Performance Evaluation 121–122 (2018) 48–66

Contents lists available at ScienceDirect

Performance Evaluation

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

Structural properties and exact analysis of energy-aware multiserver queueing systems with setup times

V.J. Maccio [*], D.G. Down

McMaster University, Hamilton, Ontario, Canada

ARTICLE INFO

Article history:
Received 31 January 2017
Received in revised form 13 February 2018
Accepted 1 March 2018
Available online 17 March 2018

Keywords:
Continuous time Markov chain
Markov decision process
Green computing
Energy-aware server
Queueing analysis

ABSTRACT

Energy consumption of today's datacenters is a constant concern from the standpoints of monetary and environmental costs. An intuitive solution to address these immense energy demands is to turn servers off to incur less costs. As such, many different authors have modelled this problem as an $M/M/C$ queue where each server can be turned on, with an exponentially distributed setup time, or turned off instantaneously. What policy the model should employ, or rather when each server should be turned on and off is far from a trivial question. A specific policy is often examined, but determining which policy to study can be a difficult process and is often a product of intuition. Moreover, while a specific policy may do comparatively well against another, in general it may be far from optimal. This problem is further accentuated when one considers the case that a policy may do well or even be optimal under a specific cost function, but far from optimal under another. To address this issue we study the structural properties of the optimal policy under linear cost functions, allowing for a significant reduction in the search space. We then leverage these structural properties to intelligently select two policies for further study. Using the recursive renewal reward technique, we offer an exact analysis of these policies alongside offering insights, observations, and implications for how these systems behave. In particular, we provide insight into the question of the number of servers that should remain on at all times under a general cost function.

1. Introduction

Immense energy consumption of datacenters has become a fact of modern life. The United States spends on the order of billions of dollars powering these systems each year [1,2]. Google alone pays an annual energy bill on the order of tens of millions of dollars [3]. While some may see this as an obligatory cost, the truth is many of these servers spend a significant amount of time idle. Moreover, an idling server uses a large percentage of the energy it would if it were busy [4]. To conserve costs, servers often have a lower energy state they can be switched to (off, sleep, etc.). However, the choice of if and when to make such a switch for each server is far from trivial. That is, while turning a server off *may* increase system efficiency, it *will* decrease system efficacy.

Such concerns have driven an interest in queueing systems where individual servers can be turned on to improve performance, and turned off to save on costs. This interest has led to different authors studying the same, or similar, queueing models. However, due to the complexity of the problem, i.e. the choice of cost function, policy implemented, model details, etc., different conclusions can be drawn from similar underlying problems. One consequence of the variety in the problems studied and the corresponding variety of insights is that it is difficult to confidently draw conclusions which are overarching

[*] Corresponding author.
E-mail address: macciov@mcmaster.ca (V.J. Maccio).

https://doi.org/10.1016/j.peva.2018.03.001
0166-5316/Crown Copyright © 2018 Published by Elsevier B.V. All rights reserved.

across the problem domain. To address this issue, this work presents a two-pronged approach. Firstly, we derive several structural properties of the optimal policy. These properties allow one to discount policies which are in turn known to exhibit sub-optimal behaviours, as well as gain confidence in previously studied policies which adhere to this structure. Secondly, we leverage these structural properties to intelligently select two policies to analyse further. We perform an exact analysis of these policies which grants insights into how these systems behave and how they should be provisioned. Specifically, we provide insight on how many servers should always remain on under any reasonable cost function.

To the best of our knowledge, Chen et al. [5] were the first to use queueing theory to tackle the problem of energy-aware provisioning in server farms. Around the same time Slegers et al. [6] studied the problem with varying traffic rates where servers are allocated dynamically and presented heuristics to conserve energy. Since then, several variations on previously studied vacation models [7] have been developed, where vacations can be viewed as the setup time for a given server. Gandhi et al. began to study these systems in [8] and provided several analytical results for the single server case, as well as some rules of thumb for the multiserver case. They continued their research in [9,10] in which they modelled a server farm as a continuous time Markov chain (CTMC) employing the *staggered setup* policy, where the number of servers in setup equals the number of jobs waiting to be served and servers shut down as soon as they idle. As will be seen, employing a two-dimensional CTMC model is a common and convenient way to view these systems. As such, in [11] Gandhi et al. introduced a method to derive moments of metrics associated with these CTMCs (such as the expected number of jobs in the system) called the recursive renewal reward (RRR) technique, where they also applied their method of analysis to the *delayed off* policy, an extension of staggered setup where servers spend some exponentially distributed period of time idle before being switched off. Phung-Duc [12] gave a comprehensive side by side comparison of RRR and other traditional methods for analysing these CTMCs. If the steady state distribution of these CTMCs is also of interest, methods introduced by Doroudi et al. in [13] may be employed.

Other authors have studied the same model as Gandhi et al. but under different policies (when servers turn on and off). Mitrani [14] studied this model where a reserved set of servers are brought into setup when the number of jobs in the system exceeds a threshold, and then shuts those servers off once the number of jobs drops below another threshold. This policy was further studied in [15]. Xu and Tian [16] studied the set of policies where e servers are turned off when there are d servers idle. Kuehn and Mashaly [17] analysed policies which wait for a threshold number of jobs to accumulate in the queue before a server starts its setup and turns servers off when they idle, under the presence of a finite buffer. Lastly, Ren et al. [18] analysed a finite two-dimensional CTMC similar to Kuehn and Mashaly in the context of virtual networks, which allows for a number of servers to always remain operational, but omits the use of turn on thresholds.

Limiting study to the single server case grants an even greater understanding of these systems. Artalejo [19] was one of the first to look at this case under general processing time distributions. However, his work focused on particular vacation models which do not fully capture the behaviour of a server which can be switched on and off. In [20] we adapted these models to better suit the domain of green computing, and were able to derive the optimal policy for the single server case under complete generality with regards to the underlying distributions and cost function. Gebrehiwot et al. [21] extended the analysis of the single server case by allowing multiple sleep states, and more recently looked at the model under the *processor sharing* service discipline [22]. Hyytiä et al. [23,24] also studied this model under processor sharing in addition to *last come first serve*, and different routing configurations. This model also has applications to or is studied in problems which arise in other fields such as manufacturing, logistics, and vacation models [25–27]. For other, non-queueing theoretic approaches to this problem see [28–32].

While the contributions of the previously mentioned works are substantial, a gap in knowledge still remains. While optimal control in the single server case is well understood, it offers less practical application than corresponding multiserver models. However, when studying the multiserver case, complexity has constrained researchers to focus on specific policies, which in general may be far from optimal. Moreover, when evaluating one of these policies it may do well for a specific cost function, but poorly under another. Therefore, saying one policy is strictly better than another, or any general statements, can be difficult. Reiterating, to address these issues this work includes but is not limited to the following contributions:

1. An examination of the optimal policy and behaviour under linear cost functions, including formal proofs of several key structural properties.
2. The description and exact analysis of two distinct policies which leverage the aforementioned structural properties, *bulk setup* and *staggered threshold*.
3. A range of numerical experiments which yield exact values for metrics of interest alongside several insights into how these systems behave, specifically with respect to the question of the number of servers one should always leave on.

For further details and discussion on the results presented here, we direct the reader to the corresponding conference publications and technical reports [33–35].

2. Model

The model under study is an $M/M/C$ queue where each server can be switched on and off, and where turn-offs are instantaneous, but turn-ons take an exponentially distributed setup time. This is described formally as follows. Jobs arrive to