J. Appl. Prob. **49,** 1052–1071 (2012)
Printed in England
© *Applied Probability Trust* 2012

ON THE VALUE FUNCTION OF THE M/G/1 FCFS AND LCFS QUEUES

ESA HYYTIÄ,* **
SAMULI AALTO,*
ALEKSI PENTTINEN * AND
JORMA VIRTAMO,* *Aalto University*

Abstract

We consider a single-server queue with Poisson input operating under first-come–first-served (FCFS) or last-come–first-served (LCFS) disciplines. The service times of the customers are independent and obey a general distribution. The system is subject to costs for holding a customer per unit of time, which can be customer specific or customer class specific. We give general expressions for the corresponding value functions, which have elementary compact forms, similar to the Pollaczek–Khinchine mean value formulae. The results generalize earlier work where similar expressions have been obtained for specific service time distributions. The obtained value functions can be readily applied to develop nearly optimal dispatching policies for a broad range of systems with parallel queues, including multiclass scenarios and the cases where service time estimates are available.

Keywords: M/G/1; FCFS; LCFS; value function; sojourn time; mean delay

2010 Mathematics Subject Classification: Primary 60K25
Secondary 90C40

1. Introduction

The value function of a system describes the relative value of a state with respect to a long-term cost or revenue rate. This depends on the stochastic model and the costs associated with it. For a single-server queue with an infinite number of system places, no customer is rejected and the natural objective is to study the mean sojourn time, i.e. the mean delay. This is equivalent to defining unit holding costs for customers present in the system. The knowledge of the value function gives important insight into a particular queueing discipline and allows one to compare different states in terms of the chosen objective. For example, the availability of the value function allows one to develop efficient and theoretically sound dispatching rules for the corresponding system of parallel queues. In particular, starting from an arbitrary state-independent policy such as Bernoulli splitting, the so-called first policy iteration (FPI) step of the Markov decision processes can be carried out.

In this paper we first consider a general single-server queue with *first-come–first-served* (FCFS), and both preemptive and nonpreemptive *last-come–first-served* (LCFS) queueing disciplines, a Poisson arrival process, and a general service time distribution. The arriving customers have arbitrary holding costs and can also be tagged with some additional information such as the customer's class. In the multiclass scenario, the service times are assumed to be class

Received 5 January 2012; revision received 25 April 2012.

* Postal address: Department of Communications and Networking, Aalto University, School of Electrical Engineering, PO Box 13000, 00076 Aalto, Finland.
** Email address: esa@netlab.tkk.fi

https://doi.org/10.1239/jap/1354716657 Published online by Cambridge University Press

specific. For any given state, the remaining service times are conditional random variables. In this rather general setting, we are able to give exact expressions for the corresponding value functions.

Throughout the paper, we consider three specific levels of state information: (i) the *size-aware* system, where the exact (remaining) service times of the existing customers are available, (ii) the *time-aware* system, where the service each customer has received (in time) is available, and (iii) the *number-aware* system, where only the number in the system is available. For the number-aware system, the actual state is as seen by a random observer, e.g. an arriving customer from a Poisson process. The size-aware setting with a unit holding cost has been treated in [17], which we first generalize to customer- or class-specific holding costs. These results are then utilized in deriving general expressions for value functions, from which the time-aware and number-aware results follow as special cases. The analytical results are further elaborated by considering the important case of a single class with unit holding cost, which can be related to the sojourn time (i.e. response time).

The analytical results can be applied to a broad range of dispatching problems where service time distributions can also be server specific. The dispatcher is assumed to be aware of the underlying service time distributions, queueing discipline, service rate, and the number of customers in each queue (number-aware). Optionally, the amount of service each customer has received may be available (time-aware), or the exact remaining service times can be known (size-aware). The information on the service received in the time-aware case becomes relevant when the service times have general distributions. For exponentially distributed service times and identical servers, Winston [37] has already shown that the *join the shortest queue* (JSQ) minimizes the mean sojourn time.

The rest of the paper is organized as follows. First, in the following section, we briefly review earlier work on the value function for queueing systems and dispatching problems. Section 2 contains the necessary definitions. Then we derive the value functions for FCFS in Section 3, and for preemptive and nonpreemptive LCFS in Sections 4 and 5. In Section 6 we discuss the application of the value function to dispatching problems. Section 7 concludes the paper.

1.1. Prior work

Prior work has studied the value function in order to find optimal dispatching rules in different settings. In particular, value functions enable FPI [30], where the idea is to improve a given basic policy by choosing an alternative action leading to an alternative system state such that the expected future costs in that state are lower for an infinite time horizon. The information is provided by the value function.

Several results already exist for minimizing the mean delay in queueing systems. In the context of routeing packets in a data network, Krishnan [23] obtained the value function for the M/M/s system. The M/M/1 queue was considered in [1]. For Markovian systems, a sufficient state description is the number in the system. Bhulai [6] derived the value function for the M/Cox(r)/1-FCFS queue, with a state description of (n, k), where n denotes the number in the system and k denotes the service phase of the customer being processed (if any). A state description with service phases can sometimes be impractical. Sassen *et al.* [32] considered the M/G/1-FCFS queue. In order to avoid dealing with the remaining service times, they resorted to an approximation and assumed that the remaining service time of the service in progress (if any) obeys the equilibrium excess distribution. In a size-aware system, the service times become known upon arrival. The value functions for a size-aware M/G/1 queue with FCFS, LCFS, SPT, and SRPT are given in [17], while M/D/1-PS and M/M/1-PS are treated in [18]

https://doi.org/10.1239/jap/1354716657 Published online by Cambridge University Press

and [19], respectively. To the best of our knowledge, no expression for M/G/1-PS is known. In contrast to the above work, we give, without resorting to any approximations, exact expressions for the value function in M/G/1-FCFS and LCFS queues under the time-aware assumption that only the number in the queue and the service received (in time) are available.

For blocking systems, so as to minimize the blocking probability, Krishnan and Ott [24] obtained the value functions of an M/M/s/s system in order to route calls in the telephone network. This result was generalized by Leeuwaarden *et al.* [34] to an M/M/s/k system, where the example application was the assignment of telephone calls to base stations in a mobile network.

For the dispatching problem in a system of parallel queues, which is the prime application of the value function, a wealth of results exist in the literature. First, Winston [37] showed the optimality of JSQ for identical servers, the Poisson arrival process, and exponentially distributed service times when the number in the queue is available. Weber [35] showed that JSQ is also optimal for identical servers with a more general arrival process and service times having a nondecreasing hazard rate. Ephremides *et al.* [10] also showed the optimality of JSQ for identical servers and exponentially distributed service times with a more general arrival process, and additionally argued for the optimality of the *round-robin* policy, when the only information is that the queues were initially in the same state. The latter result was later generalized by Liu and Towsley [26] to independent and identically distributed (i.i.d.) service times with increasing hazard rate and an arbitrary arrival process, and further, by Liu and Righter [25] to a setting of distributed homogeneous unreliable servers capable of redirecting only some tasks to other servers. However, the optimality results are available only for specific settings (e.g. identical servers and the FCFS discipline, as above).

Situations where a heterogeneous dispatching problem arises are numerous. Becker *et al.* [5] considered static Bernoulli routeing motivated by routeing telephone calls in call centres to different experts. Schwartz [33] and Conolly [8] mentioned toll booths in highways. Ansell *et al.* [4] additionally mentioned grid computing. Bonomi [7] also considered task assignment in distributed systems and compared JSQ with FCFS and processor-sharing (PS) scheduling disciplines, where PS is motivated by CPU scheduling in time-shared computer systems. Recently, Gupta *et al.* [13] also considered the JSQ policy in the setting of web server farms and PS. Similarly, Crovella *et al.* [9] and Harchol-Balter *et al.* [14] considered distributed server systems such as web server farms and distributed batch computing systems under the assumptions that only the size of a new customer is available and that each queue is served according to FCFS. They proposed an efficient service requirement (job size) based policy referred to as *size interval task assignment* or SITA for short, which Feng *et al.* [12] later proved to be the optimal policy in the given setting (see also [15]). Ephremides *et al.* [10], and Aalto and Virtamo [1], formulated a dispatching problem in the context of routeing packets in data networks. Additionally, dispatching problems have been studied in, e.g. [21] and [36].

2. Preliminaries

We consider queueing systems and, in particular, single-server FCFS and LCFS queues with a Poisson arrival process, general service time distributions, and a *general cost model*, where customer i has a certain holding cost denoted by a_i. Let z denote the information regarding the queue's state, which includes the number in the system, n, the customers' holding costs a_i, and optional information regarding their (remaining) service times Y_i. The distributions for the initial service times X_i are also known. In the size-aware case, z describes the exact remaining service times of the customers, but, in general, z stands for an arbitrary amount of

https://doi.org/10.1239/jap/1354716657 Published online by Cambridge University Press

information regarding the system's state. Thus, for a given z, the system accrues costs at rate $r(z) := \sum_{i=1}^{n} a_i$. Let the random variable $\mathbf{Z}_z(t)$ denote the state of the system at time t, where z defines the initial state, and let the random variable $V_z(t)$ denote the costs incurred during time $(0, t)$:

$$V_z(t) := \int_0^t r(\mathbf{Z}_z(s))\,\mathrm{d}s.$$

A stable queue incurs costs at an average cost rate of r,

$$r := \lim_{t\to\infty} \frac{1}{t}\,\mathrm{E}[V_z(t)],$$

which is independent of the initial state information z (ergodic system), $r = \mathrm{E}[r(\mathbf{Z})]$. The *value function* characterizes the deviation from the average cost rate r for a given state information z in the infinite time horizon:

$$v_z := \lim_{t\to\infty} \mathrm{E}[V_z(t) - rt].$$

3. FCFS queue

In this section we give exact expressions for the value function in the single-server FCFS queue. FCFS is the most common queueing discipline and is often the default assumption in the literature unless otherwise specified. It is fair in the sense that the customers leave in the same order as they arrive.

Size-aware queues have been considered by Hyytiä *et al.* [17]. The state description is a vector $(\Delta_1, \ldots, \Delta_n)$, where Δ_i denotes the known (remaining) service time of customer i. We use the convention that customer n is currently receiving service (if any). In what follows, we rely on the size-aware results given in [17], which we first generalize to arbitrary holding costs and then, by conditioning on the service times, obtain elegant expressions for the value function of FCFS with unknown service requirements (and later in Sections 4 and 5 also for the LCFS disciplines). For the FCFS queue, the size-aware result from [17] is as follows.

Lemma 1. (Size-aware M/G/1-FCFS.) *The size-aware relative value of state $z = (\Delta_1, \ldots, \Delta_n)$ with respect to the sojourn time in an M/G/1-FCFS queue is given by*

$$v_{(\Delta_1,\ldots,\Delta_n)} - v_0 = \frac{\lambda u_z^2}{2(1-\rho)} + \sum_{i=1}^{n} i\,\Delta_i, \tag{1}$$

where Δ_i denotes the known (remaining) service time of customer i, $u_z = \Delta_1 + \cdots + \Delta_n$ is the backlog, λ is the Poisson arrival rate, $\rho = \lambda\,\mathrm{E}[X]$ is the offered load, with $\mathrm{E}[X]$ *the mean service time, and customer n is served first.*

We consider a more general setting with arbitrary i.i.d. customer specific holding costs. Let a_i denote the known holding cost of customer i, and let $\mathrm{E}[A]$ denote the mean holding cost of an arbitrary customer. Result (1) generalizes straightforwardly to this setting.

Lemma 2. (Size-aware M/G/1-FCFS with arbitrary holding costs.) *For the size-aware value function of M/G/1-FCFS with arbitrary holding costs a_i, it holds that*

$$v_{(\Delta_1,\ldots,\Delta_n)} - v_0 = \frac{\lambda\,\mathrm{E}[A] u_z^2}{2(1-\rho)} + \sum_{i=1}^{n} \left(\Delta_i \sum_{j=1}^{i} a_j\right), \tag{2}$$

where $\mathrm{E}[A]$ *denotes the mean cost rate of a random customer (i.i.d.).*

https://doi.org/10.1239/jap/1354716657 Published online by Cambridge University Press

Proof. The proof is essentially the same as in [17] for the unit holding cost, and has been omitted for brevity.

In general, the remaining service times of the present n customers are conditional random variables that depend on the available information. Let the random variables Y_i denote the remaining service times, and let the a_i denote the corresponding holding cost rates, $i = 1, \dots, n$. Customers $1, \dots, n-1$ are waiting in the queue and customer n is currently receiving service (if any). The random variables Y_i may depend on the customer's class, the service received (in time), or any other additional information available. However, they are assumed to be pairwise independent. For example, the formulation includes the cases where one has additional information regarding the service time of particular customers (e.g. an estimate for the mean). We let z denote an arbitrary state and, for an empty system, we write $z := 0$.

In this more general setting, by conditioning, the size-aware result (2) gives the following proposition.

Proposition 1. *The value function for a multiclass M/G/1-FCFS with arbitrary holding costs is given by*

$$v_z - v_0 = \frac{\lambda\,\mathrm{E}[A]}{2(1-\rho)}\,\mathrm{E}\biggl[\biggl(\sum_{i=1}^{n} Y_i\biggr)^2\biggr] + \sum_{i=1}^{n}\biggl(\mathrm{E}[Y_i]\sum_{j=1}^{i} a_j\biggr), \tag{3}$$

where the remaining service times Y_i are pairwise independent but may depend on the service received, the customer's class, or any other additional information available, $\mathrm{E}[A]$ *is the mean holding cost rate of an arriving customer, and the a_i denote the known holding cost rates of the present n customers.*

Proof. Let $U = Y_1 + \cdots + Y_n$ denote the amount of unfinished work (backlog). From (2), by conditioning on the service times, we have

$$\begin{aligned} v_z - v_0 &= \mathrm{E}[\mathrm{E}[v_z - v_0 \mid (Y_1, \dots, Y_n)]] \\ &= \mathrm{E}\biggl[\frac{\lambda\,\mathrm{E}[A]U^2}{2(1-\rho)} + \sum_{i=1}^{n}\biggl(Y_i\sum_{j=1}^{i} a_j\biggr)\biggr] \\ &= \frac{\lambda\,\mathrm{E}[A]}{2(1-\rho)}\,\mathrm{E}\biggl[\biggl(\sum_{i=1}^{n} Y_i\biggr)^2\biggr] + \sum_{i=1}^{n}\biggl(\mathrm{E}[Y_i]\sum_{j=1}^{i} a_j\biggr). \end{aligned}$$

Corollary 1. *For the unit holding cost rate, $a_i = \mathrm{E}[A] = 1$, the value function reads*

$$v_z - v_0 = \frac{\lambda}{2(1-\rho)}\,\mathrm{E}\biggl[\biggl(\sum_{i=1}^{n} Y_i\biggr)^2\biggr] + \sum_{i=1}^{n} i\,\mathrm{E}[Y_i]. \tag{4}$$

We note that (3) and (4) hold in a very general setting, e.g. when the set of customer classes is noncountable (e.g. when the initial service time defines the class), or when several customers have received service due to a different queueing discipline executed earlier, as long as the *future evolution* of the queue is governed by FCFS.

3.1. Remaining service time

In an FCFS queue, only the first customer (if any) has received service. For a multiclass setting, a sufficient state description is $z = ((k_1, a_1), \dots, (k_n, a_n))$ for a number-aware system, and $z = ((k_1, a_1), \dots, (k_n, a_n); s)$ for a time-aware system, where k_i denotes the class of

https://doi.org/10.1239/jap/1354716657 Published online by Cambridge University Press

customer i, a_i denotes the holding cost of customer i, and s is the service (if any) customer n has received so far. Let X_k denote the initial service time of a class-k customer, so that $Y_i \sim X_{k_i}$ for $i = 1, \dots, n-1$ and Y_n is the remaining service time of the customer currently receiving service. For number-aware systems with a common service time distribution, we have the following result.

Lemma 3. *For a number-aware M/G/1-FCFS queue, the mean remaining service time* $\mathrm{E}[Y]$ *of the customer currently receiving service on condition that there are $N = n$ customers is*

$$\mathrm{E}[Y] = \frac{1-\rho}{\lambda \pi_n} \sum_{i=n+1}^{\infty} \pi_i,$$

where π_i denotes the steady-state probability that there are i customers in the system.

The proof of Lemma 3 is given in [11] and [27]. The complete distribution of the conditional random variable $(Y \mid N = n)$ is derived in [2], which provides $\mathrm{E}[Y^2]$. Similarly, for time-aware systems, let $Y_{n,s}$ denote the remaining service time of the customer whose attained service time is s.

Lemma 4. *For a time-aware system, the conditional remaining service time of a customer whose attained service time is s, is, by definition, $Y_{n,s} \sim (X_{k_n} - s \mid X_{k_n} > s)$, giving*

$$\mathrm{E}[(Y_{n,s})^p] = \frac{1}{\mathrm{P}\{X_{k_n} > s\}} \int_s^{\infty} (x-s)^p f_{k_n}(x)\, \mathrm{d}x, \tag{5}$$

where $f_{k_n}(x)$ denotes the probability density function of X_{k_n}.

These can be substituted into (3) and (4), yielding exact expressions for the value function. Next we will consider an important special case related to the mean sojourn time.

3.2. Single class and sojourn time

In the elementary case of a single class with unit holding cost, corresponding to the sojourn time, the natural state description for a number-aware system is $z = (n)$, and, for a time-aware system, $z = (n, s)$, where n denotes the number in the system and s denotes the amount of service (if any) customer n has received (in a time-aware system). Let X denote the common initial service time, $Y_i \sim X$ for $n = 1, \dots, n-1$, and let Y denote the remaining service time of customer n currently receiving service (if any), $Y \sim (X - s \mid X > s)$. As the service times are also independent, substituting the Y_i into (4) gives

$$\frac{\lambda}{2(1-\rho)}((n-1)\,\mathrm{E}[X^2] + \mathrm{E}[Y^2] + (n-1)(n-2)\,\mathrm{E}[X]^2 + 2(n-1)\,\mathrm{E}[X]\,\mathrm{E}[Y]) \\ + \frac{n(n-1)}{2}\,\mathrm{E}[X] + n\,\mathrm{E}[Y],$$

which, after some manipulation, gives the following corollary.

Corollary 2. (M/G/1-FCFS with unit holding cost.) *The value function for an M/G/1-FCFS with unit holding costs is given by*

$$v_z - v_0 = \frac{(n-1)(n-2\rho)\,\mathrm{E}[X] + 2(n-\rho)\,\mathrm{E}[Y] + \lambda(n-1)\,\mathrm{E}[X^2] + \lambda\,\mathrm{E}[Y^2]}{2(1-\rho)}, \tag{6}$$

where $n = 1, 2, \dots$ denotes the number in the system, X denotes the initial service time, and Y denotes the remaining service time of the customer being served (if any).

https://doi.org/10.1239/jap/1354716657 Published online by Cambridge University Press

For a time-aware system, Lemma 4 gives $\mathrm{E}[Y]$ and $\mathrm{E}[Y^2]$, which can then be substituted into (6). For a number-aware system, (6) can be developed further with the aid of Lemma 3 and the discussion thereafter.

Remark 1. For $X \sim \mathrm{Exp}(\mu)$, the memoryless property implies that $X \sim Y$, the number in the queue n describes the state, and the above, in agreement with [1, 6], reduces to

$$v_n - v_0 = \frac{n(n+1)}{2(\mu - \lambda)}.$$

The obtained expressions for the value function have intriguing forms. Similar to the Pollaczek–Khinchine M/G/1 formula [22], [31], the value function of a time-aware M/G/1-FCFS queue depends on the first two moments of the service time distribution(s). Additionally, the value function also depends on the first two moments of the conditional remaining service time Y. Moreover, the single-class expression is quadratic with respect to the number in the queue, n. We note that the result of Bhulai [6] for the Coxian service time distribution follows as a special case where Y depends on the phase of the service that is assumed to be observable.

4. Preemptive LCFS queue

In an LCFS queue, arriving customers enter at the head of the queue. Although this may sound unfair, preemptive LCFS is actually a better option than FCFS with respect to, e.g. the slowdown criterion [16]. Analysis of LCFS with a general holding cost structure turns out to be straightforward due to the fact that the current state has no effect on the sojourn times of future arrivals. From [17], we have again the size-aware result.

Lemma 5. (Size-aware preemptive M/G/1-LCFS.) *The size-aware relative value with respect to the sojourn time in a preemptive M/G/1-LCFS queue for state* $(\Delta_1, \dots, \Delta_n)$ *is given by*

$$v_{(\Delta_1,\dots,\Delta_n)} - v_0 = \frac{1}{1-\rho} \sum_{i=1}^{n} i\Delta_i,$$

where customer n is served first and customer 1 last.

With arbitrary customer specific holding costs, we have only a slightly more complicated result.

Lemma 6. (Size-aware preemptive M/G/1-LCFS with arbitrary holding costs.) *It holds that*

$$v_{(\Delta_1,\dots,\Delta_n)} - v_0 = \frac{1}{1-\rho} \sum_{i=1}^{n} \left(\Delta_i \sum_{j=1}^{i} a_j \right), \tag{7}$$

where a_j denotes the holding cost of customer j.

Proof. As the current state has no effect on future arrival sojourn times, the value function depends on the mean sojourn times $\mathrm{E}[T_i]$ of the present n customers in state z,

$$v_{(\Delta_1,\dots,\Delta_n)} - v_0 = \sum_{i=1}^{n} a_i \, \mathrm{E}[T_i].$$

The mean remaining sojourn time of customer i in a size-aware preemptive LCFS queue is $\mathrm{E}[T_i] = (1-\rho)^{-1} \sum_{j=i}^{n} \Delta_j$ [17], which then yields (7).

https://doi.org/10.1239/jap/1354716657 Published online by Cambridge University Press

With preemptive LCFS, all customers in the system have received some amount of service (Poisson arrivals). Therefore, a convenient state description in this case for a time-aware system is $((k_1, a_1, s_1), \dots, (k_n, a_n, s_n))$, where the triple (k_i, a_i, s_i) denotes the ith customer's class, holding cost, and amount of service received, respectively. For a number-aware system, the corresponding state information is $z = ((k_1, a_1), \dots, (k_n, a_n))$. Let Y_i again denote the remaining service time of customer i, and let X_k denote the initial service time of a class-k customer. Then we have the following general result.

Proposition 2. *The value function for a multiclass preemptive M/G/1-LCFS queue with arbitrary customer specific holding costs is given by*

$$v_z - v_0 = \frac{1}{1-\rho} \sum_{i=1}^{n} \biggl(\mathrm{E}[Y_i] \sum_{j=1}^{i} a_j \biggr). \tag{8}$$

Proof. The result follows immediately from (7) by conditioning on the service times.

Corollary 3. *For unit holding costs, $a_j = 1$, the value function of a multiclass preemptive M/G/1-LCFS queue reduces to*

$$v_z - v_0 = \frac{1}{1-\rho} \sum_{i=1}^{n} i\, \mathrm{E}[Y_i]. \tag{9}$$

4.1. Remaining service time

For the time-aware case, the customers' service time distributions and the amounts of service they have received are known, and, thus, the moments of the remaining service times are determined according to Lemma 4. For the number-aware case, we have the following result.

Lemma 7. *For the remaining service time Y_i of a class-k_i customer i in a preemptive M/G/1-LCFS queue at a random time instance, it holds that*

$$\mathrm{E}[Y_i^p] = \frac{\mathrm{E}[X_{k_i}^{p+1}]}{(p+1)\,\mathrm{E}[X_{k_i}]}, \tag{10}$$

where $X = X_{k_i}$ denotes the initial service time of the class-k_i customer, independent of the current number of customers.

Proof. The result is well known and holds for an arbitrary *symmetric queue* [20], such as the preemptive LCFS and PS queues. Alternatively, an elementary result from the renewal theory for residual/excess life(time) states that the pth moment of the conditional remaining service time Y of a class-k customer n *currently* receiving service, e.g. in a preemptive LCFS queue at a random time instance, is given by [22], [31]

$$\mathrm{E}[Y^p] = \frac{\mathrm{E}[X_k^{p+1}]}{(p+1)\,\mathrm{E}[X_k]}. \tag{11}$$

As the service of the customers waiting in a preemptive LCFS queue has been interrupted at a random time instance, the same random observer property also holds for them.

We note that (11) also gives the remaining service time for an M/G/1-FCFS queue under the condition that the server is busy. However, the number of customers waiting under FCFS provides additional information about Y, yielding Lemma 3. For LCFS, (5) still holds and (5) or (10), depending on the available information (time or number aware), can be substituted

https://doi.org/10.1239/jap/1354716657 Published online by Cambridge University Press

into (8) or (9), yielding compact expressions for the particular value functions. Next we again consider the important special case of a system with unit holding cost.

4.2. Single class and sojourn time

Consider next the case of a single class, $X_i \sim X$, with unit holding costs (i.e. sojourn time). A sufficient state description for a time-aware system is $z = (s_1, \dots, s_n)$, where s_i denotes the amount of service customer i has received and n is the number of customers present. For a number-aware system, the state description is simply $z = (n)$. Result (9) gives the following corollary.

Corollary 4. (Time-aware preemptive M/G/1-LCFS with unit holding cost.) *The value function for a single-class preemptive M/G/1-LCFS queue is given by*

$$v_z - v_0 = \frac{1}{1-\rho} \sum_{i=1}^{n} i \, \mathrm{E}[X - s_i \mid X > s_i]. \tag{12}$$

Corollary 5. (Number-aware preemptive M/G/1-LCFS with unit holding cost.) *It holds that*

$$v_n - v_0 = \frac{n(n+1)\,\mathrm{E}[X^2]}{4(1-\rho)\,\mathrm{E}[X]}. \tag{13}$$

Remark 2. For $X \sim \mathrm{Exp}(\mu)$, the lack-of-memory property implies that $\mathrm{E}[X - s_i \mid X > s_i] = 1/\mu$, the number in queue n describes the state, and both (12) and (13) reduce again to $v_n - v_0 = n(n+1)/2(\mu - \lambda)$, in agreement with [1] and [6].

Finally, we note that the time-aware value functions of the preemptive M/G/1-LCFS queue involve the mean and the conditional expectation of the service time X, while the value function of a number-aware system depends on the first two moments of the service time.

5. Nonpreemptive LCFS queue

In a nonpreemptive LCFS queue, a customer receiving service will not be replaced by later arrivals, and, hence, at most one customer has received service. That is, a new customer is given the first waiting place in the queue if the system is nonempty; otherwise, the service starts immediately. The size-aware value function is again available from [17] and given as follows.

Lemma 8. (Size-aware nonpreemptive M/G/1-LCFS.) *The size-aware relative value of state* $(\Delta_1, \dots, \Delta_n)$ *with respect to the sojourn time in an M/G/1 queue with a nonpreemptive LCFS discipline is given by*

$$v_{(\Delta_1,\dots,\Delta_n)} - v_0 = \frac{1}{1-\rho} \sum_{i=1}^{n} (i - \rho)\Delta_i + \frac{\lambda}{2(1-\rho)} \sum_{i=1}^{n} \Delta_i^2, \tag{14}$$

where task Δ_n *is served first and task* Δ_1 *is served last,* $\rho = \lambda\,\mathrm{E}[X]$ *with* λ *denoting the Poisson arrival rate, and* $\mathrm{E}[X]$ *is the mean service size.*

Similarly, the corresponding value function can also be deduced for arbitrary holding costs.

Lemma 9. (Size-aware nonpreemptive M/G/1-LCFS with arbitrary holding costs.) *It holds that*

$$v_{(\Delta_1,\dots,\Delta_n)} - v_0 = \frac{1}{1-\rho} \sum_{i=1}^{n} \biggl(\Delta_i \biggl(\sum_{j=1}^{i} a_j - \rho a_i \biggr) \biggr) + \frac{\lambda\,\mathrm{E}[A]}{2(1-\rho)} \sum_{i=1}^{n} \Delta_i^2,$$

https://doi.org/10.1239/jap/1354716657 Published online by Cambridge University Press

where $\mathrm{E}[A]$ *again denotes the mean holding cost of a customer (assumed to be i.i.d.) and the* a_i *denote the known holding costs of the present n customers.*

Let $Y_1, \ldots, Y_n$ denote the remaining service times of the n customers present in state z. A convenient state description for a time-aware system is $z = ((k_1, a_1), \ldots, (k_n, a_n), s)$, where n denotes the number in the system, k_i denotes the class of customer i, a_i denotes the holding cost of customer i, and s is the amount of service (if any) customer n has received. For a number-aware system, $z = ((k_1, a_1), \ldots, (k_n, a_n))$. For an empty system, we write $z := 0$. Thus, for customer i, $i = 1, \ldots, n-1$, currently waiting in the queue, the remaining service time is the original, $Y_i \sim X_{k_i}$, and Y_n denotes the remaining service time of the customer currently receiving service. The mean and the second moment of Y_n depend on the available information, i.e. whether the system is number or time aware (or something else).

Proposition 3. *The value function for a multiclass nonpreemptive M/G/1-LCFS queue with arbitrary holding costs is given by*

$$v_z - v_0 = \frac{1}{1-\rho} \sum_{i=1}^{n} \left(\mathrm{E}[Y_i] \left(\sum_{j=1}^{i} a_j - \rho a_i \right) \right) + \frac{\lambda\, \mathrm{E}[A]}{2(1-\rho)} \left(\sum_{i=1}^{n} \mathrm{E}[Y_i^2] \right). \tag{15}$$

Proof. Equation (15) is obtained from (14) by conditioning on the (remaining) service times.

Corollary 6. *For unit holding costs, the value function of a multiclass nonpreemptive M/G/1-LCFS queue is given by*

$$v_z - v_0 = \frac{1}{1-\rho} \sum_{i=1}^{n} (i-\rho)\, \mathrm{E}[Y_i] + \frac{\lambda}{2(1-\rho)} \left(\sum_{i=1}^{n} \mathrm{E}[Y_i^2] \right).$$

5.1. Remaining service time

The remaining service time in the time-aware case is again given by Lemma 4. For the number-aware system, the queueing order has no bearing on the remaining service time of the customer currently receiving service, and, hence, Lemma 3 also holds in this case.

5.2. Single class and sojourn time

The queueing discipline has no bearing on the value function in a single-class nonpreemptive system as all customers waiting in the queue are statistically equivalent and incur costs in the same manner. Thus, the results already given for the single-class FCFS queue in Section 3.2 also hold for the nonpreemptive LCFS. In fact, those results are valid for the general class of all nonpreemptive queueing disciplines.

6. Dispatching problem

Let us next consider dispatching problems, which are a prime application of single-server queues' value functions. In a basic dispatching problem, all customers arrive to a dispatcher which assigns them to one of the available servers upon arrival. The dispatching decision is irrevocable. The customers in each queue are served according to a given queueing discipline. In this paper we consider FCFS and LCFS disciplines.

The well-known dispatching policies that we consider in the numerical examples are as follows.

(i) Join the shortest queue (JSQ), which assigns a customer to the queue with the least number of customers. The ties are broken in favour of a queue with a 'smaller id'.

https://doi.org/10.1239/jap/1354716657 Published online by Cambridge University Press

(ii) Bernoulli splitting policies (RND) assign customers randomly using some probability distribution. RND-ρ uses a probability distribution that balances the load.

(iii) Size-interval-task-assignment (SITA) policies assign customers based on the service requirement, which is thus assumed to be available. SITA-e uses such service time intervals that balance the load across the servers.

(iv) Least-work-left (LWL) policies assign a customer to the queue with the least amount of outstanding work (backlog). LWL− considers the backlog before the assignment, and LWL+ afterwards.

(v) The myopic policy assigns a new customer to the queue that minimizes the costs under the assumption that no additional customers arrive. Under FCFS, it is equivalent to LWL+.

Policy iteration. A dispatching policy is considered to be state independent if a decision is independent of the state of the queues; otherwise, it is state dependent. The analysis of a complete system is difficult in general. However, when a dispatching policy is state independent, the arrival process to each queue remains Poisson and the queues can be analyzed independently. In particular, we obtain the value function of the complete system as a sum of the queue specific value functions,

$$v_z = v_{z_1} + \cdots + v_{z_m}.$$

This enables the FPI of the Markov decision process framework. In FPI, the idea is to improve a given *basic policy* for which the value function is available. For each decision, one deviates from the default action of the basic policy if the expected cost with some other action is smaller, thereby decreasing the expected cumulative cost globally in the infinite time horizon. The expected cost of choosing queue i, referred to as the *admittance cost*, is

$$w_i := v_{z_i \oplus \boldsymbol{x}} - v_{z_i},$$

where $\boldsymbol{x}$ denotes the information regarding the new customer, z_i denotes the information of the current state of queue i, and $z_i \oplus \boldsymbol{x}$ is the resulting state with the new customer added to queue i. The value functions are determined by some basic policy. Hence, w_i represents the mean additional cost if a given new customer is assigned to queue i.

This general approach can be traced back to Norman [29], and has since been applied in numerous papers (see the prior work in Section 1.1).

Server specific service times. We consider a *heterogeneous system* comprising m queues having service rates c_i. The total capacity of the system is $c_{\text{tot}} = \sum_{i=1}^{m} c_i$. Otherwise, the servers are assumed to have *no special traits*, i.e. the service time of a customer with size x in queue i is simply x/c_i. In passing, we note that the present framework also lends itself to cases where some servers can process certain customers more efficiently than others. Moreover, the holding costs could be server specific.

6.1. FCFS servers

Consider first a system of parallel FCFS queues. For a single queue ($c_i = c$), the admittance cost is

$$w = \frac{\lambda\,\mathrm{E}[A]}{2c^2(1-\rho)}(\mathrm{E}[X^2] + 2\,\mathrm{E}[X]\,\mathrm{E}[U]) + \frac{a}{c}(\mathrm{E}[X] + \mathrm{E}[U]),$$

where λ, $\mathrm{E}[A]$, and ρ denote the arrival rate, the mean holding cost, and the offered load for a given queue according to a state-independent basic policy, respectively, and c denotes the

https://doi.org/10.1239/jap/1354716657 Published online by Cambridge University Press

service rate, U denotes the amount of unfinished work, $U = Y_1 + \cdots + Y_n$, and (X, a) denotes the service requirement and holding cost of the new customer, so that U/c corresponds to the current backlog in time, and X/c corresponds to the service time of the new customer. Note that here X refers to the 'size', not to the service time as in the previous sections. The first term corresponds to the expected extra cost future arrivals incur because of a new customer, and the latter term corresponds to the cost a given new customer incurs.

Single class and unit holding cost. With unit holding cost, $A = a = 1$ and the objective is to minimize the mean sojourn time. In this case, the admittance cost reduces to

$$w = \frac{\lambda}{2c^2(1-\rho)}(\mathrm{E}[X^2] + 2\,\mathrm{E}[X]\,\mathrm{E}[U]) + \frac{1}{c}(\mathrm{E}[X] + \mathrm{E}[U]).$$

Special care must be taken when applying the single-queue FCFS results. As mentioned, the number-aware result is valid for a single M/G/1-queue, and, thus, can be utilized to compute the relative value of a state-independent policy. However, if the earlier decisions were not taken according to the basic policy, then the *remaining service times*, or even the *initial distribution of customers entering a given queue*, may be different. For time-aware cases, the knowledge of the service received is sufficient to avoid such problems. Similarly, in size-aware cases, the current state is exactly known and it does not matter how one has ended up there. Therefore, in this section we consider (i) time-aware FPI-RND-ρ, (ii) size-aware FPI-RND-ρ, and (iii) size-aware FPI-SITA-e. These, and the chosen reference policies, are summarized in Table 1.

Numerical example. For the numerical results, we consider two systems: (i) a two-server system having rates $c_1 = 1$ and $c_2 = 0.5$, and (ii) a three-server system having rates $c_1 = 1$ and $c_2 = c_3 = 0.25$, so that the total capacity in both cases is 1.5. We consider three service requirement distributions: uniform $\mathrm{U}(0.5, 1.5)$, exponential with mean 1, and a bounded Pareto distribution with the complementary cumulative distribution function

$$\mathrm{P}\{X > x\} = \frac{(a/x)^{\alpha} - (a/b)^{\alpha}}{1 - (a/b)^{\alpha}}, \qquad 0 < a < b \text{ and } a \le x \le b,$$

where $a = 0.339\,59$, $b = 1000$, and $\alpha = 1.5$, so that the mean is also about 1.

Simulation results are depicted in Figure 1. The state-independent RND-ρ and SITA-e are suboptimal with light-tailed service times, but with heavy-tailed service times, SITA-e starts to shine. LWL+ (myopic) shows remarkably good performance overall (and is easy to implement), while the size-aware FPI policy based on SITA-e achieves the lowest mean sojourn time almost everywhere.

Table 1: Sufficient state information for FCFS policies assuming a single class and unit holding cost.

	Arrival rate λ	Service time distribution	Number in the system	Service received	Backlog U	New customer
RND-ρ						
SITA-e		√				Size
JSQ			√			
LWL−					√	
LWL+ / myopic					√	Size
Time-aware FPI-RND-ρ	√	√	√	√		Mean size
Size-aware FPI-RND-ρ	√	Mean			√	Size
Size-aware FPI-SITA-e	√	Mean			√	Size

https://doi.org/10.1239/jap/1354716657 Published online by Cambridge University Press

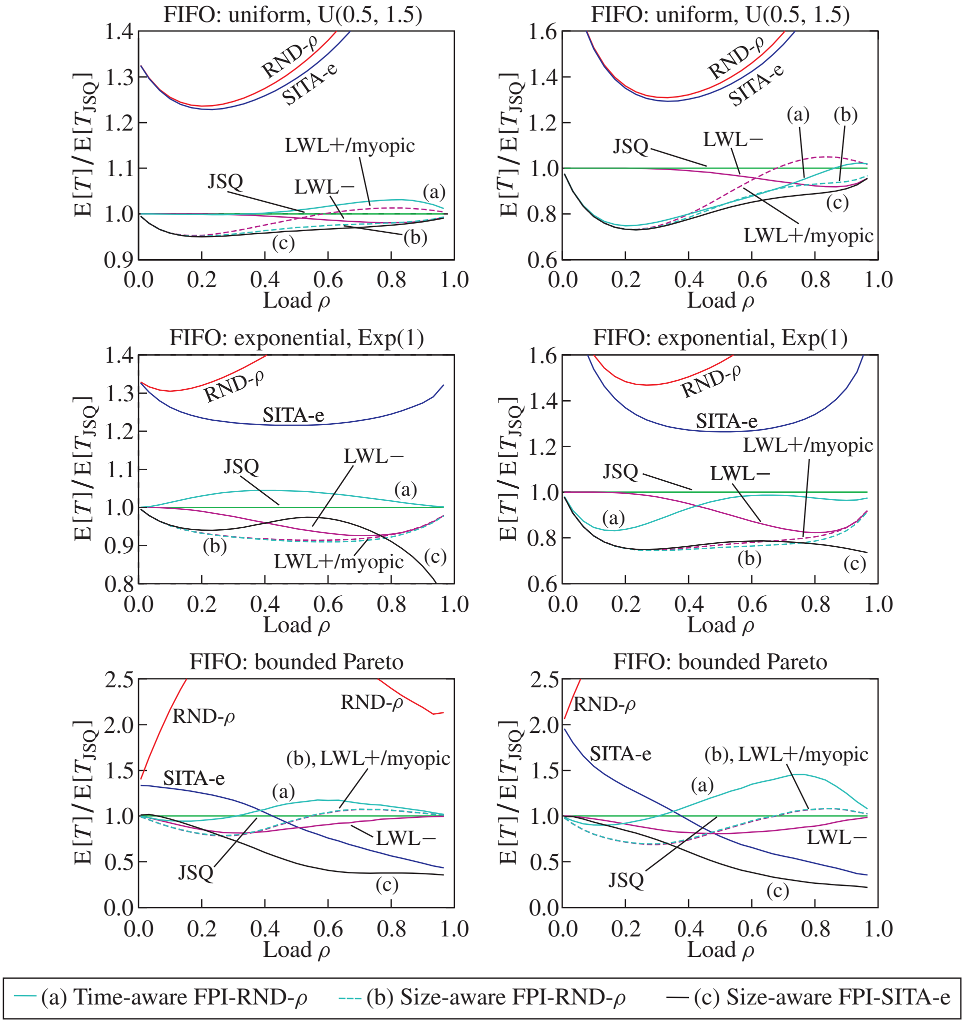

FIGURE 1: Simulation results with FCFS. *Left*: two servers with rates $c_1 = 1$ and $c_2 = 0.5$. *Right*: three servers with rates $c_1 = 1$ and $c_2 = c_3 = 0.25$.

6.2. Preemptive LCFS servers

The second example is basically the same, but here the queues follow the preemptive LCFS discipline. For a single preemptive LCFS queue, the conditional mean sojourn time for a customer with service time x is

$$\mathrm{E}[T(x)] = \frac{x}{1-\rho},$$

and, thus, RND-ρ and SITA-e, both balancing the load among the servers, achieve the same performance with respect to the sojourn time. The admittance cost to a single LCFS queue is given by Proposition 2 as

$$w = \frac{a_1 + \cdots + a_n + a}{c(1-\rho)}\,\mathrm{E}[X],$$

https://doi.org/10.1239/jap/1354716657 Published online by Cambridge University Press

where $\mathrm{E}[X]$ in our case (no special traits) is a common constant to all queues. Consequently, the (expected) size of the new customer has no bearing on the FPI decision.

Furthermore, the admittance cost depends solely on the number in the queue (together with the individual holding cost rates), which is assumed to be known in all three cases (number, time, and size aware), making the FPI policy for the preemptive LCFS robust in this sense. With a load balancing policy (e.g. RND-ρ and SITA-e), the offered load to each queue is equal, $\rho_i = \rho_j$. Therefore, the number-, time-, and size-aware FPI policies for an arbitrary load balancing basic policy are given by

$$\alpha_z = \underset{i=1,\dots,m}{\operatorname{argmin}} \frac{a_1 + \cdots + a_{n_i} + a}{c_i},$$

which is the same decision that the myopic policy would make.

Single class and unit holding cost. For a single class, unit holding cost and an arbitrary load balancing basic policy (e.g. RND-ρ and SITA-e), the above reduces to

$$\alpha_z = \underset{i=1,\dots,m}{\operatorname{argmin}} \frac{n_i + 1}{c_i}.$$

In practice, it is beneficial to favour an empty queue in case of ties. If the basic policy were, e.g. optimal Bernoulli split, then the faster queue i would have a higher load than a slower queue j, i.e. $\rho_i > \rho_j$, and for states where $(n_i + 1)/c_i = (n_j + 1)/c_j$, the FPI decision would choose the slower queue j. Hence, we also resolve the ties in favour of a slower server in the case of RND-ρ and SITA-e basic policies that balance the load. For identical servers, the above is equivalent to JSQ.

Numerical example. Here we consider the same server settings and size distributions as with FCFS. The chosen dispatching policies are given in Table 2. The simulation results are depicted in Figure 2. The results are somewhat mixed in the two-server scenario. With uniform service requirements, all state-dependent policies behave similarly. With heavy-tailed service requirements, the performance of the LWL policies also drops significantly. The FPI-based approach yields only a slightly better performance than JSQ and another variant of the myopic.

In the setting of three servers, the situation is somewhat more asymmetric as the secondary servers are four times slower than the primary server. In this case, the FPI-based approach (i.e. myopic) is a clear winner, even if the ties are resolved suboptimally by choosing a faster server. JSQ has a steady performance with a maximum deviation of about 20% from FPI/myopic.

Note that a system with nonpreemptive LCFS queues in an otherwise similar setting is effectively the same as a system with FCFS queues, and is thus omitted here.

Table 2: Sufficient state-information for LCFS policies assuming a single class and unit holding cost.

	Service time distribution	Number in the system	Backlog U	New customer
RND-ρ				
SITA-e	$\surd$			Size
JSQ		$\surd$		
LWL$-$			$\surd$	
LWL+			$\surd$	Size
Myopic / FPI (equal load)		$\surd$		

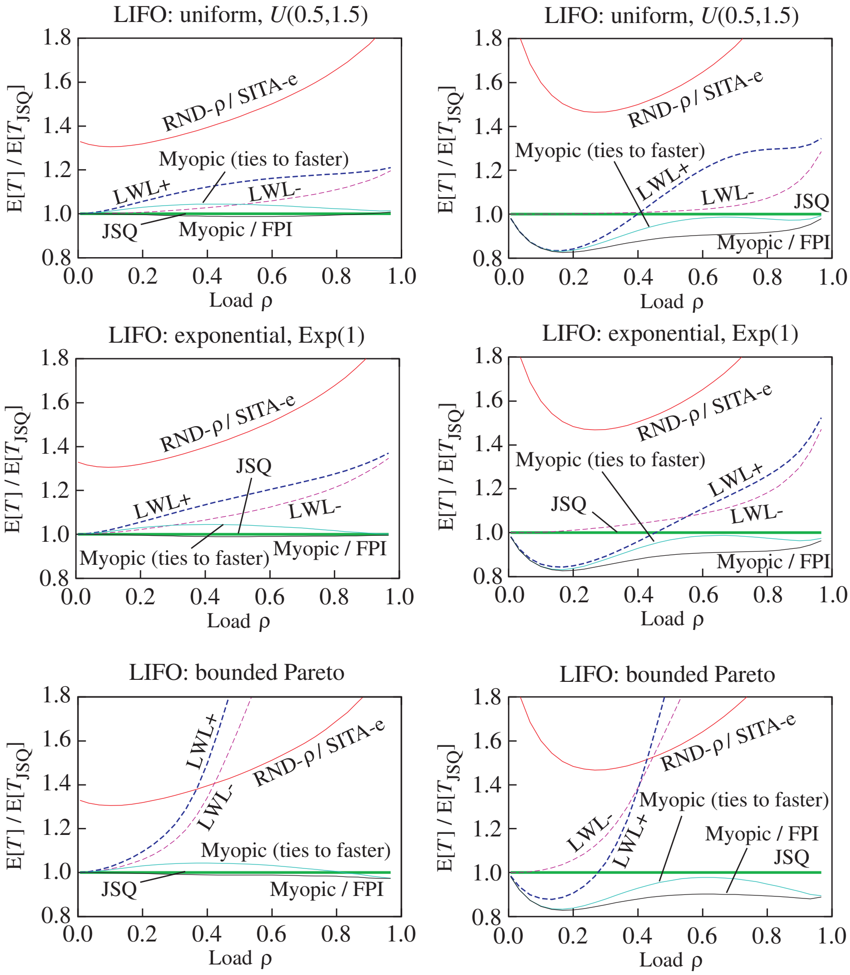

FIGURE 2: Simulation results with preemptive LCFS. *Left*: two servers with rates $c_1 = 1$ and $c_2 = 0.5$. *Right*: three servers with rates $c_1 = 1$ and $c_2 = c_3 = 0.25$.

6.3. Multiclass FCFS servers with dedicated input

In our final example, we consider a more complicated two-server system, where, in addition to *flexible* customers arriving at rate λ, both servers also receive a dedicated stream of inflexible customers arriving at rates λ_1 and λ_2, respectively. The two servers are identical, $c_1 = c_2 = 1$. However, service requirements in each arrival stream are assumed to obey exponential distributions with means μ_1, μ_2, and μ, respectively. Due to the lack-of-memory property of the exponential distribution, the time- and number-aware systems are equivalent. The system is illustrated in Figure 3. The same setting has been considered in [3] in the context of 'power-of-two', which again refers to the situation where, for each customer, the choice has to be made between two randomly chosen servers [28].

One reasonable state-independent basic policy for the Figure 3 system balances the load between the two servers. Let ρ_1, ρ_2, and ρ denote the corresponding offered loads, $\rho_i = \lambda_i/\mu_i$ and $\rho = \lambda/\mu$. If $\rho_1 < \rho_2 + \rho < \rho_1 + 2\rho$ then an exact load balancing can be achieved

https://doi.org/10.1239/jap/1354716657 Published online by Cambridge University Press

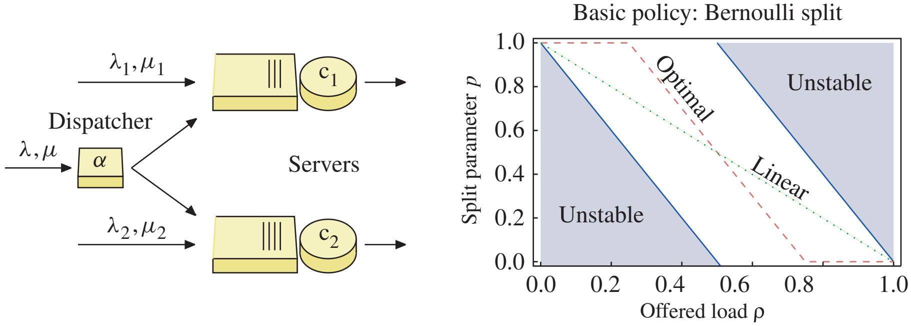

FIGURE 3: *Left*: a multiclass system where servers receive dedicated streams and only for a certain fraction of customers a server can be selected. *Right*: the feasible region for a random split probability P{queue 1} $= p_1$ assuming $c_1 = c_2 = 1$, $\rho_1 + \rho_2 = 1$, and $\rho = 0.5$. The dashed line corresponds to the optimal random split probability.

by an appropriate Bernoulli split. Otherwise, the basic policy assigns the flexible customers categorically to the server with a lower load. In either case, let p_1 and p_2 denote the fraction of flexible customers assigned to server 1 and server 2, respectively, and let ρ_i^* denote the resulting total load in server i, $\rho_i^* = \rho_i + p_i \rho$.

Assuming a number-aware system (knowledge of the service received would provide no additional information in this case) where one knows the class of each customer, the value function for queue i is given by (4) and reads

$$v_z - v_0 = \frac{\lambda_i + p_i \lambda}{1 - \rho_i^*} \Bigg(\sum_{j=1}^{n_i} \bigg(\frac{1}{\mu_{i,j}} \bigg)^2 + \sum_{j=1}^{n_i - 1} \sum_{k=j+1}^{n_i} \frac{1}{\mu_{i,j}} \frac{1}{\mu_{i,k}} \Bigg) + \sum_{j=1}^{n} j \frac{1}{\mu_{i,j}}.$$

Therefore, the admittance cost of a new customer at the dispatcher to queue i is

$$w_i = \Bigg(\frac{1}{\mu} + \sum_{j=1}^{n_i} \frac{1}{\mu_{i,j}} \Bigg) \Bigg(1 + \frac{\lambda_i + p_i \lambda}{1 - \rho_i^*} \frac{1}{\mu} \Bigg), \tag{16}$$

where the first factor corresponds to the expected backlog in queue i when the new task has been included. If the arriving tasks have the same mean, $\mu_{i,j} = \mu$, and $\lambda_1 = \lambda_2$, then $w_i = A(n_i + 1)$, where A is some common constant. That is, FPI yields JSQ, which is the optimal policy in this case. Next we will evaluate by means of simulations the FPI policy (16) in genuine multiclass situations.

Assume first that all service requirements obey the same exponential distribution with unit mean, $\mu_1 = \mu_2 = \mu = 1$. Furthermore, $\rho_1 + \rho_2 = 1$ and $\rho = 0.5$, so that the total offered load is 0.75 per server. The balance between the dedicated streams is varied so that $0 < \rho_1 < 1$. The right diagram of Figure 3 illustrates the feasible region for the random split probability p_1 and also the optimal value of p_1. Only when $0.25 \leq \rho_1 \leq 0.75$ can the load between the two servers be evenly distributed. In those cases, FPI yields JSQ when only the number in a queue is available. In the size-aware case, the (remaining) service requirements are also available and we obtain LWL. In symmetric settings, $\rho_1 = \rho_2$, it is irrelevant which queue JSQ chooses in the case of a tie. However, if, for example, $\rho_1 > \rho_2$, then the intuition says that the expected cost of choosing queue 1 is higher than with queue 2. This information can be conveyed to FPI by biasing the random split probabilities a bit in the appropriate direction. The 'linear' choice for p_1 depicted in the right diagram of Figure 3 does exactly that.

https://doi.org/10.1239/jap/1354716657 Published online by Cambridge University Press

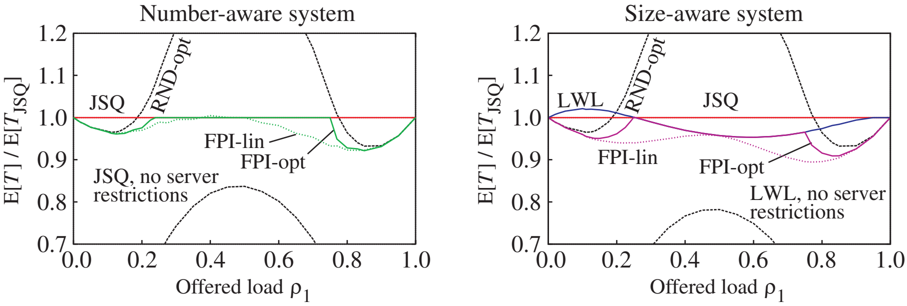

FIGURE 4: Simulation results with identical service requirement distributions. On the left, the dispatcher is aware of the number of customers in each queue, and on the right, the remaining service requirements are also available.

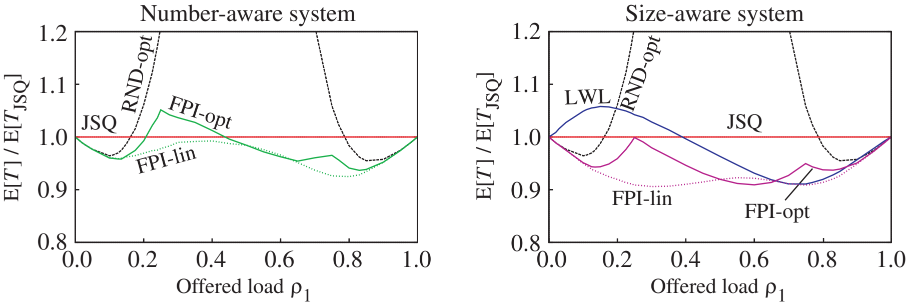

FIGURE 5: Simulation results with $1/\mu_1 = 4$ and $1/\mu_2 = 1/\mu = 1$, i.e. class-1 tasks have a four times longer mean service requirement. On the left, the dispatcher is aware of the number of tasks and their classes in each queue, and on the right, the (remaining) service requirements are also available.

The simulation results are depicted in Figure 4. On the x-axis is the load offered directly to queue 1, $\rho_1 = \lambda_1/\mu_1$, and on the y-axis the sojourn time performance against JSQ, $\mathrm{E}[T]/\mathrm{E}[T_{\mathrm{JSQ}}]$. The FPI-lin and FPI-opt policies are the FPI policies obtained for the linear and the optimal basic policies; see the right diagram of Figure 3.

Note that if $\rho_1 = \rho_2 = 0$, i.e. all customers are flexible and subject to a dispatching decision, then both JSQ and LWL are the optimal policies in the corresponding settings. The JSQ and LWL curves labelled with 'no server restrictions' correspond to such *relaxed* scenarios with $\rho = 1.5$ and, thus, serve as lower bounds. We can observe that when the dedicated streams are approximately equal, $\rho_1 \approx \rho_2$, the difference to the relaxed scenario is reasonably small. Even though the server system itself is symmetric in the sense that the dedicated arrival streams can be interchanged, the complete system, e.g. with JSQ, is not due to the fact that it chooses queue 1 in the case of ties. This explains the lack of symmetry with respect to $\rho_1 = 0.5$ in Figure 4.

When the dedicated arrival rates are highly nonuniform, the FPI policies achieve the highest gains. As explained earlier, when FPI is applied to the optimal Bernoulli split and $0.25 \leq \rho_1 \leq 0.75$, we obtain JSQ and LWL depending on the information available to the dispatcher. In contrast, the 'linear' basic policy consistently yields a better FPI policy, as expected.

In Figure 5 we present the simulation results in the same setting except that tasks arriving directly to queue 1 have a mean service requirement of 4. One can observe that in both

https://doi.org/10.1239/jap/1354716657 Published online by Cambridge University Press

number-aware and size-aware cases an FPI-based approach yields a slightly higher improvement than in the previous case. Only one FPI-based policy, FPI-opt in the number-aware case, is worse than JSQ with some values of ρ_1.

6.4. Higher moments

In our setting, the remaining service times are random variables. Hence, a decision which is optimal on average can still turn out to be wrong once the actual service times are realized. This raises the natural question of whether one can have some control over the uncertainties different decisions involve. It turns out that the same approach also lends itself to results in this direction.

Let the random variable C_i denote the actual additional costs for choosing queue i, where the random component is due to the uncertainty about the current state (remaining service times), while the effect of future arrivals has already been included in the expectations. Choosing the queue with the smallest mean $\mathrm{E}[C_i]$ gives the optimal decision in the long term. However, one can also easily obtain an arbitrary moment $\mathrm{E}[C_i^p]$, and, thus, e.g. the variance for the expected increase in the costs.

To elaborate on this, consider a single-class M/G/1-FCFS queue with unit holding cost, so that

$$C = (2hX + 1)U + hX^2 + X,$$

where $h = \lambda/(2(1-\rho))$, the random variable X is the service time of the new customer, and U is the backlog, $U = X_1 + \cdots + X_n + Y$. Thus, e.g. $\mathrm{V}[C]$ can be easily computed. For preemptive M/G/1-LCFS, the situation is even simpler, and we obtain

$$C = \frac{n+1}{1-\rho} X.$$

Hence, the cost obeys the same distribution as the service time of the new customer, and, e.g. $\mathrm{E}[C] = (n+1)/(1-\rho)\,\mathrm{E}[X]$ and $\mathrm{V}[C] = (n+1)^2/(1-\rho)^2\mathrm{V}[X]$.

7. Conclusions

In this paper we have given value functions for the M/G/1 queue with FCFS and LCFS disciplines with arbitrary holding costs. The state description is assumed to comprise the number in the queue and some additional information regarding the remaining service times. The obtained results are compact closed-form expressions, and have similarities with the Pollaczek–Khinchine mean value formula. These results generalize several earlier results, and enable efficient dispatching rules for a broad range of dispatching problems, where arriving customers are assigned to parallel queues so as to minimize the mean cost rate. In particular, the results enable one to develop dispatching policies for multiclass scenarios and for cases with different types of additional information regarding the remaining service times.

Acknowledgements

The authors would like to thank the anonymous referee for valuable comments and suggestions. This work was supported in part by the Academy of Finland in the RESMAN project (grant number 134363).

https://doi.org/10.1239/jap/1354716657 Published online by Cambridge University Press

References

[1] Aalto, S. and Virtamo, J. (1996). Basic packet routing problem. In *The Thirteenth Nordic Teletraffic Seminar* (Trondheim, Norway, August 1996), pp. 85–97.

[2] Adan, I. and Haviv, M. (2008). Conditional ages and residual service times in the M/G/1 queue. Tech. Rep. 2008–023, EURANDOM.

[3] Akgun, O., Righter, R. and Wolff, R. (2011). The power of partial power of two choices. In *ACM SIGMETRICS*, ACM, New York, pp. 46–48.

[4] Ansell, P. S., Glazebrook, K. D. and Kirkbride, C. (2003). Generalised 'join the shortest queue' policies for the dynamic routing of jobs to multiclass queues. *J. Operat. Res. Soc.* **54,** 379–389.

[5] Becker, K. J. *et al.* (2000). Allocation of tasks to specialized processors: a planning approach. *Europ. J. Operat. Res.* **126,** 80–88.

[6] Bhulai, S. (2006). On the value function of the M/Cox(r)/1 queue. *J. Appl. Prob.* **43,** 363–376.

[7] Bonomi, F. (1990). On job assignment for a parallel system of processor sharing queues. *IEEE Trans. Comput.* **39,** 858–869.

[8] Conolly, B. W. (1984). The autostrada queueing problem. *J. Appl. Prob.* **21,** 394–403.

[9] Crovella, M. E., Harchol-Balter, M. and Murta, C. D. (1998). Task assignment in a distributed system: improving performance by unbalancing load. In *Proc. SIGMETRICS '98*, ACM, New York, pp. 268–269.

[10] Ephremides, A., Varaiya, P. and Walrand, J. (1980). A simple dynamic routing problem. *IEEE Trans. Automatic Control* **25,** 690–693.

[11] Fakinos, D. (1982). The expected remaining service time in a single server queue. *Operat. Res.* **30,** 1014–1018.

[12] Feng, H., Misra, V. and Rubenstein, D. (2005). Optimal state-free, size-aware dispatching for heterogeneous M/G/-type systems. *Performance Evaluation* **62,** 475–492.

[13] Gupta, V., Harchol-Balter, M., Sigman, K. and Whitt, W. (2007). Analysis of join-the-shortest-queue routing for web server farms. *Performance Evaluation* **64,** 1062–1081.

[14] Harchol-Balter, M., Crovella, M. E. and Murta, C. D. (1999). On choosing a task assignment policy for a distributed server system. *J. Parallel Distributed Comput.* **59,** 204–228.

[15] Harchol-Balter, M., Scheller-Wolf, A. and Young, A. R. (2009). Surprising results on task assignment in server farms with high-variability workloads. In *Proc. of SIGMETRICS '09*, ACM, New York, pp. 287–298.

[16] Harchol-Balter, M., Sigman, K. and Wierman, A. (2002). Asymptotic convergence of scheduling policies with respect to slowdown. *Performance Evaluation* **49,** 241–256.

[17] Hyytiä, E., Penttinen, A. and Aalto, S. (2012). Size- and state-aware dispatching problem with queue-specific job sizes. *Europ. J. Operat. Res.* **217,** 357–370.

[18] Hyytiä, E., Penttinen, A., Aalto, S. and Virtamo, J. (2011). Dispatching problem with fixed size jobs and processor sharing discipline. In *Proc. 23rd Internat. Teletraffic Congress*, pp. 190–197.

[19] Hyytiä, E., Virtamo, J., Aalto, S. and Penttinen, A. (2011). M/M/1-PS queue and size-aware task assignment. *Performance Evaluation* **68,** 1136–1148.

[20] Kelly, F. P. (1979). *Reversibility and Stochastic Networks*. John Wiley, Chichester.

[21] Kim, J. H., Ahn, H.-S. and Righter, R. (2011). Managing queues with heterogeneous servers. *J. Appl. Prob.* **48,** 435–452.

[22] Kleinrock, L. (1975). *Queueing Systems*, Vol. I. Wiley-Interscience, New York.

[23] Krishnan, K. R. (1987). Joining the right queue: a Markov decision rule. In *Proc. 28th Conf. Decision and Control*, pp. 1863–1868.

[24] Krishnan, K. R. and Ott, T. J. (1986). State-dependent routing for telephone traffic: theory and results. In *Proc. 25th Conf. Decision Control*, pp. 2124–2128.

[25] Liu, Z. and Righter, R. (1998). Optimal load balancing on distributed homogeneous unreliable processors. *Operat. Res.* **46,** 563–573.

[26] Liu, Z. and Towsley, D. (1994). Optimality of the round-robin routing policy. *J. Appl. Prob.* **31,** 466–475.

[27] Mandelbaum, A. and Yechiali, U. (1979). The conditional residual service time in the M/G/1 queue. Unpublished manuscript. Available at http://www.math.tau.ac.il/∼uriy/Papers/conditional.pdf.

[28] Mitzenmacher, M. (2001). The power of two choices in randomized load balancing. *IEEE Trans. Parallel Distributed Systems* **12,** 1094–1104.

[29] Norman, J. M. (1972). *Heuristic Procedures in Dynamic Programming*. Manchester University Press.

[30] Puterman, M. L. (2005). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley, New York.

[31] Ross, S. M. (2000). *Introduction to Probability Models*, 7th edn. Academic Press, Burlington, MA.

[32] Sassen, S. A. E., Tijms, H. C. and Nobel, R. D. (1997). A heuristic rule for routing customers to parallel servers. *Statist. Neerlandica* **51,** 107–121.

https://doi.org/10.1239/jap/1354716657 Published online by Cambridge University Press

[33] Schwartz, B. L. (1974). Queuing models with lane selection: a new class of problems. *Operat. Res.* **22,** 331–339.

[34] Van Leeuwaarden, J., Aalto, S. and Virtamo, J. (2001). Load balancing in cellular networks using first policy iteration. Tech. Rep., Networking Laboratory, Helsinki University of Technology.

[35] Weber, R. R. (1978). On the optimal assignment of customers to parallel servers. *J. Appl. Prob.* **15,** 406–413.

[36] Whitt, W. (1986). Deciding which queue to join: some counterexamples. *Operat. Res.* **34,** 55–62.

[37] Winston, W. (1977). Optimality of the shortest line discipline. *J. Appl. Prob.* **14,** 181–189.

https://doi.org/10.1239/jap/1354716657 Published online by Cambridge University Press