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

Esa Hyytiä
University of Iceland
esa@hi.is

Rhonda Righter
UC Berkeley
rrighter@ieor.berkeley.edu

Olivier Bilenne
Aalto University
olivier.bilenne@aalto.fi

Xiaohu Wu
Aalto University
xiaohu.wu@aalto.fi

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.

Keywords

Dispatching problem; Parallel computing; Deadlines; M/D/1

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.

We consider a cost structure based on (firm) deadlines. Each job has a certain deadline for the 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" [1]. Similarly, service level agreements (SLAs) are often defined in terms of acceptable waiting times [3].

We use our results for the M/D/1 queue and policy iteration to obtain efficient dispatching heuristics. This basic setting has been studied recently in [2] 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, we derive exact closed-form expressions 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, we allow multiple deadlines each with a unique (additive) violation cost.

VALUETOOLS 2016, October 25-28, Taormina, Italy

DOI 10.4108/eai.25-10-2016.2267093

2. MODEL AND RESULTS

Let λ denote the arrival rate and d the constant service time of a job 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. The mean cost rate is $r = \lambda \mathrm{P}\{W \geq \tau\}$, and $\mathrm{P}\{W \geq \tau\}$ is available, e.g., from [4]. 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. First we 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 system is stable and ergodic, and the above limit is well-defined.

The value function for the M/D/1 queue with a single deadline takes the form of a double sum with a finite number of terms. This result is then generalized for multiple job classes, each having its own target deadline and violation cost. The value function enables policy iteration when developing cost-aware dispatching strategies for parallel servers. This is illustrated in numerical examples.

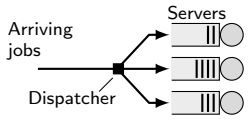

Figure 1: Dispatching system with parallel servers.

Acknowledgements

This work was supported by the Academy of Finland in the FQ4BD project (grant nos. 296206).

3. REFERENCES

[1] J. Dean and L. A. Barroso. The tail at scale. *Commun. ACM*, 56(2):74–80, Feb. 2013.

[2] E. Hyytiä and R. Righter. Routing jobs with deadlines to heterogeneous parallel servers. *Operations Research Letters*, 44(4):507–513, 2016.

[3] Z. Liu, M. S. Squillante, and J. L. Wolf. On maximizing service-level-agreement profits. In *Proc. of the 3rd ACM Conference on Electronic Commerce*, NY, USA, 2001.

[4] R. Syski. *Introduction to congestion theory in telephone systems*. North-Holland, 1986.