Operations Research Letters 44 (2016) 507–513

Contents lists available at ScienceDirect

Operations Research Letters

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

Routing jobs with deadlines to heterogeneous parallel servers

Esa Hyytiä [a,*], Rhonda Righter [b]

[a] Department of Computer Science, University of Iceland, Iceland

[b] Department of Industrial Engineering and Operations Research, University of California, Berkeley, United States

ARTICLE INFO

Article history:
Received 13 November 2015
Received in revised form
27 May 2016
Accepted 27 May 2016
Available online 5 June 2016

Keywords:
Job dispatching
Deadline
QoE
Parallel processing
SLA
Heavy-traffic approximation

ABSTRACT

We consider a dispatching system, where jobs with deadlines for their waiting times are assigned to FCFS servers immediately upon arrival. The dispatching problem is to choose a server for each job so as to minimize the probability of deadline violation. We derive an efficient deadline-aware policy in the MDP framework by means of policy improvement, analyze it, and evaluate its performance with simulations. We find that the new policy offers significant improvements over traditional heuristic policies.

1. Introduction

In a job dispatching problem, the arriving jobs or customers are routed to parallel servers immediately upon arrival. The routing decision is irrevocable, i.e., a job cannot be moved to another queue later. This setting arises in supercomputing and in many applications run in cloud systems. The dispatching problem itself is a very general dynamic decision problem, where the user must take into account both the current state of the system, as well as the anticipated future requests.

Several well-known dispatching policies can be found in the literature. The join-the-shortest-queue (JSQ) policy was studied by Haight [5] and has since then been shown to be optimal with respect to mean delay in several settings, see, e.g., [13,4]. When the remaining service times are also available, but not the size of the arriving job, the least-work-left (LWL) policy, that assigns the new job to the server with the shortest backlog, is often optimal [3]. The elementary Bernoulli split (RND) chooses the server independently at random, and thus requires no state or any other information. A better policy, that just needs only the last routing decision, is round-robin [10]. Most of the past work has focused on minimizing the (weighted) mean delay.

In contrast, we focus on a cost structure where jobs have certain deadlines. In particular, we assume a *service level agreement* in the form of maximum waiting time in queue. Jobs which end up waiting longer than that incur a fixed penalty cost, and still must be served. The *quality of experience* (QoE) is a closely related notion. Typical cost or reward functions in the QoE context have a sigmoid form: small delays yield a good experience, but above a certain threshold (deadline), the experience suddenly falls to a low quality category [9,1]. For interactive applications, such as large-scale online services provided by Google, Facebook and Amazon, subsecond response times are a common objective [11]. The deadline cost structure assumed here, defined by a single threshold τ for the waiting time, corresponds to a first level approximation for QoE.

Our main contributions are: (i) we derive new theoretical results for the M/G/1 queue subject to a non-linear cost structure (the so-called value function with respect to deadline violations) that enable efficient dynamic dispatching policies; (ii) we observe that the system may become unstable when blindly minimizing the number of deadline violations even when the offered load is $\rho < 1$; and (iii) we show how this can be avoided by a minor modification in the cost structure, and how the resulting extra term shows up in the policy. The new deadline-aware dispatching policies are further evaluated numerically.

2. Model and notation

We consider n parallel FCFS servers, where server i processes jobs at rate θ_i. In the uniform case, we can assume unit service rates, $\theta_i = 1$ for all i. Jobs arrive according to a Poisson process with

* Corresponding author.
E-mail addresses: esa@hi.is (E. Hyytiä), rrighter@ieor.berkeley.edu (R. Righter).

http://dx.doi.org/10.1016/j.orl.2016.05.009

rate λ and have i.i.d. sizes $X_j \sim X$. Thus, the service time of job j if assigned to server i is X_j/θ_i. We assume a service level agreement (SLA) in the form of a maximum waiting time in queue τ, referred to as the (soft) *deadline*. When this deadline is exceeded, $W > \tau$, a fixed cost of 1 is incurred, and the job remains in the system and must be served. In other words, the cost function is a step-function of waiting time w, $c_\tau(w) = I(w > \tau)$, and the mean cost rate is $r_\tau = \lambda\,\mathrm{P}\{W > \tau\}$. This is in contrast to $r_W = \lambda\,\mathrm{E}[W]$, the cost rate with respect to mean waiting time. The dynamic optimization task is to assign each job immediately upon arrival to one of the n servers so as to minimize the long term incurred costs (i.e., the fraction of jobs violating SLA). Jobs cannot be rejected.

3. Analysis of M/G/1-FCFS with deadlines

We start by analyzing a single M/G/1 queue, for which the Pollaczek–Khinchine mean value result gives the mean waiting time. In general, the distribution of the waiting time cannot be expressed in simple terms except for the M/M/1-FCFS,

$$\mathrm{P}\{W > t\} = \rho\, e^{-(\mu-\lambda)t}. \tag{1}$$

3.1. Value functions

As the cost function is non-linear, also the so-called value function becomes non-trivial. Formally, the value function is defined as the expected difference in costs between a system that is initially in a given state $\mathbf{z}$ and a system in equilibrium,

$$v(\mathbf{z}) = \lim_{t\to\infty} \mathrm{E}[V_t(\mathbf{z}) - r\,t],$$

where the random variable $V_t(\mathbf{z})$ denotes the costs the system incurs during $(0, t)$ when initially in state $\mathbf{z}$, and r is the long-run mean cost rate. The value function for the (mean) waiting time is already available from [7],

$$v_W(u) - v_W(0) = \frac{\lambda u^2}{2(1-\rho)}. \tag{2}$$

However, we are interested in the deadline violations. Let us start with the following monotonicity result:

Lemma 1. *The value function $v_\tau(u)$ w.r.t. deadlines is a strictly increasing function of the backlog u.*

Proof. Consider two systems, system 1 initially in state u and system 2 initially in state $u+\delta$, $\delta > 0$. Suppose the two systems receive the same jobs during $(0, t)$. For each such sample path, $V_t(u) \le V_t(u+\delta)$, and for some non-negligible set $V_t(u) < V_t(u+\delta)$. Therefore $v_\tau(u) < v_\tau(u+\delta)$. □

In passing, we note that the same holds also for every cost structure where arriving jobs incur a cost that is a non-decreasing function of the backlog u, the waiting time, obtained with $c(u) = u$, leading to (2).

For M/G/1-FCFS, we have the following result.

Proposition 1. *For $u \ge \tau$, the value function is a linear function of the backlog u,*

$$v_\tau(u) - v_\tau(\tau) = \frac{\lambda - r_\tau}{1-\rho}(u-\tau) = \frac{\lambda\,\mathrm{P}\{W \le \tau\}}{1-\rho}(u-\tau). \tag{3}$$

Proof. By the definition of the value function, $v_\tau(u) = \mathrm{E}[N_A - B_A r_\tau] + v_\tau(\tau)$, where N_A denotes the number of jobs that arrive before the workload in the queue is τ, B_A denotes the length of the corresponding time interval, and r_τ is the mean cost rate. By using the result for the mean busy period in M/G/1 with an initial backlog of u, $\mathrm{E}[B(u)] = u/(1-\rho)$, we have

$$\mathrm{E}[B_A] = \frac{u-\tau}{1-\rho},$$

and from PASTA, $\mathrm{E}[N_A] = \lambda(u-\tau)/(1-\rho)$, which together yield the desired result. □

Referring to (2) and (3), we first note that the tail of $v_\tau(u)$ is linear, whereas with $v_W(u)$ it is quadratic. Moreover, $v_\tau(u)$ is sensitive to the shape of the job size distribution unlike $v_W(u)$, as the mean cost rate $r_\tau = \lambda\,\mathrm{P}\{W > \tau\}$ depends on it.

Corollary 2. *When $\tau \to 0^+$, the value function reduces to a straight line*

$$v_0(u) - v_0(0) = \lambda\, u. \tag{4}$$

Proof. At the limit $\tau \to 0$, the mean cost rate $r \to \lambda\rho$, and (4) follows from (3). □

For $u < \tau$, the situation unfortunately is more complex.

Proposition 2. *For $0 < u < \tau$, the value function satisfies the differential equation*

$$v'_\tau(u) = -r_\tau + \lambda\,\mathrm{E}[v_\tau(u+X) - v_\tau(u)]. \tag{5}$$

Proof. For a differential time interval δ, $0 < \delta < u \le \tau$,

$$v_\tau(u) = (0 - r_\tau)\delta + (1-\lambda\delta)v_\tau(u-\delta) + \lambda\delta\,[0 + \mathrm{E}[v_\tau(u+X)]]\,,$$

which gives

$$\frac{v_\tau(u) - v_\tau(u-\delta)}{\delta} = -r_\tau - \lambda v_\tau(u-\delta) + \lambda\mathrm{E}[v_\tau(u+X)],$$

and as $\delta \to 0$, $v'_\tau(u) = -r_\tau + \lambda\,(\mathrm{E}[v_\tau(u+X)] - v_\tau(u))$. □

We note that the constant term, $v_\tau(\tau)$, that appears in (3), does not appear in (5). Therefore, as is customary with value functions, we can set, e.g., $v_\tau(0) = 0$ or $v_\tau(\tau) = 0$.

Assuming $v_\tau(u)$ is continuous at $u = \tau$, as it is with Poisson arrivals, we obtain from (3) and (5),

$$v'_\tau(\tau^-) = -r_\tau + \lambda\frac{\mathrm{E}[X]}{1-\rho}(\lambda - r_\tau) = \frac{\lambda\rho - r_\tau}{1-\rho}. \tag{6}$$

On the other hand, (3) immediately yields

$$v'_\tau(\tau^+) = \frac{\lambda - r_\tau}{1-\rho} \tag{7}$$

and thus $v'_\tau(\tau^+) - v'_\tau(\tau^-) = \lambda$ is the difference in the cost rates in states $u = \tau^+$ and $u = \tau^-$. Similarly, one can consider an empty queue, which gives (details omitted for brevity)

$$v'_\tau(0) = 0. \tag{8}$$

This implies the identity $\mathrm{P}\{W > \tau\} = \mathrm{E}[v_\tau(X)] - v_\tau(0)$. Fig. 1 illustrates a value function and its slope at these specific points.

3.2. M/M/1 queue in heavy-traffic

Let us next consider the classical M/M/1 queue with arrival rate λ and service rate μ. Note that the mean cost rate converges to the arrival rate, $r_\tau \to \lambda$, as $\lambda \uparrow \mu$. However, given a finite initial state u, there is some positive probability that some jobs manage to start their service before their deadline expires. In particular, it turns out that the relative values are well-defined and remain finite at this limit.