Probability in the Engineering and Informational Sciences, **23**, 2009, 385–408. Printed in the U.S.A.
doi:10.1017/S0269964809000217

A STATE-DEPENDENT POLLING MODEL WITH *k*-LIMITED SERVICE

E. M. M. WINANDS
Department of Mathematics
VU University
1081 HV Amsterdam
The Netherlands
E-mail: emm.winands@few.vu.nl

I. J. B. F. ADAN
Department of Mathematics and Computer Science
Technische Universiteit Eindhoven
5600 MB Eindhoven, The Netherlands
E-mail: i.j.b.f.adan@tue.nl

G. J. VAN HOUTUM
Department of Industrial Engineering and Innovation Sciences
Technische Universiteit Eindhoven
5600 MB Eindhoven, The Netherlands
E-mail: g.j.v.houtum@tue.nl

D. G. DOWN
Department of Computing and Software
McMaster University
Hamilton, Ontario L8S 4L7, Canada
E-mail: downd@mcmaster.ca

We consider a two-queue model with state-dependent setups, in which a single server alternately serves the two queues. The high-priority queue is served exhaustively, whereas the low-priority queue is served according to the *k*-limited strategy. A setup at a queue is incurred only if there are customers waiting at the polled queue. We obtain the transforms of the queue length and sojourn time distributions under the assumption of Poisson arrivals, generally distributed service times, and generally distributed setup times. The interest for this model is fueled by an application in the field of logistics. It is shown how the results of this analysis can be applied in the evaluation of a stochastic two-item single-capacity production system. From these

© 2009 Cambridge University Press 0269-9648/09 $25.00
https://doi.org/10.1017/S0269964809000217 Published online by Cambridge University Press

results we can conclude that significant cost reductions are possible by bounding the production runs of the low-priority item, which indicates the potential of the k-limited service discipline as priority rule in production environments.

1. INTRODUCTION

The present article considers a two-queue *state-dependent* polling model, in which a setup is incurred for a queue only when it is nonempty. In this model, the single server serves the high-priority queue *exhaustively* and serves the low-priority queue according to the *k-limited* service strategy. Exhaustive means that the queue must be empty before the server moves on, whereas under the k-limited strategy, the server continues working until either a predefined number of k customers is served or until the queue becomes empty, whichever occurs first.

The motivation for the present study is twofold. The first one is application oriented. Although in the past the k-limited strategy proved its merit in communication systems (see, e.g., [5,7]), the specific application that attracted our attention is in the field of logistics. In particular, in many stochastic multiproduct single-capacity make-to-stock production systems considerable setup times are incurred [i.e., the so-called *stochastic economic lot scheduling problem* (SELSP)]. For surveys on the SELSP, see Sox, Jackson, Bowman, and Muckstadt [28] and Winands, Adan and van Houtum [35]. The SELSP is a common problem in practice—for example, in glass and paper production, injection molding, metal stamping and semicontinuous chemical processes but also in bulk production of consumer products such as detergents and beers.

The presence of these setup times in combination with the stochastic environment are the key complicating factors of the SELSP. On the one hand, one aims for short cycle lengths, and thus frequent production opportunities for the various products, in order to be able to react to the stochasticity in the system. On the other hand, short cycle lengths will increase the setup frequency, which has a negative influence on the amount of capacity available for production. Consequently, this effect will hinder the timely fulfillment of demand. An important issue is the fact that in the SELSP—and in various other applications of polling systems—the objective function typically depends not only on the *mean* queue lengths but also on the *complete* marginal queue length distributions. The main interest of the present article is, therefore, in the marginal queue length distributions in the aforementioned queuing model.

In the context of the SELSP, the *exhaustive* service discipline with *state-independent* setups [i.e., the machine sets up irrespective of whether the polled product has a positive shortfall (i.e., has outstanding production orders)] has been studied by Federguen and Katalan [14,15]. Their strategy has two major drawbacks. First, in the exhaustive policy, one single product, for which a high demand arrives in a certain period of time, can occupy the machine for quite a while. The impacts of this phenomenon on the other products are stock outs, highly variable cycle lengths, and high costs. The k-limited policy circumvents this drawback and offers the possibility to the manager to control both the setup frequencies and the production runs (and,

https://doi.org/10.1017/S0269964809000217 Published online by Cambridge University Press

thus, the cycle lengths). Second, Federguen and Katalan [14,15] assumed that the machine incurs a setup for a certain product even when there is no shortfall for this product, which is, of course, suboptimal, as argued by Sox et al. [28]. This observation makes the practical relevance of the inclusion of state-dependent setups in the studied queuing model evident.

From the application point of view, the present article offers, albeit in an idealized mathematical setting, a preliminary exploration of an important managerial issue: What is the gain in performance of bounding production runs—at the expense of higher setup frequencies—in multi-item production-inventory systems. In order to answer this question analytically and not to be diverted by other effects, the present article focuses on a basic occurrence of the SELSP. The two-queue model is, however, also of interest in its own right: Production applications, in which only two items have to be produced on one single production facility, are certainly not inconceivable. In such settings, it is quite natural to apply a limited-service mechanism to provide different service to the different items in order to improve system performance.

A second motivation for the present work is the fact that we have a theoretical interest in the proposed queuing model; that is, the focus of attention is on the *exact* evaluation of this model in order to get better fundamental insights. However, so far, hardly any exact results for polling systems with the k-limited service policy have been obtained; even mean performance measures are, in general, not known. This can be explained by the fact that the k-limited strategy does not satisfy a well-known branching property for polling systems independently discovered by Fuhrmann [17] and Resing [26]. For general k, an exact evaluation for the queue length distribution is, therefore, only available for very few special two-queue cases (see Lee [22] and Ozawa [24,25]). In these models, one (low-priority) queue is served by the k-limited service strategy, whereas the other (high-priority) queue is served by the exhaustive policy. Furthermore, all of these articles make the restrictive assumption of *zero setup times*. In many applications such as the SELSP, however, the setup times might be substantial and the presence of these setup times might be crucial for the operation of the system.

A second feature of the proposed model seriously complicating the analysis is the presence of the state-dependent setups (the server sets up for a queue only when it is nonempty). Nearly all of the existing literature on polling systems makes the assumption of state-independent setups. Notable exceptions are the recent studies of Altman, Blanc, Khamisy and Yechiali [3], Günalay and Gupta [18], Gupta and Srinivasan [19], and Singh and Srinivasan [27], where exhaustive-type and gated-type service disciplines are explored in combination with state-dependent setups. The choice of modeling state-independent setups is generally not motivated by an application, but by the tractability of the resulting analysis. The present work is the first study combining the limited service discipline and state-dependent setups. Finally, we refer to Takagi [30–32] and to Levy and Sidi [23] for extensive general surveys on polling models and their applications.

We think that it is appropriate that we also bring the work of Borst, Boxma, and Levy [5] to the reader's attention, in which approximate optimal values of the service limits with respect to a weighted sum of mean waiting times are obtained for general

https://doi.org/10.1017/S0269964809000217 Published online by Cambridge University Press

k-limited polling systems with state-independent setup times. Of particular interest to the present article is the fact that they derived a (partially conjectured) rule stating that for optimal operation of these systems, the queues with the highest priority (i.e., the queue with the highest value of the ratio between cost factor and mean service time) must have their service limit set at infinity. In a two-queue system, this would result into the priority rule of the present article, providing additional evidence for the significance of the present study.

Before leaving the literature review, we should stress that one can come up with other policies that, theoretically speaking, are capable of dealing with the sharing of production capacity in the context of the SELSP; for example, one could think of a longest-queue policy (see, e.g., Cohen [9] or Flatto [16]) or a threshold policy (see, e.g., Boxma and Down [6]). These policies decide, however, on switching based on *global* information (i.e., the queue length of all queues and the state of the server), which is a serious drawback in production environments. The k-limited strategy is, however, easy implementable and can effortlessly be monitored with minimal informational requirements; that is, only *local* information of the queue currently served is needed. Furthermore, the exhaustive and k-limited policies are natural generalizations of the rotation policies advocated for deterministic cyclic problems (see also Federgruen and Katalan [14,15]), which adds to the willingness of production managers to actually implement them on the work floor. For a detailed overview on policies for the SELSP, where the policies are actually classified according to the amount of information used, we again refer to Winands et al. [35].

In sum, the main contribution of the present article is twofold. First, the model in [22] is generalized by including state-dependent setups. In particular, we obtain the transforms of the queue length and sojourn time distributions under the assumption of Poisson arrivals, generally distributed service times, and generally distributed setup times. Second, we demonstrate how the results of the analysis can be applied in the evaluation of a stochastic two-item single-capacity production system. We observe significant cost reductions by application of the k-limited policy, compared to the standard exhaustive policies, in such settings, indicating the potential of the k-limited service discipline as a priority rule in production environments.

The rest of the present article is organized as follows. In Section 2, we present the model description including the stability conditions and the balance equations. Section 3 derives the probability generating functions (p.g.f.s) of the joint and marginal queue length distributions both at service completion epochs and at arbitrary instants. The penultimate section is devoted to the application of the analysis in the field of the SELSP. Some concluding remarks are presented in Section 5.

2. MODEL DESCRIPTION

We start this section with the description of the notation and assumptions used throughout the present article. Then Sections 2.2 and 2.3 present the stability

https://doi.org/10.1017/S0269964809000217 Published online by Cambridge University Press

condition and the state description together with the corresponding balance equations, respectively.

2.1. Notation and Assumptions

We study a two-queue model, in which a single server alternately serves the two queues. The high-priority queue 1 is served *exhaustively*, whereas the low-priority queue 2 is served according to the *k-limited* strategy (see Fig. 1). The combination of these service policies obviously creates a preferential treatment of type-1 customers. Customers arrive at queue i according to a Poisson process with rate $\lambda_i > 0$. The service times at queue i are independent and identically distributed (i.i.d.) random variables with mean $\beta_i > 0$ and Laplace–Stieltjes transform (LST) $B_i(\cdot)$. When the server starts service at queue i, a setup time is incurred with mean $\tau_i \geq 0$ and LST $T_i(\cdot)$. These setup times are i.i.d. random variables. The occupation rate ρ_i at queue i is defined by $\rho_i = \lambda_i \beta_i$ and the total occupation rate ρ is given by $\rho = \rho_1 + \rho_2$. In the next subsection, we give a stability condition for the system in terms of the total occupation rate and the service parameter $k \in \{1, 2, \ldots\}$.

The setup times are assumed to be *state dependent* (i.e., the server incurs a setup for a queue only when it is nonempty). When both queues are empty, the server stops working. He starts again upon arrival of a new customer and, then, he has to set up irrespective of the type of the last customer served before the idle time. Moreover, if the server has served k customers of the low-priority queue and the high-priority queue is empty, the server starts a new sequence up to k customers of this low-priority class after a new setup time of the low-priority queue. However, it is important to note that, in this case, no setup is incurred for the high-priority queue (which is the standard assumption in polling systems with state-independent setup times).

Finally, we define $S_i(z_1, z_2)$ and $R_i(z_1, z_2)$ as the p.g.f.s of the number of type-1 and type-2 arrivals during a service time and a setup time at queue i, respectively; that is,

$$S_i(z_1, z_2) = B_i(\lambda_1(1 - z_1) + \lambda_2(1 - z_2)), \qquad i = 1, 2, \tag{1}$$

$$R_i(z_1, z_2) = T_i(\lambda_1(1 - z_1) + \lambda_2(1 - z_2)), \qquad i = 1, 2. \tag{2}$$

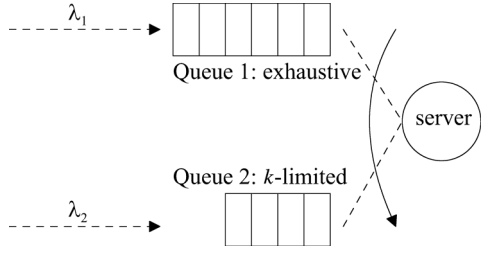


FIGURE 1. The model.

https://doi.org/10.1017/S0269964809000217 Published online by Cambridge University Press

The quantities

$$r_1 = \frac{\lambda_1}{\lambda_1 + \lambda_2} \quad \text{and} \quad r_2 = \frac{\lambda_2}{\lambda_1 + \lambda_2} \tag{3}$$

denote the probabilities that the server switches to queue 1 and 2 after an idle period, respectively.

Remark 2.1: It is important to note that the analysis of the present article also fully holds in the case of state-independent setups, *mutatis mutandis*, which are observed occasionally in telecommunication applications. However, we envision production systems as the main application for the present article (see also Section 4) and, in such applications, state-independent setups are definitely not a realistic assumption. In particular, the specific setup rule assumed here is quite common in, for example, the chemical industry. After the cleaning of the machine and delivery of raw materials, the machine is able to produce a batch with a certain maximum size. At the moment that this maximum batch has been produced, the machine has to be cleaned again and raw materials have to be replenished.

2.2. Stability Conditions

In this subsection, we derive the stability conditions for the two queues by deploying (heuristic) arguments similar to those used by Ibe and Cheng [20], who studied the stability conditions for a variety of polling systems without state-dependent setups.

Since the k-limited service policy bounds the time the server spends at queue 2, queue 2 will not affect the stability of queue 1. Hence, the stability condition of the latter simply reads

$$\rho_1 < 1. \tag{4}$$

If this condition is not satisfied, queue 2 is unstable as well: The server will serve queue 1 for an infinite amount of time and, in the meantime, the queue length of queue 2 grows to infinity.

For queue 2, we define a cycle time as the time between the start of two successive setups of the server at this queue. In a long period of time T, the expected number of customers arriving at queue 2 equals $\lambda_2 T$. Each customer brings, on average, β_2 units of work into the system. We approximately need $\lambda_2 T/k$ cycles to serve these customers in heavy traffic, which is the regime of interest for the derivation of the stability condition. In each cycle, the server always has to set up for queue 2 and has to set up for queue 1 in case queue 1 is nonempty at the instant that the server has served k customers at queue 2. The latter occurs with the probability r that the number of type-1 arrivals during a setup time plus k successive service times at queue 2 is not equal to zero; that is,

$$r = 1 - R_2(0,1)(S_2(0,1))^k. \tag{5}$$

Hence, the mean total setup time in each cycle equals $\tau_2 + r\tau_1$. Moreover, a fraction ρ_1 of time is used to serve customers at queue 1. Since queue 2 is stable if

https://doi.org/10.1017/S0269964809000217 Published online by Cambridge University Press

the server manages to serve all customers arrived in T in an amount of time less than T, we have

$$\rho_1 T + \rho_2 T + \frac{\lambda_2 T}{k}(\tau_2 + r\tau_1) < T, \tag{6}$$

which gives us, after division by T,

$$\hat{\rho} = \rho_1 + \rho_2 + \frac{\lambda_2}{k}(\tau_2 + r\tau_1) < 1. \tag{7}$$

When queue 2 satisfies (7), ρ_1 is also smaller than 1 and, thus, in this case, both queues are stable. These findings are summarized in the following theorem.

THEOREM 2.2: *The two-queue system is stable if and only if*

$$\hat{\rho} = \rho_1 + \rho_2 + \frac{\lambda_2}{k}(\tau_2 + r\tau_1) < 1. \tag{8}$$

Throughout the present article, we assume that (8) in Theorem 2.2 is satisfied, as we restrict ourselves to steady-state behavior. The above heuristic derivation of the stability condition in Theorem 2.2 can be proved rigorously by using fluid limit methodology described by Dai [10] and applied to polling models by Down [13] (see the Appendix). Note that as this is a two-queue model, alternatively we can adapt the results of Dai and Meyn [12].

2.3. State Description and Balance Equations

We study the system at embedded epochs of service completions of customers. The state of the system $\mathbf{Q}(n)$ just after the nth departure from the system can be described by the following three variables:

1. $Q_1(n)$: the number of customers in queue 1
2. $Q_2(n)$: the number of customers in queue 2
3. $C(n)$: zero when the nth departure is a type-1 customer, whereas it equals the number of type-2 departures since the last setup when the nth departure is a type-2 customer

The associated stochastic process,

$$\mathbf{Q}(n) = \{\big(Q_1(n), Q_2(n), C(n)\big), n = 1, 2, \ldots\}, \tag{9}$$

is an aperiodic and irreducible three-dimensional Markov chain. Let

$$\pi(q_1, q_2, c) = \lim_{n\to\infty} P[\big(Q_1(n), Q_2(n), C(n)\big) = (q_1, q_2, c)] \tag{10}$$

https://doi.org/10.1017/S0269964809000217 Published online by Cambridge University Press

be the limiting probability and define the corresponding generating functions for this Markov chain as follows:

$$p_1(z_1,\ z_2) = \sum_{q_1=0}^{\infty}\sum_{q_2=0}^{\infty} \pi(q_1,q_2,0) z_1^{q_1} z_2^{q_2}, \tag{11}$$

$$p_{2,j}(z_1,z_2) = \sum_{q_1=0}^{\infty}\sum_{q_2=0}^{\infty} \pi(q_1,q_2,j) z_1^{q_1} z_2^{q_2}, \qquad j=1,2,\ldots,k. \tag{12}$$

Now, the following set of $k+1$ balance equations for equally many unknowns $p_1(z_1,z_2)$ and $p_{2,j}(z_1,z_2), j=1,2,\ldots,k$, holds:

$$p_1(z_1,z_2) = \frac{S_1(z_1,z_2)}{z_1}\Bigg\{p_1(z_1,z_2) - p_1(0,z_2) + \Bigg[c_0 r_1 z_1 + \sum_{j=1}^{k-1}[p_{2,j}(z_1,0) - p_{2,j}(0,0)] + p_{2,k}(z_1,z_2) - p_{2,k}(0,z_2)\Bigg] R_1(z_1,z_2)\Bigg\}, \tag{13}$$

$$p_{2,1}(z_1,z_2) = \frac{S_2(z_1,z_2)R_2(z_1,z_2)}{z_2}\alpha(z_2), \tag{14}$$

$$p_{2,j}(z_1,z_2) = \frac{S_2(z_1,z_2)}{z_2}\left\{p_{2,j-1}(z_1,z_2) - p_{2,j-1}(z_1,0)\right\}, \qquad j=1,2,\ldots,k, \tag{15}$$

with

$$c_0 = p_1(0,0) + \sum_{j=1}^{k} p_{2,j}(0,0), \tag{16}$$

the probability that the system is left idle at a departure epoch and

$$\alpha(z_2) = c_0 r_2 z_2 + p_1(0,z_2) - p_1(0,0) + p_{2,k}(0,z_2) - p_{2,k}(0,0). \tag{17}$$

These balance equations are formulated by considering all of the possible states at the previous departure epoch from which we can reach the current state. We explain (13), which describes the case that the current departure is a type-1 customer. First, the previous departure could be a type-1 departure that did not leave the system idle. This event corresponds to the term $p_1(z_1,z_2) - p_1(0,z_2)$. Second, the term $c_0 r_1 z_1$ represents the event that the previous departure left the system idle and that the first new arriving customer is of type 1. Third, the term $\sum_{j=1}^{k-1}[p_{2,j}(z_1,0) - p_{2,j}(0,0)]$ represents the event that the last departure was a type-2 customer that was not the kth in the sequence and that left queue 2, but not the complete system, idle. Finally, $p_{2,k}(z_1,z_2) - p_{2,k}(0,z_2)$ corresponds to the event that the last departure was a type-2 customer that was the kth in the sequence and that queue 1 was not empty. The explanations of (14) and (15) are similar.

https://doi.org/10.1017/S0269964809000217 Published online by Cambridge University Press

The function $\alpha(\cdot)$ is recognized as the generating function of the number of type-2 customers at moments that a setup for queue 2 is initiated. Since, at these specific points in time, Q_1 and C are both equal to zero, the state description can be reduced to one single dimension represented by the function $\alpha(\cdot)$. The analysis of the next section is oriented toward relating the unknown generating functions $p_1(\cdot)$ and $p_{2,j}(\cdot,\cdot)$ to this function $\alpha(\cdot)$.

3. EXACT ANALYSIS OF QUEUE LENGTHS

In Section 3.1, we present the derivation of the generating functions of the joint queue length distributions at service completion epochs, which is a generalization of the method used by Lee [22] for the model without setup times. In Section 3.2 these results are used to derive expressions for the p.g.f.s of the marginal queue size distributions at arbitrary instants.

3.1. Joint Queue Lengths at Service Completion Epochs

To derive the generating functions of the joint queue length distributions at service completion epochs, we successively substitute (15) into itself and, then, into (14), which yields, for $j = 1, 2, \ldots, k$,

$$p_{2,j}(z_1, z_2) = \frac{S_2^j(z_1, z_2)R_2(z_1, z_2)\alpha(z_2) - \sum_{l=1}^{j-1} z_2^{j-l} S_2^l(z_1, z_2) p_{2,j-l}(z_1, 0)}{z_2^j}. \quad \textbf{(18)}$$

Notice that (18) gives an expression of $p_{2,j}(\cdot,\cdot)$, $j = 1, 2, \ldots, k$, as a function of the unknown functions $\alpha(\cdot)$ and $p_{2,l}(\cdot, 0)$, $l = 1, 2, \ldots, j-1$.

Now, we turn our attention to $p_1(\cdot,\cdot)$. Substituting (18) for $j = k$ into (13) and using (16) and (17) gives us, after some straightforward manipulations,

$$\begin{aligned}
&(z_1 - S_1(z_1, z_2))\, p_1(z_1, z_2) \\
&\quad = S_1(z_1, z_2) \Bigg\{ c_0 r_1 z_1 R_1(z_1, z_2) + (R_1(z_1, z_2) - 1) p_1(0, z_2) \\
&\qquad + \left(\left(\frac{S_2(z_1, z_2)}{z_2} \right)^k R_2(z_1, z_2)\alpha(z_2) - \alpha(z_2) + c_0 r_2 z_2 \right. \\
&\qquad \left. + \sum_{j=1}^{k-1} \left[1 - \left(\frac{S_2(z_1, z_2)}{z_2} \right)^j \right] p_{2,k-j}(z_1, 0) - c_0 \right) R_1(z_1, z_2) \Bigg\}. \quad \textbf{(19)}
\end{aligned}$$

https://doi.org/10.1017/S0269964809000217 Published online by Cambridge University Press

We eliminate $p_1(0,z_2)$ from the above equation by rewriting (17) as

$$\begin{aligned} p_1(0,z_2) &= \alpha(z_2) - c_0 r_2 z_2 + p_1(0,0) - p_{2,k}(0,z_2) + p_{2,k}(0,0) \\ &= \alpha(z_2) + c_0(1 - r_2 z_2) - \left(\frac{S_2(0,z_2)}{z_2}\right)^k R_2(0,z_2)\alpha(z_2) \\ &\quad - \sum_{j=1}^{k-1}\left[1 - \left(\frac{S_2(0,z_2)}{z_2}\right)^j\right] p_{2,k-j}(0,0), \end{aligned} \tag{20}$$

which yields

$$\begin{aligned} (z_1 - S_1(z_1,z_2))\, p_1(z_1,z_2) = S_1(z_1,z_2) \Bigg\{ &\left(\frac{\beta(z_1,z_2)}{z_2^k} - 1\right)\alpha(z_2) \\ &+ R_1(z_1,z_2)\sum_{j=1}^{k-1}\frac{\delta_j(z_1,z_2)}{z_2^k} p_{2,k-j}(z_1,0) \\ &+ D(z_1,z_2) - (R_1(z_1,z_2) - 1)\sum_{j=1}^{k-1}\frac{\delta_j(0,z_2)}{z_2^k} p_{2,k-j}(0,0)\Bigg\}, \end{aligned} \tag{21}$$

where

$$D(z_1,z_2) = c_0\,[r_1 z_1 R_1(z_1,z_2) + r_2 z_2 - 1]\,, \tag{22}$$

$$\beta(z_1,z_2) = S_2^k(z_1,z_2)R_1(z_1,z_2)R_2(z_1,z_2) - S_2^k(0,z_2)(R_1(z_1,z_2) - 1)R_2(0,z_2), \tag{23}$$

$$\delta_j(z_1,z_2) = z_2^k - z_2^{k-j}S_2^j(z_1,z_2). \tag{24}$$

It is again important to notice that, via (21), $p_1(\cdot,\cdot)$ is also expressed as a function of the unknown functions $\alpha(\cdot)$ and $p_{2,j}(\cdot,0), j = 1,2,\ldots,k-1$.

It is well known that for each (fixed) $|z_2| \le 1$, the term $z_1 - S_1(z_1,z_2)$ in (21) has exactly one zero $z_1 = \xi(z_2)$ with $|z_2| \le 1$ if $\rho_1 < 1$. More specifically,

$$z_1 = \xi(z_2) = \gamma_1[\lambda_2(1 - z_2)], \tag{25}$$

where $\gamma_1(\cdot)$ is the LST of the busy period of a standard $M/G/1$ queue with arrival rate λ_1 and LST of the service time $B_1(\cdot)$ (see, e.g., Takács [29]). Thus, $\xi(\cdot)$ can be seen as the p.g.f. of the distribution of the number of type-2 arrivals during such an $M/G/1$ busy period.

Remark 3.1: It is interesting to note that the function $\beta(\xi(z),z)$ is the p.g.f. of the number of type-2 customers arriving in a cycle for queue 2 in which the maximum of k customers is served.

https://doi.org/10.1017/S0269964809000217 Published online by Cambridge University Press

By analyticity of $p_1(z_1, z_2)$, the right-hand side of (21) should vanish when $z_1 = \xi(z_2)$. Hence,

$$\alpha(z) = \Bigg(D(\xi(z), z)z^k + R_1(\xi(z), z) \sum_{j=1}^{k-1} \delta_j(\xi(z), z) p_{2,k-j}(\xi(z), 0) - (R_1(\xi(z), z) - 1) \sum_{j=1}^{k-1} \delta_j(0, z) p_{2,k-j}(0, 0) \Bigg) [z^k - \beta(\xi(z), z)]^{-1}, \quad \textbf{(26)}$$

and $\alpha(z)$ is formulated as a function of the unknown functions $p_{2,j}(\cdot, 0)$, $j = 1, 2, \ldots, k-1$.

To eliminate these unknown functions, we differentiate the numerator and denominator of (18) j times with respect to z_2 and, by l'Hospital's rule, we obtain the following recursion, for $j = 1, 2, \ldots, k-1$:

$$p_{2,j}(x, 0) = \sum_{l=1}^{j} \frac{c_l}{(j-l)!} \frac{d^{j-l}}{dy^{j-l}} \left[S_2^j(x, y) R_2(x, y) \right] \Big|_{y=0} - \sum_{l=1}^{j-1} \frac{p_{2,j-l}(x, 0)}{l!} \frac{d^l}{dy^l} \left[S_2^l(x, y) \right] \Big|_{y=0}, \quad \textbf{(27)}$$

where c_l, $l = 1, 2, \ldots, k-1$, represent the probabilities that l type-2 customers are present at the start of a setup for this queue; that is,

$$c_l = \frac{1}{l!} \frac{d^l}{dy^l} [\alpha(y)] \Big|_{y=0}, \qquad l = 1, 2, \ldots, k-1. \quad \textbf{(28)}$$

From this interpretation of c_l one can easily deduce a probabilistic interpretation of (27) as well; that is, the left-hand side and right-hand side clearly represent the p.g.f.s of the number of customers in queue 1 when the server leaves queue 2 due to the fact that the latter queue is empty after j customers are served, $j = 1, 2, \ldots, k-1$. Of course, we could have derived (27) directly by using this probabilistic interpretation and, hence, avoid the use of l'Hospital's rule.

By (27) we can write $p_{2,j}(\cdot, 0)$ as a function of the unknown probabilities c_j, $j = 0, 1, \ldots, k-1$. Moreover, with the help of (18), (21), and (26), the generating functions $p_{2,j}(\cdot, \cdot)$, $p_1(\cdot)$, and $\alpha(\cdot)$ can be expressed in terms of these constants as well. The problem of finding these generating functions is thus reduced to finding the unknown probabilities c_j, which can be computed as follows. Given that (8) holds, the following theorem states that the denominator of (26) has exactly k zeros on or within the unit circle.

THEOREM 3.2: *Under (8), it holds that $z^k = \beta(\xi(z), z)$ has k roots on or within the unit circle.*

https://doi.org/10.1017/S0269964809000217 Published online by Cambridge University Press

PROOF: The derivative of $\beta(\xi(z),z)$ at $z=1$ equals

$$\beta'(\xi(1),1) = \frac{\lambda_2}{1-\rho_1}(k\beta_2+\tau_1+\tau_2) - \frac{\lambda_2}{1-\rho_1}S_2^k(0,z_2)R_2(0,z_2)\tau_1$$
$$= \frac{k\rho_2+\lambda_2(r\tau_1+\tau_2)}{1-\rho_1}, \tag{29}$$

where q is defined by (5). Because we have assumed (8), we obtain $\beta'(\xi(1),1) < k$ and the result follows from Theorem 3.2 of Adan, van Leeuwaarden and Winands [2]. ■

Since $\alpha(z)$ is bounded in $|z| \leq 1$, the zeros in the numerator must be canceled by corresponding zeros in the denominator. One of the zeros equals one and leads to a trivial equation. However, the normalization condition provides an additional equation and, therefore, we have a set of k linear equations. By making the assumption that the k roots of $z^k = \beta(\xi(z),z)$ on or within the unit circle are all distinct (see Remark 3.2), this set of equations has a unique solution for $c_j, j = 0,1,\ldots,k-1$. This completes the determination of the generating functions of the queue length distributions at service completion epochs and, hence, in the remainder of the present article, we assume that these generating functions are known.

Remark 3.2: If one or more roots of $z^k = \beta(\xi(z),z)$ on or within the unit circle coincide, our reasoning needs to be slightly modified; that is, for $\alpha(z)$ to be bounded in $|z| \leq 1$, the numerator of (26) should still have the same zeros as the denominator of (26) *and* with the same multiplicity. Additional equations can, therefore, be obtained by requiring that the derivative(s) of the numerator should also vanish where the denominator has a zero of higher multiplicity.

3.2. Marginal Queue Lengths at Arbitrary Instants

From the results of the previous subsection, we can obtain expressions for the p.g.f. $q_i(\cdot)$ of the marginal queue size distributions of queue i at type-i departure epochs; that is,

$$q_1(z) = \frac{p_1(z,1)}{r_1}, \quad \text{and} \quad q_2(z) = \frac{\sum_{j=1}^{k} p_{2,j}(1,z)}{r_2}. \tag{30}$$

By using a standard level crossing argument, in combination with PASTA, it can be shown that the marginal queue length distribution of queue i at type-i departure epochs and at arbitrary instants in time are the same. Hence, the p.g.f.s for these marginal distributions are given by (30).

From (30) we can easily obtain the LST $W_i(\cdot)$ of the sojourn time distribution of a type-i customer. Since the number of type-i customers left behind by a tagged type-i customer equals the number of customers arrived during the sojourn time of

https://doi.org/10.1017/S0269964809000217 Published online by Cambridge University Press

this tagged customer, we have

$$W_i(z) = q_i \left(1 - \frac{z}{\lambda_i}\right), \qquad i = 1, 2, \tag{31}$$

which is known as the distributional form of Little's law (see, e.g., Keilson and Servi [21]).

4. APPLICATION

In this section we use the analysis of the present article to evaluate a stochastic two-item single-capacity production system. After the description of this application in Section 4.1, we present a numerical illustration in Section 4.2.

4.1. Stochastic Economic Lot Scheduling Problem

Consider a system with one single production capacity for two products, in which there is an infinite stock space for each product and raw material is always available. Demands for the two products arrive according to stationary and mutually independent unit Poisson processes with rate $\lambda_i > 0$. Demand that cannot be satisfied directly from stock is backlogged until the product becomes available after production. The individual products are produced make-to-stock in a cyclical order with stochastic production times with mean $\beta_i > 0$ and LST $B_i(\cdot)$. Stochastic setup times with mean $\tau_i \geq 0$ and LST $T_i(\cdot)$ occur *before* the start of the production of a product. For an arbitrary number of products, this setting is often referred to as the *stochastic economic lot scheduling problem* (see Winands et al. [35], for a recent survey).

For product 1, a standard *base-stock* policy is implemented; that is, when production is commenced, the machine will continue production until a predefined target stock level b_1 has been reached. For product 2, a *quantity-limited base-stock policy* is used; that is, when the machine starts production, it will continue production until either the base-stock level b_2 has been reached or a maximum number k of products has been produced. A production order is placed immediately after demand for the corresponding product has arrived. These production orders queue up at the production facility.

The number of outstanding production orders of product i at this facility is indicated as the *shortfall* of this product. The setup times are assumed to be state dependent; that is, no setup for a product is incurred when there is no *shortfall* (no outstanding production orders). When both products have no shortfall, the machine is turned off and is turned on again upon demand arrival. When a type-2 batch of size k has been produced and product 1 has no shortfall, a new type-2 batch up to k products is started after a new setup time.

For given values of the base-stock levels, b_1 and b_2, and the quantity limit, k, the steady-state net stock level distribution I_i for product i is given by (see Fig. 2)

$$I_i = b_i - L_i, \qquad i = 1, 2, \tag{32}$$

https://doi.org/10.1017/S0269964809000217 Published online by Cambridge University Press

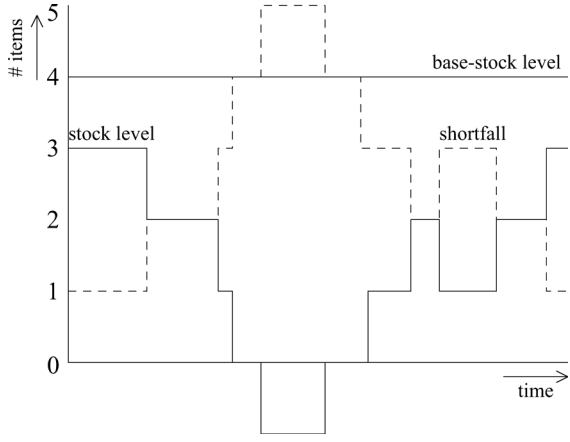


FIGURE 2. Example of the relation among base-stock level, net stock level, and shortfall.

where L_i denotes the steady-state shortfall of product i. Notice that the net stock level becomes negative when the shortfall of a certain product is larger than its base-stock level.

It is easily seen that the shortfall of a product is independent of the base-stock levels. Moreover, it is easily verified that the shortfall distribution of product i at the production facility is identical to the queue length distribution of queue i in the queuing model of the article paper. Hence, by the procedure presented in Section 3, in combination with (32), the steady-state net stock level for both products can be computed. With these distributions, various performance measures of interest can be computed.

In the remainder of the article, we consider the total expected costs C (i.e., the sum of holding and backlogging costs). For these individual cost components, we assume a simple linear structure; that is, the cost rate function for product i as a function of the net stock level x equals

$$c_i(x) = \begin{cases} h_i x, & x \geq 0 \\ -p_i x, & x < 0, \end{cases} \tag{33}$$

where h_i and p_i represent the holding and penalty cost coefficients for product i, respectively. With the help of (32), the total expected costs C can be written as

$$C = C_1 + C_2 = \mathbb{E}[c_1(I_1)] + \mathbb{E}[c_2(I_2)] = \mathbb{E}[c_1(b_1 - L_1)] + \mathbb{E}[c_2(b_2 - L_2)]. \tag{34}$$

For a given value of k, it can be shown that the optimal base-stock levels b_i^* are given by

$$b_i^* = \min\left\{ n \in \mathbb{N}_0 | P[L_i \leq n] \geq \frac{p_i}{p_i + h_i} \right\}, \qquad i = 1, 2, \tag{35}$$

https://doi.org/10.1017/S0269964809000217 Published online by Cambridge University Press

which is recognized as the solution of a standard newsboy problem. Finally, we note that we do not consider the optimization of the quantity limit k here.

4.2. Numerical Illustration

We now present some cases, which illustrate the value of the procedure of Section 3 in the evaluation of the described production system. Of course, a whole plethora of cases can be studied: different values of the quantity limit, choice of service time distributions and their parameters, choice of setup distributions and their parameters, different (ratios between) cost factors, and so forth. However, the aim of the present section is to present some illustrative cases that show the potential of the k-limited service policy in the context of the SELSP.

As described earlier, the p.g.f.s derived in the previous section have to be finished off with a number of zeros, which are numerically computed by using the Chaudhry QROOT software package [8]. We then use the method presented by Abate and Whitt [1] to numerically invert these p.g.f.s. Unfortunately, for large quantity limits, numerical problems have been encountered in the procedure. Therefore, we confine ourselves to cases with small quantity limits (see also the end of the present subsection).

In the numerical evaluation we also present results for the special case of $k = \infty$, which amounts to a two-product model with an exhaustive base-stock policy for both products. In this case, we do not use the procedure of Section 3, but we have implemented a discrete event simulation. Each simulation run is sufficiently long such that the widths of the 95% confidence intervals of the performance measures of interest are smaller than 1% of the estimated value.

The examination starts with an initial case, which is subsequently being perturbed into six cases to study the effect of (1) (ratio between) the loads, (2) (ratio between) cost factors, and (3) setup times.

Case 1 (*Initial Case*): Suppose that product 1 is a product with high costs, whereas product 2 is of secondary importance compared to the first product. Table 1 shows the detailed specifications for these two products (where the numbers in parentheses are the rates of the corresponding distributions). Table 2 shows the costs C_i per product, the total costs C, and the optimal base-stock levels b_i^* as a function of the quantity limit k. First, we observe that the marginal costs for product 2 are decreasing in the quantity

TABLE 1. Product information for Case 1

Parameter	Product 1	Product 2
Demand	Poisson(0.375)	Poisson(0.375)
Processing times	Exp(1.0)	Exp(1.0)
Setup times	Exp(0.25)	Exp(0.25)
Holding costs	10	1
Backlogging costs	90	9

https://doi.org/10.1017/S0269964809000217 Published online by Cambridge University Press

Table 2. Output for Case 1

k	b_1^*	C_1	b_2^*	C_2	C
1	3	27.5	13	13.9	41.4
2	*3*	*28.8*	*9*	*9.8*	*38.5*
3	3	30.5	7	8.7	39.2
4	3	32.5	7	8.1	40.6
5	3	34.6	6	7.7	42.3
6	4	36.5	6	7.4	43.9

limit, whereas the marginal costs for product 1 increase with this limit. Second, the optimal value of the quantity limit with respect to total costs is equal to 2. For smaller limits, the amount of capacity available for production is too low, whereas larger limits lead to more variable cycle lengths. Third, the optimal base-stock levels for product 2 are nonincreasing in the quantity limit, whereas the optimal base-stock levels for product 1 are nondecreasing in this limit. Fourth, the optimal quantity limit equals 2 in this case, and the optimal costs are equal to 38.5. In Table 2, the line with the results for the optimal quantity limit is placed in italic (and similarly for the Cases 2–7 in the Tables 3–8). Finally, it is interesting to compare Table 2 with the total costs equalling 59.2 that would be incurred if a standard exhaustive policy were implemented for both products. Via a k-limited policy for product 2, we might thus save 35.0% compared to the latter policy, clearly showing the advantage of the k-limited policy in a production environment.

Case 2 and 3 (*Effect of the Load*): These cases are similar to the initial case, except that the demand rates are perturbed; that is, in Case 2 the demand rates of products 1 and 2 equal 0.15 and 0.6, whereas in Case 3, products 1 and 2 have demand rates 0.6 and 0.15, respectively. Comparing the results in Tables 3 and 4 with the costs for exhaustive base-stock policies equaling 49.3 and 61.7, respectively, once more, significant cost reductions are observed by application of the k-limited policy. Of course, the advantages of the k-limited policy are much more pronounced in the case that the low-priority product 2 has the highest demand rate.

Case 4 and 5 (*Effect of the Cost Factors*): In the fourth and fifth case, we consider systems similar to that of the initial case and perturb the cost factors; that is, the

Table 3. Output for Case 2

k	b_1^*	C_1	b_2^*	C_2	C
1	1	13.8	24	24.9	38.7
2	1	15.4	12	12.5	27.8
3	*1*	*17.1*	*10*	*10.6*	*27.7*
4	2	18.6	9	9.8	28.4
5	2	19.2	8	9.3	28.5
6	2	19.9	8	9.0	28.8

https://doi.org/10.1017/S0269964809000217 Published online by Cambridge University Press

TABLE 4. Output for Case 3

k	b_1^*	C_1	b_2^*	C_2	C
1	*5*	*47.6*	*5*	*6.6*	*54.2*
2	5	48.7	4	5.8	54.5
3	5	49.9	4	5.4	55.3
4	5	51.0	4	5.2	56.2
5	5	52.0	4	5.0	57.0
6	5	53.0	3	4.9	57.9

TABLE 5. Output for Case 4

k	b_1^*	C_1	b_2^*	C_2	C
1	3	13.7	13	13.9	27.6
2	3	14.4	9	9.8	24.2
3	*3*	*15.2*	*7*	*8.7*	*24.0*
4	3	16.2	7	8.1	24.3
5	3	17.3	6	7.7	25.0
6	4	18.3	6	7.4	25.6

TABLE 6. Output for Case 5

k	b_1^*	C_1	b_2^*	C_2	C
1	3	5.5	13	13.9	19.4
2	3	5.8	9	9.8	15.5
3	3	6.1	7	8.7	14.8
4	*3*	*6.5*	*7*	*8.1*	*14.6*
5	*3*	*6.9*	*6*	*7.7*	*14.6*
6	4	7.3	6	7.4	14.7

cost factors for product 2 remain unaltered, whereas the holding and penalty costs for product 1 are decreased to 5 and 45 for Case 4 and to 2 and 18 for Case 5, respectively. Although the same conclusions as in Case 1 can be drawn from Tables 5 and 6, it should be observed that the advantages of the k-limited discipline are a bit less pronounced in these cases due to the leveling of the costs among the products. In fact, application of exhaustive policies for both products, which is normally done, would lead to total costs of 32.3 and 16.1, respectively.

Case 6 and 7 (*Effect of the Setup Times*): In Cases 6 and 7 we examine the effects of the sizes of the setup times. We therefore study two cases similar to the initial case, but in which the mean setup times equal 0.5 and 0.1, respectively. See Tables 7 and 8 for the results. Notice that for Case 6 the system is not stable when the quantity limit is chosen to be equal to 1. In the case that we implement the exhaustive base-stock policy for both products, the total costs are given by 62.7 and 56.3. This leads to the intuitively appealing conclusion that the advantages of the k-limited service discipline

https://doi.org/10.1017/S0269964809000217 Published online by Cambridge University Press

Table 7. Output for Case 6

k	b_1^*	C_1	b_2^*	C_2	C
1	–	–	–	–	–
2	3	30.2	16	16.9	47.1
3	*3*	*32.4*	*11*	*12.1*	*44.5*
4	4	34.7	9	10.4	45.0
5	4	36.0	8	9.4	45.4
6	4	37.6	7	8.8	46.4

Table 8. Output for Case 7

k	b_1^*	C_1	b_2^*	C_2	C
1	*2*	*26.7*	*8*	*8.8*	*35.4*
2	3	28.3	7	7.9	36.2
3	3	29.8	7	7.5	37.2
4	3	31.5	7	7.2	38.6
5	3	33.3	6	7.0	40.3
6	3	35.3	5	6.7	42.0

slightly increase in the case the setup times vanish, but we stress that even in the case of large setup times, the k-limited strategy still shows its superiority over the exhaustive policy.

4.3. Conclusion

In this section we have presented some cases, by using the results of the exact analysis of the present article, that aim to answer the question: What is the gain in performance of bounding production runs — at the expense of higher setup frequencies — in multi-item production-inventory systems. The results of the present section lead to the conjectures that the widely used exhaustive policy is not the most effective strategy in (frequently encountered) asymmetric production situations as well as that it might be desirable that production runs of low-priority products are bounded in these environments. We touch upon the possibilities of a large-scale (approximate) study to support this statement in more realistic systems in the next section.

The numerical problems encountered might be the reason for a skeptical reader to remark that simulation would be a more appropriate tool to analyze the model of the present article. We would disagree with this opinion for two reasons. First, such a remark passes over one of the main motivations of the present article: to study the borders of tractability of the k-limited service discipline. In that sense, it is important to observe that one of the key steps of our approach [the reduction of the multidimensional stochastic process to a single dimension represented by the function $\alpha(\cdot)$] already breaks down in the simplest extension of the model (i.e., a two-queue case in which both queues are served according to the k-limited service discipline).

https://doi.org/10.1017/S0269964809000217 Published online by Cambridge University Press

This observation unearths the fact that the present study has indeed reached the borders of tractability.

Second, in the past, people have observed an inefficiency of simulation techniques for (k-limited) polling systems (see, e.g., Blanc [4] or Van Vuuren and Winands [34]), especially, if one is interested in tail probabilities as needed in the present study for the computation of the optimal costs and optimal base-stock levels. In that respect, the generating functions derived in the present article provide an excellent breeding ground for the development of simple (closed-form), accurate, and efficient approximations for these tail probabilities along the lines of the dominant pole approximation as described in Tijms [33], for example.

5. CONCLUSIONS

The present article has presented an exact analysis of a two-queue state-dependent polling system with k-limited service extending the polling literature on both the k-limited service discipline and on state-dependent setups, containing the nonpreemptive priority model and the model of Lee [22] as special cases. Moreover, the results of the analysis have been applied to a make-to-stock production setting, which provides us with theoretical evidence that the k-limited strategy leads to considerable cost reductions compared to widely used (standard) exhaustive policies.

As stated in Section 1, the k-limited policy violates the branching property for polling systems (see Fuhrmann [17] and Resing [26]), implying that extensions of the analysis of the present article to more realistic systems are, in most likelihood, outside the borders of possibility and that one has to resort to approximations in these cases. Recently, Van Vuuren and Winands [34] developed an efficient and accurate approximate decomposition approach for k-limited polling systems under the assumption of generally distributed arrival, service, and setup distributions.

It is our hope that the theoretical evidence on the potential of the k-limited discipline given by the present article in combination with the efficient approximate algorithm for the k-limited strategy of [34] tempts people to a large-scale study comparing the k-limited and exhaustive strategies in the context of the (make-to-stock) production-inventory environments. Such a study would get us a second step closer to answering the question raised in Section 1 of the present paper: What is the gain in performance of bounding production runs – at the expense of higher setup frequencies – in multi-item production-inventory systems?

Acknowledgement

The authors would like to thank Onno Boxma for valuable comments.

References

1. Abate, J. & Whitt, W. (1992). Numerical inversion of probability generating functions. *Operations Research Letters* 12(4): 245–251.

https://doi.org/10.1017/S0269964809000217 Published online by Cambridge University Press

2. Adan, I.J.B.F., van Leeuwaarden, J.S.H. & Winands, E.M.M. (2006). On the application of Rouché's theorem in queueing theory. *Operations Research Letters* 34(3): 355–360.
3. Altman, E., Blanc, H., Khamisy, A. & Yechiali, U. (1994). Gated-type polling systems with walking and switch-in times. *Stochastic Models* 10: 741–764.
4. Blanc, J.P.C. (1992). An algorithmic solution of polling models with limited service disciplines. *IEEE Transactions on Communications* 40(7): 1152–1155.
5. Borst, S.C., Boxma, O.J. & Levy, H. (1995). The use of service limits for efficient operation of multistation single-medium communication systems. *IEEE/ACM Transactions on Networking* 3(5): 602–612.
6. Boxma, O.J. & Down, D.G. (1997). Dynamic server assignment in a two-queue model *European Journal of Operational Research* 103: 101–115.
7. Charzinski, J., Renger, T. & Tangemann, M. (1994). Simulative comparison of the waiting time distributions in cyclic polling systems with different service strategies. In *Proceedings of the 14th International Teletraffic Congress*, Antibes Juan-les-Pins, pp. 719–728.
8. Chaudhry, M.L. (1994). QROOT software package. Kingston: A&A Publications, Ontario, Canada.
9. Cohen, J.W. (1987). A two-queue, one-server model with priority for the longer queue. *Queueing Systems* 2(3): 261–283.
10. Dai, J.G. (1995). On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Annals of Applied Probability* 5: 49–77.
11. Dai, J.G. (1999). Stability of fluid and stochastic processing networks. Publication No. 9, Centre for Mathematical Physics and Stochastics, University of Aarhus, Denmark, Available from http://www.maphysto.dk/.
12. Dai, J.G. & Meyn, S.P. (1995). Stability and convergence of moments for multiclass queueing networks via fluid models. *IEEE Transactions on Automatic Control* 40: 1889–1904.
13. Down, D. (1998). On the stability of polling models with multiple servers. *Journal of Applied Probability* 35: 925–935.
14. Federgruen, A. & Katalan, Z. (1996). The stochastic economic lot scheduling problem: cyclical base-stock policies with idle times. *Management Science* 42(6): 783–796.
15. Federgruen, A. & Katalan, Z. (1998). Determining production schedules under base-stock policies in single facility multi-item production systems. *Operations Research* 46(6): 883–898.
16. Flatto, L. (1989). The longer queue model. *Probability in the Engineering and Informational Sciences* 3: 537–559.
17. Fuhrmann, S.W. (1981). Performance analysis of a class of cyclic schedules. Bell Laboratories Technical Memorandum 81-59531-1.
18. Günalay, Y. & Gupta, D. (1997). A polling system with a patient server and state-dependent setup times. *IIE Transactions* 29: 469–480.
19. Gupta, D. & Srinivasan, M.M. (1996). Polling systems with state-dependent setup times. *Queueing Systems* 22: 403–423.
20. Ibe, O.C. & Cheng, X. (1988). Stability conditions for multiqueue systems with cyclic service. *IEEE Transactions on Automatic Control* 33(1): 102–103.
21. Keilson, J. & Servi, L.D. (1990). The distributional form of Little's law and the Fuhrmann–Cooper decomposition. *Operations Research Letters* 9(4): 239–247.
22. Lee, D.-S. (1996). A two-queue model with exhaustive and limited service disciplines. *Stochastic Models* 12(2): 285–305.
23. Levy, H. & Sidi, M. (1990). Polling systems: applications, modeling and optimization. *IEEE Transactions on Communications* 38(10): 1750–1760.
24. Ozawa, T. (1990). Alternating service queues with mixed exhaustive and K-limited services. *Performance Evaluation* 11: 165–175.
25. Ozawa, T. (1997). Waiting time distribution in a two-queue model with mixed exhaustive and gated-type K-limited services. In *Proceedings of the International Conference on the Performance and Management of Complex Communication Networks*, Tsukuba, pp. 231–250.
26. Resing, J.A.C. (1993). Polling systems and multitype branching processes. *Queueing Systems* 13: 409–426.

https://doi.org/10.1017/S0269964809000217 Published online by Cambridge University Press

27. Singh, M.P. & Srinivasan, M.M. (2002). Exact analysis of the state dependent polling model. *Queueing Systems* 41: 371–399.
28. Sox, C.R., Jackson, P.L., Bowman, A. & Muckstadt, J.A. (1999). A review of the stochastic lot scheduling problem. *International Journal of Production Economics* 62: 181–200.
29. Takács, L. (1968). Two queues attended by a single server. *Operations Research* 16(3): 639–650.
30. Takagi, H. (1990). Queueing analysis of polling models: an update. In H. Takagi (ed.), *Stochastic analysis of computer and Communication systems*. Amsterdam: North-Holland, pp. 267–318.
31. Takagi, H. (1997). Queueing analysis of polling models: progress in 1990–1994. In J.H. Dshalalow (ed.), *Frontiers in Queueing: Models, Methods and Problems*. Boca Raton, FL: CRC Press, pp. 119–146.
32. Takagi, H., (2000). Analysis and application of polling models. In G. Haring, C. Lindemann and M. Reiser (eds.), *Performance evaluation: Origins and directions*. Lecture Notes in Computer Science Vol. 1769. Berlin: Springer, pp. 423–442.
33. Tijms, H.C. (1994). Stochastic models: An algorithmic approach. New York: Wiley.
34. Vuuren, M. van & Winands, E.M.M. (2006). Iterative approximation of k-limited polling systems. *Queueing Systems* 55(3): 161–178.
35. Winands, E.M.M., Adan, I.J.B.F. & van Houtum, G.J. (2005). The stochastic economic lot scheduling problem: a survey. Report BETA WP-133, Beta Research School for Operations Management and Logistics, Eindhoven.

APPENDIX A

Proof of Theorem 2.2

The intuition behind our proof of stability is as follows:

1. The first queue is stable if $\rho_1 < 1$, due to the exhaustive service policy.
2. The stability of the overall system is determined by queue 2, and as we are concerned with stability, we need to show that the queue length at 2 tends to decrease when it is "sufficiently large."
3. If queue 2 always has at least k customers, then no arrivals occur to queue 1 during a busy period at queue 2 plus a setup time for 2 with probability $R_2(0,1)(S_2(0,1))^k$. This determines whether a setup for queue 1 is needed and, in turn, determines the proportion of time spent on setups.
4. Stability is then determined by determining if the proportion of time taken up by serving queue 1 (ρ_1) plus the proportion of time spent on setups is sufficiently small for queue 2 to accommodate its demand.

This program can be made precise by using fluid limit methodology described by Dai [10] and applied to polling models by Down [13]. (Note that as this is a two-queue model, we can also adapt the results of Dai and Meyn [12].)

As we wish to adapt the results of [13], we need to define so-called *fluid limits* for the system. Let $Q(t) = (Q_1(t), Q_2(t))$, where $Q_i(t)$ is the number of customers in queue i at time t, $T(t) = (T_1(t), T_2(t))$, where $T_i(t)$ is the cumulative time by time t that is spent servicing customers at queue i, and $T^0(t) = (T_1^0(t), T_2^0(t))$, where $T_i^0(t)$ is the cumulative time by time t that the server spends in switching from queue i. Let $q = Q_1(0) + Q_2(0)$. Suppose that $(\bar{Q}(\cdot), \bar{T}(\cdot), \bar{T}^0(\cdot)) = (\bar{Q}_1(\cdot), \bar{Q}_2(\cdot), \bar{T}_1(\cdot), \bar{T}_2(\cdot), \bar{T}_1^0(\cdot), \bar{T}_2^0(\cdot))$ is a limit point of

$$\left(\frac{1}{q}Q(q\cdot), \frac{1}{q}T(q\cdot), \frac{1}{q}T^0(q\cdot)\right) \tag{A.1}$$

https://doi.org/10.1017/S0269964809000217 Published online by Cambridge University Press

when $q \to \infty$. This scaling allows us to look at a scaled version of the state when at least one of the queue lengths is very large. From [10], any fluid limit satisfies the following set of equations, for $i = 1, 2$:

$$\begin{aligned} &\bar{Q}_i(t) = \bar{Q}_i(0) + \lambda_i t - \mu_i \bar{T}_i(t), \\ &\bar{Q}_i(t) \geq 0, \\ &\bar{T}_i(0) = 0 \text{ and } \bar{T}_i(t) \text{ is nondecreasing,} \end{aligned} \tag{A.2}$$

$$\frac{d}{dt} \sum_{i=1}^{2} (\bar{T}_i(t) + \bar{T}_i^0(t)) = 1 \quad \text{whenever } \bar{Q}_1(t) + \bar{Q}_2(t) > 0, \tag{A.3}$$

where $\mu_i = 1/\beta_i$ is the service rate at queue i, $i = 1, 2$. Although every fluid limit must satisfy these equations, it is not the case that any solution of the above equations is a valid fluid limit. Further properties are derived in the proof of Theorem A.1.

Define

$$\hat{\rho} = \rho_1 + \rho_2 + \frac{\lambda_2}{k}(\tau_2 + r\tau_1),$$

where r has previously been defined as

$$r = 1 - R_2(0, 1)(S_2(0, 1))^k.$$

(Note that r is the probability that the number of arrivals during a setup time plus k successive service times at queue 2 is not equal to zero.) We then have the following result.

THEOREM A.1:

(i) If $\hat{\rho} < 1$, then the distribution of the queue length process $\{Q(t)\}$ converges to a steady-state distribution φ as $t \to \infty$.

(ii) If $\hat{\rho} > 1$, as $t \to \infty$,

$$P(|Q(t)| \longrightarrow \infty) \longrightarrow 1.$$

PROOF: At several points in the proof, we will need to take the time derivatives of $\bar{T}_i(t)$ and $\bar{T}_i^0(t)$. As these functions are Lipschitz, their derivatives exist almost everywhere, so we assume throughout that any derivatives are taken at time points where the derivatives exist.

From Lemma 3.1 of [13], $\rho_1 < 1$ implies that for some finite t_1 and all $t \geq t_1$, $\bar{Q}_1(t) = 0$. Using (A.2), this, in turn, implies

$$\frac{d}{dt}\bar{T}_1(t) = \rho_1 \tag{A.4}$$

for all $t \geq t_1$. Intuitively, this means that the queue length at 1 for the unscaled process eventually becomes finite and remains so (see (A.1)). For $t \geq t_1$, suppose that $\bar{Q}_2(t) > 0$. By the continuity of $\bar{Q}_2$, there exists an $h > 0$ such that $\min_{s\in[t,t+h]} \bar{Q}_2(s) > 0$. Let $c = \min_{s\in[t,t+h]} \bar{Q}_2(s)$. Then, by Theorem 4.1 of Dai [10], there is a sequence $\{q_n\}$ such that $\{q_n^{-1} Q_2(q_n s)\}$ uniformly converges to $\bar{Q}_2(s)$ for s in $[t, t + h]$. Thus, for large enough n, we have $q_n^{-1} Q_2(q_n s) \geq c/2$ for $s \in [t, t + h]$ and $q_n c/2 \geq 1$. So,

$$Q_2(u) > 0 \quad \text{for } u \in [q_n t, q_n(t + h)] \tag{A.5}$$

that is, for a large interval of time, queue 2 does not empty. We will be concerned with the derivatives of $\bar{T}_i$ at t, for which one needs to calculate $\bar{T}_i(t + h) - \bar{T}_i(t) = \lim_{n\to\infty} q_n^{-1}(T_i(q_n(t +$

https://doi.org/10.1017/S0269964809000217 Published online by Cambridge University Press

$h)) - T_i(q_n(t))$, but by (A.5), the intervals on the right-hand side correspond to those where queue 2 never empties. However, this, in turn, implies that if we model the server as always moving to queue 1, but incurring a zero setup time and performing no service if the queue is empty, then the mean time spent in switching from queue 1 to queue 2 is τ_2, and switching from queue 2 to queue 1, it is $r\tau_1$ (with probability $1 - r$ that no setup time is incurred). We can now directly apply Lemma 3.2 of [13] to get

$$\frac{d}{dt}\bar{T}_2(t)/(k/\mu_2) = \frac{d}{dt}\bar{T}_2^0(t)/(r\tau_1), \tag{A.6}$$

which makes precise the notion that when queue 2 never empties, the number of busy periods at queue 2 (each of mean length k/μ_2) is equal to the number of setups for queue 1 (including those of zero time, so the mean is $r\tau_1$). In the same manner, it follows that

$$\frac{d/dt\,(\bar{T}_1^0(t))}{\tau_2} = \frac{d/dt\,(\bar{T}_2^0(t))}{r\tau_1}. \tag{A.7}$$

Similar to (A.6), this says that the number of setups for queue 1 (including those of zero time) is equal to the number of setups for queue 2. Also, if $\bar{Q}_1(s) + \bar{Q}_2(s) > 0$ for all $s \le t$, from (A.3),

$$\bar{T}_1(t) + \bar{T}_2(t) + \bar{T}_1^0(t) + \bar{T}_2^0(t) = t; \tag{A.8}$$

that is, the server never idles. Differentiating (A.8) and using (A.7),

$$\frac{d}{dt}\bar{T}_1(t) + \frac{d}{dt}\bar{T}_2(t) = 1 - (\tau_2 + r\tau_1)\left(\frac{d/dt\,(\bar{T}_2^0(t))}{r\tau_1}\right). \tag{A.9}$$

Define

$$\begin{aligned}\bar{W}(t) &= \frac{\bar{Q}_1(t)}{\mu_1} + \frac{\bar{Q}_2(t)}{\mu_2} \\ &= \bar{W}(0) + (\rho_1 + \rho_2)t - \bar{T}_1(t) - \bar{T}_2(t),\end{aligned} \tag{A.10}$$

where the second equality follows from (A.2). Differentiating (A.10) and using (A.9), we have

$$\frac{d}{dt}\bar{W}(t) = (\tau_2 + r\tau_1)\left(\frac{d/dt\,(\bar{T}_2^0(t))}{r\tau_1}\right) - (1 - \rho_1 - \rho_2). \tag{A.11}$$

By rearranging terms in (A.9) and using (A.4) and (A.6),

$$\left(\tau_2 + r\tau_1 + \frac{k}{\mu_2}\right)\left(\frac{d/dt\,(\bar{T}_2^0(t))}{r\tau_1}\right) = 1 - \rho_1.$$

Now, substituting for $d/dt\,(\bar{T}_2^0(t))/(r\tau_1)$ in (A.11) gives

$$\begin{aligned}\frac{d}{dt}\bar{W}(t) &= \frac{(\tau_2 + r\tau_1)(1 - \rho_1)}{\tau_2 + r\tau_1 + k/\mu_2} - (1 - \rho_1 - \rho_2) \\ &= \frac{(\tau_2 + r\tau_1)\rho_2 - (1 - \rho_1 - \rho_2)(k/\mu_2)}{\tau_2 + r\tau_1 + k/\mu_2}.\end{aligned}$$

https://doi.org/10.1017/S0269964809000217 Published online by Cambridge University Press

So, $d/dt\,(\bar{W}(t)) < 0$ if and only if

$$(\tau_2 + r\tau_1)\rho_2 < (1 - \rho_1 - \rho_2)(k/\mu_2)$$

or, by rearranging terms,

$$\hat{\rho} = \rho_1 + \rho_2 + \frac{\lambda_2}{k}(\tau_2 + r\tau_1) < 1.$$

Thus, we conclude that $d/dt\,(\bar{W}(t)) < 0$ if and only if $\hat{\rho} < 1$, so we would then have for some $t_2 > 0$ and all $t \geq t_2$, $\bar{Q}_1(t) + \bar{Q}_2(t) = 0$. Theorem 4.2 of Dai [10] then yields part (i). If $\hat{\rho} > 1$, $d/dt\,(\bar{W}(t)) > 0$ and part (ii) follows from Theorem 2.5.1 of Dai [11]. ■

https://doi.org/10.1017/S0269964809000217 Published online by Cambridge University Press