

Until a few years ago, the routine testing was based only on ELISA tests that detect virus-specific antibodies in the blood. Such an ELISA test has high sensitivity and specificity but has a lower analytic detection limit which affects the identification of positive samples very soon after HIV seroconversion, as it takes time to develop a high concentration of antibodies. The latter drawback of ELISA is related to the effect of the *window period* which causes serious problems when testing for viral diseases. The window period of a given virus is defined as the period elapsing between the time a person is infected by the virus and the time antibodies are developed and can be detected. The window period varies for different types of viruses, and will influence the effectiveness of group testing. The importance of this issue depends greatly on the extent of the epidemic in a given population and the incidence rate of new infections. Examples of average window periods (based on ELISA) for some viruses are: 22 days for HIV, 70 days for HCV and 60 for HBV, but in individual cases window periods can be substantially longer. Note that antibodies are not developed in a non-live organism, implying that once a blood sample is donated, no antibodies are further produced. Consequently, the ELISA test cannot detect a viral contaminated blood sample if the individual who donated the sample was

infected prior to making the donation, but yet had not ”completed” the respective viral window period, and thereby not producing a sufficiently high concentration of antibodies to be detected by the ELISA test.

The relatively new PCR test can detect viral genetic material in the blood and has a much higher sensitivity and specificity than the ELISA test. PCR testing is especially advantageous in the window period soon after seroconversion when the virus starts multiplying but antibodies are not yet at high levels. For these blood samples, the ELISA test will be negative while the PCR test is likely to be positive. However, the PCR test is considerably more expensive than the ELISA test. Therefore, blood banks in the USA and some countries in Europe have established a new protocol with a two-stage procedure comprising of individual and pooled testing. By such a procedure blood samples are first screened individually for HIV (or any other viruses) with ELISA. Those samples that are found positive (i.e., contaminated) are discarded. All the negative samples are then pooled in groups and tested for the same viruses with PCR techniques. All donations in the negative pools are accepted while the positive pools are discarded and a resolution testing is performed to identify the individual positive donation. Pool sizes are typically 6-24.

Accordingly, we face here a service facility with two testing stages (service stations): Blood samples arrive at the first station for an ELISA test and are individually tested, and if they are found HIV positive they leave the system. Otherwise, they are forwarded to the second station which involves PCR testing, where the tests are processed in groups (batches) of size K, say. The groups then leave the system either as a contaminated group (HIV positive) or a clean group (HIV negative). It should be noted that a similar routine is also applied to detect HBV or HCV in blood testing.

However, at least one more substantial feature related to the above testing process should be taken into account, namely, *aging* (or *expiration date)*. This relates to the fact that each blood donation is basically ”divided” into at least three components that are used separately, depending on the needs of the patients: a) red blood cells - which can be used for up to 42 days (red

cells of 35-42 days old are less viscous); b) plasma and cryoprecipitate concentrations - which can be kept frozen for about one year; and c) platelets ('thromobotzitim') - which are usable for at most 5 days. The time constraints due to these expiration dates must be taken into account when considering the processing times of the two-stage testing procedure. Consequently, at each service station, one faces impatient customers.

1.2 Various queueing systems for modeling the two-stage blood screening procedure

Based on the previous subsection it can be seen that we are dealing with a queueing system consisting of two service stations in tandem with impatient customers, where items (customers) are served individually at the first service station and in batches at the second. Accordingly, our basic assumptions are made as follows. Having realized that the ELISA test is relatively cheap and that the blood samples can be tested individually in parallel, we may assume that the first service station is an infinite-server queue. As opposed to the ELISA test, the PCR test performed at the second station is excessively expensive, much more costly than the ELISA test. Accordingly, we shall assume that the second service station is a finite-server queue, with service in batches. The customers have some overall patience; we shall see later that we can decompose the tandem queue, studying each station in isolation as a queue with customer impatience. The queueing analysis of station 1 is easy. However, the combination of bulk service, multiple servers and customer impatience leaves us with little hope of obtaining an exact analytic solution of station 2; even the single server case already seems prohibitively difficult, cf. Section 3.

Batch (or bulk) service has also been extensively studied in the queueing literature. We refer to the book of Chaudhry and Templeton [16] for an extensive overview of this field. Relevant literature on queueing systems in batches can be found in Neuts [27], Nair and Neuts [26], Abolnikov and Dukhovny [1, 2], Bar-Lev et al. [10].

Impatience is also a very natural and important concept in queueing models. There is a wide range of situations in which customers may become impatient when they do not receive service fast enough. Next to blood screening, one may think of call centers and health care centers.

A pioneering paper on queueing models with impatience is [11]; it studies the $M/M/s+D$ model for the case that impatience refers to the waiting time, and the $M/M/1+D$ model for the case that the impatience refers to the sojourn time. Here the symbol D denotes deterministic (im-)patience. Baccelli et al. [3, 4] provide necessary and sufficient conditions for the existence of the virtual waiting time distribution in the $G/G/1+G$ queue. The latter distribution is subsequently obtained for $M/G/1+M$ and $M/G/1+E_k$. Finch [18] derives the waiting time distribution in the $G/M/1+D$ queue. Stanford [30] relates the waiting time distribution of the (successful) customers and the workload seen by an arbitrary arrival in $G/G/1+G$. He also considers the distribution of the number of customers, obtaining explicit results for various special cases. Stanford [31] contains a brief literature review, and [22] provides a useful approximation for the waiting time distribution in $M/G/N+G$ and several additional references on multiserver queues with impatience.

1.3 Aim and overview of the paper

Our aim is to study the two-stage group testing queueing system from a purely economic (cost-decreasing) point of view. The objective is to maximize the expected profit of running the system, which is composed of the amount earned for items which pass the tests minus costs; these costs include penalty costs for long delays, fixed daily costs per server and costs for testing bulks. By an appropriate choice of the decision variables, namely, the batch sizes and the number of servers at the second service station, one wants to optimize the various trade-offs involved.

The paper is organized as follows. In Section 2 we present the suggested tandem queue in detail. In Section 3 we try to get a feeling for the complexity of the last of the two queues in series, viz., the $M/G/S+G$ queue with batch services. It will turn out that the arguably simplest version of this bulk service queue with impatience, the $M/M/1+M$ queue with batch services, already gives rise to serious analytical complications for any batch size larger than one. We discuss these complications in Section 3.

In view of the fact that we aim to solve realistic optimization problems for blood screening

procedures, these analytical complications force us to take recourse to either a numerical approach or to approximations. We have chosen the first option. Section 4 presents a numerical solution for the queue length distribution in the $M/M/S+M$ queue with batch service. In Section 5 we then formulate and solve several optimization problems. In an appendix we discuss the power-series algorithm, which is an alternative numerical-analytic method of analyzing the $M/M/S+M$ queue with batch services.

2 The tandem queue

In this section we present a detailed description of the tandem queueing model for the blood screening procedure.

Station 1

It is quite natural to assume that the arrival process to the first service station is Poisson, with some rate Λ. The service times at the first station are assumed to be generally distributed. The resulting $M/G/\infty$ model has been extensively studied. Its queue length distribution is known to be Poisson with intensity ΛET, where ET denotes the mean service time [33]. It is also well-known [23] that the output process of the resulting $M/G/\infty$ queue is a Poisson process. Part of this output will leave the system because it has failed the ELISA test. As each item fails the test with a fixed probability p, independently of other items, the departure process of items which have passed the test is also Poisson. Part of that output has surpassed the expiration date. As this occurs for each item independently of all other items with a fixed probability (viz., the probability that the service time is larger than the patience time), the resulting input to station 2 still is Poisson.

Station 2

We conclude from the above that the arrival process at the second station is again Poisson, with rate $\lambda := \Lambda(1-p)P(T < G_{tot})$, where G_{tot} is a generic random variable that indicates the

total patience of an arbitrary customer (the time until the expiration date is exceeded). The customers/items which arrive in station 2 still have some time left until their expiration date: this is the difference between their overall patience G_{tot} and the service (=sojourn) time in station 1. We assume for the moment that the remaining patience times are independent and identically distributed, with some general distribution $G(\cdot)$ with mean $1/\gamma$. In most of our analysis, however, patience will be assumed to be exp(γ) distributed. This patience in principle refers to the *sojourn* time in station 2, in the sense that blood samples are no longer of use after a certain amount of time. However, an outdated blood sample won't be removed during an ongoing service. Hence, in our analysis, we shall let patience relate to the *waiting* time in the queue. The probability that a customer who has been taken into service remains "patient" during that service equals the probability that the service time is shorter than the remaining patience time – which in the case of exponential patience is still exp(γ). One may take that into consideration in an optimization study.

The number of servers at station 2 is S. Service in station 2 is in batches. If a server is free and there are less than k customers waiting, the free server does not yet start a service. If there are at least $K \geq k$ customers waiting, then it takes a batch of size K into service. If the number of customers waiting is m with $k \leq m < K$, then the free server takes a batch of size m into service. We denote the resulting queueing model by $M/G^{[k,K]}/S+G$. The parameters S, k and K are decision variables. There are obvious trade-offs here: e.g., there are costs involved with having a higher number of servers S, but this leads to a speed-up so that fewer customers will become impatient.

3 Queue lengths in the case of exponential patience

To get a feeling for the complexity of the $M/G^{[k,K]}/S+G$ model, we consider in this section the possibly simplest version: the $M/M^{[k,K]}/1+M$ model. The patience refers to *waiting* time in the queue. It will turn out that the determination of the queue length distribution of this model already gives rise to complicated analytical problems. We assume that service times

of successive batches are independent, exponentially distributed random variables with mean $1/\mu$, regardless of the size of the batch. We also assume that patience times are independent, exponentially distributed random variables with mean $1/\gamma$, independent of the service times and interarrival times. Let

$$\begin{aligned} p(n,0) &:= P(n\ waiting,\ server\ idle), \quad n = 0,1,\ldots,k-1, \\ p(n,1) &:= P(n\ waiting,\ server\ busy), \quad n = 0,1,\ldots. \end{aligned}$$

The global balance equations are:

$$\begin{aligned} p(0,0)\lambda &= p(1,0)\gamma + p(0,1)\mu, && (3.1) \\ p(n,0)(\lambda + n\gamma) &= p(n-1,0)\lambda + p(n+1,0)(n+1)\gamma + p(n,1)\mu, \quad n = 1,\ldots,k-2, \\ p(k-1,0)(\lambda + (k-1)\gamma) &= p(k-2,0)\lambda + p(k-1,1)\mu, \\ p(0,1)(\lambda + \mu) &= p(1,1)\gamma + \sum_{j=k}^{K} p(j,1)\mu + p(k-1,0)\lambda, && (3.2) \\ p(n,1)(\lambda + \mu + n\gamma) &= p(n-1,1)\lambda + p(n+1,1)(n+1)\gamma + p(n+K,1)\mu, \quad n = 1,2,\ldots. \end{aligned}$$

Remark 1.

Next to these global balance equations, one immediately sees that the following balance equation between the '0' states and the '1' states should hold:

$$p(k-1,0)\lambda = \sum_{j=0}^{k-1} p(j,1)\mu. \tag{3.3}$$

Equation (3.3) follows from the above global balance equations by taking an appropriate summation.

Now let us introduce the generating function

$$P(z) := \sum_{n=0}^{\infty} p(n,1) z^n, \quad |z| \leq 1. \tag{3.4}$$

It readily follows from (3.2) that

$$\begin{aligned} P(z)(\lambda+\mu) &+ \gamma z\frac{d}{dz}P(z) = \lambda z P(z) + \gamma\frac{d}{dz}P(z) \\ &+ \mu\sum_{j=k}^{K} p(j,1) + \lambda p(k-1,0) + \mu z^{-K}\sum_{n=1}^{\infty} z^{n+K}p(n+K,1), \end{aligned} \tag{3.5}$$

so, using (3.3),

$$\frac{d}{dz}P(z) = (\frac{\lambda}{\gamma} + \frac{\mu}{\gamma}\frac{1-z^{-K}}{1-z})P(z) + \frac{\mu}{\gamma}\sum_{j=0}^{K-1} p(j,1)\frac{z^{j-K}-1}{1-z}. \tag{3.6}$$

The solution of the homogeneous differential equation reads:

$$P(z) = D\mathrm{e}^{\frac{\lambda}{\gamma}z}z^{-\frac{\mu}{\gamma}}\mathrm{e}^{\frac{\mu}{\gamma}[\sum_{j=1}^{K-1}\frac{1}{jz^j}]}, \quad |z| \leq 1,$$

with D a constant. Variation of constants subsequently gives the complete solution of (3.6):

$$P(z) = D\mathrm{e}^{A_1(z)} + \int_0^z A_2(u)\mathrm{e}^{A_1(z)-A_1(u)}\mathrm{d}u, \tag{3.7}$$

where

$$A_1(z) := \frac{\lambda}{\gamma}z - \frac{\mu}{\gamma}\ln(z) + \frac{\mu}{\gamma}\sum_{j=1}^{K-1}\frac{1}{jz^j}, \tag{3.8}$$

and

$$A_2(z) := \frac{\mu}{\gamma}\sum_{j=0}^{K-1} p(j,1)\frac{z^{j-K}-1}{1-z}. \tag{3.9}$$

Letting $z \to 0$ in (3.7) and taking into account the behaviour of $A_1(z)$ for $z \to 0$, shows that we must take $D = 0$ in order for $P(0)$ to be finite. We still have several unknown constants. The three equations (3.1) allow us to express the unknown $p(n,0)$, $n = 0, 1, \ldots, k-1$ into the k unknowns $p(j,1)$, $j = 0, 1, \ldots, k-1$.

We now consider a few special cases. The main purpose is to reveal the mathematical difficulties which occur in determining the $K+1$ unknown constants.

Case 1: $k = K = 1$; i.e., no batches.

Now (3.7) simplifies to:

$$\begin{aligned} P(z) &= \frac{\mu}{\gamma} p(0,1) \int_0^z \frac{1}{u} (\frac{u}{z})^{\mu/\gamma} \mathrm{e}^{\frac{\lambda}{\gamma}(z-u)} \mathrm{d}u \\ &= \frac{\mu}{\gamma} p(0,1) \int_0^1 v^{\frac{\mu}{\gamma}-1} \mathrm{e}^{\frac{\lambda}{\gamma} z(1-v)} \mathrm{d}v \\ &= \frac{\mu}{\gamma} p(0,1) \sum_{n=0}^{\infty} z^n \frac{(\frac{\lambda}{\gamma})^n}{n!} \int_0^1 v^{\frac{\mu}{\gamma}-1} (1-v)^n \mathrm{d}v \\ &= p(0,1) \sum_{n=0}^{\infty} z^n (\frac{\lambda}{\gamma})^n \frac{\Gamma(\frac{\mu}{\gamma}+1)}{\Gamma(\frac{\mu}{\gamma}+1+n)}. \end{aligned} \tag{3.10}$$

Here we have used that $\int_0^1 v^{x-1}(1-v)^{y-1}\mathrm{d}v =: B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$.

This shows that

$$p(n,1) = p(0,1)(\frac{\lambda}{\gamma})^n \frac{\Gamma(\frac{\mu}{\gamma}+1)}{\Gamma(\frac{\mu}{\gamma}+1+n)}.$$

Indeed, for $K=1$ we have detailed balance equations $\lambda p(n-1,1) = (\mu + n\gamma) p(n,1)$, leading to the above relation between $p(n,1)$ and $p(0,1)$.

The unknown constants $p(0,1)$ and $p(0,0)$ are obtained by using (3.3) – which reduces to $p(0,0)\lambda = p(0,1)\mu$ – and the normalization condition $p(0,0) + P(1) = 1$.

Case 2: $k=1,\ K=2$.

We discuss this case to point out a mathematical difficulty in studying $P(z)$ for $K \geq 2$. Equation (3.7) reduces to:

$$\begin{aligned} P(z) &= p(0,1) \int_0^1 v^{\frac{\mu}{\gamma}-1} (1 + \frac{1}{zv}) \mathrm{e}^{\frac{\lambda}{\gamma} z(1-v) + \frac{\mu}{\gamma}\frac{1}{z}(1-\frac{1}{v})} \mathrm{d}v \\ &+ p(1,1) \int_0^1 v^{\frac{\mu}{\gamma}-1} \mathrm{e}^{\frac{\lambda}{\gamma} z(1-v) + \frac{\mu}{\gamma}\frac{1}{z}(1-\frac{1}{v})} \mathrm{d}v. \end{aligned} \tag{3.11}$$

Notice that, because of the $1/z$ in the exponent, it becomes harder to pick out the z^n coefficient; also notice that we need one more equation to determine the three unknowns $p(0,0)$, $p(0,1)$ and $p(1,1)$, next to (3.3) and the normalization condition. To handle the $1/z$ in the exponent, one might use the substitution $y = -\frac{\lambda}{\gamma} z(1-v) - \frac{\mu}{\gamma}\frac{1}{z}(1-\frac{1}{v})$, leading to

$$v = \frac{\gamma}{2\lambda z}[y + \frac{\lambda}{\gamma} z + \frac{\mu}{\gamma}\frac{1}{z} - ((y + \frac{\lambda}{\gamma} z + \frac{\mu}{\gamma}\frac{1}{z})^2 - 4\frac{\lambda}{\gamma}\frac{\mu}{\gamma})^{1/2}].$$

Expression (3.11) now transforms into

$$P(z) = p(0,1)\int_0^\infty \mathrm{e}^{-y} H_1(y,z)\mathrm{d}y + p(1,1)\int_0^\infty \mathrm{e}^{-y} H_2(y,z)\mathrm{d}y, \tag{3.12}$$

where $H_1(y,z)$ and $H_2(y,z)$ involve powers of the above square root term. The missing third equation might be obtained by using the analyticity of $P(z)$ inside the unit circle, via a careful study of the singularities of $H_1(y,z)$ and $H_2(y,z)$; we leave this problem to a further study.

4 The second station: Balance equations

Due to the analytical difficulties pointed out in the previous section, we discuss a numerical solution to the more general $M/M^{[k,K]}/S + M$ model in this section. As before, patience relates to the *waiting* time in the queue. First, we generalize the global balance equations (3.1) and (3.2) to the multiserver case with the batch sizes restricted to multiples of a kit size b. We assume that k and K are multiples of b, $k = mb$, $K = Mb$. Note that all batch sizes between k and K are allowed if $b = 1$ so that the model of the previous section is included. Then, we solve these equations by truncation and iterations, and express various performance measures of interest in terms of the state probabilities. The resulting algorithm forms the basis for the optimization study in the next section. In Appendix B an alternative approach using power-series expansions is described which did not work out well for the present model.

Define $p(n,s)$ as the probability that n customers are waiting in the queue, $n = 0,1,2,\ldots,$ while s servers are occupied with a bulk of customers of an unspecified size, $s = 0,1,\ldots,S$. For states with $s = 0$ (all servers idle) the global balance equations read (cf. (3.1)): for $n = 0,1,\ldots,k-1$,

$$(\lambda + n\gamma)p(n,0) = \lambda I_{\{n>0\}} p(n-1,0) + (n+1)\gamma I_{\{n<k-1\}} p(n+1,0) + \mu p(n,1); \tag{4.1}$$

here $I_{\{\cdot\}}$ denotes an indicator function. For states with $s = 1,\ldots,S-1$, the global balance equations read (cf. (3.2)): for $n = 0,1,\ldots,k-1$,

$$\begin{aligned}(\lambda + n\gamma + s\mu)p(n,s) = {} & \lambda I_{\{n>0\}} p(n-1,s) + \lambda I_{\{n=0\}} p(k-1,s-1) \\ & + (n+1)\gamma I_{\{n<k-1\}} p(n+1,s) + (s+1)\mu p(n,s+1);\end{aligned} \tag{4.2}$$

while for $s = S$ (all servers occupied) the global balance equations read: for $n = 0$,

$$(\lambda + S\mu)p(0, S) = \lambda p(k-1, S-1) + \gamma p(1, S) + S\mu \sum_{h=m}^{M} p(hb, S); \tag{4.3}$$

for $n = 1, 2, \ldots, b-1$,

$$(\lambda + n\gamma + S\mu)p(n, S) = \lambda p(n-1, S) + (n+1)\gamma p(n+1, S) + S\mu \sum_{h=m}^{M} p(n+hb, S); \tag{4.4}$$

and for $n = b, b+1, \ldots,$

$$(\lambda + n\gamma + S\mu)p(n, S) = \lambda p(n-1, S) + (n+1)\gamma p(n+1, S) + S\mu p(n+K, S). \tag{4.5}$$

Summing the global balance equations for fixed s, $s = 0, 1, \ldots, S-1$, over n, $n = 0, 1, \ldots, k-1$, yields with induction the following balance equations for transitions between the levels s and $s+1$, cf. (3.3):

$$\lambda p(k-1, s) = (s+1)\mu \sum_{j=0}^{k-1} p(j, s+1), \quad s = 0, 1, \ldots, S-1. \tag{4.6}$$

The average size of a service batch B can be computed as the quotient of the total rate at which batches of various size start over the rate at which services start. Batch services start when the queue length process is in states $(k-1, s)$, $s = 0, 1, \ldots, S-1$, with rate λ at which arrivals occur and then have the minimum size k, in states (n, S) with $n = hb + j$, $h = m, \ldots, M-1$, $j = 0, \ldots, b-1$, with rate $S\mu$ at which a server becomes available and then have size hb, and in states (j, S), $j = K, K+1, \ldots$, with rate $S\mu$ at which a server becomes available and then have the maximum size K. Hence,

$$E\{B\} = \frac{\lambda k \sum_{s=0}^{S-1} p(k-1, s) + S\mu \sum_{h=m}^{M-1} hb \sum_{j=0}^{b-1} p(hb+j, S) + S\mu K \sum_{j=K}^{\infty} p(j, S)}{\lambda \sum_{s=0}^{S-1} p(k-1, s) + S\mu \sum_{j=k}^{\infty} p(j, S)}. \tag{4.7}$$

Using the balance equation (4.6) between levels s and $s+1$, $s = 0, 1, \ldots, S-1$, the first terms in the numerator and the denominator of (4.7) can be replaced, and the factor μ can be canceled:

$$E\{B\} = \frac{k \sum_{s=1}^{S} s \sum_{j=0}^{k-1} p(j, s) + S \sum_{h=m}^{M-1} hb \sum_{j=0}^{b-1} p(hb+j, S) + SK \sum_{j=K}^{\infty} p(j, S)}{\sum_{s=1}^{S} s \sum_{j=0}^{k-1} p(j, s) + S \sum_{j=k}^{\infty} p(j, S)}. \tag{4.8}$$

The denominator represents the average number of busy (occupied) servers, to be denoted by $E\{O\}$:

$$E\{O\} = \sum_{s=1}^{S} s \sum_{j=0}^{k-1} p(j,s) + S \sum_{j=k}^{\infty} p(j,S).$$

The rate at which customers renege is

$$\gamma \sum_{s=0}^{S} \sum_{n=1}^{\infty} np(n,s) = \gamma E\{Q\};$$

here, $E\{Q\}$ denotes the average number of customers waiting in the queue for service. Hence, the fraction of customers that is lost due to reneging is

$$P_{\text{loss}} = \frac{\gamma}{\lambda} E\{Q\}. \tag{4.9}$$

The rate at which served customers leave the system is $\lambda_{\text{out}} = \lambda[1 - P_{\text{loss}}]$. Since $P_{\text{loss}} \leq 1$, (4.9) implies that for all parameter values,

$$E\{Q\} \leq \frac{\lambda}{\gamma}. \tag{4.10}$$

By Little's law, the mean time in queue of the customers is $E\{W\} = E\{Q\}/\lambda$, but this includes both customers who renege and those who are ultimately served. For the mean number of customers in the system, $E\{N\}$, Little's law implies $E\{N\} = \lambda E\{R\}$, with $E\{R\}$ the mean time in system. The difference of these relations yields

$$E\{N\} - E\{Q\} = \lambda E\{R - W\} = \frac{\lambda}{\mu}[1 - P_{\text{loss}}],$$

since $R - W$ represents the time in service, and this has a mean of $1/\mu$ if a customer gets served and of 0 for a reneging customer. Hence, with (4.9),

$$E\{N\} = \frac{\lambda}{\mu} + \left(1 - \frac{\gamma}{\mu}\right) E\{Q\}. \tag{4.11}$$

Note that if $\gamma = \mu$ we have for all S, k and K,

$$E\{N\} = \frac{\lambda}{\mu}.$$

From (4.11) and (4.9) it follows that the mean sojourn time can be written as

$$E\{R\} = \frac{1}{\mu} + \left(\frac{1}{\gamma} - \frac{1}{\mu}\right) P_{\text{loss}}. \tag{4.12}$$

Table 1: A case study with bulk sizes between k and K; $\lambda = 12$, $\mu = 2$, $\gamma = 0.2$, $p_b = 0.001$.

S	k	K	$E\{Q\}$	$E\{N\}$	P_{loss}	$E\{R\}$	$E_{\mathcal{S}}\{R\}$	$E\{B\}$	$E\{O\}$	λ_{out}	λ_{good}
1	6	6	9.75	14.78	0.1626	1.232	1.325	6.00	0.84	10.05	9.99
1	6	12	5.37	10.83	0.0894	0.902	0.946	7.56	0.72	10.93	10.84
1	6	18	4.90	10.41	0.0817	0.868	0.907	7.81	0.71	11.02	10.92
1	6	24	4.80	10.32	0.0800	0.860	0.899	7.87	0.70	11.04	10.93
1	12	12	6.59	11.93	0.1098	0.994	1.060	12.00	0.45	10.68	10.55
1	12	18	6.31	11.68	0.1052	0.974	1.036	12.46	0.43	10.74	10.60
1	12	24	6.25	11.63	0.1042	0.969	1.031	12.56	0.43	10.75	10.61
2	6	6	2.93	8.63	0.0488	0.720	0.744	6.00	0.95	11.43	11.36
2	6	12	2.75	8.48	0.0459	0.707	0.730	6.20	0.92	11.45	11.38
2	6	18	2.74	8.47	0.0457	0.706	0.729	6.21	0.92	11.45	11.38
2	12	12	5.62	11.06	0.0937	0.922	0.977	12.00	0.45	10.88	10.75
3	6	6	2.53	8.28	0.0421	0.690	0.711	6.00	0.96	11.49	11.42
3	6	12	2.52	8.27	0.0420	0.689	0.711	6.01	0.96	11.50	11.43
3	12	12	5.61	11.05	0.0935	0.921	0.976	12.00	0.45	10.88	10.75
4	6	6	2.51	8.26	0.0418	0.688	0.709	6.00	0.96	11.50	11.43

The computation of other characteristics of the sojourn time distribution than its mean is more involved. This is illustrated in Appendix A, where also the computation of the conditional mean sojourn time $E_{\mathcal{S}}\{R\}$, given that the customer is eventually served, is explained.

It is assumed that each item (customer) is of bad quality with probability p_b. The result of testing a bulk of items is either good — if all items in the bulk are good — or bad otherwise. Hence, a bulk of hb items passes the test with probability $(1-p_b)^{hb},\ h = m, \ldots, M$. The rate at which items, which have passed the test, leave the system will be denoted by λ_{good}. This rate is found in a similar way as (4.7):

$$\lambda_{\text{good}} = \lambda k(1-p_b)^k \sum_{s=0}^{S-1} p(k-1,s) + S\mu \sum_{h=m}^{M-1} hb(1-p_b)^{hb} \sum_{j=0}^{b-1} p(hb+j,S) + S\mu K(1-p_b)^K \sum_{j=K}^{\infty} p(j,S). \tag{4.13}$$

Table 1 contains a few cases evaluated by truncating the number of waiting customers at 100. Here, the arrival rate of $\lambda = 12$, the service rate of $\mu = 2$ and the reneging rate $\gamma = 0.2$ are fixed, and the number of servers S, the minimum bulk size k and the maximum bulk size K are varied. We take $b = 1$ but restrict ourselves to k, K values which are multiples of 6 as kits of size 6 are often used in practice. The finite set of balance equations is solved by iteration,

Table 2: A case study with bulk sizes multiples of 6; $\lambda = 12$, $\mu = 2$, $\gamma = 0.2$, $p_b = 0.001$.

S	k	K	$E\{Q\}$	$E\{N\}$	P_{loss}	$E\{R\}$	$E_{\mathcal{S}}\{R\}$	$E\{B\}$	$E\{O\}$	λ_{out}	λ_{good}
1	6	12	6.00	11.40	0.1000	0.950	1.002	7.14	0.76	10.80	10.71
1	6	18	5.61	11.05	0.0936	0.921	0.970	7.30	0.75	10.88	10.78
1	6	24	5.53	10.98	0.0922	0.915	0.963	7.33	0.74	10.89	10.80
1	12	18	6.42	11.78	0.1070	0.981	1.046	12.28	0.44	10.72	10.58
1	12	24	6.38	11.74	0.1063	0.979	1.042	12.34	0.43	10.72	10.59
2	6	12	2.83	8.55	0.0472	0.712	0.737	6.09	0.94	11.43	11.36
2	6	18	2.82	8.54	0.0471	0.712	0.736	6.10	0.94	11.44	11.36
3	6	12	2.53	8.27	0.0421	0.689	0.711	6.00	0.96	11.49	11.43

until the sum of the absolute differences between the jth and the $(j-1)$st iterant is smaller than $\epsilon = 10^{-10}$. The truncation error is smaller than this ϵ with the foregoing truncation level. To compute a value for λ_{good}, the probability of a bad item is taken as $p_b = 0.001$. Note that in the case $S = 1$, $k = K = 6$, the system would be unstable if customers did not renege. To increase λ_{out} and λ_{good} starting from this boundary case, adding a second server has a stronger influence on the performance of the system than changing the minimum and maximum bulk sizes, but adding a third server only leads to a minor increase. Further, it seems best in this example to set the minimum batch size at $k = 6$. For $S \geq 2$ and $k = 6$, increasing the maximum batch size K only leads to minor improvements of λ_{good}.

Table 2 considers the same cases as Table 1 but with the restriction that the bulk sizes can only be multiples of the kit size $b = 6$. Of course, if $k = K$ there is no difference between the systems. It turns out that this more restricted system performs worse (more congestion and more reneging) with comparable parameter settings, as could be expected.

Remark 2

The approach of the present section can be extended to the case of phase-type service times, at the expense of added complexity and a state-space explosion.

Table 3: Optimal values for S, k and K for the case $\mu = 4$, $\gamma = 0.3$, $p_b = 0.001$, $G = 100$, $C_p = 32$, $C_s = 50$, $C_b = 5$, $C_i = 1$, $b = 1$, for various values of λ.

λ	S	k	K	$C(S,k,K)$	$G\lambda_{\text{good}}$	$C_p\lambda_{\text{good}}E\{R\}$	C_B	C_sS
30	2	6	24	2418.42	2888.05	318.12	51.52	100.00
60	4	6	24	5019.47	5877.33	551.39	106.47	200.00
300	9	12	24	26056.45	29454.45	2527.97	420.03	450.00
600	15	12	24	52542.58	59041.96	4911.79	837.59	750.00
3000	48	18	24	264231.45	294276.73	23825.40	3819.88	2400.00
6000	92	18	24	529226.15	588863.71	47391.45	7646.12	4600.00

5 Performance optimization

The M/M/S system with bulk services and deadlines will be optimized with respect to the number of servers S, the minimum bulk size k and the maximum bulk size K. The latter are restricted to the values $6 \leq k \leq K \leq 24$, while k and K are six folds but $b = 1$. This means that for each value of S, 10 combinations of k and K will be considered. The objective is to maximize profit per day. It is assumed that an amount G is earned for each item in a bulk that passes the test. To compensate for long responses and, consequently, rather useless items, a penalty cost of C_p is included per expected number of days in the system per item. By taking $C_p \approx G/3.5$ nothing will be earned if $E\{R\} \geq 3.5$ days (leaving 0.5 day for the first phase of testing). Further costs are a fixed daily cost C_s per server, and a cost $C_b + C_i b$ for testing a bulk of size b, $k \leq b \leq K$. Note that the average cost of a bulk is $C_b + C_i E\{B\}$, while $\lambda_{\text{out}}/E\{B\}$ bulks are tested per day, on the average. Hence, $C_B \doteq (C_b + C_i E\{B\})\lambda_{\text{out}}/E\{B\}$ is the average daily cost of testing the bulks. Summarizing, we have the following daily profit:

$$C(S,k,K) = G\lambda_{\text{good}} - C_p\lambda_{\text{good}}E\{R\} - C_sS - (C_b + C_iE\{B\})\lambda_{\text{out}}/E\{B\}. \qquad (5.1)$$

Table 3 shows the optimal values of S, k and K, and the corresponding maximum profit, for the case $\mu = 4$, $\gamma = 0.3$, $p_b = 0.001$, $G = 100$, $C_p = 32$, $C_s = 50$, $C_b = 5$, $C_i = 1$, for various values of the arrival rate λ. For higher values of λ a higher truncation level is required to keep the truncation error below, say, 10^{-8}, up to 1000 for $\lambda = 6000$. This last case took more than 8 minutes cpu time to evaluate about 300 parameter settings for models with up to 3200 states.

Table 4: Sensitivity analysis on base case with $\lambda = 600$, $\mu = 4$, $\gamma = 0.3$, $p_b = 0.001$.

λ	S	k	K	$C(S,k,K)$	$G\lambda_{\text{good}}$	$C_p\lambda_{\text{good}}E\{R\}$	C_B	C_sS
base case	15	12	24	52542.58	59041.96	4911.79	837.59	750.00
$p_b = 0.01$	31	6	6	49128.35	56387.32	4610.94	1098.02	1550.00
$C_s = 200$	10	18	24	50721.59	58512.15	5035.21	755.35	2000.00
$C_b = 25$	12	18	24	51676.88	58651.08	4954.10	1420.10	600.00
$C_i = 10$	15	12	24	47160.05	59041.96	4911.79	6220.12	750.00
$C_p = 16$	15	12	24	54998.47	59041.96	2455.89	837.59	750.00
$G = 50$	14	12	24	23032.52	29493.10	4929.53	831.05	700.00
$G = 25$	11	18	24	8364.01	14651.82	4978.18	759.63	550.00
$\mu = 2$	26	12	24	47200.39	58967.75	9637.83	829.53	1300.00
$\gamma = 0.1$	15	12	24	52646.01	59168.44	4933.22	839.21	750.00
multiple 6	15	12	24	52535.16	59040.16	4916.65	838.34	750.00
deadline 3	14	12	24	52698.08	59198.13	4966.40	833.65	700.00

We restrict the search to combinations with $SK \geq \lambda/\mu$ because models which would not be stable without reneging are expected to be far from optimal. Further, the maximum profit over all k and K for fixed S seems to be a concave function of S. We stopped the search when S is four more than the current best S to be safe. It seems that taking K as large as possible is best in all cases. Further, the best values for k and, of course, S are increasing with the arrival rate λ.

Table 4 contains a sensitivity analysis of the optimum for the case $\lambda = 600$ with respect to each of the other parameters of the system. If the fraction of bad items p_b increases, the optimal number of servers strongly increases while the maximum bulk size decreases to reduce the waste of good items. If the cost per server C_s increases, the optimal number of servers decreases while the minimum bulk size increases to let the servers handle larger bulks. Note that nevertheless the bulk cost per day C_B decreases due to a larger fraction of reneging customers. If the fixed cost per bulk C_b increases, the optimal minimum bulk size increases while the number of servers decreases. If the item-dependent cost C_i increases, the optimal strategy does not change, only the maximum profit becomes lower. If the penalty cost for congestion C_p decreases, the optimal strategy does not change, only the maximum profit becomes higher. If the gain per item G decreases to 50, the optimal number of servers slightly decreases. If it

further decreases to 25, the optimal number of servers more strongly decreases because the gain no longer dominates the costs as strongly as in other cases. If the mean service time $1/\mu$ increases, the optimal number of servers strongly increases. If the average deadline $1/\gamma$ increases from $3\frac{1}{3}$ day to 10 days, the optimal strategy does not change while the maximum profit slightly increases since less customers renege. If the bulk sizes are restricted to multiples of the kit size 6, the optimal number of servers is the same and the maximum profit is slightly less. Finally, if the (exponential) reneging rate is set equal to 0, but a finite buffer is chosen of $3KS\mu$ — so that the time in queue of an item that arrives when the queue is almost full is 3 days with a small variance, mimicking a constant deadline —, the optimal number of servers slightly decreases (but the profit with 15 servers is very close: 52697.97). In comparison to the base case, the difference in performance stems from changing γ; the loss probability is negligible in both cases.

Remark 3

The factor $E\{R\}$ in the objective function (5.1) includes items that have reneged. The mean sojourn time $E_{\mathcal{S}}\{R\}$ of items that are actually tested could be computed as indicated in Appendix A. However, the computation of $E_{\mathcal{S}}\{R\}$ takes much more time than the numerical solution of the global balance equations, and this burden increases with λ. It might be expected that the optimum will not differ since it is quite insensitive to the penalty cost C_p. Indeed, replacing $E\{R\}$ by $E_{\mathcal{S}}\{R\}$ in (5.1) leads for $\lambda = 30$ up to $\lambda = 600$ to the same optimal values of S, k and K with a slightly smaller maximum profit than in Table 3.

Remark 4

It should be noticed that station 1 plays no direct role in the optimization problem, due to its behaviour as a delay system (infinite server system). Its role is confined to a reduction of the arrival rate from Λ to λ (at station 2), and an adaptation of the patience at station 2 (by subtracting the time spent in station 1).

Remark 5

We have tried to apply successive overrelaxation (cf. Tijms [35], Appendix D) to speed up the computation of the steady-state distribution. A factor 1.2 gave divergence in all experiments, a factor 1.1 gave a speed-up in some cases but divergence in other cases. So for reliability's sake we have used plain Gauss-Seidel iteration to generate the tables.

Remark 6

In this paper we have mainly concentrated on the case of exponential patience. If patience at the second station were deterministic (which is a reasonable assumption in the case of blood testing), then we'd like to suggest the following approximation, which is based on an idea that was developed in [15] to approximate loss probabilities in multiserver queues with impatience. Let us assume that patience is deterministic D, and that D is considerably larger than the mean service time $1/\mu$; otherwise a large percentage of the customers would even become impatient in light traffic. Our first observation is that we may assume that customers who eventually become impatient, are rejected immediately. Consider a tagged customer C who finds j customers waiting upon arrival. C's waiting time is the sum of $\lfloor \frac{j}{K} \rfloor + 1$ exp(μ) service times, which is Erlang($\lfloor \frac{j}{K} \rfloor + 1, \mu$) distributed. Our second observation is that such an Erlang distribution can be quite accurately approximated by a deterministic (degenerate) distribution if $\lfloor \frac{j}{K} \rfloor + 1$ is not too small. This implies that there will be a sharp distinction between $(\lfloor \frac{j}{K} \rfloor + 1)\frac{1}{\mu} < D$ and $(\lfloor \frac{j}{K} \rfloor + 1)\frac{1}{\mu} \geq D$. In the former case, C is likely to remain patient, and in the latter case he is likely to become impatient (i.e., be rejected immediately). Hence we propose to approximate the queue length distribution in the $M/M^{[k,K]}/1 + D$ system by that in the $M/M^{[k,K]}/1/R$ system with finite waiting room R, where $R = K\lceil \mu D - 1 \rceil$. The latter system has been studied in [10], which paper allows the service time distribution to be general and depending on the batch size. As a final remark, we'd like to refer to [22] for approximations for performance measures of the $M/G/N + G$ queue. It may be worthwhile to adapt their approach to the case of batch service.

A Sojourn times for served customers

The distribution of the waiting time W (time in queue) can be obtained both for customers who are eventually served and for those who renege by conditioning on the state in which they find the system upon arrival (using PASTA). The conditional waiting time given the state upon arrival is equivalent to the time until absorption into either the service state $\mathcal{S}$ or the reneging state $\mathcal{R}$ of a possibly infinite-state Markov process, cf. e.g. Neuts [28, Section 3.9]. Let Q_A^+ be the number of customers in the queue and let S_A^+ be the number of busy servers just *after* the arrival of a tagged customer. Then, the joint probability that this tagged customer has a waiting time less than t and is served can be written as

$$P\{W \leq t, \mathcal{S}\} = \sum_{s=0}^{S-1} p(k-1,s) + \sum_{s=0}^{S-1}\sum_{n=0}^{k-2} p(n,s)\tilde{q}_{n+1}^{\mathcal{S}}(t) + \sum_{n=0}^{\infty} p(n,S) q_{n+1}^{\mathcal{S}}(t); \qquad \text{(A.1)}$$

with

$$\tilde{q}_i^{\mathcal{S}}(t) = P\{W \leq t, \mathcal{S} | Q_A^+ = i, S_A^+ < S\}, \quad q_i^{\mathcal{S}}(t) = P\{W \leq t, \mathcal{S} | Q_A^+ = i, S_A^+ = S\}. \qquad \text{(A.2)}$$

Similarly, the joint probability that this tagged customer has a waiting time less than t and reneges can be written as

$$P\{W \leq t, \mathcal{R}\} = \sum_{s=0}^{S-1}\sum_{n=0}^{k-2} p(n,s)\tilde{q}_{n+1}^{\mathcal{R}}(t) + \sum_{n=0}^{\infty} p(n,S) q_{n+1}^{\mathcal{R}}(t); \qquad \text{(A.3)}$$

with

$$\tilde{q}_i^{\mathcal{R}}(t) = P\{W \leq t, \mathcal{R} | Q_A^+ = i, S_A^+ < S\}, \quad q_i^{\mathcal{R}}(t) = P\{W \leq t, \mathcal{R} | Q_A^+ = i, S_A^+ = S\}. \qquad \text{(A.4)}$$

Case 1: $S_A^+ < S$

The situation when a customer finds at least one server idle is relatively simple. In this case the number of busy servers is irrelevant, because the new arrival will either renege or become part of the first formed batch (of size k). The Markov process to be considered for the conditional waiting time has transient states $\{1, 2, \ldots, k-1\}$, representing the number of customers present in the queue, and absorbing states $\mathcal{S}$, entered with rate λ from state $k-1$, and $\mathcal{R}$, entered with rate γ from all transient states $\{1, 2, \ldots, k-1\}$. Let $\hat{Q}(t)$ be the number of customers in

the queue and let $\hat{S}(t)$ be the number of busy servers at time t after the arrival of the tagged customer. The auxiliary functions for the transient states,

$$\tilde{q}_i(t,n) = P\{\hat{Q}(t) = n,\, \hat{S}(t) < S | \hat{Q}(0) = i,\, \hat{S}(0) < S\}, \quad i,n = 1,\ldots,k-1, \tag{A.5}$$

satisfy the forward differential equations: for $i,n = 1,\ldots,k-1$,

$$\frac{d}{dt}\,\tilde{q}_i(t,n) = \lambda I_{\{n>1\}}\tilde{q}_i(t,n-1) + n\gamma I_{\{n<k-1\}}\tilde{q}_i(t,n+1) - (\lambda + n\gamma)\tilde{q}_i(t,n). \tag{A.6}$$

The conditional waiting time probabilities are then determined by: for $i = 1,\ldots,k-1$,

$$\frac{d}{dt}\,\tilde{q}_i^{\mathcal{S}}(t) = \lambda\tilde{q}_i(t,k-1); \quad \frac{d}{dt}\,\tilde{q}_i^{\mathcal{R}}(t) = \gamma\sum_{n=1}^{k-1}\tilde{q}_i(t,n). \tag{A.7}$$

The conditional mean waiting times can be computed as follows. First, write the forward differential equations (A.6) for the transient states in the matrix-vector form

$$\frac{d}{dt}\,\tilde{\mathbf{q}}_i(t) = A\tilde{\mathbf{q}}_i(t). \tag{A.8}$$

Then, solve $\boldsymbol{\omega}_i$ from $A^2\boldsymbol{\omega}_i = \mathbf{e}_i$, with $\mathbf{e}_i$ the ith unit vector of length $k-1$. Finally,

$$E\{W,\mathcal{S}|Q_A^+ = i, S_A^+ < S\} = \lambda\omega_i(k-1), \quad E\{W,\mathcal{R}|Q_A^+ = i, S_A^+ < S\} = \gamma\sum_{n=1}^{k-1}\omega_i(n). \tag{A.9}$$

Case 2: $S_A^+ = S$

The situation when an arriving customer finds all servers busy is more complicated. In this case we have to keep track of the total number of customers in the queue to determine whether or not a batch can be formed, and of the rank in the queue of the tagged customer to determine whether this customer becomes part of a batch or not. So the Markov process has transient states $\{(r,n), 1 \le r \le n\}$ with all servers busy plus $\{1,2,\ldots,k-1\}$ with at least one server idle (here, the number of idle servers and the rank of the tagged customer become irrelevant again), and absorbing states $\mathcal{S}$ and $\mathcal{R}$. The service state $\mathcal{S}$ is entered with rate λ from state $k-1$ with an idle server and with rate $S\mu$ from states (r,n) with all servers busy, if $r \le K$ and $n \ge K$, or if $n = hb+j$, $h = m,\ldots,M-1$, $j = 0,\ldots,b-1$ and $r \le hb$. The reneging state $\mathcal{R}$ is entered with rate γ from all transient states. Let $\hat{R}(t)$ be the rank in the queue of the tagged customer

at time t after the arrival of this customer. The auxiliary functions for the transient states, for $i = 1, 2, \ldots,\ r = 1, \ldots, i,\ n = r, r+1, \ldots,$

$$q_i(t, r, n) = P\{\hat{R}(t) = r,\ \hat{Q}(t) = n,\ \hat{S}(t) = S | \hat{R}(0) = \hat{Q}(0) = i,\ \hat{S}(0) = S\}, \tag{A.10}$$

and for $i = 1, 2, \ldots, n = 1, \ldots, k-1$,

$$q_i(t, n) = P\{\hat{R}(t) \leq \hat{Q}(t) = n,\ \hat{S}(t) < S | \hat{R}(0) = \hat{Q}(0) = i,\ \hat{S}(0) = S\}, \tag{A.11}$$

satisfy the forward differential equations: for $i = 1, 2, \ldots,\ r = 1, \ldots, i,\ n = r, r+1, \ldots,$

$$\begin{aligned}\frac{d}{dt}\, q_i(t, r, n) &= \lambda I_{\{n>r\}} q_i(t, r, n-1) + S\mu I_{\{n<b\}} \sum_{h=m}^{M-1} q_i(t, r+hb, n+hb) \\ &\quad + S\mu I_{\{r+K\leq i\}} q_i(t, r+K, n+K) + r\gamma I_{\{r<i\}} q_i(t, r+1, n+1) \\ &\quad + (n+1-r)\gamma q_i(t, r, n+1) - (\lambda + n\gamma + S\mu) q_i(t, r, n);\end{aligned} \tag{A.12}$$

and for $i = 1, 2, \ldots,\ n = 1, \ldots, k-1$,

$$\frac{d}{dt}\, q_i(t, n) = \lambda I_{\{n>1\}} q_i(t, n-1) + n\gamma I_{\{n<k-1\}} q_i(t, n+1) + S\mu \sum_{r=1}^{n} q_i(t, r, n) - (\lambda + n\gamma) q_i(t, n). \tag{A.13}$$

The conditional waiting time probabilities (A.4) are then determined by: for $i = 1, 2, \ldots,$

$$\frac{d}{dt}\, \tilde{q}_i^{\mathcal{S}}(t) = \lambda q_i(t, k-1) + S\mu \sum_{h=m}^{M-1} \sum_{j=0}^{b-1} \sum_{r=1}^{hb} q_i(t, r, hb+j) + S\mu \sum_{n=K}^{\infty} \sum_{r=1}^{K} q_i(t, r, n); \tag{A.14}$$

$$\frac{d}{dt}\, \tilde{q}_i^{\mathcal{R}}(t) = \gamma \sum_{n=1}^{k-1} q_i(t, n) + \gamma \sum_{r=1}^{m} \sum_{n=r}^{\infty} q_i(t, r, n). \tag{A.15}$$

The conditional mean waiting times can be computed in a similar way as (A.8) and (A.9), but to write the forward differential equations (A.12) and (A.13) for the transient states in a matrix-vector form the states have to be put on a row: first the states $\{1, 2, \ldots, k-1\}$ with at least one server idle, and then the states $(1,1), (1,2), \ldots, (1,T_1), (2,2), (2,3), \ldots, (2,T_2), \ldots, (i,i), (i,i+1), \ldots, (i,T_i)$ with all servers busy; here, T_r are suitable truncation levels, $r = 1, \ldots, i$. Further details are left to the reader.

Once the conditional mean waiting times like (A.9) have been computed, the conditional mean waiting times $E_{\mathcal{S}}\{W\}$, given that a customer is eventually served, and $E_{\mathcal{R}}\{W\}$, given

that a customer reneges, can be computed with the aid of (A.1) and (A.3). The conditional mean sojourn times then follow from $E_{\mathcal{S}}\{R\} = E_{\mathcal{S}}\{W\} + 1/\mu$ and $E_{\mathcal{R}}\{R\} = E_{\mathcal{R}}\{W\}$. As a check, it should hold that $(1 - P_{\rm loss})E_{\mathcal{S}}\{R\} + P_{\rm loss}E_{\mathcal{R}}\{R\} = E\{R\}$, cf. (4.12).

B Power-series expansions

An option for numerical analysis of the M/M/S queue with batch services and deadlines that we have considered is the power-series algorithm (PSA) as summarized in Blanc [13]. The PSA has proven to be a very useful analytic-numerical method for the analysis of multi-dimensional Markov chains. In principle, it can be applied to any Markov chain that satisfies some mild regularity conditions; see also Hooghiemstra and Koole [20]. Its main limitations are the storage requirements for the coefficients of the power-series expansions of the joint queue-length probabilities (but this is not an issue in this one-dimensional model) and the fact that some technical parameters have to be tuned by trial and error for proper convergence of the power series and for avoiding numerical inaccuracy.

Because arrivals occur one by one, the stationary state probabilities for the current model allow the following power-series expansions with $\theta = \lambda$:

$$p(n,s) = \theta^{n+sk} \sum_{\ell=0}^{\infty} \theta^{\ell}\, b(\ell;n,s), \quad n = 0,1,\ldots,k-1, \;\; s = 0,1,\ldots,S-1; \tag{B.1}$$

and

$$p(n,S) = \theta^{n+Sk} \sum_{\ell=0}^{\infty} \theta^{\ell}\, b(\ell;n,S), \quad n = 0,1,\ldots. \tag{B.2}$$

Substitution of these power-series expansions into the global balance equations leads after some rearrangements to a set of recursions for the coefficients $b(\ell;n,s)$. The coefficients $b(\ell;0,0)$ which are not included in the above scheme because the factor in front of this coefficient vanishes, are determined by the requirement that the probabilities sum to one. The efficiency of the algorithm is further enhanced in [13]. A peculiarity of the current system with reneging customers is that it is stable for all $\lambda > 0$. If a system is stable for all positive values of the arrival rate, then the bilinear mapping discussed in [12] for obtaining convergence of the series

Table 5: Some test results with the PSA.

λ	μ	S	k	K	γ	$E\{Q\}$	$p(0,0)$	P_{loss}	$E\{B\}$	$E\{O\}$	G	H	L
0.95	1.0	1	1	1	1.0	0.3367	0.3867	0.3545	1.00	0.61	0.00	1.0	20
0.95	1.0	1	1	1	0.1	1.6376	0.2138	0.1724	1.00	0.79	0.28	1.0	25
0.95	0.5	1	2	2	0.5	0.7856	0.2319	0.4135	2.00	0.56	1.00	0.54	25
0.95	0.5	1	2	2	0.1	2.0689	0.0990	0.2178	2.00	0.74	1.00	0.1	40
0.95	0.5	1	1	2	0.5	0.8475	0.1783	0.4460	1.28	0.82	0.00	1.0	25
0.95	0.5	2	1	1	0.5	0.4833	0.1496	0.2544	1.00	1.42	0.00	1.0	20
0.95	1/8	2	2	4	0.1	2.9192	0.0086	0.3073	2.88	1.83	1.00	0.1	60
0.95	1/12	2	6	6	1/12	3.5336	0.0050	0.3100	6.00	1.31	1.00	0.07	120

(B.1) and (B.2) is not suitable. In such cases a bilinear mapping of the following form may be useful:

$$\theta = \frac{\lambda}{H + G\lambda}, \qquad \lambda = \frac{H\theta}{1 - G\theta}. \tag{B.3}$$

This maps $\lambda = \infty$ to $\theta = 1/G$. Table 5 displays some results obtained with the PSA; here, G and H are the parameters of the mapping (B.3) and L is the level (power of θ) at which the summations in (B.1) and (B.2) have been truncated. As is illustrated in this table, suitable values for these parameters, which have to be found by trial and error, vary strongly with the parameters of the model. And in contrast to other applications of the PSA, minor deviations of the stated values of G and H make the PSA unstable in the sense that adding more and more terms does not improve the convergence of the series due to loss of accuracy, even in conjunction with the epsilon algorithm, cf. [13]. The reason for the problem may lie in an interplay of the effects of reneging customers and batch services. On the one hand, entire functions like $e^{-\lambda}$ (in the case $\gamma = \mu = 1$, $k = K = S = 1$) may appear in the queue-length distribution of systems with reneging customers. Rational functions produced by the PSA with the epsilon algorithm will all fail to be good approximations for large enough values of λ. In this case, the version of the PSA with postponed normalization (PSA/N) as introduced in [20] suffers the same problem since it has to approximate the function $e^{+\lambda}$. On the other hand, the queue-length distributions of M/M/1 systems with batch services are known to possess branch points as function of λ, which may lie close to the origin $\lambda = 0$. Branch points can be approximated by rational functions, but their presence close to the origin influences good choices for G and H. Unfortunately, this

makes the PSA in its present forms unsuitable for optimization of the current system whereby the performance has to be evaluated for a variety of parameter settings, in contrast to the successful optimization in, e.g., Blanc and Van der Mei [14].

Acknowledgment

The research of Shaul Bar-Lev and David Perry was supported by a visitor grant from the Dutch Science Foundation NWO.

References

[1] Abolnikov, L. and Dukhovny, A. (1992). Markov chains with transition delta-matrix; ergodicity conditions, invariant probability measures and applications. *Journal of Applied Mathematics and Stochastic Analysis* 5 (1), 83–98.

[2] Abolnikov, L. and Dukhovny, A. (1999). Queueing processes and optimization problems in quality control systems with a group-individual testing procedure. *Engineering Simulation* 16, 165–178.

[3] Baccelli, F. and Hébuterne, G. (1981). On queues with impatient customers. In: F.J. Kylstra (ed.). *Performance '81* (North-Holland Publ. Cy., Amsterdam), pp. 159-179.

[4] Baccelli, F., Boyer, P. and Hébuterne, G. (1984). Single-server queues with impatient customers. *Adv. Appl. Probab.* 16, 887–905.

[5] Bar-Lev, S.K., Boneh, A. and Perry, D. (1990). Incomplete identification models for group testable items. *Naval Research Logistics* 37, 647–659.

[6] Bar-Lev, S.K., Stadje, W. and van der Duyn Schouten, F.A. (2003). Hypergeometric group testing models with incomplete information. *Probability in the Engineering and Informational Sciences* 17, 335–350.

[7] Bar-Lev, S.K., Stadje, W. and van der Duyn Schouten, F.A. (2004). Optimal group testing with processing times and incomplete identification. *Methodology and Computing in*

Applied Probability 6, 55–72.

[8] Bar-Lev, S.K., Stadje, W. and van der Duyn Schouten, F.A. (2005). Multinomial group testing models with incomplete identification. *Journal of Statistical Planning and Inference* 135, 384-401.

[9] Bar-Lev, S.K., Stadje, W. and van der Duyn Schouten, F.A. (2006). Group testing procedures with incomplete identification and unreliable testing results. *Applied Stochastic Models in Business and Industry* 22, 281-296.

[10] Bar-Lev, S.K., Parlar, M., Perry, D., Stadje, W. and van der Duyn Schouten, F.A. (2007). Applications of bulk queues to group testing models with incomplete identification. *European Journal of Operational Research* 183, 226–237.

[11] Barrer, D.Y. (1957). Queuing with impatient customers and ordered service. *Oper. Res.* 5, 650–656.

[12] Blanc, J.P.C. (1987). A note on waiting times in systems with queues in parallel, *Journal of Applied Probability* 24, 540–546.

[13] Blanc, J.P.C. (1993). Performance analysis and optimization with the power-series algorithm. In: L. Donatiello, R. Nelson (eds.), *Performance Evaluation of Computer and Communication Systems* (Springer, Berlin), pp. 53–80.

[14] Blanc, J.P.C. and Van der Mei, R.D. (1995). Optimization of polling systems with Bernoulli schedules, *Performance Evaluation* 22, 139–158.

[15] Boxma, O.J. and De Waal, P.R. (1994). Multiserver queues with impatience customers. In: J. Labetoulle and J.W. Roberts (eds.), *Proceedings ITC-14* (North-Holland Publ. Cy., Amsterdam), pp. 743-756.

[16] Chaudhry, M.L. and Templeton, J.G.C. (1983). *A First Course in Bulk Queues* (Wiley, New York).

[17] Du, Ding-Zhu and Hwang, F.K. (2000) *Combinatorial Group Testing and its Applications* (2nd ed.; World Scientific, Singapore).

[18] Finch, P.D. (1960). Deterministic customer impatience in the queueing system $GI/M/1$. *Biometrika* 47, 45–52.

[19] Gastwirth, J. and Johnson, W. (1994). Screening with cost-effective quality control: potential applications to HIV and drug testing. *J. Amer. Statist. Assoc.* 89, 972–981.

[20] Hooghiemstra, G. and Koole, G. (2000). On the convergence of the power-series algorithm, *Performance Evaluation* 42, 21–39.

[21] Hourfar, M.K., Jork, C., Schottstedt, V., Weber-Schehl, M., Brixner, V., Busch, M.P., Geusendam, G., Gubbe, K., Mahnhardt, C., Mayr-Wohlfart, U., Pichl, L., Roth, W.K., Schmidt, M., Seifried, E. and Wright, D.J. (2008). Experience of German Red Cross blood donor services with nucleic acid testing: results of screening more than 30 million blood donations for human immunodeficiency virus-1, hepatitis C virus, and hepatitis B virus. *Transfusion* 48(8), 1558-1566.

[22] Iravani, F. and Balcioğlu, B. (2008). Approximations for the $M/GI/N + GI$ type call center. *Queueing Systems* 58, 137-153.

[23] Kleinrock, L. (1975). *Queueing Systems, Vol. I: Theory* (Wiley, New York).

[24] Li, C.H. (1962). A sequential method for screening experimental variables. *J. Amer. Statist. Assoc.* 57, 455–477.

[25] Litvak, E., Tu, X.M. and Pagano, M. (1994). Screening for the presence of a disease by pooling sera samples. *J. American Statistical Association* 89, 424–434.

[26] Nair, S.S. and Neuts, M.F. (1972). Distribution of occupation time and virtual waiting time of a general class of bulk queues, *Sankhya*, Series A 34, 17–22.

[27] Neuts, M.F. (1967). A general class of bulk queues with Poisson input. *Annals of Mathematical Statistics* 38, 759–770.

[28] Neuts, M.F. (1981). *Matrix Geometric Solutions in Stochastic Models: an Algorithmic Approach.* (Johns Hopkins Univ. Press, Baltimore).

[29] Schottstedt, V., Tuma, W., Bünger, G., and Lefevre, H. (1998). PCR for HVC and HIV-1 experiences and firs results from routine screening. *Biologicals* 26, 101-104.

[30] Stanford, R.E. (1979). Reneging phenomenon in single channel queues. *Math. Oper. Res.* 4, 162-178.

[31] Stanford, R.E. (1990). On queues with impatience. *Adv. Appl. Probab.* 22, 768-769.

[32] Stramer, S.L., Glynn, S.A., Kleinan, S.H., Strong, D.M., Caglioti, S., Wright, D.J., Dodd, R.Y. and Busch, M.P. (2004). Detection of HIV-1 and HCV infections among antibody-negative blood donors by nucleic acid-amplification. *The New England Journal of Medicine* 351(8), 760-768.

[33] Takács, L. (1962). *Introduction to the Theory of Queues* (Oxford University Press, Oxford).

[34] Tu, X.M., Litvak, E., Pagano, M. (1995). On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: application to HIV screening. *Biometrika* 82, 287–297.

[35] Tijms, H.C. (1986). *Stochastic Modelling and Analysis: A Computational Approach* (Wiley, Chichester).

[36] Uhl, G., Liu, Q.,Walther, D., Hess, J. and Naiman, D. (2001). Polysubstance abuse-vulnerability genes: genome scans for association using 1,004 subjects and 1,494 single-nucleotide polymorphisms. *Amer. J. Human Genet.* 69, 1290–1300.

[37] Wein, L.M. and Zenios, S.A. (1996). Pooled testing for HIV screening: capturing the dilution effect. *Oper. Res.* 44, 543–569.

[38] Wolf, J. (1985). Born again group testing: multiaccess communications. *IEEE Trans. Inform. Theory* 31, 185–191.

[39] Xie, M., Tatsuoka, K., Sacks, J. and Young, S. (2001). Group testing with blockers and synergism. *J. Amer. Statist. Assoc.* 96, 92–102.

[40] Zhu, L., Hughes-Oliver, J. and Young, S. (2001). Statistical decoding of potent pools based on chemical structure. *Biometrics* 57, 922–930.