Performance Evaluation 75–76 (2014) 17–35

Contents lists available at ScienceDirect

Performance Evaluation

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

Task assignment in a heterogeneous server farm with switching delays and general energy-aware cost structure

Esa Hyytiä [a,*,1], Rhonda Righter [b], Samuli Aalto [a]

[a] Department of Communications and Networking, Aalto University, Finland

[b] University of California Berkeley, Ind. Eng. and Opns. Res. Dept., Berkeley, CA 94720-1777, United States

HIGHLIGHTS

- We model a server farm with switching delays and a general cost structure.
- We derive value functions for $M/G/1$ queues with switching delay.
- Switching (setup) delay shows up as an additional term in the value function.
- We develop energy-aware policies that also control the set of active servers.
- Our heuristics outperform standard policies over a wide range of parameters.

ARTICLE INFO

Article history:
Received 27 March 2013
Received in revised form 9 January 2014
Accepted 27 January 2014
Available online 3 March 2014

Keywords:
Task assignment
$M/G/1$-queue
FCFS
Switching delay
Energy-aware
MDP

ABSTRACT

We consider the task assignment problem to heterogeneous parallel servers with switching delay, where servers can be switched off to save energy. However, switching a server back on involves a constant server-specific delay. We will use one step of policy iteration from a starting policy such as Bernoulli splitting, in order to derive efficient task assignment (dispatching) policies that minimize the long-run average cost. To evaluate our starting policy, we first analyze a single work-conserving $M/G/1$ queue with a switching delay and derive a value function with respect to a general cost structure. Our costs include energy related switching and processing costs, as well as general performance-related costs, such as costs associated with both means and variability of waiting time and sojourn time. The efficiency of our dispatching policies is illustrated with numerical examples.

1. Introduction

In task assignment (dispatching) problems, arriving jobs are assigned to available servers immediately upon arrival. In the basic setting, these servers are modeled as work-conserving queues with a constant service rate. Task assignment problems arise, e.g., in transportation (e.g., street toll booths), manufacturing systems, (data) traffic routing, super computing, cloud computing, data centers (e.g., web server farms) and other distributed computing systems. A typical objective is to minimize the mean waiting time or the mean sojourn time (i.e., response time, latency or delay, where the terminology varies with the application). Recently, energy consumption, e.g., in data centers and distributed computing in general, has become an important consideration. Therefore, in this paper we assume that the general cost structure includes energy consumption

* Corresponding author. Tel.: +358 50 4354971; fax: +358 9 4512474.
E-mail addresses: esa.hyytia@aalto.fi (E. Hyytiä), rrighter@ieor.berkeley.edu (R. Righter), samuli.aalto@aalto.fi (S. Aalto).

[1] The main part of the research was carried out while Esa Hyytiä was visiting Professor Rhonda Righter at UC Berkeley.

http://dx.doi.org/10.1016/j.peva.2014.01.002
0166-5316/

related costs in addition to the usual performance metrics such as the mean sojourn time. Moreover, we consider a model where a server can be switched off in order to save energy. However, switching a server back on is assumed to take a certain setup time d during which no job can be served, and may incur a cost. The optimal dispatching policy finds an appropriate trade-off between the criteria of minimizing energy costs while maximizing customer satisfaction.

We approach this challenging problem in the framework of Markov decision processes (MDP) [1–3]. First we consider an isolated $M/G/1$ queue with a switching delay and derive the so-called size-aware value functions with respect to different terms in the cost structure. We consider two types of costs: (i) costs defined by a rate (e.g., energy consumption) and (ii) immediate costs (e.g., switching the server on or off). For some quantities such as the waiting time, it is possible to define the cost either as a cost rate (i.e., the number of waiting jobs) or as an immediate cost (the waiting time becomes known upon an arrival to a size-aware FCFS queue). We will utilize both conventions in this paper.

The value functions characterize the difference in the long-term mean costs between initial states. Using the value functions as inputs to the MDP-based heuristics of first policy iteration (FPI) and lookahead, we derive energy- and state-aware dispatching policies, which also take into account the switching delays in the servers. The resulting policies are evaluated and illustrated with numerical examples.

In this paper, we extend earlier results [4,5] for the $M/G/1$ queue in several respects. First, we include a switching delay, which is an important system characteristic when, e.g., a server in a data center is switched off in order to save energy. Second, we also consider general holding cost functions, and derive explicit expressions for the second moments of the waiting time and sojourn time, which allow one to improve fairness in the system. Moreover, the energy consumption model also includes switching costs and is thus more comprehensive than in [6]. The new results are the *size-aware value functions* for the $M/G/1$ queues, when the objective is to minimize the (mean) backlogs, waiting times, sojourn times, or any combination thereof (as well as higher moments). These results also yield the corresponding value functions for a system of parallel servers (our model for a server farm) operating under any static policy, which are prerequisites for the efficient dynamic policies via the FPI and Lookahead approaches.

1.1. Related work

Dispatching problems have been studied extensively in the literature. Within each queue, first-come-first-served (FCFS) scheduling is usually assumed, but other scheduling disciplines have also been studied. Only a few optimality results are known for dispatching problems, and these generally require homogeneous servers. In general, the optimal decision depends on the available information.

Round-Robin (RR), followed by FCFS, is the optimal policy when it is only known that the queues were initially in the same state, and the dispatching history is available [7–9].

Crovella et al. [10] and Harchol-Balter et al. [11] assume that the dispatcher is aware of the size of a new job, but not of the state of the FCFS queues, and propose *Size-Interval-Task-Assignment* (SITA) policies, where each queue receives the jobs within a certain size interval (e.g., short jobs to one queue, and the rest to another). Feng et al. [12] later showed that SITA is the optimal size-aware static policy for homogeneous servers.

When the number of tasks per server is available, the intuitive *Join-the-Shortest-Queue* (JSQ) policy assigns a new job to the server with the fewest tasks. Assuming exponentially distributed interarrival times and job sizes, and homogeneous servers, [13] shows that JSQ with FCFS minimizes the mean waiting time. Since then the optimality of JSQ has been shown in many other settings [14,7,15–18]. Recently, Akgun et al. [19], have shown the optimality of JSQ for $G/M/1$ queues under very general assumptions. In contrast, Gupta et al. consider JSQ with the *processor-sharing* (PS) scheduling discipline in [20]. Whitt [21], on the other hand, provides several counterexamples with non-exponential services where the JSQ/FCFS policy is not optimal, even when servers are homogeneous.

If the queue-specific backlogs (unfinished work, workload) are available, then the *Least-Work-Left* (LWL) policy chooses the queue with the smallest backlog, i.e., it is the greedy (myopic) policy minimizing the waiting time of the new job. Interestingly, the $M/G/k$ system with a central queue is equivalent to LWL, which means that at no time instance, a server is idle at the same time when a job is waiting in some queue. Daley [22], based on Foss's work [23], has shown that $G/G/k$ (i.e., LWL with general inter-arrival times) stochastically minimizes both the maximum and total backlog with identical servers at an arbitrary arrival time instance. In contrast, the counterexample given by Stoyan [24] shows that pathwise RR can yield both a lower waiting time and a lower total backlog (at arrival times).

In general, simple policies such as RR, JSQ and LWL will no longer be optimal when there is switching delay, even in the case of homogeneous servers. Using one step of policy iteration (FPI) in the MDP framework is a promising approach to developing good heuristics for the dispatching problems with heterogeneous servers. Krishnan [25] has utilized it for minimizing mean sojourn time with parallel $M/M/s$-FCFS servers. See also Aalto and Virtamo [26]. Recently, FCFS, LCFS, SPT and SRPT queues were analyzed in [4,5] with a general service time distribution. Similarly, PS is considered in [27,28]. The key idea with the above work is to start with an arbitrary static policy, and then carry out the first policy iteration (FPI) step utilizing the value functions (relative values of states). This general approach is due to Norman [29]. See also [30,31] for $M/G/1$ and $M/\mathrm{Cox}(r)/1$, and [32,33] for blocking systems.

Server farms with switching delays have only been considered recently, and only for homogeneous servers. Artalejo et al. [34] give steady-state results for an $M/M/k$ with setup delays, where idle servers are switched off and at most one server can be in an exponentially distributed setup phase. Gandhi and Harchol-Balter [35] consider an $M/G/k$ with an