1992

P. Wartenhorst

N Parallel queueing systems with server breakdown and repair

Department of Operations Research, Statistics, and System Theory Report BS-R9236 December

CWI is the research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a non-profit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands organization for scientific research (NWO).

N Parallel Queueing Systems with Server Breakdown and Repair

Pieter Wartenhorst
CWI
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

Abstract

A model is presented that can be used to study the influence of machine breakdown and limited repair capacity on the performance of a system that has to provide service continuously. We consider a system consisting of N single server queueing stations, each serving its own stream of customers. The servers of the stations are subject to breakdown. Broken servers are repaired by a joint repair facility consisting of K parallel repairmen. Whenever $K < N$, this repair facility is causing interference between the N stations.

We study the behaviour of the queue length at a particular station (say station 1) in the long run. By modelling station 1 as an *M/M/1 queue in a Markovian environment* we obtain an exact matrix-geometric solution for the marginal queue length distribution and for the distributions of the queue length at the beginning of an up- (down-) period of the server. Since the matrix-geometric solution becomes rather time- and memory-consuming when N is large, we also develop approximating closed form expressions. By assuming the lengths of subsequent down-periods to be independent, station 1 is approximated by an *M/M/1 queue with independent interruptions.* Stochastic decompositions are employed to obtain approximations for the queue length distributions.

With this model several design issues can be investigated such as the number of repairmen that is needed to maintain a certain pool of machines, or the number of machines that can be assigned to a certain crew of repairmen. Several numerical examples illustrate how the queue at station 1 is affected by breakdowns of server 1 and by the other stations due to congestion at the repair facility.

1980 Mathematics Subject Classification (1991 Revision): 60K10, 60K25, 90B22, 90B25.
Keywords & Phrases: reliability, queueing, maintenance, repair, vacation, matrix-geometric solution.

1 Introduction

The influence of machine breakdown on the performance of continuous production- (or service-) systems is often not well understood or not taken into account at all. However, the performance of a system may be heavily affected by machine breakdown, and limited repair capacity. The overall production capacity will be affected, but also severe build-up of jobs may occur at a

Report BS-R9236
ISSN 0924-0659
CWI
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

broken machine. Similar phenomena occur in the area of computer and communication systems, where failure and repair of processors have a major impact on the flow of jobs that have to be handled by those processors.

In this paper we present a queueing model that can be used to study such influences. The model consists of N single server queueing stations, each serving its own stream of customers (jobs). The servers (machines or processors) of the N stations are subject to breakdown. Broken servers are repaired by a joint repair facility consisting of K parallel repairmen. We study the influence of server breakdown and limited repair capacity ($K < N$) on the number of customers present at a particular station (say station 1) in the long run.

We consider both the marginal queue length distribution and the conditional distributions of the queue length at the beginning of an up- (down-) period of server 1. Severe queue build-up during down-periods is best expressed by the conditional queue length distributions. An exact solution of the queue length distributions is obtained by modelling station 1 as an *M/M/1 queue in a Markovian environment* which has a matrix-geometric solution (see Neuts [1981] or Nelson [1991] for a general explanation of the matrix-geometric method). We also present useful expressions from which the moments of the queue length distributions can be calculated. Since the matrix-geometric solution becomes rather time- and memory-consuming when N is large, we develop an approximating model for which closed form expressions are derived. By assuming the lengths of subsequent down-periods to be independent, station 1 is approximated by an *M/M/1 queue with independent interruptions*. Stochastic decompositions are employed to obtain approximations for the queue length distributions. The results of this paper have been implemented in an interactive computer program that is used to analyze the model and its extensions. The accuracy of the approximation has been tested during extensive numerical experiments. A general conclusion is that the approximation performs extremely well as long as the repair facility is not too heavily loaded. If the repairmen are working most of their time, then the quality of the approximation depends on several model parameters. The approximation is especially sensitive with respect to the speed at which customers pass station 1. For a given traffic intensity the accuracy decreases when customers move faster (i.e. both shorter interarrival times and shorter service times). From a comparison with an alternative approximation we conclude that it is very important to take down-periods of the server explicitly into account, even when approximating the model. Several numerical examples illustrate how the queue length at station 1 is affected by breakdowns of server 1 and by other stations due to congestion at the repair facility. An important conclusion from those examples is that the queue length of a queueing station subject to server breakdown is heavily influenced by the speed at which customers pass the station, whereas the queue length of a queueing station without server breakdown is not.

This research is located on the border of reliability theory and queueing theory. A popular model in reliability theory is the machine repair model, where a number of machines (the N servers in our case) is maintained by a number of repairmen (K in our case). With respect to the machines one is often only interested in whether they are operational or not. Jobs (our customers) that have to be processed by those machines are not considered. On the contrary, in queueing theory one is usually interested in the performance of systems where either the servers are never failing or the down-time of a broken server is not affected by possible breakdowns of other servers. Our model is a two level combination of reliability theory and queueing theory. At level 1 the N servers and the K repairmen constitute an ordinary machine repair model

(actually broken servers can be seen as customers of the repair facility). At level 2 customers are served by servers that are subject to breakdown. Whenever $K < N$ interaction between broken servers at the repair facility (level 1) influences the behaviour of the queue lengths at the various stations (level 2).

In queueing theory literature two more or less disjoint streams of papers exist that both consider queueing systems subject to breakdown of the servers. *Purely analytically solved* models are known as vacation models, where every time period during which the server is not working is called a vacation. For an extensive survey of such models we refer to Doshi [1990]. Most of the earlier results in this area are presented in terms of Laplace transforms and moment generating functions. Recently elegant and insightful results have been obtained by employing stochastic decompositions of the measures of interest (cf. Fuhrmann & Cooper [1985]). These explicit expressions are very efficient to solve a model numerically. A main drawback of those purely analytical solutions is their restricted modelling flexibility. Several minor model extensions are not straightforward to analyze.

The *matrix-geometric method* does not suffer from this drawback. It offers a unified approach to solve a wide variety of queueing problems. Because of its great modelling flexibility this method lends itself for implementation in an interactive computer program that can be used to study several model variations (Haverkort et al. [1992] have made a first attempt to build a generally applicable software tool to analyze queueing models that have a matrix-geometric solution; however, until now their tool is not able to analyze queueing systems subject to server breakdown). Unfortunately a matrix-geometric solution provides less direct insight than a purely analytical one. Furthermore the desired expressions are given in terms of matrices that have to be solved recursively, which may become a time- and memory-consuming task for large problem instances.

Models that are related to ours and that have been solved by the matrix-geometric method, are presented by Neuts & Lucantoni [1979], Vinod [1985] and Colard & Latouche [1980]. Neuts & Lucantoni [1979] consider a single queue of customers, each served by one of N parallel servers. These servers are subject to breakdown and are repaired by one of K parallel repairmen. Vinod [1985] considers the same model with ample repair ($K=N$). For $N=1$ he imposes some restrictions on the server down-periods (either independent of the queue length or only occurring when the server is active). Colard & Latouche [1980] consider a computer system serving both batch (low priority) and interactive (high priority) jobs. The interactive jobs are generated by N terminals. Note that there exists a strong connection between models of queueing systems subject to server breakdown, and models of queueing systems serving two types of customers (low and high priority). In the latter model the high priority customers have the same influence on the queue of low priority customers, as breakdowns of the server have on the queue length in the first model. In all these papers only a single queue is studied that is interrupted either by server breakdowns or by arrivals of high priority customers. In this paper, however, we are interested in the mutual influence of different queueing systems via the limited repair capacity.

This paper is organized as follows. In Section 2 we give a detailed model description and obtain some preliminary results such as the stationary distribution of the number of broken servers in the long run. In Section 3.1 we obtain an exact solution of the queue length distributions by modelling station 1 as an *M/M/1 queue in a Markovian environment* which has a matrix-geometric solution. We also present useful expressions from which the moments of

the queue length distributions can be calculated. In Section 3.2 the approximating model is presented for which closed form expressions are derived. By assuming the lengths of subsequent down-periods to be independent, station 1 is approximated by an *M/M/1 queue with independent interruptions*. Stochastic decompositions are employed to obtain approximations for the queue length distributions. The accuracy of the approximation is tested in Section 4.1. Several numerical examples are presented in Section 4.2. Appendix A presents the distribution of the length of an arbitrary down-period and some related measures. A self-contained treatment of the matrix-geometric method applied to an M/M/1 queue in a Markovian environment is given in Appendix B.

2 Model and preliminary results

We consider a multiple queueing system, consisting of N parallel single server queueing stations, each with infinite buffer capacity. At station 1 customers arrive according to a Poisson process with rate λ. Customers are served according to a FCFS service discipline. The amounts of service required by customers are independent exponentially distributed stochastic variables with mean μ^{-1}. At service completion customers leave the system. Since we are only interested in the queue length behaviour of station 1, we do not specify the arrival and service processes at the remaining $N-1$ stations (although it is not unreasonable to assume them to be identical to those at station 1; cf. Remark 2.1).

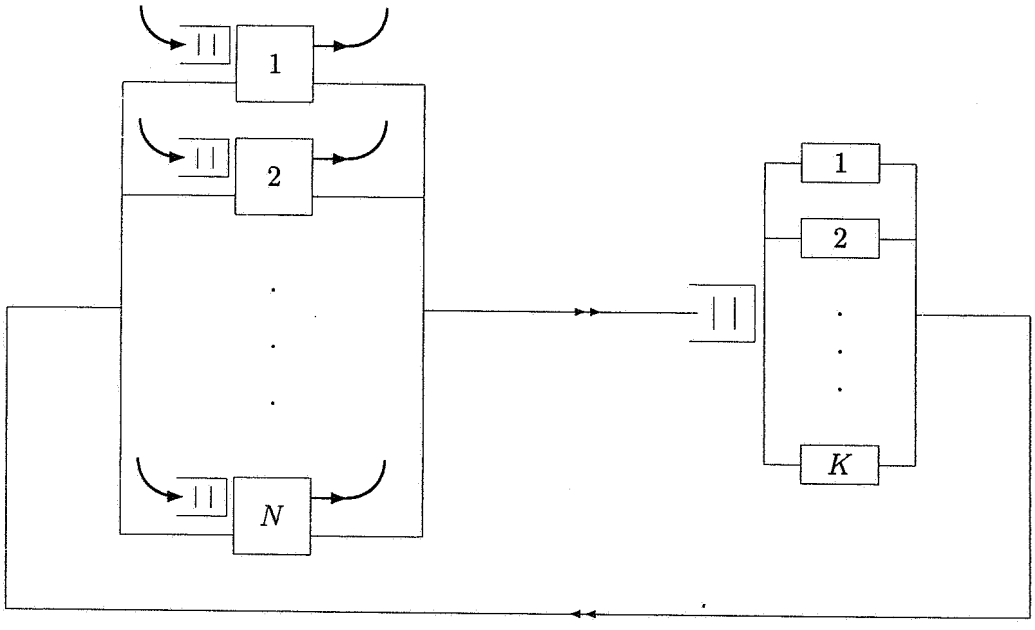

Figure 1: N parallel queueing stations and K repairmen.

A special feature of this system is that servers are subject to breakdown. The lifetimes of

the servers are i.i.d. according to an exponential distribution with mean σ^{-1}. The repairshop consists of K identical parallel repairmen, each repairing one broken server at a time. There is a single queue of broken servers, which are repaired by one of the repairmen in FCFS order. The repairtimes of broken servers are i.i.d. according to an exponential distribution with mean ν^{-1}. After a repair on a server is completed, this server immediately returns to its working position to serve (possibly waiting) customers. Figure 1 depicts the structure of the model. The thick lines denote the streams of customers visiting the N stations. The thin lines denote the routing of broken servers to the repair facility and back again to their working positions.

It is important to note that whenever $K < N$, the repairshop is causing interdependence between the separate queueing stations.

Remark 2.1 This model can extended into several directions. The number of repairmen K_i and the repair rate ν_i may depend on the number of broken servers i ($i = 0, ..., N$). In this way threshold-policies can be incorporated in the model (with $K_i \leq K_{i+1}$, $\nu_i \leq \nu_{i+1}$, $i = 0, ..., N-1$). Such policies are useful since most of the time only a few repairmen are needed to maintain the system properly. Also repair of broken servers need not always be carried out at full (expensive) speed. Furthermore the N servers need not be identical. Their failure rates σ_j and repair rates ν_j (or $\nu_{i,j}$; $j = 1, ..., N$) may differ.

For such more general models results can be obtained, which are similar to the ones presented in this paper. These extensions, however, increase the complexity of the expressions found, without significantly increasing the insight in the problem. That is why we do not make these assumptions. □

Several performance measures can be obtained quite easily. Let

$$\begin{aligned} N_b &:= \text{number of broken servers in the long run,} \\ p_i &:= P(N_b = i), \\ \gamma &:= \sigma/\nu. \end{aligned} \tag{2.1}$$

The stationary probability distribution of the number of broken servers is given by the following lemma.

Lemma 2.1

$$p_i = \begin{cases} \binom{N}{i} \gamma^i \, p_0 \,, & 0 \leq i \leq K, \\ \binom{N}{i} \dfrac{i!\,\gamma^i}{K!\,K^{i-K}} \, p_0 \,, & K \leq i \leq N, \end{cases} \tag{2.2}$$

$$p_0 = \left\{ \sum_{i=0}^{K} \binom{N}{i} \gamma^i + \sum_{i=K+1}^{N} \binom{N}{i} \frac{i!\,\gamma^i}{K!\,K^{i-K}} \right\}^{-1}. \tag{2.3}$$

Proof: The servers together with the repair facility form a special case of the general machine repair model. The presented result is known from literature (cf. Kleinrock [1975]). We present a short proof that indicates how similar results can be obtained for more general models (cf. Remark 2.1).

Consider the number of failed servers i $(0 \leq i \leq N)$ as the state of a birth-death process with birth rates α_i and death rates β_i:

$$\alpha_i := (N-i)\,\sigma, \qquad 0 \leq i \leq N, \tag{2.4}$$

$$\beta_i := \min\{i, K\}\,\nu, \qquad 0 \leq i \leq N. \tag{2.5}$$

The steady state distribution of a general birth-death process is known (cf. Kleinrock [1975]):

$$p_i = \frac{\alpha_0\,\alpha_1\,\alpha_2\,\ldots\,\alpha_{i-1}}{\beta_1\,\beta_2\,\beta_3\,\ldots\,\beta_i}\,p_0, \quad i \geq 0. \tag{2.6}$$

Substitution of (2.4) and (2.5) into (2.6) yields (2.2) and (2.3). □

Remark 2.2 One should not compute p_i explicitly from (2.2) and (2.3). Instead p_i is implemented using the following recurrence relationship.

$$p_i = \frac{\alpha_{i-1}}{\beta_i}\,p_{i-1}, \quad i \geq 1,$$

with normalization equation:

$$\sum_{i=0}^{N} p_i = 1.$$

□

Insight in the functioning of the repair facility can be obtained, for instance by computing the first two moments of the number of servers waiting for, or under, repair (straightforward from Lemma 2.1).

A first impression of the behaviour of station 1 is obtained by considering the server's lifetime distribution (which is given), and the distribution of an arbitrary down-time (i.e. sojourn time of a broken server at the repair facility). Let

$$D := \text{length of an arbitrary down time of server 1.} \tag{2.7}$$

The distribution of D and its moments $E[D^i]$, $i \geq 1$, are computed in Appendix A. The state of the servers can be described by a stochastic process that regenerates itself every time when server 1 breaks down and the remaining $N-1$ servers are up (cf. (3.5)). Employing Theorem 3.4 in Tijms [1986] we conclude that the long run fraction of time that server 1 is up (i.e. the availability) is given by

$$P_{up} := \frac{E[\text{lifetime}]}{E[\text{lifetime}] + E[D]} = \frac{1}{1+\sigma E[D]}. \tag{2.8}$$

Similarly the long run fraction of time that the server is down is given by

$$P_{down} := 1 - P_{up} = \frac{\sigma E[D]}{1+\sigma E[D]}.$$

The traffic intensity ρ (which is equal to the long run fraction of time that server 1 is busy) is given by

$$\rho = \frac{\lambda}{\mu}.$$

The effective traffic intensity ρ_{eff} (which is equal to the conditional long run fraction of time that server 1 is busy, given that this server is up) is given by

$$\rho_{eff} = \frac{\rho}{P_{up}},$$

which is assumed to be smaller than 1 (cf. (3.11)).

3 Queue length distribution

In this section we present both an exact solution, and a much simpler approximate expression for the queue length distribution at station 1 in the long run. Let

$$L := \text{number of customers in queue 1 (either in service or waiting to be served).} \tag{3.1}$$

The distribution of L is called the *marginal queue length distribution.* This is not the most relevant distribution to be studied, because severe queue build-up during down-periods may be obscured. Therefore, we also study *conditional queue length distributions.* The most useful ones are the distributions of

$$L_{|beginDown} := \text{number of customers in queue 1 at the beginning of an arbitrary down-period,} \tag{3.2}$$

$$L_{|beginUp} := \text{number of customers in queue 1 at the beginning of an arbitrary up-period.} \tag{3.3}$$

The distribution of $L_{|beginUp}$ is known as soon as the distribution of $L_{|beginDown}$ is known, since

$$L_{|beginUp} \stackrel{d}{=} L_{|beginDown} + \text{number of arrivals during an arbitrary down-period,} \tag{3.4}$$

where the latter two quantities are independent, and ' $\stackrel{d}{=}$ ' denotes 'is equal in distribution'.

In Section 3.1 we present exact solutions for these distributions by modelling station 1 as an *M/M/1 queue in a Markovian environment.* This model is solved using the matrix-geometric method, successfully elaborated by Neuts [1981] (cf. Appendix B).

If a large number of stations N is involved, then the matrix-geometric method tends to be rather time- and memory-consuming. Therefore, in Section 3.2, we develop an approximation by assuming successive down-times of server 1 to be independent. Thus, in the approximation station 1 is modeled as an *M/M/1 queue with independent interruptions.* Using stochastic decompositions of the various queue length distributions, we obtain approximations for the distributions of L, $L_{|beginDown}$ and $L_{|beginUp}$.

3.1 Exact solution: M/M/1 queue in a Markovian environment

In general a Markovian environment is modeled by an irreducible continuous time Markov chain $\{X(t);\ t \geq 0\}$ with a finite state space $\mathcal{S} = \{1, ..., |\mathcal{S}|\}$ and infinitesimal generator $Q = \{q_{i \to j};\ i, j \in \mathcal{S}\}$. When the environmental process is in state j $(X(t) = j;\ j \in \mathcal{S})$, customers arrive to the single server queue according to a Poisson process with rate λ_j and service is carried out at rate μ_j. By $\underline{\lambda}$ we denote the vector $(\lambda_1, ..., \lambda_{|\mathcal{S}|})$. The vector $\underline{\mu}$ is defined similarly. For an extensive treatment of queues in a random environment we refer to Neuts [1981], [1978a], and [1978b].

To study the behaviour of queue 1, the environment is sufficiently well described by the number of failed servers and by the position of server 1 in the system (due to the memoryless property of the exponential distribution). Thus the environment at time t is given by the following state vector.

$$X(t) := (I, N_1, N_2), \qquad t \geq 0, \tag{3.5}$$

where

$$I := \begin{cases} 0 & \text{if server 1 is up,} \\ 1 & \text{if server 1 is down,} \end{cases} \tag{3.6}$$

$$N_1 := \text{number of broken servers succeeding server 1 in queue waiting for (or under) repair,} \tag{3.7}$$

$$N_2 := \text{number of broken servers preceding server 1 in queue waiting for (or under) repair.} \tag{3.8}$$

Server 1 is down when $N_1 > 0$, and server 1 is either up *or* down when $N_2 > 0$. $\{X(t);\ t \geq 0\}$ contains all information that is needed to analyze the influence of possible breakdowns of any of the N servers upon the behaviour of the queue at station 1. The environmental state space is given by:

$$\mathcal{S} := \{(0, 0, n_2) \mid n_2 = 0, ..., N-1\} \cup \{(1, n_1, n_2) \mid n_1 = 0, ..., N-1;\ n_2 = 0, ..., N-n_1-1\}.$$

Note that

$$N_b := I + N_1 + N_2.$$

Furthermore, the repair rate of the entire repairshop is given by:

$$\min\{N_b, K\}\ \nu.$$

The non-zero entries of the matrix with transition intensities $Q = \{\ q_{i \to j}\ ;\ i, j \in \mathcal{S}\ \}$ are given by:

$$q_{(0,0,n_2) \to (1,0,n_2)} = \sigma \qquad n_2 = 0, ..., N-1,$$

$$
\begin{aligned}
q_{(0,0,n_2)\to(0,0,n_2+1)} &= (N-n_2-1)\,\sigma && n_2 = 0,...,N-2,\\
q_{(0,0,n_2)\to(0,0,n_2-1)} &= \min\{n_2,K\}\,\nu && n_2 = 1,...,N-1,\\
q_{(1,n_1,n_2)\to(1,n_1+1,n_2)} &= (N-n_1-n_2-1)\,\sigma && n_2 = 0,...,N-2;\ n_1 = 0,...,N-n_2-2,\\
q_{(1,n_1,n_2)\to(0,0,n_1+n_2)} &= \nu && n_2 = 0,...,\min\{N-1,K-1\};\\
& && n_1 = 0,...,N-n_2-1,\\
q_{(1,n_1,n_2)\to(1,n_1,n_2-1)} &= \min\{n_2,K\}\,\nu && n_2 = 1,...,N-1;\ n_1 = 0,...,N-n_2-1,\\
q_{(1,n_1,n_2)\to(1,n_1-1,n_2)} &= \min\{n_1,K-n_2-1\}\,\nu && n_2 = 0,...,\min\{N-2,K-2\};\\
& && n_1 = 1,...,N-n_2-1,\\
q_{i\to i} &= -\textstyle\sum_{j\neq i} q_{i\to j} && i \in \mathcal{S}.
\end{aligned}
$$

Now our particular station 1 is an M/M/1 queue in Markovian environment $\{X(t);\ t \geq 0\}$, with arrival rate λ_j and service rate μ_j when the environment $X(t) = j,\ j \in \mathcal{S}$, where

$$
\begin{aligned}
\lambda_j &= \lambda \qquad \forall j \in \mathcal{S},\\
\mu_j &= \begin{cases} \mu & \text{if } X(t) = (0,0,n_2), \quad n_2 = 0,...,N-1,\\ 0 & \text{if } X(t) = (1,n_1,n_2), \quad n_1 = 0,...,N-1,\ n_2 = 1,...,N-n_1-1. \end{cases}
\end{aligned}
$$

To obtain the desired queue length distributions, we need the stationary probability vector of Q.

$$\pi := \{\pi_x;\ x \in \mathcal{S}\},$$

where

$$\pi_x := \lim_{t\to\infty} P(X(t) = x), \qquad x \in \mathcal{S}.$$

Since the state space of the environmental Markov chain is finite, and every state can be reached from every other state, $\{X(t);\ t \geq 0\}$ is irreducible and positive recurrent. Hence these limiting probabilities exist (cf. Ross [1983], p. 152), and are given by the following theorem.

Theorem 3.1

$$\pi_{0,0,i} = \frac{N-i}{N}\,p_i, \qquad i = 0,...,N-1, \tag{3.9}$$

$$\pi_{1,i,j} = \frac{1}{N}\,p_{i+j+1}, \qquad i = 0,...,N-1,\ j = 0,...,N-i-1. \tag{3.10}$$

Proof: (3.9) and (3.10) are obtained by conditioning on the number of failed servers. Since all servers are equal, they are equally likely to be in a failed state at a certain epoch in time. Thus, given that i servers are at the repair facility, the probability that server 1 is not among them is $(N-i)/N$. Similarly, given that $i+j+1$ servers form a queue at the repair facility, each server has the same probability $1/N$ of occupying the $(i+1)$st position in this queue. □

The stationary joint probability distribution of the number of customers in queue and the environmental state is denoted by vectors x_k, $k \geq 0$, such that

$$x_{k,j} := \lim_{t\to\infty} P(k \text{ customers in queue at time } t \text{ and } X(t) = j), \qquad k \geq 0,\ j \in \mathcal{S},$$

where the number of customers in queue includes the one in service. Queue 1 is stable, and thus x_k exists, if and only if (cf. Neuts [1981])

$$\rho_{eff} < 1 \quad \Leftrightarrow \quad \pi\underline{\lambda} < \pi\underline{\mu}. \tag{3.11}$$

Application of the matrix-geometric method leads to the following theorem which is proved by Neuts [1981] (for an explanation of this method, and for a definition of the matrix R that is employed in Theorem 3.2, we refer to Appendix B).

Theorem 3.2

$$x_k = \pi(I-R)R^k, \qquad k \geq 0.$$

□

Remark 3.1 x_k is calculated using the following recurrence relationship.

$$\begin{aligned} x_{k+1} &= x_k R, \qquad k \geq 0, \\ x_0 &= \pi(I-R). \end{aligned}$$

□

Now, using Theorem 3.2, we are able to derive the *marginal queue length distribution* which is given by:

$$x_k e, \quad k \geq 0,$$

where the vector e denotes $(e_1, ..., e_{|\mathcal{S}|})$ with $e_i = 1$, $\forall i$.

The *conditional queue length distribution*, given that the environmental state is in $\mathcal{A}$, is obtained from

$$\sum_{j\in\mathcal{A}} x_{k,j} / \sum_{j\in\mathcal{A}} \pi_j, \qquad k \geq 0, \text{ for any subset } \mathcal{A} \text{ of } \mathcal{S}.$$

By properly choosing $\mathcal{A}$ one can obtain various interesting conditional queue length distributions. For instance the conditional distribution of the number of customers in queue 1 given that the server is *up*, $L_{|Up}$, is obtained by choosing

$$\mathcal{A} = \{(0, 0, n_2) \mid n_2 = 0, ..., N-1\}.$$

The conditional distribution of the number of customers in queue given that the server is *down and under repair* $L_{|underRepair}$ is obtained with

$$\mathcal{A} = \{(1, n_1, n_2) \mid n_2 = 0, ..., \min\{N-1, K-1\};\ n_1 = N - n_2 - 1\}. \tag{3.12}$$

Now, knowing the distributions of $L_{|Up}$ and $L_{|underRepair}$, the desired distributions of $L_{|beginDown}$ (3.2) and $L_{|beginUp}$ (3.3) are obtained by employing the following two lemma's.

Lemma 3.1

$$L_{|beginDown} \stackrel{d}{=} L_{|Up}. \tag{3.13}$$

Proof: Let

$N(t)$:= number of breakdowns of server 1 that occur in $(0, t]$.

$N(t)$ is a doubly stochastic Poisson process: if server 1 is up then breakdowns occur according to a Poisson process with rate σ; if server 1 is down then breakdowns occur with rate 0 (no breakdowns occur). For this newly defined 'arrival' process a generalization of the PASTA property (Poisson Arrivals See Time Averages) is valid (cf. Van Doorn & Regterschot [1988]). The Conditional PASTA property implies that the queue length distribution in continuous time given that server 1 is up equals the queue length distribution at epochs where an 'arrival' takes place (i.e. a breakdown of server 1 occurs) given that server 1 is up. □

Lemma 3.2

$$L_{|beginUp} \stackrel{d}{=} L_{|underRepair}. \tag{3.14}$$

Proof: Similar to the proof of Lemma 3.1. In this case we have to consider

$N'(t)$:= number of repair completions of server 1 that occur in $(0, t]$.

A repair completion of server 1 occurs according to a Poisson process with rate ν if server 1 is under repair, and with rate 0 otherwise. Now again, (3.14) is implied by the Conditional PASTA property. □

An intuitive explanation of (3.13) is obtained as follows. Suppose the system has been running for a long time. Fix some time t (t large) at which server 1 is up. From Ross [1983] Proposition 3.4.5 we know that the age of the server at time t is $exp(\sigma)$ distributed (after ignoring down-periods, the remaining process is a renewal process with $exp(\sigma)$ distributed interrenewal times). So, for both $L_{|beginDown}$ and $L_{|Up}$ we consider the queue length at a point in time where the server has been up for an $exp(\sigma)$ distributed amount of time since the end of the preceding down-period. The state of the system at the beginning of this exponentially distributed amount of time is the same for both situations. Hence, (3.13) is valid. In a similar way the validity of (3.14) can be argued, since for both $L_{|beginUp}$ and $L_{|underRepair}$ we consider the queue length

at a point in time that is preceded by an $exp(\nu)$ distributed amount of time during which the server was under repair.

To obtain the moments of the queue length distributions we define:

$$u^{(i)} := \sum_{k=0}^{\infty} k^i x_k.$$

In general $u^{(i)}$ $(i \geq 1)$ can be obtained from Lemma 2 in Colard & Latouche [1980]. For $i = 1, 2$ we have:

$$\begin{aligned} u^{(1)} &= \pi R(I-R)^{-1}, \\ u^{(2)} &= \pi R(I-R)^{-2}(I+R). \end{aligned}$$

All conditional moments are obtained from

$$E[L^i \mid X(\infty) \in \mathcal{A}] = \sum_{j \in \mathcal{A}} u_j^{(i)} / \sum_{j \in \mathcal{A}} \pi_j, \qquad i \geq 1.$$

Remark 3.2 If the distributions of D and $L_{|beginDown}$ are known, then the distribution of $L_{|beginUp}$ follows from (3.4). So, we do not need (3.12) in that situation, because the distribution of $L_{|underRepair}$ follows from Lemma 3.2. □

3.2 Approximation: M/M/1 queue with independent interruptions

The matrix-geometric solution tends to be rather expensive with respect to CPU and memory-requirements, especially because of the matrix R which has to be numerically solved by successive substitution (cf. Appendix B). In this section we develop an approximation for our model. For this approximation explicit simple expressions are derived for the distributions of the approximating quantities $\hat{L}$, $\hat{L}_{|beginDown}$ and $\hat{L}_{|beginUp}$ (which correspond to L, $L_{|beginDown}$ and $L_{|beginUp}$ in the exact model). The accuracy of the approximation is investigated in Section 4.1.

Consider station 1 in isolation. This is an M/M/1 queue subject to server breakdown. The server alternates between 'up'- and 'down'-state. Successive up-times are i.i.d. according to the exponential lifetime distribution. Successive down-times are *not* i.i.d., because of interaction with other failed servers at the repair facility. In the approximation we *assume* that successive down-times are i.i.d. as well, and are equally distributed as D – the length of an arbitrary down-time in the exact model (cf. Appendix A). Since the epochs at which the server breaks down do not depend on the number of customers in queue, station 1 is thus approximated by an *M/M/1 queue with independent interruptions.* This approximating model is a so-called *vacation model.* For an extensive survey of single server queueing models with vacations we refer to Doshi [1990].

For the approximating M/M/1 queue with independent interruptions we define

$\hat{L}$:= number of customers in queue,

$\hat{L}_{|beginDown}$:= number of customers in queue at the beginning of an arbitrary down-period,

$\hat{L}_{|beginUp}$:= number of customers in queue at the beginning of an arbitrary up-period,

$\hat{L}_{|Up}$:= number of customers in queue when the server is up,

$\hat{L}_{|Down}$:= number of customers in queue when the server is down.

The M/M/1 queue with independent interruptions has been studied by Gaver [1962], who derives the probability generating function of the marginal queue length distribution, from which the moments can be obtained by differentiation. However, the expression for the variance in Gaver [1962] is incorrect. Gaver considers a compound Poisson(λ) arrival process, i.e. batches of customers arrive according to a Poisson(λ) process. When all batch sizes are equal to one (i.e. ordinary Poisson(λ) arrivals) a term $\lambda E[C]$ should be added to the right hand side of Formula (8.9) in Gaver [1962].

Instead of finding the correct expression by differentiating Gaver's generating function (which is a tedious task), we present a new derivation which is based on a stochastic decomposition of $\hat{L}$ (Theorem 3.3). In Theorem 3.4 we show that the distribution of $\hat{L}_{|beginDown}$ follows from results for the ordinary M/G/1 queue (no server breakdown). As in the previous section, the distribution of $\hat{L}_{|beginUp}$ is obtained from (3.4).

In Theorem 3.3 and 3.4 we make use of the notion of *completion time*. The completion time of a customer at queue 1 is the duration of the period that elapses between the start and completion of the service of that customer. This period is simply the customer's service time if there are no interruptions (i.e. server breakdowns). Otherwise it is the sum of the customer's service time, and the durations of the interruptions occurring during that time. By $M/G/1_{|C}$ we denote an ordinary M/G/1 queue with Poisson(λ) arrivals and service times that are equal in distribution to the completion times in the original model. We define

$$L_{M/G/1_{|C}} := \text{number of customers present in the } M/G/1_{|C} \text{ queue.} \tag{3.15}$$

Furthermore, let

L_{PD} := number of Poisson arrivals during a time interval that is distributed as the equilibrium backward recurrence time of a down-period. (3.16)

The distribution and the moments of both $L_{M/G/1_{|C}}$ and L_{PD} are discussed in Appendix A. Now we are able to prove the following theorem.

Theorem 3.3

$$\hat{L} \stackrel{d}{=} L_{M/G/1_{|C}} + X\, L_{PD}, \tag{3.17}$$

where X is a Bernoulli stochastic variable, independent of L_{PD} with

$$P(X = 1) = P_{down}. \tag{3.18}$$

X and L_{PD} are independent of $L_{M/G/1_{|C}}$.

Proof: After replacing service times by completion times we have an $M/G/1_{|C}$ vacation system with *exhaustive service* (i.e. interruptions may only occur when the queue is empty). A time period during which the server is inactive (either idle or down) is called a *vacation*. The queue length distribution of our initial approximation is identical to the queue length distribution of this vacation system. The following M/G/1 Decomposition Property now holds (cf. Fuhrmann & Cooper [1985]).

$$\hat{L} \stackrel{d}{=} L_{M/G/1_{|C}} + L_{PV}, \tag{3.19}$$

where $L_{M/G/1_{|C}}$ and L_{PV} are independent, and

L_{PV} := number of customers in the queue at a random point in time when (given that) the server is on vacation.

Employing the PASTA property (Poisson Arrivals See Time Averages) L_{PV} is distributed as the number of customers in the system found by an arriving customer in a vacation (tagged customer). The distribution of L_{PV} is now obtained by conditioning on the state of the server as found by the tagged customer.

The system is empty at the start of every vacation (exhaustive service). The tagged customer finds the server either up or down. If the server is up, then the system is still empty and service is immediately started after arrival of the tagged customer. If the server is down, then the tagged customer finds those customers that have arrived earlier during the same down-period. Since down-times are i.i.d. and customers arrive according to a Poisson process, this number of customers is equal in distribution to L_{PD} (cf. (3.16)).

The probabilities that the tagged customer finds the server either up or down are obtained by ignoring all busy periods (a busy period starts with the service of the first customer that has arrived in the preceding vacation, and ends when the system becomes empty again). The remaining process is an alternating renewal process with $exp(\sigma)$ distributed up-times and down-times that are distributed as D (cf. Ross [1983], pp. 66-67). Now, again employing the PASTA property, the probability that the tagged customer finds the server up (down) is equal to P_{up} (P_{down}). Thus

$$L_{PV} \stackrel{d}{=} X\, L_{PD}, \tag{3.20}$$

where X is the indicator function defined in (3.18). □

Perhaps somewhat less intuitively clear, Equation (3.20) can also be derived directly (without employing the PASTA property and without appeal to the tagged customer), by conditioning to the state of the server at a random point in time within a vacation.

In the next theorem we show that the distribution of $\hat{L}_{|beginDown}$ is equal to the queue length distribution of an ordinary M/G/1 queue where service times have been replaced by completion times (i.e. the $M/G/1_{|C}$ queue).

Theorem 3.4

$$\hat{L}_{|beginDown} \stackrel{d}{=} L_{M/G/1_{|C}}. \tag{3.21}$$

Proof: As in Lemma 3.1 we can show that

$$\hat{L}_{|Up} \stackrel{d}{=} \hat{L}_{|beginDown}. \tag{3.22}$$

Furthermore it is easy to see that

$$\hat{L}_{|Down} \stackrel{d}{=} \hat{L}_{|beginDown} + L_{PD}. \tag{3.23}$$

The distribution of $\hat{L}$ can be obtained by conditioning to the state of the server at an arbitrary point in time. Hence

$$\hat{L} \stackrel{d}{=} (1-X)\, \hat{L}_{|Up} + X\, \hat{L}_{|Down}, \tag{3.24}$$

where X is defined by (3.18). From (3.22), (3.23) and (3.24) we conclude that

$$\hat{L} \stackrel{d}{=} \hat{L}_{|beginDown} + X\, L_{PD}. \tag{3.25}$$

Comparison of (3.17) and (3.25) yields (3.21). □

Remark 3.3 A direct (but more intricate) proof of Theorem 3.4 can be obtained by considering the number of customers found both by an arriving customer in the approximation given that the server is up, and by an arbitrary arriving customer in the $M/G/1_{|C}$ queue. In both situations an arriving customer finds the server idle with probability $1-\rho_{eff}$. With probability $\rho_{eff}P_{up}$ the number of customers found is distributed as the number of customers in the $M/G/1_{|C}$ queue given that the server is up and busy. With probability $\rho_{eff}P_{down}$ the number of customers found is distributed as the number of customers in the $M/G/1_{|C}$ queue given that the server is down. From this observation, together with (3.22), the result follows. □

Remark 3.4 The M/G/1 Decomposition Property (3.19), that was employed in Theorem 3.3, holds for any M/G/1 vacation system without interrupted services (under some regularity conditions, cf. Fuhrmann & Cooper [1985]). If services are interrupted due to independent server breakdown, then such a decomposition property holds for the distribution of the amount of work in system (i.e. the workload; cf. Boxma [1989]). However, it does not generally hold for the distribution of the number of customers in system. By employing similar arguments as Boxma [1989] we can show that the decomposition property does hold when service times are exponentially distributed. Hence a second decomposition of $\hat{L}$ can be obtained by considering any period during which the server is inactive as a vacation (including interrupted services). Similar to Theorem 3.3 we can prove

$$\hat{L} \stackrel{d}{=} L_{M/M/1} + Y\, (\hat{L}_{|beginDown} + L_{PD}),$$

where $L_{M/M/1}$ denotes the number of customers present in the ordinary M/M/1 queue, L_{PD} is defined in (3.16), $\hat{L}_{|beginDown}$ is obtained from Theorem 3.4, and

$$P(Y=1) = 1 - P(Y=0) = P_{down}/(1-\rho).$$

A decomposition for $L_{M/G/1_{|C}}$ can be obtained by first realizing that $L_{M/G/1_{|C}}$ is distributed as the queue length of an M/M/1 queue where server breakdown may only occur when the server

is busy. Then, by considering all periods during which the server is idle or down (interrupted service) as a vacation we can prove:

$$L_{M/G/1_{|C}} \stackrel{d}{=} L_{M/M/1} + Z\,(L_{|beginDown;M/G/1_{|C}} + L_{PD}),$$

where $L_{beginDown;M/G/1_{|C}}$ (which is still unknown) denotes the number of customers in queue at the beginning of an arbitrary down-period (i.e. interruption) in the $M/G/1_{|C}$ queue, and

$$P(Z=1) = 1 - P(Z=0) = \rho_{eff} P_{down}/(1-\rho).$$

□

All moments of $\hat{L}$, $\hat{L}_{|beginDown}$ and $\hat{L}_{|beginUp}$ can now be obtained from (3.17), (3.21) and (3.4) respectively. The moments of $\hat{L}$ follow easily from Theorem 3.3, since $L_{M/G/1_{|C}}$ and L_{PV} in (3.19) are independent. Thus

$$\begin{aligned} E[\hat{L}] &= E[L_{M/G/1_{|C}}] + P_{down}\,E[L_{PD}], \\ Var[\hat{L}] &= Var[L_{M/G/1_{|C}}] + P_{down}\,E[L_{PD}^2] - (P_{down}\,E[L_{PD}])^2. \end{aligned}$$

The moments of $\hat{L}_{|beginDown}$ are equal to the moments of $L_{M/G/1_{|C}}$, due to Theorem 3.4. The first two moments of $\hat{L}_{|beginUp}$ are obtained by employing Equation (3.4) and Theorem 3.4.

$$\begin{aligned} E[\hat{L}_{|beginUp}] &= E[L_{M/G/1_{|C}}] + \lambda\,E[D], \\ Var[\hat{L}_{|beginUp}] &= Var[L_{M/G/1_{|C}}] + \lambda^2\,E[D^2] + \lambda\,E[D](1 - \lambda\,E[D]). \end{aligned}$$

The moments of D, L_{PD} and $L_{M/G/1_{|C}}$ are presented in Appendix A.

4 Numerical experiments

Both the exact and the approximate solution of our model have been implemented in an interactive computer program. The accuracy of the approximation has been tested during extensive numerical experiments. Section 4.1 contains general conclusions from those experiments that are illustrated by some numerical results. In Section 4.2 the influence of server breakdown and limited repair capacity on the queue at station 1 is investigated.

Both computation time- and memory-requirements of the exact method strongly depend on N. For instance the order of the matrix R is equal to the dimension of the environmental state space:

$$|\mathcal{S}| = \tfrac{1}{2}\,N\,(N+3).$$

So, e.g. for $N = 30$, R has $|\mathcal{S}|^2 = 245025$ entries. To obtain the moments of the queue length distributions several data structures of similar dimensions are needed. Furthermore, R is solved recursively from Equation B.2 in Appendix B. If a large number of iterations is required (which mostly depends on ρ_{eff}), this becomes a rather time-consuming task. For instance the exact solution of the case $N = 10$; $\rho_{eff} = 0.9$ in Table 4 took 1146 iterations and more than 6 minutes on a SUN Sparc station, whereas the approximation took only a fraction of a second. The

examples in Section 4.2 have been generated by employing the exact method of Section 3.1, with an exception for the case $N = 40$ in the last example where the approximation was used, because of excessive memory-requirements. On a simple PC the computations for the exact method may become infeasible to be carried out. Then the approximation will be the only way to analyze the system.

4.1 Accuracy of the approximation

Before giving general conclusions we present illustrative test results in Table 1 upto Table 4. We use the following abbreviated notation. E_L, E_{LbD} and E_{LbU} denote the exact values of $E[L]$, $E[L_{|beginDown}]$ and $E[L_{|beginUp}]$. c_L^2, c_{LbD}^2 and c_{LbU}^2 denote the corresponding squared coefficients of variation, where the squared coefficient of variation of a stochastic variable X with mean E_X and variance σ_X^2 is defined as

$$c_X^2 := \frac{\sigma_X^2}{[E_X]^2}.$$

By % we denote the relative error of the approximation. For instance the error of $E[\hat{L}]$ is defined as

$$\frac{E[\hat{L}] - E[L]}{E[L]} \times 100\% .$$

In each table there is a single repairman ($K = 1$), and the servers are identical with mean lifetime

$$\frac{1}{\sigma} = 10,$$

and mean repairtime

$$\frac{1}{\nu} = 1.$$

In Table 1 and Table 2 five stations are considered ($N = 5$), resulting in only minor interaction between broken servers at the repair facility ($P(N_b > K) = 0.15$; $E[N_b] = 0.64$). In Table 3 and Table 4 $N = 10$, resulting in increased interaction between broken servers at the repair facility ($P(N_b > K) = 0.57$; $E[N_b] = 2.2$). The difference between Table 1 and Table 2 is the speed at which customers pass station 1. In Table 2 customers are ten times as fast as in Table 1, e.g. $\lambda = 4.4$, $\mu = 10$ versus $\lambda = 0.44$, $\mu = 1$, resulting in the same traffic intensity. The same difference exists between Table 3 and Table 4. In each table results are presented for an effective traffic intensity of 0.5, 0.7 and 0.9 .

Note that the only difference between the exact solution and the approximation is that in the approximation consecutive down-periods are assumed to be independent, whereas in the exact situation they are not. So the quality of the approximation depends on the correlation between successive down-periods in the exact model.

This correlation is caused by interaction between broken servers at the repair facility. Such interaction only occurs when the number of servers exceeds the number of repairmen available.

ρ_{eff}	E_L	%	c_L^2	%	E_{LbD}	%	c_{LbD}^2	%	E_{LbU}	%	c_{LbU}^2	%
0.5	1.2	-0.4	1.98	-0.5	1.1	-0.2	2.13	-0.5	1.8	-0.1	1.19	-1.0
0.7	2.7	-0.4	1.45	-0.5	2.6	-0.3	1.54	-0.6	3.5	-0.2	0.99	-1.0
0.9	10.5	-0.5	1.12	-0.3	10.3	-0.4	1.15	-0.3	11.5	-0.4	0.95	-0.4

Table 1: $N = 5$; $K = 1$; slow customers: $\lambda = 0.44,\ 0.61,\ 0.785$; $\mu = 1$; $P(N_b > K) = 0.15$.

ρ_{eff}	E_L	%	c_L^2	%	E_{LbD}	%	c_{LbD}^2	%	E_{LbU}	%	c_{LbU}^2	%
0.5	2.7	-1.5	4.0	-3.7	1.9	-1.1	5.1	-3.6	8.3	-0.2	0.9	-4.2
0.7	6.1	-1.8	2.8	-4.1	5.0	-1.6	3.3	-4.0	13.9	-0.6	0.9	-5.1
0.9	23.7	-2.1	1.6	-2.8	22.2	-2.0	1.7	-2.8	33.7	-1.3	0.9	-4.1

Table 2: $N = 5$; $K = 1$; fast customers: $\lambda = 4.4,\ 6.1,\ 7.85$; $\mu = 10$; $P(N_b > K) = 0.15$.

ρ_{eff}	E_L	%	c_L^2	%	E_{LbD}	%	c_{LbD}^2	%	E_{LbU}	%	c_{LbU}^2	%
0.5	1.4	-2.3	2.05	-3.3	1.2	-1.4	2.43	-3.0	2.3	-0.7	1.08	-5.2
0.7	3.4	-2.9	1.51	-3.6	3.1	-2.2	1.71	-3.6	4.6	-1.5	0.94	-5.7
0.9	13.2	-3.3	1.15	-1.9	12.8	-3.1	1.21	-2.1	14.7	-2.7	0.94	-3.1

Table 3: $N = 10$; $K = 1$; slow customers: $\lambda = 0.39,\ 0.55,\ 0.708$; $\mu = 1$; $P(N_b > K) = 0.57$.

ρ_{eff}	E_L	%	c_L^2	%	E_{LbD}	%	c_{LbD}^2	%	E_{LbU}	%	c_{LbU}^2	%
0.5	5.1	-5.5	3.56	-8.7	3.1	-4.6	6.04	-8.7	13.7	-1.0	0.86	-12.8
0.7	12.4	-7.4	2.56	-10.1	9.4	-6.8	3.56	-9.8	24.4	-2.6	0.88	-15.9
0.9	49.4	-8.8	1.53	-7.0	45.4	-8.6	1.74	-6.9	64.8	-6.0	0.94	-11.8

Table 4: $N = 10$; $K = 1$; fast customers: $\lambda = 3.9,\ 5.5,\ 7.08$; $\mu = 10$; $P(N_b > K) = 0.57$.

This immediately explains why the approximation is exact if $K \geq N$, since then every server has its own repairman, and no interference of broken servers will occur. Usually the servers will hardly ever be down all at the same time. So, for K close to N the approximation still performs extremely well. For the case $K < N$ the approximation performs well as long as *the number of broken servers does not exceed the number of repairmen too often*, irrespective of the other model parameters. Table 1 and Table 2 confirm this conclusion for a situation where the number of broken servers (N_b) exceeds the number of repairmen ($K = 1$) only 15 percent of the time.

If there are broken servers waiting at the repair facility much of the time, then there exists some dependency between consecutive down-periods. As a result, in such a situation the quality of the approximation becomes more sensitive to the arrival and service process at station 1 itself. An *increased effective traffic intensity* ρ_{eff} causes the queue length to become more sensitive to variations in the down-times, since severe queue build-up is more likely to occur. Therefore correlation between down-periods affects the queue length distribution, and thus the accuracy of the approximation (illustrated by all four tables). However, the largest influence on the

accuracy of the approximation comes from the speed at which customers pass station 1. At a given traffic intensity *the accuracy decreases when customers move faster* (i.e. both shorter interarrival times and shorter service times). This can be explained quite easily, since in case of fast customers a small extension of a down-period causes a large number of additional customers to enter station 1 during this down-period (thus intensifying queue build-up at station 1). In case of slow customers small variations of down-times may not even be noticed because of the larger interarrival times. This is best illustrated by comparison of Table 3 and Table 4 where the number of broken servers (N_b) exceeds K during 57 percent of the time.

In all our numerical experiments with $K < N$ the approximation gives *lower bounds* on the exact values of the mean and variance of the various queue length distributions L, $L_{|beginDown}$, and $L_{|beginUp}$. This suggests the existence of positive correlation between successive down-times in the exact model. We do have intuitive arguments to explain this. For $K = 1$ and $N > 1$, suppose a particular repair of server 1 takes extra time to be finished. During this time other servers will break down without being repaired. So, at departure from the repair facility, server 1 leaves more broken servers behind than it usually does. It takes the repairman quite some time to get through the extra large number of broken servers. In the meantime servers are breaking down and eventually server 1 breaks down again. Then it is likely that server 1 finds an increased amount of broken servers at the repair facility (and thus server 1 will experience a longer down-period again), since the extra build-up of broken servers that started during its previous prolonged stay at the repair facility may still not have faded away. This in turn will cause an additional build-up of the number of customers at station 1.

For $K > 1$ such build-up at the repair facility is less severe since the repair facility can never be blocked by a single broken server. This explains why *the approximation performs better for larger* K. This is illustrated by Table 5 where all model parameters are identical to Table 4 except for the number of repairmen. $K = 2$ in this case (and λ is increased to get results for $\rho_{eff} = 0.5,\ 0.7$, and 0.9 again).

ρ_{eff}	E_L	%	c_L^2	%	E_{LbD}	%	c_{LbD}^2	%	E_{LbU}	%	c_{LbU}^2	%
0.5	2.1	-0.4	3.64	-1.1	1.5	-0.3	4.26	-1.0	6.8	-0.1	0.88	-1.2
0.7	4.8	-0.5	2.50	-1.2	4.1	-0.4	2.86	-1.2	11.4	-0.1	0.80	-1.6
0.9	18.4	-0.6	1.49	-0.8	17.5	-0.5	1.60	-0.8	26.8	-0.3	0.80	-1.2

Table 5: $N = 10$; $K = 2$; fast customers: $\lambda = 4.5,\ 6.3,\ 8.07$; $\mu = 10$; $P(N_b > K) = 0.10$.

A remarkable result from our experiments is that the approximation *never produces 'bad' results* (the error in the approximated mean queue lengths never exceeds 10%; the error in the approximated squared coefficients of variation never exceeds 20%). We have two arguments to explain this. The first argument is that successive up- and down-times are independent (cf. Theorem 1 in Schassberger & Daduna [1987]; an intuitive way to see this is the following: consider servers and repairmen only. Server 1 sees the system upon his arrival at station 1 in equilibrium with $N - 1$ servers (Lavenberg & Reiser [1980]). During his stay at station 1, server 1 does not influence the remainder of the system. Hence upon his arrival at the repair facility he still sees the system in equilibrium with $N - 1$ servers. Thus an up-time of server 1 and its successive down-time are independent. Obviously the next up-time is independent of the

length of the current down-period). So, there is only an indirect dependency possible between successive down-times, caused by accumulation of broken servers at the repair facility. This dependency only occurs when $K < N$. The second argument is that for fixed K and $N \to \infty$, the repair facility becomes the bottleneck in the system. Servers spend most of their time at the repair facility and $c^2[D] \to 0$. So, down-times become deterministic and hence independent of each other (we refer to Kleinrock [1976], pp. 206-209 for a detailed explanation of the behaviour of D when $N \to \infty$; see also the last example of Section 4.2).

Remark 4.1 Another very simple approximation that is sometimes proposed for the original model, is the following. First compute the long run fraction of time that the server will be up. From Equation (2.8) we have

$$P_{up} = \frac{1}{1 + \sigma E[D]}.$$

Now, instead of modelling down-periods explicitly, one adjusts the service rate in the following way.

$$\mu' := \mu \; P_{up}.$$

Then the model is solved as an ordinary M/M/1 queue with arrival rate λ and adjusted service rate μ'. So, the mean and the squared coefficient of variation of the approximated queue length are given by the following expressions.

$$E[L'] := \frac{\lambda}{\mu' - \lambda}; \qquad c^2[L'] := \frac{\mu'}{\lambda}. \tag{4.1}$$

Without numerically testing this approximation we know already that this is a bad approximation. From (4.1) we see that the values of $E[L']$ and $c^2[L']$ are independent of the speed at which customers pass station 1, whereas in Table 1 upto Table 4 we have seen that this speed has a major impact on the queue length distributions. Furthermore from this simple approximation we do not get insight in the fluctuations of the queue length at the beginning and at the end of a down-period ($L_{|beginDown}$ and $L_{|beginUp}$ respectively).

We have tested the simple approximation for the situations presented in Table 1 upto Table 4. Relative errors in the mean queue lengths ranged from 13% in Table 1 up to 80% in Table 4. From this remark we conclude that *it is very important to take down-periods of the server explicitly into account*, even when approximating the model. □

4.2 Sensitivity analysis

Before presenting further numerical examples, we make the following important observation. From Table 1 upto Table 4 we see that *the speed of the customers heavily affects the queue length*, i.e. for a given traffic intensity the queue length increases when customers move faster (both shorter interarrival times and shorter service times). For the approximating M/M/1 queue with independent interruptions we can even derive the following conclusion from (3.17), (3.21) and (3.4): when varying the speed of the customers at queue 1 (i.e. when varying λ and μ both

in the same extent) the mean queue length is a linear function of λ. For instance the mean marginal queue length is given by

$$E[\hat{L}] \;=\; \alpha_1 \;+\; \lambda\alpha_2,$$

where α_1 and α_2 are constants that follow from (3.17). This is an important observation, since the queue length of a system without server breakdown is not affected by the speed of the customers at all! Employing Little's Law we see that the mean sojourn time of a customer at queue 1 is given by

$$E[\hat{W}] \;=\; E[\hat{L}]/\lambda = \alpha_1/\lambda \;+\; \alpha_2,$$

which is inverse linear in λ.

In the remainder of this section we investigate the influence of server breakdown and limited repair capacity on the various queue length distributions. In the first example there is one single repairman ($K = 1$) maintaining a number of servers that is varied from $N = 1$ to $N = 15$. The servers are identical with mean lifetime

$$\frac{1}{\sigma} = 10,$$

and mean repairtime

$$\frac{1}{\nu} = 1.$$

The arrival and service rate of customers arriving at station 1 are equal to

$$\lambda = 1.2 \quad \text{and} \quad \mu = 2.0$$

respectively.

Figure 2 depicts $E[L]$ (solid line), $E[L_{|beginDown}]$ and $E[L_{|beginUp}]$ (both dotted) as functions of N. Note that the difference between $E[L_{|beginUp}]$ and $E[L_{|beginDown}]$ is equal to

$$E[L_{|beginUp}] - E[L_{|beginDown}] = \lambda\; E[D],$$

where $E[D]$ is given in Figure 3.

First consider Figure 2 for the case $N = 1$. The influence of server breakdown on the queue length at station 1 is investigated by comparing the mean queue length $E[L]$ for $N = 1$ with the mean queue length of a similar M/M/1 queue without breakdown:

$$E[L] = 2.3; \qquad E[L_{M/M/1}] = \frac{\lambda}{\mu - \lambda} = 1.5\ .$$

Furthermore, the queue subject to server breakdown contains more fluctuation since the mean queue length varies from $E[L_{|beginUp}] = 2.25$ to $E[L_{|beginDown}] = 3.35$.

The remainder of Figure 2 ($N = 2, ..., 15$) shows the influence of limited repair capacity on the queue length distributions. The limited repair capacity causes a regular increase of $E[D]$ (see Figure 3). Even though there is a certain increase of $E[D]$ already for $N = 2, ..., 6$, this does not affect the mean queue lengths considerably. For $N = 7, ..., 10$ the influence of the limited

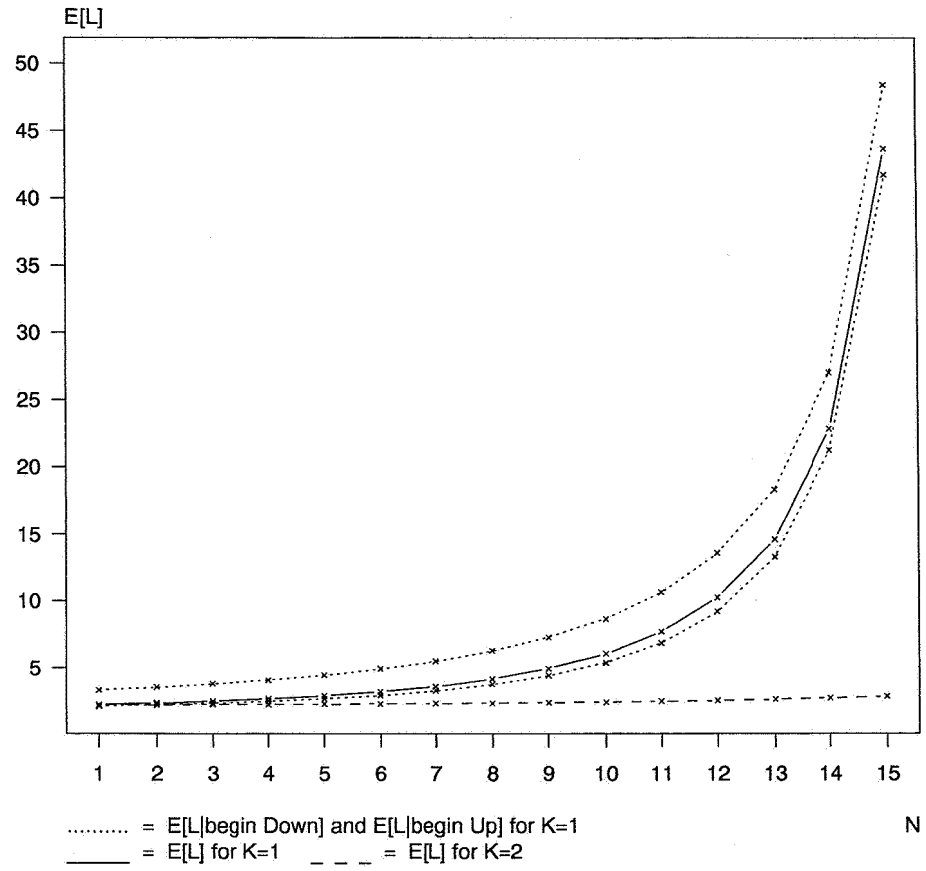

Figure 2: Influence of server breakdown and limited repair capacity on mean queue lengths.

N	1	2	3	4	5	6	7	8
ρ_{eff}	0.660	0.665	0.672	0.679	0.688	0.698	0.711	0.725
N	9	10	11	12	13	14	15	
ρ_{eff}	0.743	0.764	0.789	0.818	0.852	0.891	0.934	

Table 6: $K = 1$; effective traffic intensity.

repair capacity becomes noticeable. For $N > 10$, ρ_{eff} is approaching 0.9 (see Table 6) causing an enormous growth of the mean queue lengths.

In this example the negative influence of limited repair capacity on the queue length distributions can be canceled out for the greater part by adding a second repairman to the repair facility ($K = 2$). In Figure 3 we see that $E[D]$ (the dashed line) does not increase for greater N when $K = 2$, and thus the queue length will hardly be affected (see Figure 2; where the dashed line denotes the mean queue length $E[L]$ for $K = 2$).

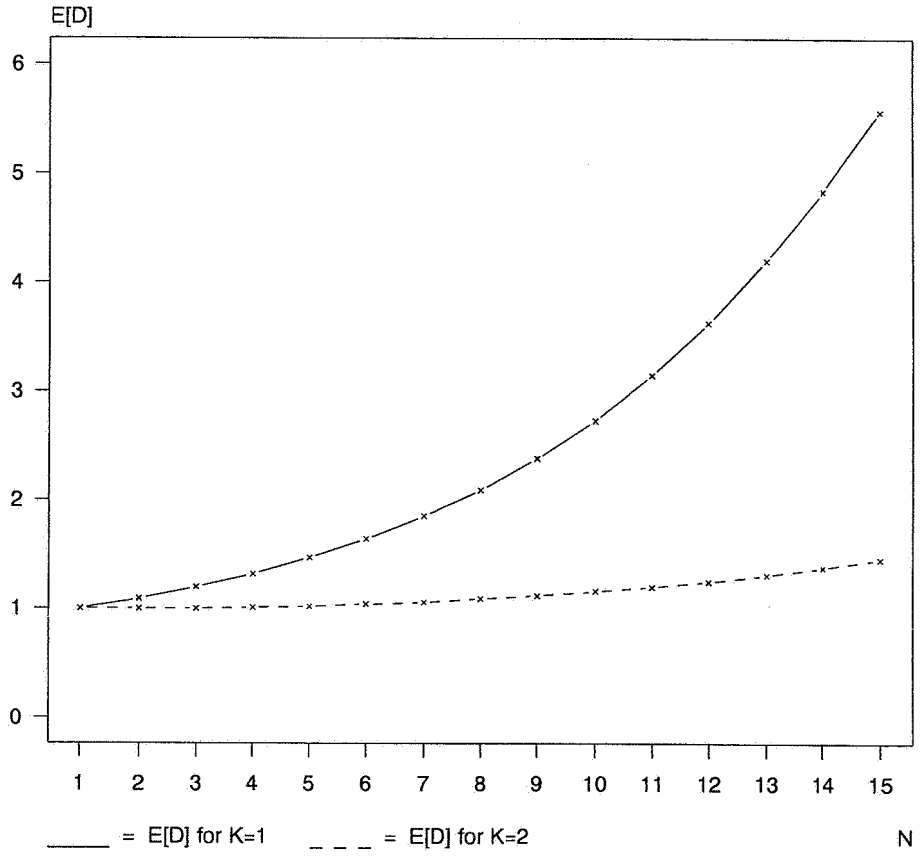

Figure 3: Influence of server breakdown and limited repair capacity on mean down-time.

In the latter example the second repairman was available all the time. However, if costs are taken into consideration, it might be cost-effective to implement some kind of threshold policy, i.e. there exists some threshold value m such that the second repairman is used if and only if the number of broken servers is equal to or exceeds m. From Remark 2.1 we know that such threshold policies can be incorporated in our model in the following way.

$$K_i := \begin{cases} 1 & \text{for } i = 1, ..., m-1 \ , \\ 2 & \text{for } i = m, ..., N \ . \end{cases}$$

For a situation with 15 stations ($N = 15$), Figure 4 shows the effect on the mean queue lengths of using the second repairman according to a threshold policy. The threshold value is varied from $m = 1$ (two repairmen available all the time) to $m = 16$ (never use the second repairman). For $m = 1$ and $m = 2$ this policy is equivalent to the case with $K = 2$ (two repairmen available all the time). For $m = 3, ..., 6$ the increase in $E[L]$ is moderate since large build-up of broken servers at the repair facility is still dealt with by two repairmen. For $m = 7, ..., 12$ addition of a second repairman considerably reduces the length of a server down-period, and thus improves the performance of station 1. There is only a small probability that the number of broken servers exceeds 12. Thus for $m = 12, ..., 15$ the extra repairman is hardly ever used, and hence there is

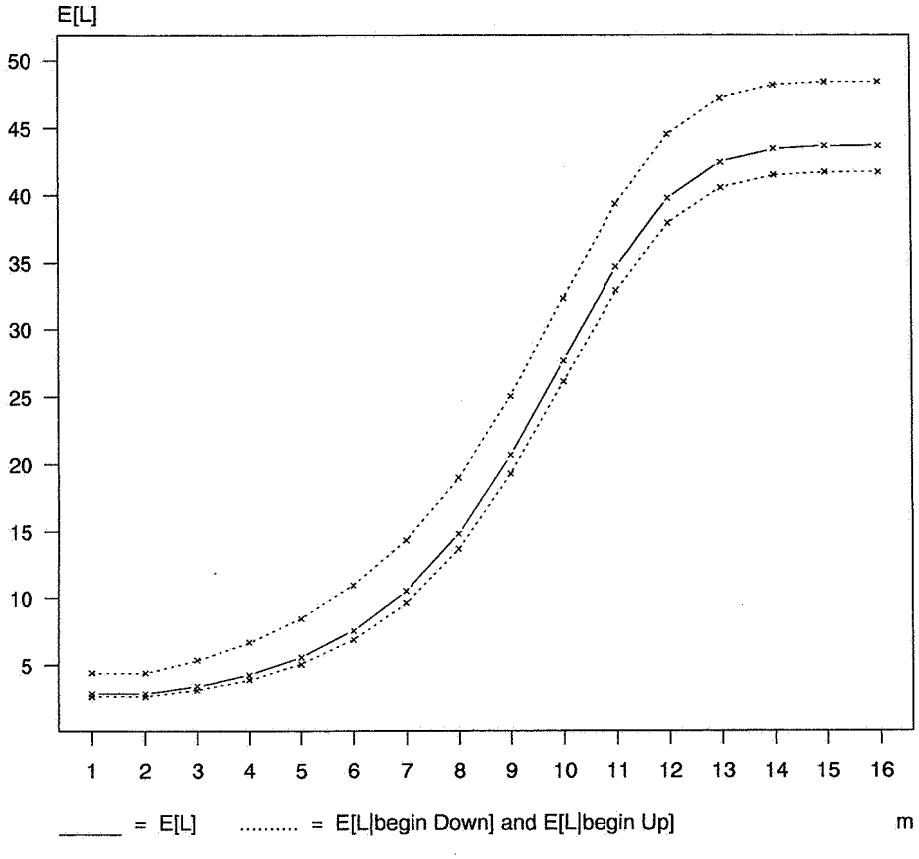

Figure 4: $N = 15$; threshold policies.

not much difference between the mean queue lengths corresponding to those values of m.

Our last example deals with the question 'is more always better ?'. Suppose customers arrive at a service system (consisting of N stations) according to a Poisson process with rate Λ. The arriving customers are randomly distributed over the N stations. So, at station 1 customers arrive according to a Poisson process with rate

$$\lambda = \frac{\Lambda}{N}.$$

Now the question is whether there exists an optimal value of N minimizing the average sojourn time of a customer in the system, or whether addition of an extra station will always improve the performance of the system. The mean sojourn time $E[W]$ is easily obtained employing Little's Law.

$$E[W] = E[L] \ / \ \lambda.$$

In Figure 5 we present $E[W]$ as a function of N, where N is varied from $N = 1$ to $N = 40$. The remaining model parameters are specified as follows.

$K = 1$; $\Lambda = 1.64$; $\mu = 2.0$; $\sigma = 0.1$; $\nu = 1.0$.

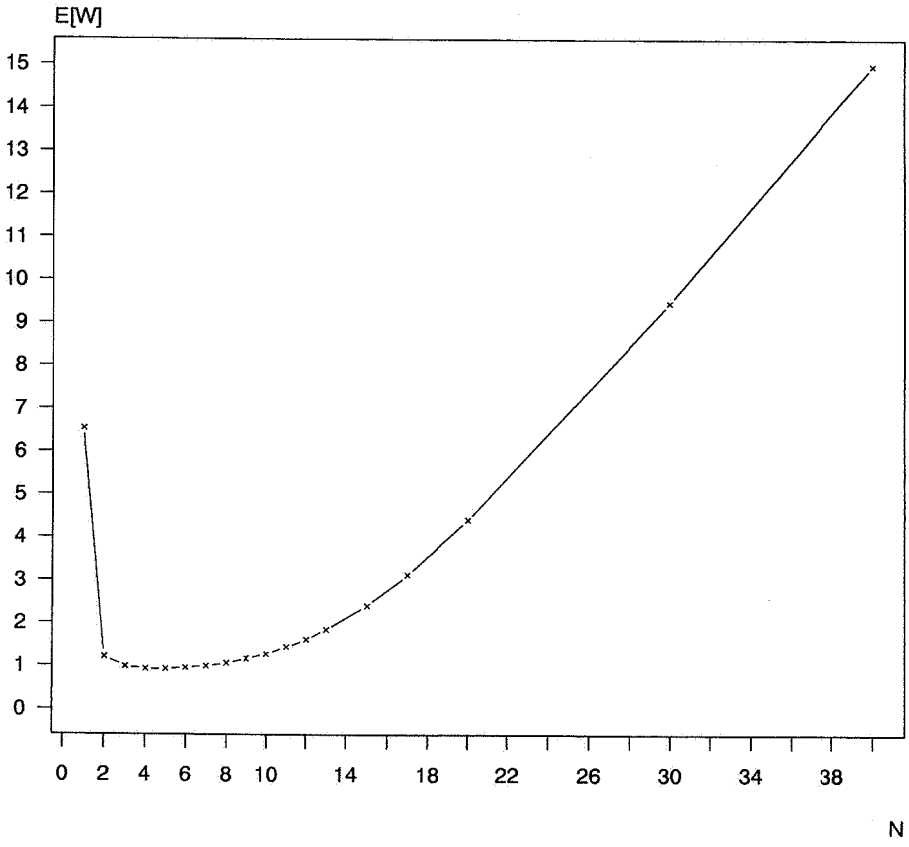

Figure 5: Mean sojourn time $E[W]$ as function of N; $\lambda = \Lambda/N$.

Clearly there exists an optimum value for N ($N^* = 5$ in this example). In Table 7 we present some illustrative results that help to explain the behaviour of $E[W]$ ($\#up$ denotes the number of servers that are up in the long run).

N	1	2	3	4	5	6	7	10	15	20	30	40
$E[D]$	1.00	1.09	1.20	1.32	1.47	1.64	1.85	2.73	5.57	10.04	20.00	30.00
$\#up$	0.91	1.80	2.68	3.53	4.34	5.15	5.91	7.85	9.64	9.98	10.00	10.00
$E[W]$	6.54	1.21	0.99	0.93	0.93	0.95	1.00	1.27	2.41	4.41	9.48	14.98

Table 7: $E[D]$, $\#up$ and $E[W]$ as functions of N.

From Table 7 we see that initially the addition of an extra station causes $E[D]$ to increase only slightly, whereas the number of operational servers initially increases almost linearly with N. So, the total capacity of the N stations increases strongly causing a decrease of $E[W]$. However, for $N > 6$ $E[D]$ keeps growing and finally becomes linear in N approaching its

asymptote:

$$D \approx \frac{N}{\nu} - \frac{1}{\sigma}, \qquad \text{for } N \to \infty. \tag{4.2}$$

For a detailed explanation of the behaviour of D we refer to a discussion of finite population models in Kleinrock [1976], pp. 206-209. Here we just note that when N increases, the repair facility becomes the bottleneck of the system. The maximum number of repaired servers that leave the repair facility per time unit is equal to ν. Servers fail with rate λ. So, when the repair facility is working at full capacity (which it is for $N \to \infty$), the maximum number of operational servers is given by

$$\frac{\nu}{\sigma} = 10,$$

which is confirmed by Table 7. So, for $N \to \infty$ the customers are still equally distributed over the N stations of which only $\frac{\nu}{\sigma}$ are up on average. Therefore arriving customers will often find the server at the repair facility, and thus they will have to wait until this server is repaired. This explains the increase of $E[W]$ for $N \to \infty$. By substitution of the asymptotic expression (4.2) for D into (3.17), it can even be shown that $E[W]$ finally becomes a linear function of N.

5 Conclusion

In this paper we present a model of a queueing station subject to server breakdown and limited repair capacity. Both exact and approximate solutions are obtained for the queue length distributions at this queueing station. The exact solution has great modelling flexibility, but becomes time- and memory-consuming for larger problem instances. The approximation is simple and gives accurate results.

The approximation performs well whenever the number of broken servers does not exceed the number of repairmen too often. If the repair facility is working at full capacity most of the time (in a situation with limited repair capacity), then the accuracy of the approximation is negatively influenced by the traffic intensity and especially by the speed at which customers pass station 1. The approximation performs better for a larger number of repairmen. Whenever the repair capacity is limited, the approximation gives lower bounds on the exact values of the mean and variance of the various queue length distributions.

A general conclusion from our sensitivity analysis is that it is important to take down-periods of the server explicitly into account, even when approximating the model. Furthermore the queue length of a service station subject to server breakdown is heavily influenced by the speed at which customers pass this station, whereas the queue length of a station without server breakdown is not.

Acknowledgements The author would like to thank Onno Boxma, Sem Borst, Frank van der Duyn Schouten and Erik van Doorn for their valuable suggestions and helpful comments, and Rob van der Horst for visualizing the numerical examples.

Appendix A. Distribution of D, L_{PD} and $L_{M/G/1_{|C}}$

The length of an arbitrary down-period D is equal in distribution to the sojourn time of a broken server at the repair facility. The distribution of D is obtained by conditioning on the number of broken servers at the repair facility found by a server that just broke down (tagged server). Let

$$r_j := P(\text{tagged server finds } j \text{ servers at the repair facility}), \quad 0 \leq j \leq N-1.$$

To determine the probabilities r_j, we consider the N stations together with the repair facility as a closed queueing network. The N servers become customers moving around the network. According to the Arrival Theorem (Lavenberg & Reiser [1980]) the stationary state probabilities at instants at which customers move from one service station to another are equal to the stationary state probabilities at a random point in time for the network with one less customer. Therefore r_j is the stationary probability that j servers are at the repair facility of a system with $N-1$ stations and K repairmen. Thus r_j follows from Lemma 2.1 with N replaced by $N-1$.

Now, suppose the tagged server finds j $(0 \leq j \leq N-1)$ servers at the repair facility. If

$$0 \leq j \leq K-1,$$

then repair on the tagged server is started immediately and the tagged server will leave the system after an $exp(\nu)$ distributed amount of time. If

$$K \leq j \leq N-1,$$

then all repairmen are busy. In this situation the tagged server has to wait until one of the repairmen becomes free, i.e. until repair on $j-K+1$ servers is completed. Due to the memoryless property of the exponential distribution this takes an $Erlang[j-K+1, K\nu]$ distributed amount of time (i.e. the sum of $j-K+1$ minima of K $exp(\nu)$ distributed stochastic variables). Thus the distribution of the sojourn time of the tagged server at the repair facility is the convolution of an $Erlang[j-K+1, K\nu]$ and an $exp(\nu)$ distributed stochastic variable in this case.

The moments of D are obtained in a straightforward way. We present $E[D]$, $E[D^2]$ and $E[D^3]$.

$$E[D] = \sum_{j=0}^{\min\{K-1,N-1\}} r_j \frac{1}{\nu} + \sum_{m=1}^{N-K} r_{m+K-1} \left[\frac{m}{K\nu} + \frac{1}{\nu}\right],$$

$$E[D^2] = \sum_{j=0}^{\min\{K-1,N-1\}} r_j \frac{2}{\nu^2} + \sum_{m=1}^{N-K} r_{m+K-1} \left[\frac{m(m+1)}{(K\nu)^2} + \frac{2}{\nu}\left[\frac{m}{K\nu} + \frac{1}{\nu}\right]\right],$$

$$E[D^3] = \sum_{j=0}^{\min\{K-1,N-1\}} r_j \frac{6}{\nu^3} + \sum_{m=1}^{N-K} r_{m+K-1} \left[\frac{m(m+1)}{(K\nu)^2}\left[\frac{m+2}{K\nu} + \frac{3}{\nu}\right] + \frac{6}{\nu^2}\left[\frac{m}{K\nu} + \frac{1}{\nu}\right]\right].$$

The distribution of L_{PD} (defined by (3.16)) is obtained by conditioning on the length of a backward recurrence time of a down-period, that is distributed according to the following distribution function (Ross [1983], Proposition 3.4.5):

$$\frac{1}{E[D]} \int_0^t P(D > x)dx.$$

The first two moments of L_{PD} depend on $E[D]$, $E[D^2]$ and $E[D^3]$ in the following way.

$$\begin{aligned} E[L_{PD}] &= \lambda \frac{E[D^2]}{2E[D]}, \\ E[L_{PD}^2] &= \lambda \frac{E[D^2]}{2E[D]} + \lambda^2 \frac{E[D^3]}{3E[D]}, \end{aligned}$$

where $\frac{E[D^2]}{2E[D]}$ and $\frac{E[D^3]}{3E[D]}$ equal the first two moments of the excess down-time.

The moments of $L_{M/G/1_{|C}}$ (3.15) follow from known results on the ordinary M/G/1 queue (cf. Kleinrock [1975] or Jain [1991]) by considering completion times instead of service times. Suppose customers arrive at a single server according to a Poisson(λ) process, where they require an amount of service time S that is generally distributed. No explicit expressions exist for the distribution of the number of customers in the system at a random point in time ($L_{M/G/1}$). The mean and the variance are given by the following expressions.

$$E[L_{M/G/1}] = \lambda E[S] + \frac{\lambda^2 E[S^2]}{2(1-\lambda E[S])}, \tag{A.1}$$

$$Var[L_{M/G/1}] = E[L_{M/G/1}] + \lambda^2 \left[E[S^2] - E[S]^2\right] + \frac{\lambda^3 E[S^3]}{3(1-\lambda E[S])} + \frac{\lambda^4 E[S^2]^2}{4[1-\lambda E[S]]^2}. \tag{A.2}$$

As described in Section 3.2, we can inflate the service times S with possible interruptions (due to server breakdown) to obtain the so-called completion times C. The mean and variance of $L_{M/G/1_{|C}}$ are obtained by substituting $E[C]$, $E[C^2]$ and $E[C^3]$ for $E[S]$, $E[S^2]$ and $E[S^3]$ respectively in (A.1) and (A.2). The moments of C are obtained from the Laplace-Stieltjes transform, which has been derived by Gaver [1962]:

$$\tilde{U}(s) = \tilde{V}(s + \sigma - \sigma\tilde{F}(s)), \tag{A.3}$$

where $\tilde{F}(s)$, $\tilde{V}(s)$ and $\tilde{U}(s)$ denote the Laplace-Stieltjes transforms of D, S and C respectively. Differentiation of (A.3) leads to the following expressions for $E[C]$, $E[C^2]$ and $E[C^3]$.

$$\begin{aligned} E[C] &= E[S][1 + \sigma E[D]], \\ E[C^2] &= E[S^2][1 + \sigma E[D]]^2 + \sigma E[S]E[D^2], \\ E[C^3] &= E[S^3][1 + \sigma E[D]]^3 + \sigma E[S]E[D^3] + 3\sigma E[S^2]E[D^2][1 + \sigma E[D]]. \end{aligned}$$

Appendix B. Matrix-geometric solution

For a general explanation of matrix-geometric solutions in stochastic models we refer to Neuts [1981] or Nelson [1991]. Here, we restrict ourselves to the solution of an M/M/1 queue in a Markovian environment (see Section 3.1 or Neuts [1981] for a definition).

Let $\Delta\underline{a}$ denote the diagonal matrix $\text{diag}(a_1, .., a_{|\mathcal{S}|})$ for some vector $(a_1, .., a_{|\mathcal{S}|})$. Then by lexicographically ordering the state space, the queue may be studied as a quasi birth-death process (Neuts [1981], p.258), with generator $\tilde{Q}$ given by

$$\tilde{Q} = \begin{pmatrix} A_1 + A_2 & A_0 & 0 & 0 & ... \\ A_2 & A_1 & A_0 & 0 & ... \\ 0 & A_2 & A_1 & A_0 & ... \\ 0 & 0 & A_2 & A_1 & ... \\ . & . & . & . & \\ . & . & . & . & \\ . & . & . & . & \end{pmatrix},$$

where

$$\begin{aligned} A_0 &:= \Delta(\underline{\lambda}), \\ A_1 &:= Q - \Delta(\underline{\lambda} + \underline{\mu}), \\ A_2 &:= \Delta(\underline{\mu}). \end{aligned}$$

The state (i, j), $i \geq 0$, $1 \leq j \leq |\mathcal{S}|$ corresponds to a queue length i and an environmental state j. Let π denote the stationary probability vector of Q (the environmental Markov chain). Then the queue is *stable* if and only if (Neuts [1981])

$$\pi\underline{\lambda} \quad < \quad \pi\underline{\mu}. \tag{B.1}$$

To obtain the stationary probability vector of the quasi birth-death process we need the matrix R, which is the minimal non-negative solution of the equation

$$R^2 A_2 + R A_1 + A_0 = 0. \tag{B.2}$$

Let

$$\begin{aligned} B_2 &= -A_2 A_1^{-1}, \\ B_0 &= -A_0 A_1^{-1}. \end{aligned}$$

Then, R is numerically solved from (B.2) by successive substitution:

$$\begin{aligned} R_{n+1} &= R_n^2 B_2 + B_0, \qquad n \geq 2, \\ R_1 &= B_0. \end{aligned}$$

The iteration is stopped when

$$\max_{i,j \in \mathcal{S}} \{R_{n+1}\,[i, j] - R_n[i, j]\} < \epsilon. \qquad (\epsilon = 1.0E - 7)$$

It can be shown that the sequence $\{R_n\}$ is entry-wise non-decreasing and converges monotonically to a non-negative matrix R which satisfies Equation (B.2) (Neuts [1981]). Provided (B.1) holds, the following relationship can be used to check the accuracy of R:

$$R\underline{\mu} = \underline{\lambda}.$$

Now, the stationary probability vector $\underline{x} = (x_0, x_1, ...) = (x_{0,1}, .., x_{0,|\mathcal{S}|}, x_{1,1}, .., x_{1,|\mathcal{S}|}, ...)$ of the stable queue is a *matrix-geometric probability vector*, and is given by (cf. Theorem 3.2, and Neuts [1981]):

$$x_k = \pi(I - R)R^k, \qquad k \geq 0.$$

The following relationship can be used to check the accuracy of x_k, $k \geq 0$:

$$\sum_{k=0}^{\infty} x_k = \pi.$$

References

[1] Boxma O.J. (1989). Workloads and waiting times in single-server systems with multiple customer classes. *Queueing Systems*, vol. 5, pp. 185-214.

[2] Colard J.P. & Latouche G. (1980). Algorithmic analysis of Markovian model for a system with batch and interactive jobs. *Opsearch*, vol. 17, no. 1, pp. 12-32.

[3] Doshi B. (1990). Single server queues with vacations. In: *Stochastic analysis of computer and communication systems*, H. Takagi (ed.), North-Holland Publishing Company, Amsterdam, pp. 217-265.

[4] Fuhrmann S.W. & Cooper R.B. (1985). Stochastic decompositions in the M/G/1 queue with generalized vacations. *Operations Research*, vol. 33, no. 5, pp. 1117-1129.

[5] Gaver D.P. (1962). A waiting line with interrupted service, including priorities. *Journal of the Royal Statistical Society*, vol. B24, pp. 73-91.

[6] Haverkort B.R., Van Moorsel A.P.A. & Dijkstra A. (1992). MGMtool: a performance modelling tool based on matrix geometric techniques. *Memoranda Informatica 92-35, University Twente.*

[7] Jain R. (1991). *The art of computer systems performance analysis.* Wiley, New York.

[8] Kleinrock L. (1975). *Queueing systems; volume I: theory.* Wiley, New York.

[9] Kleinrock L. (1976). *Queueing systems; volume II: computer applications.* Wiley, New York.

[10] Lavenberg S.S. & Reiser M. (1980). Stationary state probabilities at arrival instants for closed queueing networks with multiple types of customers. *Journal of Applied Probability*, vol. 17, pp. 1048-1061.

[11] Nelson R. (1991). Matrix geometric solutions in Markov models: a mathematical tutorial. *IBM Research Report* RC 16777.

[12] Neuts M.F. (1978). The M/M/1 queue with randomly varying arrival and service rates. *Opsearch*, vol. 15, no. 4, pp. 139-157.

[13] Neuts M.F. (1978). Further results on the M/M/1 queue with randomly varying rates. *Opsearch*, vol. 15, no. 4, pp. 158-168.

[14] Neuts M.F. (1981). *Matrix-geometric solutions in stochastic models: an algorithmic approach.* The Johns Hopkins University Press, Baltimore, Maryland.

[15] Neuts M.F. & Lucantoni D.M. (1979). A Markovian queue with N servers subject to breakdowns and repairs. *Management Science*, vol. 25, no. 9, pp. 849-861.

[16] Ross S.M. (1983). *Stochastic processes.* Wiley, New York.

[17] Schassberger R. & Daduna H. (1987). Sojourn times in queueing networks with multiserver modes. *Journal of Applied Probability*, vol. 24, pp. 511-521.

[18] Tijms H.C. (1986). *Stochastic modelling and analysis: a computational approach.* Wiley, Chichester.

[19] Van Doorn E.A. & Regterschot G.J.K. (1988). Conditional PASTA. *Operations Research Letters*, vol. 7, no. 5, pp. 229-232.

[20] Vinod B. (1985). Unreliable queueing systems. *Computers and Operations Research*, vol. 12, pp. 323-340.