Simple Bounds for Queueing Systems with Breakdowns

Nico M. Van Dijk *
Technical University of Twente, Enschede, The Netherlands

Received 20 November 1986
Revised 6 July 1987

Computationally attractive and intuitively obvious simple bounds are proposed for finite service systems which are subject to random breakdowns. The services are assumed to be exponential. The up and down periods are allowed to be generally distributed. The bounds are based on product-form modifications and depend only on means. A formal proof is presented. This proof is of interest in itself. Numerical support indicates a potential usefulness for quick engineering and performance evaluation purposes.

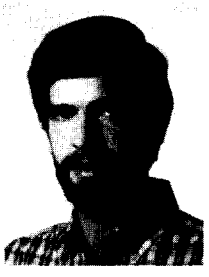
Keywords: Breakdown, Call Congestion, Job-Local-Balance, Bounding Methodology, Product Form, Bound, Insensitivity, Markov Chain.

1. Introduction and methodology

Service systems can be subject to breakdowns or service interruptions such as due to a machine failure, regular off-periods e.g. lunch or shift times, a blocked output channel, or some other external cause. This paper will be concerned with finite service systems which alternate between up and down periods during which service can and cannot be provided as resulting from a breakdown. The objectives are the following:

(i) To apply a recently developed bounding methodology.
(ii) To propose computationally attractive bounds.
(iii) To secure a measure of insensitivity.

Queueing systems with breakdowns, service interruptions or priorities do not generally exhibit the celebrated Erlang-type expression (cf. [8,9,14,18]). Closed form expressions (cf. [1,9,16,17,18]) as well as approximations (cf. [1,10,14,21]) have therefore been developed for specific situations. Especially the case of a single server with an infinite capacity has been studied in depth (cf. [8,9]), while multi-server models have received some attention in the exponential case (cf. [19,20]). The results of all these studies, however,



Nico M. van Dijk received his M.Sc. and Ph.D. in Applied Mathematics from the University of Leiden, The Netherlands in 1979 and 1983, respectively. Since then he has been with the University of British Columbia, the University of Twente, and the Free University of Amsterdam where he is currently associate professor in the Faculty of Economical Sciences and Econometrics. His current main research interests concern both exact expressions and bounds for queueing networks and their application to various areas such as computer performance evaluation, telecommunication, and flexible manufacturing.

* Present affiliation: Faculty of Economical Sciences and Econometrics, Free University, P.O. Box 7161, 1007 MC Amsterdam, The Netherlands.

North-Holland
Performance Evaluation 8 (1988) 117–128

still require a fair amount of computation and have to take into account the distributional forms of up and down periods. As for approximations, moreover, the accuracy cannot be guaranteed beforehand. For engineering or performance evaluation purposes, however, one might just be interested in robust but secure bounds which can be obtained at low computational expense, so as to get a quick impression of the performance of the system.

This paper proposes simple bounds for the call congestion (i.e., the steady-state probability that the system is saturated), and thus also for the throughput, of finite service systems with breakdowns. The services are hereby assumed to be exponential, while the up and down periods are allowed to be generally distributed. Moreover, both single- and multi-server systems are included.

The bounds are based on product-form modifications as according to a general bounding methodology for non-product-form systems. This methodology is introduced in [5]. The bounds are computationally attractive and intuitively appealing. Moreover, they are insensitive (robust) to the distributional forms of the up and down periods. That is, they depend upon these periods only through their means. A measure of insensitivity is hereby secured. Numerical support is provided for both the single-server delay and pure multi-server case. The numerical results indicate that the bounds can serve as reasonable and secure first estimates of the order of magnitude and thus qualify for quick engineering or performance evaluation purposes.

The nonexponential service case will also be addressed. Counterintuitively, for the single-server case the validity of the bounds will be questioned by means of a stochastic realization. For the pure multi-server case, it will be argued and conjectured that the bounds remain valid (that is, are insensitive).

The bounds are intuitively supported. Nevertheless, a formal proof is presented in order to justify their use in practice. (Besides, as argued in the nonexponential case, intuition might be incorrect.) The technique of this proof has already been succesful in other situations (cf. [4,5,7]) and seems a fruitful extension to known comparison techniques (cf. [6,24,25,26]).

The paper is strongly related to [3,4,5,7] in its methodology of finding bounds and its technique of proving these bounds. However, in view of the practical importance of the system studied and the special technicalities involved, the results of this paper deserve special attention.

The organization is as follows. First, in the remainder of this introduction the bounding methodology will be outlined. Section 2 first describes and discusses the system of interest. Next, the bounding methodology will be applied on a purely intuitive basis. The bounds and numerical support will thereby be included. Also, the nonexponential case and an extension to tandem queues are touched upon. Section 3 presents the formal proof of the bounds. An evaluation concludes the paper.

1.1. Bounding methodology

The bounding methodology is based upon a so-called notion of job-local-balance (JLB) which states:

"The rate into a state due to a particular job = the rate out of that state due to that job."

This notion, which is introduced in [12], can be seen as a refined version of other notions such as local-, detailed-, or partial balance (cf. [2,15,23]). In [12,13] it is demonstrated that JLB is responsible for product-type expressions and insensitivity properties. The following bounding methodology for a non-product-form queueing system is therefore suggested:

"Modify the original system such that
(i) the notion of job-local-balance is guaranteed,
(ii) bounds for a performance measure of interest are expected."

The methodology may thus lead to bounds which can be computed by product forms and which may possess insensitivity properties. The finding of appropriate modifications may itself result from the intuitive interpretation of JLB, as will appear in Section 2. The methodology has already led to simple and insensitive bounds for M/G/c/n-queues [6], overflow situations [3], and finite tandem configurations [5,7]. In this paper it will be investigated for systems with breakdowns. The performance measure of

interest will be the call congestion B defined as

$B =$ "the steady-state probability that the system is saturated".

By

$$T = \lambda[1 - B],$$

where λ denotes the arrival intensity, results for the throughput T are hereby included, while similar results for other measures such as a mean queue length or processor utilization can be given along the same lines. (It may be noted that the methodology is not related to [28] which concerns product-form systems.)

2. Model and bounds

2.1. Model

Consider a service facility which can accommodate at most N jobs at a time. Jobs arrive according to a Poisson process with parameter λ. An arriving job is rejected and lost if upon its arrival N jobs are already present. The service requirements are exponential with parameter μ.

The facility itself, however, is subject to breakdowns independently of whether it is busy or not and how long it has been busy. A breakdown renders the system inoperative for a while. The system thus alternates between operative (up) and inoperative (down) periods. These 'up' and 'down' periods are assumed to constitute an alternating renewal process with distribution functions F_1 and F_0 and means γ_1^{-1} and γ_0^{-1}, respectively.

When n jobs are present and the facility is 'up' (operative), it provides service at a rate $\phi(n)$, where $\phi(n)$ is nondecreasing in n and $\phi(0) = 0$. Special applications of our model are:

(a) $\phi(n) = 1$, $n \leqslant N$ (*Single-server case*). A prototypical example is a CPU which serves jobs in a processor-sharing, first-come first-served (FCFS), or last-come first-served (LCFS) preemptive manner. A breakdown may result from a device failure within the CPU.

(b) $\phi(n) = n$, $n \leqslant N$ (*Pure multi-server case*). This is the situation of a multi-processor computer unit or a multi-channel telephone exchange in which each job present uses one processor or channel while there is no waiting accommodation (buffer). A breakdown may reflect the blocking of a jointly used single output device or channel.

(c) $\phi(n) = n$, $n \leqslant c$, $\phi(n) = c$, $c \leqslant n \leqslant N$ (*Multi-server delay case*). This case, which includes both (a) and (b), may correspond to a multi-server station with c servers and $N - c$ waiting places. Besides the breakdown possibilities mentioned in (a) and (b), a breakdown might also occur by a failure of a control device or a stockout of a source that needs to supply raw material for services.

2.2. The stationary queue length distribution

The stationary queue length distribution of the system described above does not exhibit a simple geometric form and depends on the distributional forms of the up and down periods. For a single-server facility with an infinite capacity of accommodation ($N = \infty$) and exponential operating periods the generation function of the queue length distribution has been obtained in [18, p. 103]. This function has a non standard form and explicitly depends on the form of F_0. For the finite capacity case ($N < \infty$) a similar result can be expected but does not seem to have been reported. For example, with $N = \lambda = \mu = \gamma_0 = \gamma_1 = 1$ and F_1 exponential, the stationary probability that the facility is up and serving a job is easily shown to be equal to $\frac{3}{10}$ when F_0 is exponential, but $\frac{7}{23}$ when F_0 is an Erlang-2 distribution. As will be heuristically argued below, this lack of insensitivity and a simple explicit expression can be explained by the notion of job-local-balance.

2.3. JLB-failure

Without restriction of generality, suppose that the jobs are being served in a last-in first-out preemptive manner and consider a state in which the system is down while at least one job is present. The last entered

job from the jobs present is thus interrupted in its service. Consequently, the rate out of that state due to that job is 0. That same job, however, could have entered while the system was already down, so that the rate into that state due to that job is positive. JLB thus fails by service interrupted jobs when the system is down. According to [12], a product-form can therefore not be expected and according to [13] (or [23]) the system is not insensitive.

2.4. Upper bound

In view of the reasoning above, the following modification is suggested so as to repair JLB:

"Whenever the system is down let it reject arriving jobs".

Roughly, JLB seems hereby repaired since also the rate into a state due to jobs present when the system is down is equal to 0. Intuitively, this modification will have the effect that jobs are rejected more frequently, so that the call congestion (that is the probability that an arriving job is lost) will be enlarged. The modification thus suggests an upper bound for the call congestion of the original system. We will therefore refer to this modified model as the 'upper bound model'.

In order to calculate this upper bound, let (n, θ) denote that n jobs are present while the facility has the status θ, where $\theta = 1$ stands for 'up' and $\theta = 0$ for 'down'. Then, when both F_0 and F_1 are exponential, the following stationary distribution of the upper bound model can be concluded on the basis of JLB as according to [12]. It can also be verified easily by substitution in the global balance equations. With c a normalizing constant and $\rho = \lambda/\mu$:

$$\pi(n, \theta) = c[\gamma_\theta]^{-1}\rho^n \Big/ \left[\prod_{k=1}^{n} \phi(k)\right] \quad (n \leqslant N,\ \theta = 0, 1), \tag{1}$$

where the denominator in the right-hand side is defined as 1 for $n = 0$. Moreover, since these probabilities satisfy the notion of JLB (under appropriate disciplines such as a last-in first-out preemptive and with a down period seen as a visit of a priority job that cycles around), it follows from [13] (or indirectly [23]) that these probabilities remain valid if the exponentiality assumptions for F_0 and F_1 are dropped. That is, they depend only on the means γ_0^{-1} and γ_1^{-1} and are thus insensitive to the 'up' and 'down' periods.

Write $\tau = [\gamma_1/\gamma_0]$, which represents the fraction of time that the system is down up to a factor $(\gamma_1 + \gamma_0)/\gamma_1$. Also, replace the normalizing constant c in (1) by $\bar{c}\ \gamma_1$, where $\bar{c}$ is also a normalizing constant, so that $[\pi(\cdot)/\bar{c}] = [\pi(\cdot)/c]\gamma_1$. Then, by virtue of the 'PASTA' theorem (cf. [27]), expression (1), and the intuitive reasoning above, it is suggested that the following expression provides an upper bound B_U on the call congestion, regardless of the distributional forms of F_0 and F_1. This bound is insensitive to the 'up' and 'down' periods:

$$B_U = \frac{\left[\rho^N \Big/ \prod_{k=1}^{N} \phi(k)\right] + \tau \sum_{n=0}^{N} \left[\rho^n \Big/ \prod_{k=1}^{n} \phi(k)\right]}{(1+\tau) \sum_{n=0}^{N} \left[\rho^n \Big/ \prod_{k=1}^{n} \phi(k)\right]}. \tag{2}$$

2.5. Lower bound

In view of the JLB-failure again, the alternative modification which would repair JLB is to avoid that jobs are interrupted in their service. This will be established by naively assuming that the system never breaks down. The system is then reduced to a standard birth-death process of which the stationary distribution is also given by (1) but with $\gamma_0 = \infty$ so that $\pi(\cdot, \theta) = 0$ for $\theta = 0$. Intuitively, a system without breakdowns will have a lower call congestion than with breakdowns, regardless of the up and down time distributions. A naive lower bound B_L for the call congestion which is insensitive to the 'up' and 'down' periods is thus suggested by the right-hand side of expression (2) with $\tau = 0$.

2.6. Numerical results

Numerical examples of the above conjectured upper and lower bound are given in Table 1 for the single-server case (a) and in Table 2 for the pure multi-server case (b). Similar examples can be given for the less extreme mixed case (c). The results seem to indicate that for a wide range of parameters the bounds provide reasonable secure estimates of the order of magnitude at hardly computational expense.

For fixed downtime intensity τ the width between the lower and upper bound is rather constant throughout, so that for small call congestions (say less than 0.10) they can hardly be seen as accurate. However, they are not meant as approximations but merely as quick and robust indicators. Besides, for realistic call congestions in the order of 0.1 also the down-time intensity is likely to be quite small, say less than 2%, in which case the bounds, at least the upper bounds, are quite useful estimates. It is also noted that the bounds give qualitative insight such as in the impact of breakdowns for decreasing breakdown intensity τ.

2.1. Remark (*Active breakdowns*). The type of breakdown considered above is known in the literature as 'independent breakdown' (cf. [18, p. 101]). In contrast, when a breakdown can occur only when the system is busy it is called an 'active breakdown' (cf. [18, p. 101]). The job-local-balance arguments given above so as to obtain bounds for the call congestion can almost verbally be adopted to the active breakdown case.

Table 1
Single-server case (a)

N	ρ	τ	B_L	B_U
80	10	0.1	0.90	0.91
80	2	0.1	0.50	0.55
		0.05	0.50	0.53
20	1.5	0.1	0.33	0.40
		0.05	0.33	0.37
		0.02	0.33	0.35
4	1	0.2	0.20	0.33
		0.1	0.20	0.27
		0.05	0.20	0.24
		0.02	0.20	0.22
4	0.75	0.05	0.10	0.15
		0.02	0.10	0.13
		0.01	0.10	0.12
10	1	0.01	0.091	0.10
		0.005	0.091	0.096
15	1	0.02	0.062	0.081
		0.01	0.062	0.072
		0.005	0.062	0.068
4	0.50	0.02	0.032	0.052
		0.01	0.032	0.042
		0.005	0.032	0.037
10	0.8	0.01	0.023	0.032
		0.005	0.023	0.029
		0.001	0.023	0.025
10	0.5	0.01	0.000	0.011
		0.005	0.000	0.006
		0.001	0.000	0.002

Table 2
Multi-server case (b)

N	ρ	τ	B_L	B_U
10	100	0.1	0.90	0.91
10	20	0.1	0.53	0.58
		0.05	0.53	0.56
20	30	0.1	0.38	0.44
		0.05	0.38	0.41
		0.02	0.38	0.40
8	8	0.2	0.23	0.36
		0.1	0.23	0.30
		0.05	0.23	0.27
		0.02	0.23	0.25
20	20	0.1	0.16	0.24
		0.05	0.16	0.20
		0.02	0.16	0.18
30	25	0.05	0.052	0.098
		0.01	0.052	0.062
20	15	0.05	0.045	0.091
		0.01	0.045	0.065
		0.005	0.045	0.055
8	4	0.02	0.030	0.050
		0.01	0.030	0.040
		0.005	0.030	0.036
10	5	0.01	0.018	0.028
		0.005	0.018	0.024
		0.001	0.018	0.020
10	1	0.01	0.000	0.010
		0.005	0.000	0.005
		0.001	0.000	0.001

The lower bound will remain the same while the upper bound can be obtained from (1) provided state (0,0) is excluded. Numerical examples turned out to be of the same order.

2.2. Remark (*Nonexponential services*). Based upon the job-local-balance notion (cf. [12]) or reversibility arguments (cf. [15]), the bounding modifications are insensitive also to the service distribution (i.e., they depend only on the mean μ^{-1}) for a class of disciplines which includes the processor-sharing, last-in first-out (LIFO) preemptive and pure multi-server discipline. At first instance, therefore, one might expect that also the bounds are insensitive to the service distribution. However, one has to be most careful!

Counterexample. To shed some light on this, let us consider the extreme example of deterministic up and down periods, with respective lengths 6 and 2 as well as deterministic services of length 4. Let $N = 2$ while jobs are being served in a last-in first-out preemptive manner by a single server. Let a realization of the Poisson arrival process have successive arrivals at times: 3, 7, 10, 11, and 22. Then in Fig. 1(a) the corresponding realizations for the queueing processes are graphically indicated for the original and the upper bound model. Herein D_i denotes the departure time of the ith actually accepted job. Now observe that the second arrival is rejected in the upper bound model whereas accepted in the original model. Since however this accepted job takes over the service of the first job, the next completion in the original model requires 4 rather than the residual 1 unit of service to be completed. As a result, during the 9–12 period this leads to 2 rejections in the original model as opposed to 2 acceptances in the upper bound model. Within the regeneration cycle 3–22, which is the same for both models, we thus observe one more rejection in the original than in the upper bound model. This conflicts with the initial guess of an upper bound. Roughly speaking, the 9–12 period is responsible for this. Or, more generally, the fact that a shorter

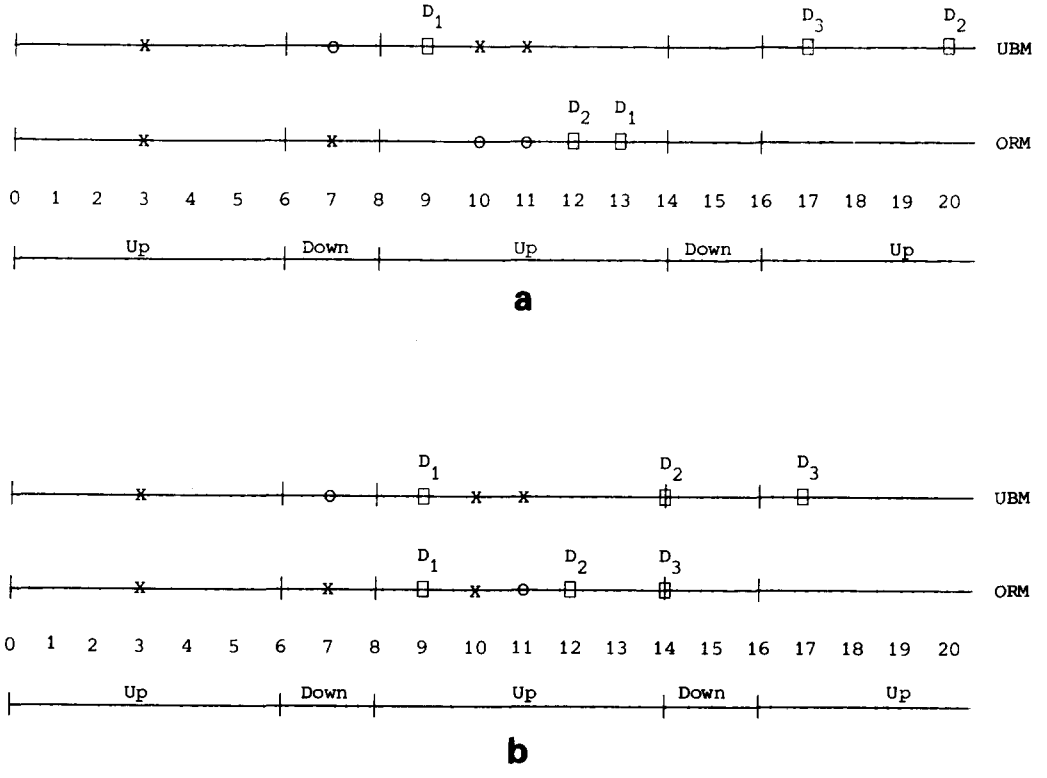


Fig. 1. (a) Single-server case (LIFO-PRE). (b) Pure multi-server case ($\phi(k) = k$). UBM: upper bound model, ORM: original model, □: completion, ×: acceptation, ○: rejection.

(mean) residual service time needs to be replaced by a longer (mean) total service time, which may enlarge the saturation time later on. Of course, the above example could also be seen as a realization with exponential services. On the average however the mean residual time in that case is not smaller than the mean service time.

For a similar feature in tandem queues a numerical counterintuitive example could be established (cf. [7]). For the present system however a numerical counterexample has not been found within an accuracy of 10^{-7} by using Erlang-2 service distributions. The *question* as to whether or not the bounds are *insensitive* for the system studied in this paper with a last-in first-out or processor sharing *single-server* discipline thus remains *open*. From the above example, however, it is at least clear that a proof by sample path arguments, such as in [6], seems impossible.

On the other hand, as illustrated by Fig. 1(b), for pure multi-server disciplines, by which each accepted job is assigned a single-server, the above conflict seems to be avoided since accepted jobs never take over service from present jobs. In this case the departure times of the upper bound model always exceed those of the original model, which corresponds to our intuition of an upper bound for the call congestion. The conjecture in Remark 2.3 seems therefore in order.

2.3. Remark (*Nonexponential pure multi-server case*). For the pure multi-server case (i.e., an accepted job is always assigned a server), it is conjectured that the bounds remain valid also for nonexponential services with mean μ^{-1}. The bounds would thus be insensitive also to the services. The proof of this conjecture can be expected with the same technique as used in Section 3, but will involve much more complex technicalities similarly to [4].

2.4. Remark (*Finite exponential tandem queues with breakdowns*). The bounds of the present paper can be combined with those in [5,7], so as to obtain simple bounds for 2-stage exponential tandem queues in which each station has a finite capacity constraint and is subject to a breakdown independently of the other station.

3. Proof of the bounds

In this section it will be shown that expression (2) is indeed an upper bound for the call congestion of the original model described in Section 2. The proof will be restricted to phase-type 'up' and 'down' periods. By standard arguments of weak convergence on so-called D-spaces, however (cf. [11]), the proof can hereby be concluded also for generally distributed up and down periods. The proof for the lower bound, moreover, can be given along the same lines.

Throughout, let a subscript U indicate the upper bound model, while a subscript (U) is used in an expression which should be read both with and without subscript U. Then, with B denoting the call congestion, it is to be proven that

$$B \leqslant B_{\mathrm{U}}, \tag{3}$$

under the assumption that for some ν_0, Q_0, p_0^k, $k = 1, \ldots, Q_0$, and ν_1, Q_1, p_1^k, $k = 1, \ldots, Q_1$, the distribution functions F_0 and F_1 are specified by

$$F_0 = \sum_{k=1}^{Q_0} p_0^k E_{\nu_0}^k \quad \text{with } [\gamma_0]^{-1} = \sum_{k=1}^{Q_0} (k/\nu_0) p_0^k,$$

$$F_1 = \sum_{k=1}^{Q_1} p_1^k E_{\nu_1}^k \quad \text{with } [\gamma_1]^{-1} = \sum_{k=1}^{Q_1} (k/\nu_1) p_1^k,$$

where E_ν^k denotes an Erlang-distribution with k exponential phases with parameter ν and where the values p_l^k denote probabilities with $\sum_{k=1}^{Q_l} p_l^k = 1$, $l = 0, 1$. To this end, let state (n, θ, l) denote that $n \geq 0$ jobs are present and that the facility is 'up' when $\theta = 1$ and 'down' when $\theta = 0$, with l residual exponential phases

with parameter ν_1 and ν_0, respectively. The corresponding queueing processes of the original and upper bound model then constitute continuous-time Markow chains. In order to deal with these chains in a recursive manner, we artificially introduce associated discrete-time Markov chains at epochs $\{nh,\ n = 0, 1, 2, \dots\}$, where h can be any fixed number such that $h \leqslant [\lambda + \mu + \gamma_0 + \gamma_1]^{-1}$.

First, let

$$\begin{aligned} &A(\theta) = 1 \quad \text{for } \theta = 0, 1, \\ &A_U(\theta) = 1 \quad \text{for } \theta = 1 \text{ but } = 0 \text{ for } \theta = 0. \end{aligned} \tag{4}$$

The one-step transition probabilities $p_{(U)}[(n, \theta, l) \to (\bar{n}, \bar{\theta}, \bar{l})]$ for a transition of these chains from a state (n, θ, l) into $(\bar{n}, \bar{\theta}, \bar{l})$ in a single-step are now defined by

$$\begin{aligned} &\lambda h A_{(U)}(\theta) && \text{for } (n, \theta, l) \to (n+1, \theta, l) && (n \leqslant N-1), \\ &\mu h \phi(n) && \text{for } (n, 1, l) \to (n-1, 1, l) && (n \geqslant 1), \\ &\nu_0 h && \text{for } (n, 0, l) \to (n, 0, l-1) && (l \geqslant 2), \\ &\nu_0 h p_1^k && \text{for } (n, 0, 1) \to (n, 1, k) && (k = 1, \dots, Q_1), \\ &\nu_1 h && \text{for } (n, 1, l) \to (n, 1, l-1) && (l \geqslant 2), \\ &\nu_1 h p_0^k && \text{for } (n, 1, 1) \to (n, 0, k) && (k = 1, \dots, Q_0), \\ &\left[1 - \lambda h 1_{\{n \leqslant N-1\}} A_{(U)}(\theta) - \mu h \phi(n) - \nu_\theta h\right] && \text{for } (n, \theta, l) \to (n, \theta, l), \end{aligned} \tag{5}$$

with $A(\theta)$ and $A_U(\theta)$ substituted for the original and upper bound model, respectively. These artificial chains have the advantage over the original continuous-time Markov chains, in that the one-step period is equidistant while no order term in h is involved. Let $G^D_{(U)}$ denote the discrete-time generator matrix of these chains as defined by

$$G^D_{(U)} = \left[P_{(U)} - I\right] h^{-1},$$

where $P_{(U)}$ is the one-step transition matrix and I the identity matrix. Then one directly concludes from (5) that

$$G^D_{(U)} = G^C_{(U)}$$

where $G^C_{(U)}$ denotes the generator or matrix of jumprates (also known as infinitesimal operator or differential matrix) for the continuous-time Markov chains corresponding to the queueing processes.

Now note that the stationary probability vector π (row-vector) of a finite irreducible Markov chain with generator G in both discrete and continuous time is uniquely determined, up to normalization, by (cf. [22, pp. 145, 247])

$$\pi G = 0.$$

As a result, for any fixed $h \leqslant [\lambda + \mu + \gamma_0 + \gamma_1]^{-1}$, the stationary distribution of the discrete-time chain defined by (5) is equal to that of the corresponding (original or upper bound) continuous-time Markov chain of the associated queueing process. The analysis of the call congestion can thus be restricted to the more convenient discrete-time Markov chains defined by (5).

Let T respectively T_U denote the one-step expectation operator of this chain for the original and upper bound model. That is, for any function f and with $p_{(U)}[(n, \theta, l) \to (\bar{n}, \bar{\theta}, \bar{l})]$ the one-step transition probabilities defined by (5) we have

$$T_{(U)} f(n, \theta, l) = \sum_{(\bar{n}, \bar{\theta}, \bar{l})} p_{(U)}\left[(n, \theta, l) \to (\bar{n}, \bar{\theta}, \bar{l})\right] f(\bar{n}, \bar{\theta}, \bar{l}).$$

Throughout, let $1_{\{A\}}$ denote the indicator of the event A, i.e. $1_{\{A\}} = 1$ if A is satisfied and 0 else. Then, from the definition of the one-step transition probabilities for the original and upper bound model by (5), where (4) has to be taken into account, it readily follows that

$$(T - T_U) f(n, \theta, l) = \lambda h 1_{\{\theta = 0\}} 1_{\{n < N\}} \left[f(n+1, \theta, l) - f(n, \theta, l)\right], \tag{6}$$

for any (n, θ, l) and function f. This difference in the one-step expectation operators is the key to the proof. To see this, let $\pi_{(U)}$ denote the stationary distribution. Now note that the system throughput in equilibrium is equal to both the mean number of accepted jobs and the mean number of departures per unit of time. Hence,

$$\lambda(1 - B_{(U)}) = \sum_{(n, \theta, L)} \pi_{(U)}(n, \theta, l) 1_{\{\theta=1\}} \phi(n) \mu.$$

Inequality (3) can thus be verified by proving

$$\sum_{(n, \theta, l)} \pi(n, \theta, l) 1_{\{\theta=1\}} \phi(n) \mu \geqslant \sum_{(n, \theta, l)} \pi_U(n, \theta, l) 1_{\{\theta=1\}} \phi(n) \mu. \tag{7}$$

To this end, let

$$g(n, \theta, l) = 1_{\{\theta=1\}} \phi(n),$$

and introduce the function $V^t_{(U)}$, $t = 0, 1, 2, \ldots,$ by

$$V^t_{(U)}(n, \theta, l) = \sum_{k=0}^{t-1} T^k_{(U)} g(n, \theta, l),$$

where $T^k_{(U)}$ denotes the kth power of $T_{(U)}$ for $k \geqslant 1$ while $T^0_{(U)}$ is the identity function. The function $V^t_{(U)}(n,\theta, l)$ thus represents the total expected reward over t periods with one-step reward function g when starting in state (n, θ, l).

Then, by virtue of the fact that $T^{k+1}_{(U)} = T_{(U)} T^k_{(U)}$, as due to the Markov property, it readily follows that

$$V^{k+1}_{(U)}(n, \theta, l) = g(n, \theta, l) + T_{(U)} V^k_{(U)}(n, \theta, l), \tag{8}$$

while the irreducibility and finiteness of the underlying Markov chains and the definition of g imply that, independently of the initial state $(\bar{n}, \bar{\theta}, \bar{l})$ within the irreducible set of states,

$$\sum_{(n, \theta, l)} \pi_{(U)}(n, \theta, l) 1_{\{\theta=1\}} \phi(n) = \lim_{t \to \infty} \frac{1}{t} V^t_{(U)}(\bar{n}, \bar{\theta}, \bar{l}). \tag{9}$$

Subtraction of the recursion relation (8) with $k + 1 = t$ for the upper bound model from that for the original model gives

$$\begin{aligned}(V^t - V^t_U)(n, \theta, l) &= (T - T_U) V^{t-1}(n, \theta, l) + T_U (V^{t-1} - V^{t-1}_U)(n, \theta, l) \\ &= \sum_{k=0}^{t-1} T^k_U (T - T_U) V^{t-k-1}(n, \theta, l) \quad (t \geqslant 0),\end{aligned}$$

where the last expression is obtained by iteration and the observation that $V^0(\cdot,\cdot,\cdot) = V^0_U(\cdot,\cdot,\cdot) = 0$. Now note that T_U is a monotone operator. By dividing both hand sides of the latter relation by t and letting t tend to infinity, we may thus conclude from (9) and the latter expression that (7) is guaranteed if for all (n, θ, l) and $t \geqslant 0$:

$$(T - T_U) V^t(n, \theta, l) \geqslant 0.$$

By now recalling (6), we have thus completed the proof by showing that for all (n, θ, l) and $t \geqslant 0$:

$$\left[V^t(n + 1, \theta, l) - V^t(n, \theta, l)\right] \geqslant 0. \tag{10}$$

The comparison of two systems is thus transformed into a monotonicity condition for one of them. This is the essence of the proof.

Inequality (10) will be proven by induction to t. Clearly, for $t = 0$ it is satisfied. Assume that it is satisfied for all (n, θ, l) and $t \leqslant m$. Then we need to verify (10) for $t = m + 1$. From the recursion relation

(8), the definiton of the expectation operator T, the one-step transition probabilities according to (3), and the one-step reward function g, we derive

$$
\begin{aligned}
&V^{m+1}(n+1,\theta,l)-V^{m+1}(n,\theta,l)\\
&=\Big\{1_{\{\theta=1\}}\phi(n+1)+\lambda h 1_{\{n+1<N\}}V^m(n+2,\theta,l)+\mu h 1_{\{\theta=1\}}\phi(n+1)V^m(n,1,l)\\
&\quad+\nu_0 h 1_{\{\theta=0\}}1_{\{l\geqslant 2\}}V^m(n+1,0,l-1)+\nu_0 h 1_{\{\theta=0\}}1_{\{l=1\}}\sum_{k=1}^{Q_1}p_1^k V^m(n+1,1,k)\\
&\quad+\nu_1 h 1_{\{\theta=1\}}1_{\{l\geqslant 2\}}V^m(n+1,1,l-1)+\nu_1 h 1_{\{\theta=1\}}1_{\{l=1\}}\sum_{k=1}^{Q_1}p_0^k V^m(n+1,0,k)\\
&\quad+\left[1-\lambda h 1_{\{n+1<N\}}-\mu h 1_{\{\theta=1\}}\phi(n+1)-\nu_0 h 1_{\{\theta=0\}}-\nu_1 h 1_{\{\theta=1\}}\right]\\
&\qquad\times V^m(n+1,\theta,l)\Big\}\\
&\quad-\Big\{1_{\{\theta=1\}}\phi(n)+\lambda h 1_{\{n<N\}}V^m(n+1,\theta,l)\\
&\qquad+\mu h 1_{\{\theta=1\}}1_{\{n>0\}}V^m(n-1,\theta,l)+\nu_0 h 1_{\{\theta=0\}}1_{\{l\geqslant 2\}}V^m(n,0,l-1)\\
&\qquad+\nu_0 h 1_{\{\theta=0\}}1_{\{l=1\}}\sum_{k=1}^{Q_1}p_1^k V^m(n,1,k)+\nu_1 h 1_{\{\theta=1\}}1_{\{l\geqslant 2\}}V^m(n,1,l-1)\\
&\qquad+\nu_1 h 1_{\{\theta=1\}}1_{\{l=1\}}\sum_{k=1}^{Q_0}p_0^k V^m(n,0,k)\\
&\qquad+\left[1-\lambda h 1_{\{n<N\}}-\mu h 1_{\{\theta=1\}}\phi(n)-\nu_0 h 1_{\{\theta=0\}}-\nu_1 h 1_{\{\theta=1\}}\right]V^m(n,\theta,l)\Big\}. \qquad (11)
\end{aligned}
$$

Consider the first and second expression between $\{\cdot\,\cdot\}$ of the right-hand side. Rewrite the second term of the second expression as

$$\lambda h\left[1_{\{n+1<N\}}+1_{\{n+1=N\}}\right]V^m(n+1,\theta,l),$$

the third term of the first expression as

$$\mu h 1_{\{\theta=1\}}\{\phi(n)+[\phi(n+1)-\phi(n)]\}V^m(n,1,l),$$

the last term of the first expression as

$$
\begin{aligned}
&\left[1-\lambda h 1_{\{n<N\}}-\mu h 1_{\{\theta=1\}}\phi(n+1)-\nu_0 h 1_{\{\theta=0\}}-\nu_1 h 1_{\{\theta=1\}}\right]V^m(n+1,\theta,l)\\
&\quad+\lambda h 1_{\{n+1=N\}}V^m(n+1,\theta,l),
\end{aligned}
$$

and finally the last term of the second expression as

$$
\begin{aligned}
&\left[1-\lambda h 1_{\{n<N\}}-\mu h 1_{\{\theta=1\}}\phi(n+1)-\nu_0 h 1_{\{\theta=0\}}-\nu_1 h 1_{\{\theta=1\}}\right]V^m(n,\theta,l)\\
&\quad+\mu h 1_{\{\theta=1\}}[\phi(n+1)-\phi(n)]V^m(n,1,l).
\end{aligned}
$$

Now, by collecting corresponding terms from the first and second expression we can transform the

right-hand side of (11) into:

$$
\begin{aligned}
&1_{\{\theta=1\}}[\phi(n+1)-\phi(n)]\\
&+\lambda h 1_{\{n+1<N\}}[V^m(n+2,\theta,l)-V^m(n+1,\theta,l)]\\
&+\lambda h 1_{\{n+1=N\}}[V^m(n+1,\theta,l)-V^m(n+1,\theta,l)]\\
&+\mu h 1_{\{\theta=1\}}1_{\{n\geq 1\}}\phi(n)[V^m(n,1,l)-V^m(n-1,1,l)]\\
&+\mu h 1_{\{\theta=1\}}[\phi(n+1)-\phi(n)][V^m(n,1,l)-V^m(n,1,l)]\\
&+\nu_0 h 1_{\{\theta=0\}}1_{\{l\geq 2\}}[V^m(n+1,0,l-1)-V^m(n,0,l-1)]\\
&+\nu_1 h 1_{\{\theta=1\}}1_{\{l\geq 2\}}[V^m(n+1,1,l-1)-V^m(n,1,l-1)]\\
&+\nu_0 h 1_{\{\theta=0\}}1_{\{l=1\}}\sum_{k=1}^{Q_1}p_1^k[V^m(n+1,1,k)-V^m(n,1,k)]\\
&+\nu_1 h 1_{\{\theta=1\}}1_{\{l=1\}}\sum_{k=1}^{Q_0}p_0^k[V^m(1,0,k)-V^m(n,0,k)]\\
&+\left[1-\lambda h 1_{\{n<N\}}-\mu h 1_{\{\theta=0\}}\phi(n+1)-\nu_0 h 1_{\{\theta=0\}}-\nu_1 h 1_{\{\theta=1\}}\right]\\
&\quad\times[V^m(1,\theta,l)-V^m(n,\theta,l)].
\end{aligned}
$$

Now note that the first term is nonnegative since $\phi(\cdot)$ is nondecreasing and observe that both the third and fifth term are equal to 0. By applying the induction hypotheses (10) for $t=m$ to the other terms, we have hereby verified (10) for $t=m+1$. Relation (10) is thus satisfied. This completes the proof of (3).

4. Evaluation

A bounding methodology for non-product-form queueing systems has been applied to finite exponential service facilities with breakdowns. This methodology is based upon modifications which satisfy the so-called notion of job-local-balance so as to guarantee a product-form expression. Computationally attractive and intuitively appealing bounds have so been obtained for the call congestion (and throughput). These bounds moreover depend on the up and down time distributions only through their means (insensitivity property). In particular, for the pure multi-server case they are also claimed to be insensitive to the service distributions. The bounds appear to be quite reasonable fast indicators and can thus be useful for quick engineering and performance evaluation purposes. The formal proof of these bounds is of interest in itself. Extensions to other breakdown systems such as finite tandem queues seem possible. The validity of the bounds for general services leaves open interesting questions.

Acknowledgment

The comments of the referees which helped to clarify the paper are greatly appreciated.

References

[1] H. Bruneel, On the behavior of buffers with random server interruptions, *Performance Evaluation* **3** (1983) 165–175.

[2] K.M. Chandy, J.H. Howard and D.F. Towsley, Product form and local balance in queueing networks, *J. ACM* **24** (1977) 250–263.

[3] N.M. van Dijk, Simple and insensitive bounds for a grading and an overflow model, *Oper. Res. Lett.* **6** (1986) 73–76.

[4] N.M. van Dijk, A formal proof for the insensitivity of simple bounds for finite multi-server non-exponential tandem queues, *Adv. Appl. Probab.* (1988) (to appear).

[5] N.M. van Dijk and B.F. Lamond, Bounds for the call congestion of finite single-server exponential tandem queues, *Oper. Res.* (1987) (to appear).

[6] N.M. van Dijk, P. Tsoucas and J. Walrand, Simple bounds and monotonicity of the call congestion of finite multi-server delay systems, *Probability in the Engineering and Informational Sciences* (1987) (to appear).

[7] N.M. van Dijk and J. van der Wal, *Simple Bounds for Finite Tandem Queues*, Res. Rept., Twente University of Technology, Enschede, The Netherlands, 1986.

[8] B.T. Doshi, Queueing systems with vacations—a survey, *Queueing Systems* **1** (1986) 29–66.

[9] A. Federgruen and L. Green, Queueing systems with service interruptions, *Oper. Res.* **34** (1986) 752–768.

[10] M.J. Fisher, Approximation to queueing systems with interruptions, *Management Sci.* **24** (3) (1977) 338–344.

[11] A. Hordijk and R. Schassberger, Weak convergence of generalized semi-Markov processes, *Stochastic Process. & Appl.* **2** (1982) 271–291.

[12] A. Hordijk and N.M. van Dijk, Networks of queues; Part I: Job-local-balance and the adjoint process, Part II: General routing and service characteristics, in: *Lecture Notes in Control and Information Sciences, Vol. 60* (Springer, Berlin, 1983) 158–205.

[13] A. Hordijk and N.M. van Dijk, Adjoint process, job-local-balance and insensitivity for stochastic networks, *Bull. 44th Session Internat. Stat. Inst.* **50** (1983) 776–788.

[14] J.S. Kaufman, Approximation methods for networks of queues with priorities, *Performance Evaluation* **4** (1984) 183–198.

[15] F.P. Kelly, *Reversibility and Stochastic Networks* (Wiley, New York, 1979).

[16] T.T. Lee, $M/G/1/N$ queue with vacation time and exhaustive service discipline, *Oper. Res.* **32** (1984) 774.

[17] Y. Levy and U. Yechiali, $M/M/S$ queue with servers vacations, *INFO* **14** (1976) 153.

[18] N.K. Jaiswal, *Priority Queues* (Academic Press, New York, 1968).

[19] I.L. Mitrany and B. Avi-Itzhak, A many server queue with server interruptions, *Oper. Res.* **16** (1967) 628–638.

[20] M. Neuts and D. Lucantoni, A Markovian queue with N servers subject to breakdowns and repairs, *Management Sci.* **25** (1979) 849.

[21] P. Nain, Queueing systems with service interruptions: An approximation model, *Performance Evaluation* **3** (1983) 123–129.

[22] S.M. Ross, *Introduction to Probability Models* (Academic Press, New York, 3rd ed., 1985).

[23] R. Schassberger, The insensitivity of stationary probabilities in networks of queues, *Adv. Appl. Probab.* **10** (1978) 906–912.

[24] D. Stoyan (D.J. Daley, ed.), *Comparison Methods for Queues and Other Stochastic Models* (Wiley, New York, 1983).

[25] R. Suri, A concept of monotonicity and its characterization for closed queueing networks, *Oper. Res.* **33** (1985) 606–624.

[26] W. Whitt, Comparing counting processes and queues, *Adv. Appl. Probab.* **13** (1981) 207–220.

[27] R.W. Wolff, Poisson arrivals see time averages, *Oper. Res.* **30** (1982) 223–231.

[28] J. Zahorjan, K.C. Sevcik, D.L. Eager and B. Galler, Balanced job bound analysis of queueing networks, *Comm. ACM* **25** (1982) 134–141.