Performance Evaluation 114 (2017) 32–44

Contents lists available at ScienceDirect

Performance Evaluation

journal homepage: www.elsevier.com/locate/peva

Dispatching fixed-sized jobs with multiple deadlines to parallel heterogeneous servers

Esa Hyytiä [a,c,*], Rhonda Righter [b], Olivier Bilenne [c], Xiaohu Wu [c]

[a] Department of Computer Science, University of Iceland, Iceland
[b] Department of Industrial Engineering and Operations Research, UC Berkeley, United States
[c] Department of Communications and Networking, Aalto University, Finland

ARTICLE INFO

Article history:
Available online 25 May 2017

Keywords:
Dispatching problem
Parallel computing
Deadlines
M/D/1
MDP

ABSTRACT

We study the M/D/1 queue when jobs have firm deadlines for waiting (or sojourn) time. If a deadline is not met, a job-specific deadline violation cost is incurred. We derive explicit value functions for this M/D/1 queue that enable the development of efficient cost-aware dispatching policies to parallel servers. The performance of the resulting dispatching policies is evaluated by means of simulations.

1. Introduction

In the dispatching problem, each arriving job is routed to one of the available servers immediately upon arrival. Even though a single fast server would often be preferred, the parallel servers are needed to match increasing capacity demands. Moreover, short latency, in the absence of preemptive scheduling, requires parallel servers.

In this paper, we consider a cost structure based on (firm) deadlines. Each job has a certain deadline for the maximum waiting time it can tolerate. If this waiting time is exceeded, a deadline violation cost is incurred, but the job must still be served. This cost structure stems from quality-of-experience metrics, where customers observe a good service level whenever the waiting time is "short", but as soon as a given customer-specific threshold is exceeded, the observed service quality drops. That is, the tails of the response time distribution are one of the most crucial performance measures [1]. Similarly, service level agreements (SLAs) are often defined in terms of acceptable waiting times [2].

This basic setting has been studied recently in [3] in the context of M/G/1 queues. However, the results given there are either asymptotic or in the form of differential equations. In contrast, here we derive exact closed-form expressions (that satisfy the aforementioned differential equations and asymptotic behavior). More specifically, the main contributions of this paper are the first exact results for the value function and admission cost for the M/D/1 queue subject to a general deadline-based cost structure. Even though the service times are assumed to be fixed, the deadlines and their violation costs can vary according to some probability distributions. Moreover, there can be multiple deadlines with added cost for each deadline that is violated.

The approach itself is general, and traditionally the objective is the minimization of the mean sojourn time (see, e.g., [4–6]), possibly combined with the energy consumption (see, e.g., [7,8]). The value function for M/G/1-FCFS then enjoys elementary closed-form expressions. However, e.g., the processor sharing (PS) scheduling makes the situation more complex and exact results are available only for M/D/1-PS and M/M/1-PS [4,9]. The approach lends itself also to minimization of blocked jobs in loss systems [10].

* Corresponding author at: Department of Computer Science, University of Iceland, Iceland.
E-mail address: esa@hi.is (E. Hyytiä).

http://dx.doi.org/10.1016/j.peva.2017.04.003

2. Basic model and notation

The basic model for a single M/D/1-FCFS queue with deadlines is as follows. We let λ denote the arrival rate and d the constant service time of a job in the M/D/1 queue so that the offered load is $\rho = \lambda d$. Jobs whose waiting time in queue, W, reach time τ, referred to as the deadline, incur a unit cost. This is equivalent to having a deadline $\tau + d$ for the sojourn time. We assume that $\rho < 1$ for stability, and the deadline must be positive to be meaningful, $\tau > 0$. The mean cost rate is

$$r = \lambda \, \mathrm{P}\{W \geq \tau\}. \tag{1}$$

In general, the distribution of the waiting time cannot be expressed in simple terms, but instead in the form of the Laplace–Stieltjes Transform (LST) [11] or an infinite sum involving convolutions [12]. However, for the M/D/1 queue the waiting time distribution is available [13–15]

$$\mathrm{P}\{W \leq \tau\} = (1-\rho) \sum_{i=0}^{\lfloor \tau/d \rfloor} \frac{(\lambda(id-\tau))^i}{i!} e^{-\lambda(id-\tau)}. \tag{2}$$

In the general case, we have multiple classes of jobs, each with its own arrival rate λ_i, target deadline τ_i and i.i.d. deadline violation cost H_i. The total arrival rate is $\lambda = \sum_i \lambda_i$, and the stability requirement is that $\lambda d = \rho < 1$. The mean cost rate in this case is

$$r = \sum_i \lambda_i \, \mathrm{E}[H_i] \, \mathrm{P}\{W \geq \tau_i\}.$$

Our first task is to derive the so-called value function with respect to the deadline cost structure. Formally, the value function is defined as

$$v(u) \triangleq \lim_{t \to \infty} \mathrm{E}[V(u,t) - rt],$$

where u is the current backlog (unfinished work) in the queue, and the random variable $V(u,t)$ denotes the deadline violation costs during time $(0,t)$ when the system is initially in state u. Given $\rho < 1$, the M/D/1 queue is stable, the system is ergodic, and the above limit is well-defined. (In fact, the limit is finite and well-defined also when $\rho \geq 1$ and the system is unstable.)

The M/G/1 queue has been analyzed in [3] in the context of the basic (single-class) cost structure. In particular, it is shown that the value function is a linear function of u for $u > \tau$, and for $0 \leq u \leq \tau$, $v(u)$ satisfies an integro-differential equation that can be solved numerically. Moreover, explicit results are given for M/G/1 when (i) $\tau < X$ and the load $\rho < 1$, and when (ii) $\tau \gg X$ and $\rho \to 1$ (the heavy-traffic regime), where X denotes the (random) service time. These two results naturally also hold for the corresponding M/D/1 queues.

In contrast, we analyze the general case when $\rho < 1$ and τ is arbitrary, and obtain an explicit closed-form expression for the value function. Moreover, we give the value function for the general multi-class case, and experiment with various dynamic dispatching policies obtained through one policy improvement step.

3. M/D/1 with single deadline

In this section, we assume a single deadline τ that applies to all jobs and a unit deadline violation cost, $h = 1$. These results are later generalized to multiple job classes with distinct deadlines in Section 4.

From [3], we know that the value function for $u > \tau$ is a linear function of u, i.e., $v(u) = v(\tau) + v_0(u)$ with

$$v_0(u) = \frac{\lambda - r}{1 - \rho}(u - \tau), \qquad u > \tau. \tag{3}$$

In general, the value function satisfies the following differential equation,

$$v'(u) = -r + \lambda \, \mathbf{1}(u \geq \tau) + \lambda \, \mathrm{E}[v(u+X) - v(u)], \quad u \geq 0, \tag{4}$$

where X denotes the random i.i.d. service time, and $\mathbf{1}(u \geq \tau)$ is 1 if the condition is true and otherwise zero. Additionally, $v'(0) = 0$. These equations can be solved *numerically* as discussed in [3]. However, exact closed-form results have not been available. For M/D/1, the differential equation (4) simplifies:

$$v'(u) + \lambda \, v(u) = -r + \lambda \, \mathbf{1}(u \geq \tau) + \lambda \, v(u+d), \quad u > 0. \tag{5}$$

In general, the mean cost rate r follows from the boundary condition $v'(0) = 0$. However, with M/D/1 we can use (2).