SWITCHING DIFFUSION APPROXIMATIONS FOR OPTIMAL POWER MANAGEMENT IN PARALLEL PROCESSING SYSTEMS

SAUL C. LEITE, MARCELO D. FRAGOSO, AND RODOLFO S. TEIXEIRA

Abstract. In this paper, we investigate optimal power management in parallel processing systems composed of one queue and several identical processing stations. Power consumption is controlled by setting some of the stations into an inactive state, where they consume less power but are unable to provide service. This way, we are faced with the conflicting objective of minimizing power consumption while maintaining a desired quality of service. A distinguishing feature here, regarding most previous literature on this subject, is that we consider systems operating a policy that may turn the reserve machines on or off with setup times and under general inter-arrival or service time distributions, subject to some conditions. When these condition fail, we also provide a model with general inter-arrival times and exponentially distributed service times. To some extent, a controlled switching diffusion obtained in this paper via heavy traffic analysis and stochastic optimal control theory are the technical underpinning of the paper. We also propose a numerical approach to the solutions of the optimal control problems based on the Markov chain approximation method. Finally, we consider some numerical experiments that illustrate the efficiency of the proposed approach.

Keywords. Heavy traffic Analysis, Switching Diffusion, Stochastic Optimal Control, Parallel Processing Queuing Systems.

1. Introduction

Large data centers (or server farms) support major Internet-based organizations. It is known that power consumption accounts for a large fraction of the cost incurred in maintaining these data centers [14]. Nudged in part by this, the problem of managing power consumption in large scale parallel computer systems composed of one queue of pending jobs and a bank of identical processing stations has been the focus of various research nowadays. Some authors have considered controlling power usage through dynamic voltage (or frequency) scaling (DVFS) (e.g. [26]). In these cases, power consumption is controlled by scaling CPU frequency either to conserve energy or to

Centro de Matemática Computação e Cognição, Universidade Federal do ABC, Santo André, SP, Brazil

Coordenação de Métodos Matemáticos e Computacionais, Laboratório Nacional de Computação Científica, Petrópolis, RJ, Brazil

Pós-Graduação em Engenharia Mecânica, Pontifícia Universidade Católica, Rio de Janeiro, RJ, Brazil

E-mail addresses: saul.leite@ufabc.edu.br, frag@lncc.br, rodolfo@aluno.puc-rio.br.

This research was partially supported by the Minas Gerais Research Foundation (FAPEMIG) under the grant APQ 00945/14 and by the Brazilian National Council for Scientific and Technological Development (CNPq) under the grants CNPq-304801-2015-1 and CNPq-421486/2016-3.

generate less heat. It has been argued, however, that energy savings improvements in modern processors may limit the potential savings via DVFS [28, 36]. The problem of controlling power usage by either shutting servers down or putting them into a sleep state are considered in [3, 12, 29, 30, 32]. From the point of view of modeling, this type of approach is more demanding, since the queuing model should account for the *setup times*, which are the times required for the system to change state. In particular, Mitrani [30] considers a system with a predefined block of servers, called reserves, which can be turned on and off. A queuing model is derived for such system, and the average cost of the system under steady state is calculated, from where optimal parameters are chosen with the help of a heuristic. Niyato et al. [32] has a different approach, modeling the optimization problem as a discrete time Markov decision process. Except for [29], it is important to mention here that a common feature on the above works is that they assume *Poisson arrival* and *exponentially distributed service time* requirements in order to derive their queuing model. In [29], the authors consider systems with general inter-arrival and service time distributions by combining their model with a probability estimate derived from a heavy traffic analysis. However, this estimate was obtained for systems in which servers cannot be turned on or off.

There are different approaches to combining the conflicting objectives of minimizing power consumption and of providing reasonable quality of service. In [12], the so-called energy-response time product is considered, where the cost of the system is measured as a product of mean power consumption and mean customer response time. In [28, 29, 30], the cost is calculated as the weighted sum of mean energy consumption and a measure of average quality of service. In both of these cases, it is not an easy task to choose weights for these measures, and the resulting policies may vary depending on the weights chosen. Instead of combining these conflicting objectives into one cost function, it seems more natural to adopt a constrained optimization approach, which guarantees a certain level of quality of service in the system whilst minimizing energy consumption. This approach, which is also adopted here, was first proposed in [32], where the optimization problem is set as one of minimizing power consumption subject to a maximum mean waiting time constraint.

In this paper, we consider *optimal* power management of parallel processing systems with *one queue* and a *large number* of identical processing stations. The control is performed by changing the state of a block of servers, called reserves, among *active* and *inactive* states. In active states, the stations are operational whereas, in inactive states, they are either sleeping or turned off. In either case, the reserve stations consume less power but are also unable to service pending tasks. The transitions from one state to the other require a *setup time*, where the stations may be actively consuming power but cannot provide service. Therefore, there is a conflicting objective of minimizing power consumption while maintaining a desired quality of service.

A distinguishing feature here, regarding previous literature on this subject, is that we consider systems operating under a policy that may turn the reserve machines *on* or *off* with setup times and *under general inter-arrival or service time distributions*. This, in turn, brings to bear new modeling issues in this scenario. Our approach starts from a novel point of view, which hinges on the following steps. Regarding the first

issue, this is sorted out in this work by appending in our model *a pure jump process* to account for the state of the reserve machines. As for the second issue, we are able to consider a system with general inter-arrival and service time distributions via heavy traffic approximation under an assumption that stations work together on pending jobs and the work lost, when the stations shut down, is negligible (with respect to the amount of work in the system). Systems that can be modeled under such assumptions are, for instance, *job-splitting systems*, where individual jobs that join the system are split into independent tasks among the available bank of processing servers. In this regard, index servers in web search engines are practical examples of this type of service [4]. When this assumption fails, we also provide a model with general inter-arrival times and exponentially distributed service times.

To some extent, the linchpin of our approach to deal with general distribution is the derivation of a heavy traffic approximation for this controlled system under a limiting condition. Out of the bent which wends most of the literature on this subject, *a novel feature* of this heavy traffic approximation is that the limit process is a *controlled switching diffusion*. The optimization problem is then set as an stochastic optimal control problem with an ergodic cost and, optionally, subject to a quality of service constraint. Since to carve out a closed analytical solution for the optimization problem is a rather difficult task, *a numerical approach is proposed* using the Markov chain approximation method (MCAM) [25] adapted to the diffusion with the jumping parameter considered here. Numerical data is also presented, where optimal control problems are solved numerically and the resulting controls applied to a queuing system simulation. Two scenarios are then considered for the numerical experiments. In the first, an ergodic cost combining the conflicting objective of power consumption and quality of service is proposed. The setup is the same as the one presented in [30]. We show that the control constructed via our approach gives better results than the heuristics presented in [30]. In the second scenario, a constrained optimization problem is proposed, where an ergodic cost measuring power consumption is subject to a constraint with respect to the system performance. These numerical experiments show that the switching diffusion approximation works well and gives interesting insights about the problem.

It is perhaps noteworthy to mention that heavy traffic (or diffusion) approximations are used here by the fact that exact queuing models that can capture transient behavior under general inter-arrival and service time distributions assumptions are considered intractable. This becomes even more challenging when one wishes to apply optimal control methods. Diffusion approximation are constructed with only the first two moments of the inter-arrival and service time distributions and are known to give good estimates for the system when it is operating under a moderate or heavy demand [23, 38]. In addition, control policies which were devised for a system under heavy traffic are often found to work quite well under moderate conditions (see [24] and references therein), since they are provisioned for the worst case.

Of highlight here also is the fact that the limiting condition assumed here was first proposed by Halfin and Whitt in their seminal paper [16], where a sequence of queuing systems with an increasing arrival rate and number of processing stations is considered. Such limit condition reflects well the scenario of large data centers, which attend a

very large population of customers and require a large cluster of processing stations. The limit system is attained as the arrival rates increase to infinity together with the number of servers, while the sequence of queuing systems approaches heavy-traffic. *A distinctive feature* of the models considered here, *which is not seen elsewhere*, is that we consider the convergence of controlled queuing systems and, as result, the limit process is a controlled switching diffusion whose *the drift parameter is subject to pure jump process*, that accounts for the different states of the reserve stations. More precisely, we show that scaled models of the queuing systems operating a given control policy converge weakly to a controlled switching diffusion with the same control policy.

A very brief résumé of the *main differences* of our approach, vis-à-vis previous approaches, goes as follows:

- Predicated on a policy that may turn the reserve machines *on* or *off*, which is modeled here by a controlled pure jump process, in conjunction with heavy traffic techniques and optimal control theory, we devise a *new modeling* for the problem, via a controlled switching diffusion, which *differs completely* from those of previous works.
- Through the adaptation of the Markov chain approximation method [25], we are able to recast the optimization problem as a Markov decision process and consider the problem of minimizing energy consumption subject to a quality of service constraint.

An outline of the content of this paper is as follows. In Section 2 we provide the bare essential of notations. The queuing models and some basic assumptions are introduced in Section 3. Section 4 introduces the pure jump process and elaborates on the classes of admissible controls. Next, in Section 5, we derive the heavy traffic approximations for the queuing models. The control problem and the proposed numerical approach, via the Markov chain approximation method, are presented in Section 6. In Section 7, some numerical experiments are considered, where the optimal control is found numerically and applied to a simulation. Finally, some conclusions are provided in Section 8. In the Appendix A we present the proof of Lemmata 3.3, 5.4 and Theorems 6.1, 6.3

2. Notations

In what follows $(\Omega, \mathcal{F}, \mathbb{P})$ stands for a probability space where the stochastic processes are to be defined and $\mathbb{E}[\cdot]$ denotes the mathematical expectation with respect to $\mathbb{P}$. The set of non-negative real numbers is denoted by $\mathbb{R}_+$ and the set of non-negative integers is denoted by $\mathbb{N}_0$. For $d \geq 1$, let $\mathbb{R}^d$ denote the d-dimensional Euclidean space. For two real numbers a and b, let $a \wedge b$ and $a \vee b$ denote the least and the greatest of the two numbers, respectively. For a set S, let $\mathbb{I}_S(\cdot)$ denote its indicator function, which takes value $\mathbb{I}_S(s) = 1$ if $s \in S$ and 0 otherwise. Sometimes it is more convenient to write $\mathbb{I}_{\{s \in S\}}$ instead of $\mathbb{I}_S(s)$, in these cases, we will use this alternative notation. In addition, for $S \subseteq \mathbb{R}$, $C_0^2(S \times E)$ denotes the set of real valued functions on $S \times E$ that are continuous with compact support and have continuous first and second partial derivatives with respect to the first argument. For a Polish space U we denote by $\mathcal{B}(U)$ its Borel σ-algebra and by $\mathcal{P}(U)$ the Polish space of probability measures on $(U, \mathcal{B}(U))$ endowed with the Prohorov topology (the topology of weak convergence [6]). Unless

mentioned otherwise, the stochastic processes considered here have sample paths that are right-continuous with left-limits. Let $D_{\mathbb{R}^d}[0,\infty)$ denote the set of such functions $\omega : [0,\infty) \to \mathbb{R}^d$, which are continuous from the right and have limits from the left, endowed with Skorohod's J1 topology. We denote by $C_{\mathbb{R}^d}[0,\infty)$ the set of continuous functions $\omega : [0,\infty) \to \mathbb{R}^d$ equipped with the topology of uniform convergence on compact sets. We say that a sequence of stochastic processes $\{X^n\}$ is tight if their corresponding probability law $\mathfrak{L}(X^n)$ form a tight sequence (see, e.g., [18, Chapter VI] for more detail). In addition, we say that $\{X^n\}$ is C-tight if it is tight and all the limit points of $\{\mathfrak{L}(X^n)\}$ assign probability 1 to the set of continuous functions $C_{\mathbb{R}^d}[0,\infty)$. Finally, more specific notation will be introduced throughout the paper according to the necessity.

3. Queuing Models and Basic Assumptions

We consider queuing systems composed of one queue and n identical processing stations. Since we are interested in the limit system, which is approached as the number of stations and the arrival rates increase, we will index the model and some of the driving stochastic processes with a superscript n. We shall refer to the queuing system with n servers as the n-th system. This way, let us denote by $\{\Delta_l^{a,n}\}_{l\in\mathbb{N}}$ the sequence of inter-arrival times for a system with n servers (or the n-th system), which are assumed to be independent and identically distributed in l. Let $A^n := \{A^n(t)\}_{t\geq 0}$ denote the counting process indicating, for each $t \geq 0$, the number of job arrivals to the n-th system by time t. That is $A^n(t) := \max\{k \in \mathbb{N}_0 : \sum_{l=1}^k \Delta_l^{a,n} \leq t\}$, for each $t \geq 0$, with the convention that $\sum_{l=1}^0 = 0$. Let $\{\Delta_l^d\}_{l\in\mathbb{N}}$ denote the sequence of service time requirements for each arriving job, which are assumed to be independent and identically distributed random variables with finite variance and which are independent of the inter-arrival times. We assume that the service time distribution is invariant as the number of stations in the system increases, that is why the sequence is not indexed with the superscript n.

As mentioned previously, in order to reduce the energy consumption, the system is controlled by setting some of the machines in the system to an inactive state when the demand is low and turning them back to an active state when the demand increases. For simplicity, we henceforth refer to the active state of the machines as the *on* state and the inactive state as the *off* state. Of the n machines present in the system, we consider that $\mathfrak{o}^n$ of them always stay *on* and the remaining $\mathfrak{r}^n := (n - \mathfrak{o}^n)$ can be turned *on* or *off*. The quantities $\mathfrak{o}^n$ and $\mathfrak{r}^n$ are indexed with the parameter n to indicate that they may change when the number of servers n change. We suppose that the stations change state together as a block, in the sense that they become ready to be used and they shut down at the same time. We call $\mathfrak{r}^n$ the number of reserve stations and, since the machines are identical, any of the n stations can be chosen to be part of the reserve block at any given time. In order to model the state of the reserve machines, let $\theta^n := \{\theta^n(t)\}_{t\geq 0}$ be a controlled pure jump process taking values in a finite set E. The precise definition of this process and the assumptions involved are given later in Section 4. This process indicates the state of the reserve machines of the n-th system at each time t. For example, we can consider the set of states E to be given

by $E = \{0, 1, 2, 3\}$, where the state 0 indicates that the reserve machines are turned off; the state 1 indicates that they are turning on; 2 indicates that they are turning off and cannot work on pending tasks; and 3 indicates that they are operational. Let $\mathbb{f}^n(i)$ denote the fraction of servers that are operational when the state of the reserve machines is i. We suppose that there is always an $i \in E$ such that $\mathbb{f}^n(i) = 1$, this corresponds to the case where every machine is turned on and operational. Using the example described above, where $E = \{0, 1, 2, 3\}$, we can define: $\mathbb{f}^n(i) := \mathfrak{o}^n/n$, if $i \in \{0, 1, 2\}$, and $\mathbb{f}^n(i) := 1$ if $i = 3$.

In order to characterize the limiting condition with increasing arrival rates and servers, let us define the scaled arrival sequence $\{\Delta_l^{s,n}\}$ as follows:

$$\Delta_l^{s,n} := n\Delta_l^{a,n}, \quad \text{for each } l \in \mathbb{N}. \tag{3.1}$$

In the spirit of the assumptions used by Halfin and Whitt [16] (*mutatis-mutandis*), we consider the following assumptions.

Assumption 3.1. The sequence $\{|\Delta_l^{s,n}|^2; n\}$ is uniformly integrable and, in addition, there are positive constants λ^s and σ_s^2 such that:

$$\lambda^{s,n} := \mathbb{E}\left[\Delta_l^{s,n}\right]^{-1} = (\bar{\Delta}^{s,n})^{-1} \to \lambda^s \quad \text{and} \quad \sigma_s^{2,n} := \mathbb{E}\left[(1 - \lambda^{s,n}\Delta_l^{s,n})^2\right] \to \sigma_s^2,$$

as $n \to \infty$.

Clearly Assumption 3.1 on $\Delta_l^{s,n}$ and its definition given by (3.1) imply that the arrivals become increasingly faster as the number of servers n increases. The condition above is quite broad. For example, it would suffice to consider sequences of random variables $\{\Delta_l^{s,n}\}_{l\in\mathbb{N}}$ that are identically distributed in n and have finite variance. In practice however, since one is interested in one queuing system with fixed arrival and service rates which is to be approximated by the limit system, this assumption serves only to give precise conditions for the convergence of the random variables and associated stochastic processes to their appropriate limits. We will discuss in more detail how the n-th system can be approximated by the limit system in Section 5.3.

For later reference and in analogy to the notation introduced by Assumption 3.1, let us introduce the following notation $\bar{\Delta}^d := \mathbb{E}\left[\Delta_l^d\right]$, $\lambda^d := (\bar{\Delta}^d)^{-1}$, $\bar{\Delta}^{a,n} := \mathbb{E}\left[\Delta_l^{a,n}\right]$, and $\lambda^{a,n} := (\bar{\Delta}^{a,n})^{-1}$. Let us define the *traffic intensity parameter* as $\rho^n := \lambda^{a,n}\bar{\Delta}^{d,n}/n$ and, in addition, let $\sigma_d^2 := \mathbb{E}\left[(1 - \lambda^d\Delta_l^d)^2\right]$ and notice that $\sigma_a^{2,n} := \mathbb{E}\left[(1 - \lambda^{a,n}\Delta_l^{a,n})^2\right] = \sigma_s^{2,n}$.

As the arrival rate and the number of processing stations increase, the system approaches a condition called "heavy-traffic," where the system is operating near its maximum processing capacity. This is introduced in the model by the assumption below. Recall that we assumed that there is one $i \in E$ such that $\mathbb{f}^n(i) = 1$, which corresponds to a state of the processing stations where every machine is turned on and operational. Therefore, the assumption below implies, in particular, that $\rho^n \to 1$, as $n \to \infty$.

Assumption 3.2. Let $\rho^n := \lambda^{a,n}\bar{\Delta}^d/n$ be the traffic intensity parameter for the n-th system. There is a function $b : E \to \mathbb{R}$ such that:

$$b^n(i) := \sqrt{n}\,(\rho^n - \mathbb{f}^n(i)) \to b(i), \tag{3.2}$$

for each $i \in E$.

We now consider two models, which are assumed to satisfy Assumptions (3.1) and (3.2) but differ in their service regime. In the first model, the machines in the system work together to complete the pending jobs and, in addition, the work lost when machines are turned off is assumed to be negligible. This is an approximation to the behavior of large parallel systems that perform task splitting or task parallelization, where the various processing stations act as if they were a single "fast" processor. For example, in index servers of web search engines, jobs are divided into many (often n) smaller tasks which are then assigned to different processors [4]. For the second model, each processing station serves one individual job at a time and the service time requirements Δ_l^d are assumed to be exponentially distributed. This is the same assumptions used in [16], however, here we are interested in a heavy traffic approximation for the controlled system. It was shown in [16] that such a system with general service time requirements cannot be approximated via heavy traffic by a Markov process. This is true unless a more general representation for the state is considered, such as the ones used in [19, 33]. For notational convenience, in the following sections, we do not distinguish the notation used for the driving processes in the two models, such as the inter-arrival and service times or the controlled pure jump process representing the state of the reserve machines.

3.1. **Model I.** In this section, we consider a workload model for the first queuing system. Let $X^n := \{X^n(t)\}_{t\geq 0}$ denote the workload process. That is, for each $t \geq 0$, $X^n(t)$ represents the sum of pending service time requirements in the system at time t or, equivalently, $X^n(t)$ is the total time that a single processing station has to work in order to complete the remaining work in the system at time t. We define X^n as follows:

$$X^n(t) := X^n(0) + \sum_{l=1}^{A^n(t)} \Delta_l^d - n\int_0^t \mathbb{ff}^n(\theta^n(s))ds + R^n(t), \quad t \geq 0, \tag{3.3}$$

where $X^n(0)$ is the initial workload for the n-th system, which is assumed to be a non-negative random variable, independent of the other driving processes. Except for the introduction of the pure jump process θ^n, this form for the workload has been used previously in the literature, see, for instance, the model of Section 5.3.1 of [23]. Notice that $n\mathbb{ff}^n(\theta^n(t))$ represents the number of active stations at time t and, therefore, the integral term in (3.3) accounts for the combined amount of processing time the queue has received by time $t \geq 0$. Clearly, the sum $\sum_{l=1}^{A^n(t)} \Delta_l^d$ represents the total amount of work that has arrived to the system by time t. The process R^n compensates the possible idle time periods of the system, it is often referred to as the reflection process. It is defined to be a non-decreasing process with initial condition $R^n(0) = 0$, which may increase only at times t such that $X^n(t) = 0$. This process prevents X^n from taking negative values. More precisely, R^n (together with X^n) is defined pathwisely as the unique solution of the Skorohod problem for the unreflected process defined analogously to X^n in (3.3), but without the reflection term (see [37]).

In order to derive the heavy traffic limit, let us rewrite (3.3). We begin by defining the processes $M^{a,n}$ and M^d as follows:

$$M^{a,n}(t) := \sum_{l=1}^{\lfloor t \rfloor}(1-\lambda^{a,n}\Delta_l^{a,n}), \text{ and } M^d(t) := \sum_{l=1}^{\lfloor t \rfloor}(1-\lambda^d\Delta_l^d), \ t \geq 0, \tag{3.4}$$

where $\lfloor t \rfloor$ denotes the largest integer not greater than t. Notice that

$$\begin{aligned}\sum_{l=1}^{A^n(t)}\Delta_l^d &= -\bar{\Delta}^d\sum_{l=1}^{A^n(t)}(1-\lambda^d\Delta_l^d) + \bar{\Delta}^d\sum_{l=1}^{A^n(t)}(1-\lambda^{a,n}\Delta_l^{a,n}) + \bar{\Delta}^d\lambda^{a,n}\sum_{l=1}^{A^n(t)}\Delta_l^{a,n} \\ &= -\bar{\Delta}^d M^d(A^n(t)) + \bar{\Delta}^d M^{a,n}(A^n(t)) + \bar{\Delta}^d\lambda^{a,n}t + \bar{\Delta}^d\lambda^{a,n}\varepsilon^{a,n}(t),\end{aligned} \tag{3.5}$$

where $\varepsilon^{a,n}(t)$ is an error term that accounts for the time difference between t and the time of last arrival. Clearly, this error term $\varepsilon^{a,n}(t)$ is bounded by $\Delta^{a,n}_{A^n(t)+1}$. Then, by (3.4) and (3.5), we can re-write (3.3) as follows:

$$X^n(t) = Y^n(t) + R^n(t) + \varepsilon_1^n(t), \tag{3.6}$$

where

$$Y^n(t) := X^n(0) + \bar{\Delta}^d\left[M^{a,n}(A^n(t)) - M^d(A^n(t))\right] + n\int_0^t\left[\rho^n - \mathbb{f}^n(\theta^n(s))\right]ds, \tag{3.7}$$

$$R^n(t) := \sup_{s\leq t}(-Y^n(s) - \varepsilon_1^n(s)) \vee 0, \tag{3.8}$$

$$\varepsilon_1^n(t) := \bar{\Delta}^d\lambda^{a,n}\varepsilon^{a,n}(t), \tag{3.9}$$

for $t \geq 0$, where ρ^n is the traffic intensity of Assumption 3.2. The form of the process R^n in (3.8) is given by the solution of the one-dimensional Skorohod problem [37].

3.2. **Model II.** Now we consider the second queuing system, where the service time requirements are exponentially distributed and the service is not parallelized (in the sense that the servers are not working together on pending jobs). Let Q^n denote the number of customers in the system with n servers (including the customers in service). Let A^n be the point process defined in the beginning of this section and let D^n denote the departure process, i.e, $D^n(t)$ represents the number of job departures by time t. Since the service times are exponentially distributed, we can represent D^n by a time-changed unit rate Poisson process as follows:

$$D^n(t) = N\left(\int_0^t\lambda^d\Big(\left[\mathbb{f}^n(\theta^n(s))n\right] \wedge Q^n(s)\Big)ds\right),$$

where $N := \{N(t)\}_{t\geq 0}$ is a Poisson process with rate 1, which is independent of the other driving processes. Recall that $\lambda^d = \mathbb{E}\left[\Delta_l^d\right]^{-1}$ denotes the service rate and, in addition, notice that $\left[\mathbb{f}^n(\theta^n(t))n\right] \wedge Q^n(t)$ indicates the number of active servers at time t. Therefore, we can represent the process $Q^n(t)$ as the solution of the following equation:

$$Q^n(t) = Q^n(0) + A^n(t) - N\left(T^n(t)\right), \ t \geq 0, \tag{3.10}$$

where $Q^n(0)$ represents initial number of customers, which is assumed to be non-negative and independent of the other driving processes, and T^n is defined as

$$T^n(t) := \int_0^t \lambda^d \Big([\mathbb{f}^n(\theta^n(s))n] \wedge Q^n(s) \Big) ds, \quad t \geq 0. \tag{3.11}$$

This time-changed representation for the departure process has been used in [27] for an uncontrolled Markovian queue (with exponential arrival and service times) with n servers. Although the point process A^n here may not be necessarily a Poisson process, this representation is still valid, in the sense that there is a unique (up to indistinguishability) solution to (3.10), by the construction given in the proof of Theorem 4.1 (a) on page 327 of [9]. Notice that, in contrast to the model of Section 3.1, this time we do not need to compensate for idle times, since the rate of the time-changed Poisson process $D^n(t)$ will be zero at times t such that $Q^n(t) = 0$.

Proceeding in similar fashion to the previous section, let $M^{a,n}$ be defined as in (3.4) and let $\hat{N}$ be the centered Poisson process $\hat{N}(t) = N(t) - t$, for $t \geq 0$. Notice that,

$$A^n(t) = \sum_{l=1}^{A^n(t)} (1 - \lambda^{a,n}\Delta_l^{a,n}) + \lambda^{a,n} \sum_{l=1}^{A^n(t)} \Delta_l^{a,n} = M^{a,n}(A^n(t)) + \lambda^{a,n} t + \lambda^{a,n}\varepsilon^{a,n}(t),$$

where $\varepsilon^{a,n}(t)$ is the error term accounting for the time difference between t and the last arrival time, which is bounded by $\Delta^{a,n}_{A^n(t)+1}$. Therefore, we can re-write (3.10) as follows:

$$\begin{aligned} Q^n(t) = Q^n(0) + M^{a,n}(A^n(t)) - \hat{N}\left(T^n(t)\right) \\ + n\lambda^d \int_0^t \left[\rho^n - \mathbb{f}^n(\theta^n(s))\right] \vee \left[\rho^n - Q^n(s)/n\right] ds + \varepsilon_2^n(t), \quad t \geq 0 \end{aligned} \tag{3.12}$$

where $\varepsilon_2^n(t) := \lambda^{a,n}\varepsilon^{a,n}(t)$, and ρ^n is the traffic intensity of Assumption 3.2.

The scaled versions of the expressions (3.6)-(3.9) and (3.12) will be used to derive the heavy traffic limit in Section 5.

3.3. **The Scaled Queuing Systems.** In this section, we present the scaled version of the queuing models presented in the last section. We begin this section by presenting some common notation for both models and a preliminary lemma that will be used in Section 5.

Having in mind the definition of $M^{a,n}$ and M^d, given by (3.4) and that of A^n, let us define the following scaled processes:

$$\begin{aligned} m^{a,n}(t) &:= n^{-1/2} M^{a,n}(nt), \quad & m^{d,n}(t) &:= n^{-1/2} M^d(nt), \\ a^n(t) \quad &:= n^{-1} A^n(t), & \tilde{m}^{d,n}(t) &:= n^{-1/2} \hat{N}(nt), \quad \text{for } t \geq 0, \end{aligned}$$

The process $\hat{N}$ is the centered Poisson process associated with the departure process of Section 3.2.

Lemma 3.3. *The sequences of processes $\{m^{a,n}\}$, $\{m^{d,n}\}$ and $\{\tilde{m}^{d,n}\}$ are C-tight and the sequence $\{a^n\}$ converges in probability to $\lambda^s(\cdot)$, where $\lambda^s(t) := \lambda^s t$, for each $t \geq 0$.*

Proof. This result is well-known but we provide a proof in the Appendix A for the sake of completeness. □

In what follows, we present the scaled version of the queuing models of Section 3.1 and 3.2, respectively.

Queuing Model I: Let $x^n(t) := n^{-1/2}X^n(t)$, for $t \geq 0$, where $X^n(t)$ satisfies (3.3). By (3.6)-(3.9), notice that the scaled process satisfies the following equation:

$$x^n(t) = y^n(t) + r^n(t) + \bar{\varepsilon}_1^n(t), \tag{3.13}$$

with y^n given by:

$$y^n(t) := x^n(0) + \int_0^t b^n(\theta^n(s))ds + \bar{\Delta}^{d,n}w^n(t), \tag{3.14}$$

where

$$w^n(t) := m^{a,n}(a^n(t)) - m^{d,n}(a^n(t)),$$

and b^n is given by (3.2), $r^n(t) := n^{-1/2}R^n(t)$, and $\bar{\varepsilon}_1^n(t) := n^{-1/2}\varepsilon_1^n(t)$. Recall that $\varepsilon_1^n(t)$ is bounded by $\Delta^d\lambda^{a,n}\Delta^{a,n}_{A^n(t)+1}$, therefore $\bar{\varepsilon}_1^n(t)$ is bounded by $n^{-1/2}\Delta^d\lambda^{a,n}\Delta^{a,n}_{A^n(t)+1} = n^{-1/2}\Delta^d\lambda^{s,n}\Delta^{s,n}_{A^n(t)+1}$. For later use, let us define $\mathcal{F}_t^n := \sigma\{x^n(s),\, \theta^n(s);\, s \leq t\}$ for each $t \geq 0$, which denotes the σ-algebra generated by $\{x^n(s),\, \theta^n(s);\, s \leq t\}$.

Queuing Model II: For this model, we use a different type of scaling, which is similar to the one used in [16]. Let us define

$$q^n(t) := n^{-1/2}(Q^n(t) - n\rho^n), \quad t \geq 0,$$

where Q^n satisfies (3.12). Let the function $g^n : \mathbb{R} \times E \to \mathbb{R}$ be given by

$$g^n(\xi, i) = \lambda^d\Big(b^n(i) \vee -\xi\Big). \tag{3.15}$$

By (3.12), we have that q^n satisfies:

$$q^n(t) = q^n(0) + \int_0^t g^n(q^n(s), \theta^n(s))ds + \tilde{w}^n(t) + \bar{\varepsilon}_2^n(t), \quad t \geq 0 \tag{3.16}$$

where

$$\tilde{w}^n(t) := m^{a,n}(a^n(t)) - \tilde{m}^{d,n}(\bar{T}^n(t)),$$

$\bar{\varepsilon}_2^n(t) := n^{-1/2}\varepsilon_2^n(t)$ and $\bar{T}^n(t) = n^{-1}T^n(t)$, $t \geq 0$. Notice that using (3.11) and the function g^n given by (3.15), $\bar{T}^n$ can be rewritten as:

$$\bar{T}^n(t) := \lambda^{s,n}t - n^{-1/2}\int_0^t g^n(q^n(s), \theta^n(s))ds. \tag{3.17}$$

Similarly to what was done in the previous paragraph, let us define $\tilde{\mathcal{F}}_t^n := \sigma\{q^n(s),\, \theta^n(s);\, s \leq t\}$.

4. On the Pure Jump Process and the Control

We have defined in Section 3 the process θ^n as a *controlled pure jump process* taking value in a finite set E, which represents the state of the reserve machines. In this section, we present the assumptions on this process θ^n and the class of admissible controls. In order to have an interchanging notation for the two queuing models, let $\{\mathcal{G}_t^n\}$ denote either the filtration $\{\mathcal{F}_t^n\}$, when considering model I, or $\{\tilde{\mathcal{F}}_t^n\}$, when considering model II. Also, let (z^n, G) denote either $(x^n, \mathbb{R}_+)$, for model I, or $(q^n, \mathbb{R})$, for model II.

We begin by describing the class of controls considered here, which are usually called randomized Markovian controls in the Markov decision process literature (see, for instance, [15, Chapter 2]). A favorable feature of this class, in the context of this paper, is that it allows us to represent the policies determined by the numerical methods of Section 6. In order to define this class of controls, let $\mathcal{U}$ be a compact subset of the real numbers and $\mathcal{P}$ be the set of probability measures on $\mathcal{U}$, endowed with the Prohorov metric. The set $\mathcal{U}$ represents the set of all possible control values. A *randomized Markovian control* is then defined as a Borel measurable function v on $G \times E$ taking values in $\mathcal{P}$ (i.e., $v : G \times E \to \mathcal{P}$) [1].

Such controls, which assign a probability measure over the set of control values for a given state of the system, are sometimes called relaxed controls in the literature (see, for instance, [7]). Here, however, we leave this terminology to the more general relaxation (as in [22, Section 3.3]) that also includes the time variable, which will be presented later in this section. Relaxed control comes up in this paper as an essential tool in the context of the heavy traffic analysis. These relaxed controls are more suitable to derive the heavy traffic limit theorems. As will be clear in the following, we simply account here for the relaxed controls which are associated with a randomized control. Essentially, we consider the relaxed control which are obtained via an integral representation of the randomized control as in (4.2) implying that this relaxed control is an admissible control, as elaborated later on.

Before we go into the issue of the relaxed control we tarry for a moment in order to introduce an assumption. Except for the convergence of the transition rates, which is necessary for the heavy traffic analysis, this assumption has been used previously in the literature (see, for instance, (2.2) of [13, p. 1188] or (2.5) of [15, p. 11] and the discussion thereafter).

Assumption 4.1. For a randomized Markovian control $v : G \times E \to \mathcal{P}$, the associated transition rate functions $\lambda_{ij}^n : \mathcal{U} \to \mathbb{R}$ are such that:

$$\lim_{\delta \downarrow 0} \frac{1}{\delta}\big\{\mathbb{P}(\theta^n(t+\delta) = j | \theta^n(t) = i, \mathcal{G}_t^n) - \delta_{ij}\big\} = \int_{\mathcal{U}} \lambda_{ij}^n(\alpha) v(z^n(t), i)(d\alpha), \quad t \geq 0, \tag{4.1}$$

for each $i, j \in E$, where δ_{ij} is Kronecker's delta function and, for each $\alpha \in \mathcal{U}$ and n, the matrix $\Lambda^n(\alpha) := \{\lambda_{ij}^n(\alpha)\}$ satisfies $\lambda_{ij}^n(\alpha) \geq 0$ and $\lambda_{ii}^n(\alpha) = -\sum_{j \neq i} \lambda_{ij}^n(\alpha)$.

[1] Note that the definition of randomized Markovian control used here relates to the definition given in [15] by $\pi_t(C, i) = v(z^n(t), i)(C)$ for a state $i \in E$ and $C \in \mathcal{B}(\mathcal{U})$. In [15], the control policy is not dependent on another process as it is the case here.

In addition, we assume that there are functions $\lambda_{ij} : \mathcal{U} \to \mathbb{R}$ such that $\lambda_{ij}^n \to \lambda_{ij}$ uniformly as $n \to \infty$ for each $i, j \in E$. This implies in particular that $\lambda_{ij}(\alpha) \geq 0$ and $\lambda_{ii}(\alpha) = -\sum_{j \neq i} \lambda_{ij}(\alpha)$, for each $\alpha \in \mathcal{U}$.

Some remarks are now in order:

Remark 4.2. Of course the convergence of the matrices Λ^n to Λ is easily satisfied when $\Lambda^n = \Lambda$ for all sufficiently large n. Since, in practice, we usually consider a fixed queuing system with a fixed size n, which is approximated by the heavy traffic limit model, this condition is only introduced here to define precisely the notion of convergence needed to show the limit theorems and to allow greater generality to the results.

Remark 4.3. Notice that one can construct a pure jump process θ^n satisfying (4.1) that is adapted to $\{\mathcal{G}_t^n\}$ for any given randomized Markovian control v and transition rate functions $\lambda_{ij}^n : \mathcal{U} \to \mathbb{R}$ satisfying the conditions of Assumption 4.1. Since z^n represents either x^n or q^n, it has piecewise linear sample paths that behave deterministically between the jump-times of (z^n, θ^n). Therefore, θ^n can be constructed as pieced together non-homogeneous continuous-time Markov chains on E between the jump times of z^n.

Relaxed controls: relaxed controls are very useful in showing the weak convergence of control sequences. This class of controls will be used in the proofs of Theorems 5.5 and 5.7. We will discuss in this section how randomized Markovian controls applied to the stochastic process of interest, i.e., $v(z^n, \theta^n)$, can be represented as an admissible relaxed control, μ^n, with respect to the filtration engendered by (z^n, θ^n). A sequence of such admissible relaxed controls $\{\mu^n\}$ are easily shown to be tight since the space of relaxed controls, as discussed below, is compact. In order to characterize a weakly converging subsequence of $\{\mu^n\}$ with limit μ, we use Lemma 5.4 (of the next section) in order to show that μ is the relaxed control associated with the same randomized Markovian control v applied to the limit processes (z, θ).

In this sense, we will briefly introduce some notation and recall a few results about this subject. As before, let $\mathcal{U}$ be compact and $\mathcal{R}(\mathcal{U} \times [0, \infty))$ denote the set of measures $\mu(\cdot)$ on $\mathcal{B}(\mathcal{U} \times [0, \infty))$ satisfying $\mu(\mathcal{U} \times [0, t]) = t$ for all $t \geq 0$, where $\mathcal{B}(\mathcal{U} \times [0, \infty))$ denotes the σ-algebra of Borel subsets of $\mathcal{U} \times [0, \infty)$. We endow $\mathcal{R}(\mathcal{U} \times [0, \infty))$ with the weak compact topology, induced by the following notion of convergence: a sequence $\{\mu^n\} \subset \mathcal{R}(\mathcal{U} \times [0, \infty))$ converges to $\mu \in \mathcal{R}(\mathcal{U} \times [0, \infty))$ if and only if

$$\int_0^\infty \int_{\mathcal{U}} \varphi(\alpha, s) \mu^n(d\alpha ds) \to \int_0^\infty \int_{\mathcal{U}} \varphi(\alpha, s) \mu(d\alpha ds), \quad \text{as } n \to \infty,$$

for all real valued continuous functions φ that have compact support. The space $\mathcal{R}(\mathcal{U} \times [0, \infty))$ is compact under this topology, since $\mathcal{U}$ is compact. This implies that any sequence has a converging subsequence (for more detail, see [22, p. 47]).

A random variable μ defined on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P})$ taking values in $\mathcal{R}(\mathcal{U} \times [0, \infty))$ is called an *admissible relaxed control* with respect to $\{\mathcal{F}_t\}$ if the function defined as $\mu(B, t) := \mu(B \times [0, t])$ is $\{\mathcal{F}_t\}$-adapted for each $B \in \mathcal{B}(\mathcal{U})$. Since $\mu(B, t)$ are non-decreasing in t, the derivative of $\mu(B, t)$ with respect to t, which we will denote by $\mu_t(B)$, exists for almost all t, for each B.

Now, let $v : G \times E \to \mathcal{P}$ be a randomized Markovian control, and let (ζ, ϑ) be a stochastic process taking values in $G \times E$ that is adapted to a filtration $\{\mathcal{F}_t\}$. Then, we can define a relaxed control μ^v associated with $v := v(\zeta, \vartheta)$ as follows:

$$\mu^v(B \times S) := \int_S v(\zeta(s), \vartheta(s))(B)ds, \tag{4.2}$$

for each $B \times S \in \mathcal{B}(\mathcal{U} \times [0, \infty))$. That is, $v(\zeta(\cdot), \vartheta(\cdot))(B)$ is the derivative of $\mu^v(B, \cdot)$. In addition, (4.2) implies that μ^v is an admissible control with respect to $\{\mathcal{F}_t\}$.

5. Switching Diffusion Approximations

In this section we present the heavy traffic limit theorems. We show that the scaled versions of the queuing models presented in Section 3.1 and 3.2 controlled (through θ^n) by a given randomized Markovian control v converge in distribution to a switching diffusion controlled by the same control function v. For model I, this limit controlled switching diffusion can be formally defined as the pair (x, θ) taking values in $\mathbb{R}_+ \times E$ satisfying the following equations:

$$x(t) = x(0) + \int_0^t b(\theta(s))ds + \sigma w(t) + r(t), \tag{5.1}$$

$$\mathbb{P}\left(\theta(t+\delta) = j \middle| \theta(t) = i, x(s), \theta(s), s \le t\right) = \int_{\mathcal{U}} \lambda_{ij}(\alpha)v(x(s), \theta(s))(d\alpha)\delta + o(\delta), \tag{5.2}$$

for $t \ge 0$, $i \ne j$, where $b : E \to \mathbb{R}$ is given by (3.2), $\sigma > 0$ the diffusion coefficient and w is a standard Brownian motion, and r is the "reflection process," which is non-decreasing, increases only at time $t \ge 0$ such that $x(t) = 0$ and satisfies $r(0) = 0$. The function $v : \mathbb{R}_+ \times E \to \mathcal{P}$ in (5.2) is the randomized Markovian control function and $\lambda_{ij} : \mathcal{U} \to \mathbb{R}$ are the controlled transition rate functions for θ, which are assumed (under Assumption 4.1) to satisfy $\lambda_{ij}(\alpha) \ge 0$, for $i \ne j$, and $\lambda_{ii}(\alpha) = -\sum_{j \ne i} \lambda_{ij}(\alpha)$, for each $i \in E$ and $\alpha \in \mathcal{U}$.

For model II, the limit controlled switching diffusion process (q, θ) is defined likewise, where q satisfies an equation similar to (5.1), except that the reflection term r in (5.1) is not present and the drift term depends on q, that is,

$$q(t) = q(0) + \int_0^t g(q(s), \theta(s))ds + \varrho w(t), \quad t \ge 0 \tag{5.3}$$

where $g : \mathbb{R} \times E \to \mathbb{R}$ is the drift function, given by

$$g(\xi, i) = \lambda^d\Big(b(i) \vee -\xi\Big), \tag{5.4}$$

$\varrho > 0$ is the diffusion coefficient, and the process θ satisfies (5.2) with q in place of x, for a control function $v : \mathbb{R} \times E \to \mathcal{P}$.

The weak convergence of each model is considered separately in the following Sections 5.1 and 5.2. In Section 5.3, we show how one can use the switching diffusion limits to approximate the behavior of these queuing systems.

5.1. **Switching Diffusion Approximation for Model I.** We begin this section by defining precisely the switching diffusion considered here. We follow the approach in [13] and write the process as a jump-diffusion. Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P})$ be a filtered probability space (satisfying the usual conditions), where the processes are to be defined. Let $v : \mathbb{R}_+ \times E \to \mathcal{P}$ be a randomized Markovian control and let $\lambda_{ij}^v : \mathbb{R}_+ \to \mathbb{R}$ be given by:

$$\lambda_{ij}^v(\xi) := \int_{\mathcal{U}} \lambda_{ij}(\alpha) v(\xi, i)(d\alpha), \quad \text{for } i, j \in E \text{ and } \xi \in \mathbb{R}_+, \tag{5.5}$$

where λ_{ij} are given by Assumption 4.1. Similarly to [13], for each $\xi \in \mathbb{R}$ and $v \in \mathcal{P}$, let $\{\Delta_{ij}^v(\xi)\}_{i,j\in E}$ be disjoint intervals of the real line, where each interval $\Delta_{ij}^v(\xi)$ has length $\lambda_{ij}^v(\xi)$. Let $h^v : \mathbb{R}_+ \times E \times \mathbb{R} \to \mathbb{R}$ be defined by

$$h^v(\xi, i, \gamma) := \sum_{j \in E} (j - i) \mathbb{I}_{\{\gamma \in \Delta_{ij}^v(\xi, i)\}}.$$

Recall that $\mathbb{I}_{\{\cdot \in S\}}$ denotes the indicator function of the set S.

Now let x_0 and θ_0 be given $\mathcal{F}_0$-measurable random variables taking values in $\mathbb{R}_+$ and E, respectively. Let w be a standard Brownian motion, which is a martingale with respect to $\{\mathcal{F}_t\}$ and independent of x_0 and θ_0, and let $\mathfrak{p}$ be a Poisson random measure relative to $\{\mathcal{F}_t\}$, independent of w, x_0, θ_0 and with intensity $dt \times m(d\gamma)$, where m denotes the Lebesgue measure in $\mathbb{R}$. The pair (x, θ) is a reflected controlled switching diffusion with initial condition (x_0, θ_0), drift function $b : E \to \mathbb{R}$ and diffusion coefficient $\sigma > 0$ if x is a continuous $\{\mathcal{F}_t\}$-adapted process; θ is a right-continuous with left-limits $\{\mathcal{F}_t\}$-adapted process; and (x, θ) satisfy the following equations:

$$x(t) = x_0 + \int_0^t b(\theta(s))ds + \sigma w(t) + r(t) \tag{5.6}$$

$$\theta(t) = \theta_0 + \int_0^t \int_{\mathbb{R}} h^v(x(s), \theta(s-), \gamma)\mathfrak{p}(ds, d\gamma) \quad \text{for } t \geq 0, \tag{5.7}$$

for some continuous non-decreasing process r, satisfying $r(0) = 0$, which increases only in times t such that $r(t) = 0$. In particular, r satisfies:

$$r(t) = \int_0^t \mathbb{I}_{\{x(s)=0\}} dr(s). \tag{5.8}$$

Theorem 5.1. *Given (x_0, θ_0), v, w, and $\mathfrak{p}$, as specified in the above paragraph, there is a unique reflected controlled switching diffusion (x, θ) satisfying* (5.6) *and* (5.7) *for a given drift function b and diffusion coefficient $\sigma > 0$.*

Proof. It is well known that there is a strong solution for a reflected diffusion of the form:

$$x(t) = x_0 + \int_0^t b(i)ds + \sigma w(t) + r(t),$$

for each $i \in E$ (see [37]). The result then follows by the same construction via an interlacing procedure and the unicity argument, which are used in Theorem 3.1 of [31, p. 2457]. □

We now define the associated infinitesimal generator to the above defined reflected controlled switching diffusion. This generator will be defined with respect to a relaxed control. This more general approach will be useful in showing weak convergence in Theorem 5.5. For a relaxed control μ, the *infinitesimal generator*, which is denoted by $\mathcal{L}^\mu$, has domain given by $\mathcal{D}_+ := \{f \in C_0^2(\mathbb{R}_+ \times E) | f_x(0,i) \geq 0, i \in E\}$, where f_x denotes the partial derivative of f with respect to the first argument and $\mathcal{L}^\mu$ is defined as follows:

$$(\mathcal{L}^\mu f)(s,\xi,i) = b(i) f_x(\xi,i) + \frac{\sigma^2}{2} f_{xx}(\xi,i) + \sum_{j \in E} \int_{\mathcal{U}} f(\xi,j) \lambda_{ij}(\alpha) \mu_s(d\alpha), \tag{5.9}$$

for $(s,\xi,i) \in [0,\infty) \times \mathbb{R}_+ \times E$ and function f in the domain $\mathcal{D}_+$ of $\mathcal{L}^\mu$, where f_{xx} denotes the second partial derivative with respect to the first argument. Whenever μ is the relaxed control associated with the randomized Markovian control v, as given by (4.2), we will use the following notation for the infinitesimal generator $\mathcal{L}^v$. Although it is an slight abuse of notation, it is very convenient and simplifies the notation considerably.

Now we define the associated submartingale problem. Recall the definition of x^n and θ^n as the scaled processes for model I, defined by (3.13)-(3.14) and (4.1). This submartingale problem will be used to characterize the weak-sense limit of the processes (x^n, θ^n) and that of the control $v(x^n, \theta^n)$ as well as to establish the desired adaptiveness property for the limit processes.

Definition 5.2. [21, p. 146] Suppose that (ζ, ϑ) is a stochastic process with sample paths that are right-continuous with left limits. Let μ be a relaxed control. The process (ζ, ϑ) is said to solve the submartingale problem for $\mathcal{L}^\mu$ if there exists a filtration $\{\mathcal{F}_t\}$ such that (ζ, ϑ) is $\{\mathcal{F}_t\}$-adapted, μ is an admissible control with respect to $\{\mathcal{F}_t\}$ and the process defined by:

$$f(\zeta(t), \vartheta(t)) - f(\zeta(0), \vartheta(0)) - \int_0^t (\mathcal{L}^\mu f)(s, \zeta(s), \vartheta(s)) ds, \quad t \geq 0$$

is an $\{\mathcal{F}_t\}$-submartingale for each $f \in \mathcal{D}_+$.

Remark 5.3. Notice that a function $f \in \mathcal{D}_+$ can be smoothly extended to have domain $\mathbb{R}^2$, so that this extension, $\bar{f}$, satisfies $\bar{f}_x(0,i) \geq 0$, for $i \in E$, $\bar{f} = f$ on $\mathbb{R}_+ \times E$ and belongs to $C_0^2(\mathbb{R}^2)$, the set of real valued functions on $\mathbb{R}^2$ that are continuous, have compact support, and have continuous first and second partial derivatives. Using Itô's formula (see, Theorem 4.57 [18, p. 57]) on this extended function $\bar{f}$, we have that a reflected controlled switching diffusion satisfying (5.6) and (5.7) solve the submartingale problem for $\mathcal{L}^v$ as stated above.

Let $\{\mathcal{F}_t^n\}$ be the filtration given by $\sigma\{x^n(s), \theta^n(s); s \leq t\}$. Given the discussion above, it is clear that the martingale property given below is useful to characterize the weak-sense limit. Using (4.1), the fact that x^n is continuous almost everywhere and under the assumption of Lemma 5.4, given below, we have that

$$M_f^n(t) := f(\theta^n(t)) - f(\theta^n(0)) - \int_0^t \sum_{j \in E} \int_{\mathcal{U}} \lambda^n_{\theta^n(s)j}(\alpha) f(j) v(x^n(s), \theta^n(s))(d\alpha) ds \tag{5.10}$$

is an $\{\mathcal{F}_t^n\}$-martingale, for any function $f : E \to \mathbb{R}$.

The following lemma will be used to characterize the weak-sense limit of the control. It says that if (x, θ) satisfies the submartingale problem for $\mathcal{L}^\mu$, for some relaxed control μ, then the law induced by x assigns zero probability to sample functions that spend more than a negligible amount of time at the set of discontinuity points of v. This implies that if (x^n, θ^n) converges weakly to (x, θ) and the latter is a solution to the martingale problem for $\mathcal{L}^\mu$, then the relaxed control μ^{v^n} associated with $v(x^n, \theta^n)$, as in (4.2), converges to μ, whose derivative has the same law as $v(x, \theta)$. The control policies determined by the numerical methods of Section 7 satisfy the piecewise constant assumption of the lemma below.

Lemma 5.4. *Let v be a randomized Markovian control function which is piecewise constant and has a finite number of discontinuity points with respect to the first argument. Suppose that μ is a relaxed control and that (x, θ) is a solution to the submartingale problem for $\mathcal{L}^\mu$. Let G_d denote the finite set of discontinuity points of v and let us define the neighborhood $N_\epsilon(G_d) = \{x \in G \mid \text{dist}(x, G_d) \leq \epsilon\}$ for $\epsilon > 0$. Then, for any $t \geq 0$ and $\delta > 0$ we have that:*

$$\lim_{\epsilon \to 0} \mathbb{P}\left(\int_0^t \mathbb{I}_{N_\epsilon(G_d)}(x(s)) ds \geq \delta \right) = 0.$$

Proof. See the Appendix A. □

Now we are ready to present the main theorem of this section.

Theorem 5.5. *Suppose that the initial condition for the n-th system $(x^n(0), \theta^n(0))$ converges weakly to (x_0, θ_0), a random variable taking values in $\mathbb{R}_+ \times E$. Let (x^n, θ^n) be the stochastic process satisfying* (3.13)-(3.14) *and* (4.1) *for a randomized Markovian control v satisfying the assumption on Lemma 5.4. Then (x^n, θ^n) converges in distribution to the reflected controlled switching diffusion having initial condition (x_0, θ_0), drift given by* (3.2)*, diffusion coefficient $\sigma := \sqrt{\lambda^s (\bar{\Delta}^d)^2 (\sigma_a^2 + \sigma_d^2)}$ and control v.*

Proof. We begin by showing tightness of the sequence $\Psi^n = \{(x^n, \theta^n)\}$. Notice that (4.1) together with the assumption that $\lambda_{ij}^n \to \lambda_{ij}$ implies that $\{\theta^n\}$ is tight by Theorem 2.7(a) of [20, p. 10]. Let $d^n(\cdot) := \int_0^\cdot b^n(\theta^n(s)) ds$, Theorem 2.7(a) of [20, p. 10] can also be used to show that $\{d^n\}$ is tight. It is also asymptotic continuous, or C-tight, since d^n is continuous (e.g., Proposition VI.3.26 of [18, p. 351]). In addition, by Lemma 3.3, $\{(m^{a,n}, m^{d,n}, a^n)\}$ is C-tight and since the composition mapping is continuous (e.g., Theorem 3.1 of [39, p. 75]), we have that $\{w^n\}$ is also C-tight. Therefore, $\{y^n\}$ is C-tight.

Recall that the error term $\bar{\varepsilon}_1^n(t) := n^{-1/2} \varepsilon_1^n(t)$, appearing in (3.13), is bounded by $n^{-1/2} \bar{\Delta}^d \lambda^{s,n} \Delta^{s,n}_{A^n(t)+1}$ and let us define the process $\eta^{s,n}$ as the one taking values $\eta^{s,n}(t) := n^{-1/2} \Delta^{s,n}_{A^n(t)+1}$, for $t \geq 0$. Now, for $T > 0$, let $p_n := \mathbb{P}\left(A^n(T) > \lambda^s n(T + \delta)\right)$, then we

have

$$
\begin{aligned}
(5.11)\qquad & \mathbb{P}\left(\sup_{t\le T}|\eta^{s,n}(t)|\ge\delta\right)\\
&\le \mathbb{P}\left(\sup_{l\le A^n(T)+1} n^{-1/2}|\Delta_l^{s,n}|\ge\delta, A^n(T)\le\lambda^s n(T+\delta)\right)+p_n\\
&\le \sum_{l=1}^{\lambda^s n(T+\delta)+1}\mathbb{P}\left(|\Delta_l^{s,n}|\ge\delta n^{1/2}\right)+p_n\\
&\le \mathbb{E}\left[\mathbb{I}_{\{|\Delta_1^{s,n}|\ge\delta n^{1/2}\}}|\Delta_1^{s,n}|^2\right]\frac{\lambda^s n(T+\delta)+1}{\delta^2 n}+p_n,
\end{aligned}
$$

where we used Chebychev's inequality in the third line and the fact that $\Delta_l^{s,n}$ are identically distributed. Since $\{\Delta_1^{s,n};n\}$ is uniformly integrable and by Lemma 3.3, we have that the right-hand side of (5.11) goes to zero as $n\to\infty$. Hence, $\eta^{s,n}$ and $\bar{\varepsilon}_1^n$ converge in probability to the "zero" process. This, together with the fact that $\{y^n\}$ is C-tight, implies that the sequences $\{x^n\}$ and $\{r^n\}$ are C-tight since the reflection map is continuous (see, e.g., [40, p. 439]). Therefore, we have shown that $\{(x^n,\theta^n)\}$ is tight.

Now we characterize the limit of any weakly converging subsequence of $\{(x^n,\theta^n)\}$ as a solution to the submartingale problem for $\mathcal{L}^\mu$, for some relaxed control μ. For this, suppose that (x,θ) is a weak-sense limit of a converging subsequence and let $\{\mathcal{F}_t\}$ be its natural filtration. Let p be an integer and $\{t_k\}_{k=1}^p$ a set of real numbers such that $0\le t_k\le t$ for each $k=1,\dots,p$. Let $h_k:\mathbb{R}_+\times E\to\mathbb{R}$ be continuous and bounded functions, for $k=1,\dots,p$. In order to show that (x,θ) is a solution to the $\{\mathcal{F}_t\}$-submartingale problem for $\mathcal{L}^\mu$, it is enough to show that it satisfies:

$$
\begin{aligned}
(5.12)\qquad & \mathbb{E}\left[\left[f(x(t+\tau),\theta(t+\tau))-f(x(t),\theta(t))\right]\prod_{k=1}^p h_k(x(t_k),\theta(t_k))\right]\\
&\qquad\ge \mathbb{E}\left[\int_t^{t+\tau}(\mathcal{L}^\mu f)(s,x(s),\theta(s))ds\prod_{k=1}^p h_k(x(t_k),\theta(t_k))\right],
\end{aligned}
$$

fore each $t,\tau\ge0$, $f\in\mathcal{D}_+$, p, $\{t_k\}_{k=1}^p$ and continuous and bounded functions $\{h_k\}_{k=1}^p$.

Now, we will show (5.12). For that, let μ^n be the relaxed control associated with $v(x^n,\theta^n)$, in the sense of (4.2). Since $\mathcal{R}(\mathcal{U}\times[0,\infty))$ is compact, the sequence $\{\mu^n\}$ is tight. Now let $\Upsilon^n:=(x^n,\,\theta^n,\,y^n,\,r^n,\,m^{a,n},\,m^{d,n},\,a^n,\,\eta^{s,n},\,\bar{\varepsilon}_1^n,\,\mu^n)$, we have already shown that $\{\Upsilon^n\}$ is tight. Let us extract a weakly converging subsequence from $\{\Upsilon^n\}$, for convenience of notation, this converging subsequence will also be indexed with the subscript n, and let $\Upsilon:=(x,\theta,y,r,m^a,m^d,a,\eta^s,\bar{\varepsilon}_1,\mu)$ denote its weak sense limit. We use the Skorohod Representation Theorem, so that we can suppose that each component of Υ^n converges almost surely in their appropriate topologies. Let $\tilde{\mathcal{G}}_t^n$ denote the minimal σ-algebra with respect to which $\{\Upsilon^n(s);s\le t\}$ is measurable.

For given $f\in\mathcal{D}_+$, $t,\tau\ge0$, define $s_{0,t}^{a,n}:=t$ and $s_{i,t}^{a,n}:=\tilde{s}_{i,t}^{a,n}\wedge(t+\tau)$, where $\{\tilde{s}_{i,t}^{a,n}\}$ are the (increasing in i) jump times of a^n after t for $i\in\{1,2,\dots\}$. Let $d_{i,t}^{a,n}:=|s_{i+1,t}^{a,n}-s_{i,t}^{a,n}|$,

$i \geq 0$, and notice that $d_{i,t}^{a,n} \leq \Delta_l^{a,n} = n^{-1}\Delta_l^{s,n}$, for some $l \geq 1$. Therefore, $\sup_i d_{i,t}^{a,n} \to 0$ a.s. as $n \to \infty$ by the a.s. convergence of $\eta^{s,n}$ to the zero process. Now, notice that

$$
\begin{aligned}
(5.13) \qquad & f(x^n(t+\tau), \theta^n(t+\tau)) - f(x^n(t), \theta^n(t)) \\
& = \sum_{i=0}^{\infty} \mathbb{I}_{\{s_{i,t}^{a,n} < t+\tau\}} \left[f(x^n(s_{i+1,t}^{a,n}), \theta^n(s_{i+1,t}^{a,n})) - f(x^n(s_{i,t}^{a,n}), \theta^n(s_{i,t}^{a,n})) \right].
\end{aligned}
$$

Considering the term inside the bracket for some i such that $s_{i,t}^{a,n} < t + \tau$ and using Taylor's Theorem, we have that

$$
\begin{aligned}
(5.14) \quad & f(x^n(s_{i+1,t}^{a,n}), \theta^n(s_{i+1,t}^{a,n})) - f(x^n(s_{i,t}^{a,n}), \theta^n(s_{i,t}^{a,n})) \\
& = f(x^n(s_{i,t}^{a,n}), \theta^n(s_{i+1,t}^{a,n})) - f(x^n(s_{i,t}^{a,n}), \theta^n(s_{i,t}^{a,n})) + \Delta_{i+1,t}^{x,n} f_x(x^n(s_{i,t}^{a,n}), \theta^n(s_{i,t}^{a,n})) \\
& \quad + (\Delta_{i+1,t}^{x,n})^2 f_{xx}(x^n(s_{i,t}^{a,n}), \theta^n(s_{i,t}^{a,n})) + \delta_{1,i}^n + \delta_{2,i}^n + \delta_{3,i}^n + \delta_{4,i}^n,
\end{aligned}
$$

where $\Delta_{i+1,t}^{x,n}$ is defined to be

$$
(5.15) \qquad \Delta_{i+1,t}^{x,n} := \frac{\bar{\Delta}^d}{\sqrt{n}} \left(\xi_{i+1,t}^{a,n} - \xi_{i+1,t}^{d} \right) + \int_{s_{i,t}^{a,n}}^{s_{i+1,t}^{a,n}} b^n(\theta^n(s)) ds + \int_{s_{i,t}^{a,n}}^{s_{i+1,t}^{a,n}} dr^n(s),
$$

$\xi_{i+1,t}^{a,n} := (1 - \lambda^{a,n}\Delta_{i+1,t}^{a,n})$ and $\xi_{i+1,t}^{d} := (1 - \lambda^{d}\Delta_{i+1,t}^{d})$, with $\Delta_{i+1,t}^{a,n}$ and $\Delta_{i+1,t}^{d}$ denoting the $(i+1)$-th inter-arrival and service time after time t, respectively. In addition, the terms $\delta_{j,i}^n$, $j = 1, \ldots, 4$, above are given by

$$
\begin{aligned}
\delta_{1,i}^n &= \Delta_{i+1,t}^{x,n} \left[f_x(x^n(s_{i,t}^n), \theta^n(s_{i+1,t}^{a,n})) - f_x(x^n(s_{i,t}^n), \theta^n(s_{i,t}^{a,n})) \right] \\
\delta_{2,i}^n &= (\Delta_{i+1,t}^{x,n})^2 \left[f_{xx}(x^n(s_{i,t}^{a,n}) + d_i^n, \theta^n(s_{i+1,t}^{a,n})) - f_{xx}(x^n(s_{i,t}^{a,n}), \theta^n(s_{i+1,t}^{a,n})) \right] \\
\delta_{3,i}^n &= (\Delta_{i+1,t}^{x,n})^2 \left[f_{xx}(x^n(s_{i,t}^{a,n}), \theta^n(s_{i+1,t}^{a,n})) - f_{xx}(x^n(s_{i,t}^{a,n}), \theta^n(s_{i,t}^{a,n})) \right] \\
\delta_{4,i}^n &= \Delta_{i+1,t}^{\bar{\varepsilon},n} \left[\Delta_{i+1,t}^{x,n} f_x(x^n(s_{i,t}^n), \theta^n(s_{i+1,t}^{a,n})) + (\Delta_{i+1,t}^{x,n})^2 f_{xx}(x^n(s_{i,t}^{a,n}), \theta^n(s_{i+1,t}^{a,n})) \right]
\end{aligned}
$$

for $d_i^n = c_i^n \Delta_{i+1}^{x,n}$, $c_i^n \in [0,1]$, and $\Delta_{i+1,t}^{\bar{\varepsilon},n} := (\bar{\varepsilon}_1^n(s_{i+1,t}^{a,n}) - \bar{\varepsilon}_1^n(s_{i,t}^{a,n}))$. By (5.15), we have that

$$
(5.16) \qquad \mathbb{E}\left[\Delta_{i+1,t}^{x,n} \middle| \tilde{\mathcal{G}}_{s_{i,t}^{a,n}}^n \right] = \mathbb{E}\left[\int_{s_{i,t}^{a,n}}^{s_{i+1,t}^{a,n}} b^n(\theta^n(s)) ds + \int_{s_{i,t}^{a,n}}^{s_{i+1,t}^{a,n}} dr^n(s) \middle| \tilde{\mathcal{G}}_{s_{i,t}^{a,n}}^n \right],
$$

and also that

$$
(5.17) \qquad \mathbb{E}\left[(\Delta_{i+1,t}^{x,n})^2 \middle| \tilde{\mathcal{G}}_{s_{i,t}^{a,n}}^n \right] = \frac{(\bar{\Delta}^d)^2 (\sigma_a^{2,n} + \sigma_d^2)}{n} + \mathbb{E}\left[\delta_{i,5}^n \middle| \tilde{\mathcal{G}}_{s_{i,t}^{a,n}}^n \right],
$$

where the term $\delta_{i,5}^n$ is such that

$$
\begin{aligned}
\delta_{5,i}^n \leq & \frac{2\bar{\Delta}^{d,n}}{\sqrt{n}} \left(\xi_{i+1,t}^{a,n} - \xi_{i+1,t}^{d} \right) \left(\int_{s_{i,t}^{a,n}}^{s_{i+1,t}^{a,n}} b^n(\theta^n(s)) ds + \int_{s_{i,t}^{a,n}}^{s_{i+1,t}^{a,n}} dr^n(s) \right) \\
& + 2 \left(\int_{s_{i,t}^{a,n}}^{s_{i+1,t}^{a,n}} b^n(\theta^n(s)) ds \right)^2 + 2 \left(\int_{s_{i,t}^{a,n}}^{s_{i+1,t}^{a,n}} dr^n(s) \right)^2.
\end{aligned}
$$

Therefore, using the expansion given by (5.14), the conditional expectation of the difference (5.13) with respect to $\tilde{\mathcal{G}}_t^n$ can be written as:

$$
\begin{aligned}
(5.18)\quad \mathbb{E}\Bigg[&\sum_{i=0}^{\infty}\mathbb{I}_{\{s_{i,t}^{a,n}<t+\tau\}}\int_{s_{i,t}^{a,n}}^{s_{i+1,t}^{a,n}}\sum_{j\in E}\int_{\mathcal{U}}\lambda_{\theta^n(s)j}^n(\alpha)\mu_s^n(d\alpha)f(x^n(s_{i,t}^{a,n}),j)ds\\
&+\sum_{i=0}^{\infty}\mathbb{I}_{\{s_{i,t}^{a,n}<t+\tau\}}\int_{s_{i,t}^{a,n}}^{s_{i+1,t}^{a,n}}b^n(\theta^n(s))f_x(x^n(s_{i,t}^{a,n}),\theta^n(s_{i,t}^{a,n}))ds\\
&+\sum_{i=0}^{\infty}\mathbb{I}_{\{s_{i,t}^{a,n}<t+\tau\}}\int_{s_{i,t}^{a,n}}^{s_{i+1,t}^{a,n}}f_x(x^n(s_{i,t}^{a,n}),\theta^n(s_{i,t}^{a,n}))dr^n(s)\\
&+\sum_{i=0}^{\infty}\mathbb{I}_{\{s_{i,t}^{a,n}<t+\tau\}}\frac{\lambda^{s,n}(\bar{\Delta}^d)^2\,(\sigma_a^{2,n}+\sigma_d^2)}{2}\int_{s_{i,t}^{a,n}}^{s_{i+1,t}^{a,n}}f_{xx}(x^n(s_{i,t}^{a,n}),\theta^n(s_{i,t}^{a,n}))ds\,\Bigg|\,\tilde{\mathcal{G}}_t^n\Bigg],
\end{aligned}
$$

module the terms involving $\delta_{j,i}^n$, for $j=1,\ldots,5$, where we used the martingale property given by (5.10) in the first line, but with the derivative $\mu_\cdot^n$ replacing $v(x^n,\theta^n)$, we used (5.16) for the second and third line and, for the last line, we used (5.17) and the fact that

$$
\frac{1}{n}=\lambda^{s,n}\bar{\Delta}^{a,n}=\lambda^{s,n}\mathbb{E}\left[(s_{i+1,t}^{a,n}-s_{i,t}^{a,n})\big|\,\tilde{\mathcal{G}}_{s_{i,t}^{a,n}}^n\right],
$$

by (3.1). Assume for now that the terms involving $\delta_{j,i}^n$ converge almost surely to zero.

Now, let us define the piecewise constant versions of x^n and θ^n as

$$
\begin{aligned}
\check{x}^n(s)&:=x^n(s_{i,t}^{a,n})\text{ for }s\in[s_{i,t}^{a,n},s_{i+1,t}^{a,n}),\\
\check{\theta}^n(s)&:=\theta^n(s_{i,t}^{a,n})\text{ for }s\in[s_{i,t}^{a,n},s_{i+1,t}^{a,n}).
\end{aligned}
$$

Therefore, we have that (5.18) can be further rewritten as:

(5.19)

$$
\begin{aligned}
\mathbb{E}\Big[&\int_t^{t+\tau}\sum_{j\in E}\int_{\mathcal{U}}\lambda_{\theta^n(s)j}^n(\alpha)\mu_s^n(d\alpha)f(\check{x}^n(s),j)+b^n(\theta^n(s))f_x(\check{x}^n(s),\check{\theta}^n(s))\\
&+\frac{\lambda^{s,n}(\bar{\Delta}^d)^2\,(\sigma_a^{2,n}+\sigma_d^2)}{2}f_{xx}(\check{x}^n(s),\check{\theta}^n(s))ds+\int_t^{t+\tau}f_x(\check{x}^n(s),\check{\theta}^n(s))dr^n(s)\Big|\tilde{\mathcal{G}}_t^n\Big].
\end{aligned}
$$

Since $\mathcal{R}(\mathcal{U}\times[0,\infty))$ is endowed with the topology of weak convergence, we have that:

$$
\begin{aligned}
\int_t^{t+\tau}\int_{\mathcal{U}}\lambda_{\theta^n(s)j}^n(\alpha)\mu_s^n(d\alpha)f(\check{x}^n(s),j)ds&=\int_t^{t+\tau}\int_{\mathcal{U}}\lambda_{\theta(s)j}(\alpha)\mu_s^n(d\alpha)f(x(s),j)ds+\tilde{\delta}_j^n\\
&\longrightarrow\int_t^{t+\tau}\int_{\mathcal{U}}\lambda_{\theta(s)j}(\alpha)\mu_s(d\alpha)f(x(s),j)ds,\quad\text{a.s.},
\end{aligned}
$$

for each $j \in E$, where $\tilde{\delta}_j^n$ is given by

$$\begin{aligned}\tilde{\delta}_j^n &= \int_t^{t+\tau} \int_{\mathcal{U}} \Big[\lambda^n_{\theta^n(s)j}(\alpha) f(\check{x}^n(s), j) - \lambda_{\theta(s)j}(\alpha) f(x(s), j)\Big] \mu_s^n(d\alpha) ds \\ &= \int_t^{t+\tau} \int_{\mathcal{U}} \Big\{ \left[\lambda^n_{\theta^n(s)j}(\alpha) - \lambda_{\theta^n(s)j}(\alpha)\right] f(\check{x}^n(s), j) \\ &\quad + \left[\lambda_{\theta^n(s)j}(\alpha) - \lambda_{\theta(\beta^n(s))j}(\alpha)\right] f(\check{x}^n(s), j) + \left[\lambda_{\theta(\beta^n(s))j}(\alpha) - \lambda_{\theta(s)j}(\alpha)\right] f(\check{x}^n(s), j) \\ &\quad + \lambda_{\theta(s)j}(\alpha) \left[f(\check{x}^n(s), j) - f(x(s), j)\right] \Big\} \mu_s^n(d\alpha) ds,\end{aligned}$$

where β^n are non-decreasing functions mapping $[0, T]$ to $[0, T]$, for $T > t + \tau$. Notice that Proposition 5.3 of [9, p. 119], together with the assumption that λ_{ij}^n converges uniformly to λ_{ij}, the fact that f is continuous with respect to its first argument and that θ has at most a finite number of discontinuity points in a bounded time interval, implies that $\tilde{\delta}_j^n \to 0$ a.s..

Therefore, multiplying (5.19) by $\prod_{k=1}^p h_k(x^n(t_k), \theta^n(t_k))$, where p, $\{t_k\}_{k=1}^p$ and $\{h_k\}_{k=1}^p$ are defined as in (5.12), taking expectation and the limit as $n \to \infty$, we get

$$\mathbb{E}\left[\Big(\int_t^{t+\tau} (\mathcal{L}^\mu f)(s, x(s), \theta(s)) ds + \int_t^{t+\tau} f_x(x(s), \theta(s)) dr(s)\Big) \prod_{k=1}^p h_k(x(t_k), \theta(t_k))\right],$$

by Lebesgue's Dominated Convergence Theorem, the fact f can be smoothly extended to $C_0^2(\mathbb{R}^2)$ as in Remark 5.3, that (x, r) is continuous a.s., and θ has a finite number of jumps in a finite time interval. Since (x, r) is the solution of the Skorohod problem for y and by the continuity of the reflection map, the process r must satisfy (5.8). This implies that

$$\int_t^{t+\tau} f_x(x(s), \theta(s)) dr(s) \geq 0,$$

since $f_x(0, \cdot) \geq 0$ and, therefore, we complete the proof of (5.12) by showing that the terms $\sum_{i=0}^\infty \mathbb{I}_{\{s_{i,t}^{a,n} < t+\tau\}} \delta_{j,i}^n$, for $j = 1, \ldots, 4$, and $\sum_{i=0}^\infty \mathbb{I}_{\{s_{i,t}^{a,n} < t+\tau\}} \delta_{5,i}^n f_{xx}(x^n(s_{i,t}^{a,n}), \theta^n(s_{i,t}^{a,n}))$ that were omitted in (5.18) converge to zero in the mean.

In order to show that, let us first consider the case $j = 1$. By Proposition 5.3 of [9, p. 119], there is a sequence of non-decreasing functions $\{\beta^n\}$ mapping $[0, T]$ to $[0, T]$, for $T > t + \tau$, such that $\sup_{s \in [t, t+\tau]} \left\{f_x(x^n(s_{i,t}^n), \theta^n(s)) - f_x(x^n(s_{i,t}^n), \theta(\beta^n(s)))\right\} \to 0$ almost surely as $n \to \infty$. Also, $\Delta_{i+1,t}^{x,n} \to 0$ almost surely as $n \to \infty$, since $m^{a,n}$, $m^{d,n}$, d^n, and r^n are asymptotically continuous processes. This together with the fact that θ has at most a finite number of jumps in finite time completes the proof for $j = 1$. For the remaining terms, where $j \neq 1$, the argument is similar, using in addition the facts that f_{xx} is continuous with respect to the first argument and $\bar{\varepsilon}_1^n$ converges a.s. to the zero process.

Now that we have shown that (x, θ) is a solution to the submartingale problem for $\mathcal{L}^\mu$, we can characterize the limit of the relaxed control sequence $\{\mu^n\}$. By Lemma 5.4, the law induced by x does not charge sets whose sample functions spend more than a negligible amount of time near the discontinuity points of $v(\cdot, i)$. With this lemma and

using the Skorohod representation, it is straight forward to show that

$$\int_0^t \int_{\mathcal{U}} \varphi(\alpha, s)\mathbb{I}_{\{\theta^n(s)=i\}}\mu^n(d\alpha ds) = \int_0^t \int_{\mathcal{U}} \varphi(\alpha, s)v(x^n(s), i)ds \longrightarrow \int_0^t \int_{\mathcal{U}} \varphi(\alpha, s)v(x(s), i)ds, \quad \text{a.s.},$$

as $n \to \infty$, for all real valued continuous functions φ that have compact support and $i \in E$. This implies that μ, which is the limit of μ^n, is the relaxed control associated with $v(x, \theta)$, in the sense of (4.2). Thus, (x, θ) is a solution to the submartingale problem for $\mathcal{L}^v$. In addition, since (x, r) is the solution to the Skorohod problem for y, where y satisfies: $y(t) = x_0 + \int_0^t b(\theta(s))ds + \sigma w(t)$, $t \geq 0$, (x, θ) is the unique solution of (5.6) and (5.7). □

5.2. **Switching Diffusion Approximation for Model II.** Now we consider the switching diffusion approximation for model II. The approach is similar, we show that $\{(q^n, \theta^n)\}$ is tight and that any converging subsequence converges to the solution of a martingale problem. This time we have a martingale problem instead of the submartingale problem considered in the previous section since the limit process for this model is not reflected.

Similarly to Section 5.1, let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P})$ be a stochastic basis under the usual assumptions where the stochastic processes are to be defined. Let q_0 and θ_0 be given $\mathcal{F}_0$-measurable random variables taking values in $\mathbb{R}$ and E, respectively. Let w be a standard Brownian motion, which is a martingale with respect to $\{\mathcal{F}_t\}$ and independent of q_0 and θ_0, and let $\mathfrak{p}$ be a Poisson random measure relative to $\{\mathcal{F}_t\}$, independent of w, q_0, θ_0 and with intensity $dt \times m(d\gamma)$, where m denotes the Lebesgue measure in $\mathbb{R}$. The pair (q, θ) is a *controlled switching diffusion* with initial condition (q_0, θ_0), drift function $g : \mathbb{R} \times E \to \mathbb{R}$, diffusion coefficient $\varrho > 0$ and randomized Markovian control $v : \mathbb{R} \times E \to \mathcal{P}$ if q is a continuous $\{\mathcal{F}_t\}$-adapted process; θ is a right-continuous with left-limits $\{\mathcal{F}_t\}$-adapted process; q satisfies the following equation:

$$q(t) = q_0 + \int_0^t g(q(s), \theta(s))ds + \varrho w(t) \tag{5.20}$$

and θ satisfies (5.7) with q replacing x there. Existence and unicity of solutions for these equations are given by Theorem 6.2.3 in [2, p. 367], since it can easily be verified that the Lipschitz and growth conditions of [2] are satisfied for this equation.

We now establish the associated martingale problem. Let μ denote a relaxed control. We denote by $\mathcal{L}^\mu$ the infinitesimal generator of the above switching diffusion, which is given by (see, e.g., [10, 11]):

$$(\mathcal{L}^\mu f)(s, \xi, i) = g(\xi, i)f_x(\xi, i) + \frac{\varrho^2}{2}f_{xx}(\xi, i) + \sum_{j \in E} \int_{\mathcal{U}} f(\xi, j)\lambda_{ij}(\alpha)\mu_s(d\alpha), \tag{5.21}$$

for $(s, \xi, i) \in [0, \infty) \times \mathbb{R} \times E$ and function f in its domain $\mathcal{D}$, given by $\mathcal{D} := C_0^2(\mathbb{R} \times E)$. It is straight forward to show that a switching diffusion (q, θ) defined above is the solution of this martingale problem, using a similar argument to that in Remark 5.3. Now we define the associated martingale problem.

Definition 5.6. Suppose that (ζ, ϑ) is a stochastic process with sample paths that are right-continuous with left limits. Let μ denote a relaxed control. The process (ζ, ϑ) is said to solve the martingale problem for $\mathcal{L}^\mu$ if there exists a filtration $\{\mathcal{F}_t\}$ such that (ζ, ϑ) is $\mathcal{F}_t$-adapted and the process defined by:

$$f(\zeta(t), \vartheta(t)) - f(\zeta(0), \vartheta(0)) - \int_0^t (\mathcal{L}^\mu f)(s, \zeta(s), \vartheta(s))ds, \quad t \geq 0$$

is an $\{\mathcal{F}_t\}$-martingale for each $f \in \mathcal{D}$.

Let v denote a randomized Markovian control. Similarly to the previous section, if μ is the relaxed control associated with v, we will denote the infinitesimal generator by $\mathcal{L}^v$. The above martingale problem for $\mathcal{L}^v$ has been shown to have an unique solution recently in [41]. Also, notice that, for a control v satisfying the Assumption 4.1, an analogous result to Lemma 5.4 is valid for the model of this section.

Now we are ready to present the main theorem of this section.

Theorem 5.7. *Suppose that the initial condition for the n-th system $(q^n(0), \theta^n(0))$ converges weakly to (q_0, θ_0), a random variable with finite expectation taking values in $\mathbb{R} \times E$. Let (q^n, θ^n) be the stochastic process satisfying* (3.16) *and* (4.1) *for a randomized Markovian control v satisfying the assumption on Lemma 5.4. Then (q^n, θ^n) converges in distribution to the controlled switching diffusion having initial condition (q_0, θ_0), drift given by* (5.4)*, diffusion coefficient $\varrho := \sqrt{\lambda^s(\sigma_a^2 + 1)}$ and control v.*

Proof. Since the drift term in (3.16) depends on q^n, we begin by verifying the following condition, which is usually known as the compact containment condition: for each $t > 0$, $\epsilon > 0$, there are $K > 0$ and $n_0 > 0$ such that

$$\mathbb{P}\left(\sup_{0 \leq s \leq t} |q^n(s)| > K\right) \leq \epsilon. \tag{5.22}$$

In order to show this, let $\tau_K^n = \inf\{s \geq 0 : |q^n(s)| > K\}$ for some constant $K > 0$. We get the following by taking absolute absolute value in (3.16) with t replaced with $t \wedge \tau_k^n$:

$$\begin{aligned} |q^n(t \wedge \tau_K^n)| \leq |q^n(0)| + \int_0^{t \wedge \tau_K^n} \lambda^d \max_{i \in E} |b^n(i)| + \lambda^d |q^n(s)| ds \\ + |m^{a,n}(a^n(t \wedge \tau_K^n))| + |\tilde{m}^{d,n}(\bar{T}^n(t \wedge \tau_K^n))| + |\bar{\varepsilon}_2^n(t \wedge \tau_K^n)|. \end{aligned} \tag{5.23}$$

First recall that $|\bar{\varepsilon}_2^n(t \wedge \tau_K^n)| \leq n^{-1/2} \lambda^{a,n} \Delta^{a,n}_{A^n(t \wedge \tau_K^n)+1}$. Therefore, since $\{\Delta_l^{a,n}\}$ are identically distributed in l and by the fact that $\Delta^{a,n}_{A^n(t \wedge \tau_K^n)+1}$ is independent of $\tilde{\mathcal{F}}^n_{t \wedge \tau_K^n}$, we have that $\mathbb{E}\left[|\bar{\varepsilon}_2^n(t \wedge \tau_K^n)|\right] \leq n^{-1/2}$. Now, notice that we have the following using Jensen's inequality:

$$\mathbb{E}[|m^{a,n}(a^n(t \wedge \tau_K^n))|] \leq \mathbb{E}\left[\sup_{s \leq t} |m^{a,n}(a^n(s))|\right] \leq \mathbb{E}\left[\sup_{s \leq t} \left(m^{a,n}(a^n(s))\right)^2\right]^{1/2}.$$

It is straight forward to show that $m^{a,n}(a^n(\cdot))$ is an $\tilde{\mathcal{F}}_t^n$ martingale. Therefore, we have by Doob's inequality (e.g., Theorem 1.43 [18, p. 11]) that $\mathbb{E}\left[\sup_{s \leq t} \left(m^{a,n}(a^n(s))\right)^2\right] \leq$

$4\mathbb{E}\left[(m^{a,n}(a^n(t)))^2\right]$. Using integration by parts for functions of bounded variation and Wald's Lemma (since $A^n(t)$ is an $\{F_l^n\}$-stopping time for each $t \geq 0$, where $F_l^n := \sigma(\Delta_k^{a,n}; k \leq l)$), we have that: $4\mathbb{E}\left[(m^{a,n}(a^n(t)))^2\right] \leq \frac{4}{n}\mathbb{E}\left[A^n(t)\right]\sigma_s^{2,n}$. By definition of A^n, we know that $\sum_{l=1}^{A^n(t)} \lambda^{a,n}\Delta_l^{a,n} \leq \lambda^{a,n}t$, taking expectation and using Wald's Lemma again implies that $\mathbb{E}[A^n(t)] \leq \lambda^{a,n}\, t$. This gives the following estimate:

$$\mathbb{E}[|m^{a,n}(a^n(t \wedge \tau_K^n))|] \leq 2\sqrt{\lambda^{s,n}t\sigma_s^{2,n}}.$$

In addition, by properties of time-changed Poisson processes, $\tilde{m}^{d,n}(\bar{T}^n(t \wedge \tau_K^n))$ is an $\tilde{\mathcal{F}}_t^a$ martingale with quadratic variation given by $\bar{T}^n(t \wedge \tau_K^n)$. By (3.17), this implies that:

$$\begin{aligned}\mathbb{E}[|\tilde{m}^{d,n}(\bar{T}^n(t \wedge \tau_K^n))|] &\leq \mathbb{E}\left[\left(\tilde{m}^{d,n}(\bar{T}^n(t \wedge \tau_K^n))\right)^2\right]^{1/2} = \mathbb{E}\left[\bar{T}^n(t \wedge \tau_K^n))\right]^{1/2} \\ &\leq \sqrt{\lambda^{s,n}t + \frac{\lambda^d}{\sqrt{n}}\max_{i\in E}|b^n(i)|t + \frac{\lambda^d}{\sqrt{n}}Kt}.\end{aligned}$$

Let us define $C_{K,n,t}$ as follows:

$$\begin{aligned}C_{K,n,t} := \mathbb{E}[|q^n(0)|] &+ \lambda^d \max_{i\in E}|b^n(i)|t \\ &+ 2\sqrt{\lambda^{s,n}t\sigma_s^{2,n}} + \sqrt{\lambda^{s,n}t + \frac{\lambda^d}{\sqrt{n}}\max_{i\in E}|b^n(i)|t + \frac{\lambda^d}{\sqrt{n}}Kt}.\end{aligned}$$

Using (5.23), we have that

$$\mathbb{E}\left[|q^n(t \wedge \tau_K^n)|\right] \leq C_{K,n,t} + \int_0^t \lambda^d \mathbb{E}\left[|q^n(s \wedge \tau_K^n)|\right] ds$$

By Gronwall's inequality (e.g. Proposition 6.1.4 of [2, p. 295]), we get: $\mathbb{E}\left[|q^n(t \wedge \tau_K^n)|\right] \leq C_{K,n,t}e^{\lambda^d t}$. Since, $\mathbb{E}\left[|q^n(t \wedge \tau_K^n)|\right] \geq K\mathbb{P}(\tau_K^n < t)$, we have that

$$\mathbb{P}(\tau_K^n < t) = \mathbb{P}\Big(\sup_{s\leq t}|q^n(s)| > K\Big) \leq \frac{C_{K,n,t}e^{\lambda^d t}}{K}, \tag{5.24}$$

which implies the compact containment condition (5.22), since the left-hand side of (5.24) gets smaller when K or n increases.

Let us define $\tilde{d}^n(\cdot) := \int_0^{\cdot} g^n(q^n(s), \theta^n(s))ds$. Using the same arguments used in the proof of Theorem 5.5, we conclude that $\{\tilde{d}^n\}$ is C-tight. In addition, the same argument used in the proof of Theorem 5.5 can be used to show that $\{\theta^n\}$ is tight. The C-tightness of $\{\tilde{d}^n\}$ implies that $\{\bar{T}^n\}$ converges in probability to a process taking values $\lambda^{s,n}t$, by (3.17). This together with Lemma 3.3 implies that $\{(m^{a,n}, \tilde{m}^{d,n}, a^n, \bar{T}^n)\}$ is C-tight and that $\{\tilde{w}^n\}$ is C-tight, since the composition map is continuous. We also have that $\{\bar{\varepsilon}_2^n\}$ converges in probability to the zero process, by the same argument used in the proof of Theorem 5.5. Therefore, we have that $\tilde{\Psi}^n = \{(q^n, \theta^n)\}$ is tight.

Now we characterize the limit of any weak-sense limit of a converging subsequence of $\{(q^n, \theta^n)\}$ as the solution to the martingale problem for $\mathcal{L}^\mu$, for some relaxed control μ, where the infinitesimal operator is given by (5.21). For this, suppose that (q, θ) is a

weak-sense limit of a converging subsequence of $\{(q^n, \theta^n)\}$ and let $\{\tilde{\mathcal{F}}_t\}$ be its natural filtration. Let p be an integer and $\{t_k\}_{k=1}^p$ a set of real numbers such that $0 \leq t_k \leq t$ for each $k = 1, \ldots, p$. Let $h_k : \mathbb{R} \times E \to \mathbb{R}$ be continuous and bounded functions, for $k = 1, \ldots, p$. In order to show that (q, θ) is a solution to the $\{\tilde{\mathcal{F}}_t\}$-martingale problem for $\mathcal{L}^\mu$ it is enough to show that it satisfies:

$$\mathbb{E}\left[\left[f(q(t+\tau), \theta(t+\tau)) - f(q(t), \theta(t))\right] \prod_{k=1}^{p} h_k(q(t_k), \theta(t_k))\right]$$
$$= \mathbb{E}\left[\int_t^{t+\tau} (\mathcal{L}^\mu f)(s, q(s), \theta(s)) ds \prod_{k=1}^{p} h_k(q(t_k), \theta(t_k))\right],$$

fore each $t, \tau \geq 0$, $f \in \mathcal{D}$, p, $\{t_k\}_{k=1}^p$ and continuous and bounded functions $\{h_k\}_{k=1}^p$. Clearly, the remainder of the proof follows analogously to the proof of Theorem 5.7 and therefore the details are omitted.

□

5.3. **Application of the Heavy Traffic Approximations.** Now that we have derived the switching diffusion limits for the two queuing models, we show how they can be used to approximate these queuing systems. In practice there is one physical system that we want to approximate with a fixed number of servers, say $\breve{n}$. Suppose that this physical system under consideration has arrival rate $\breve{\lambda}^a$ with squared coefficient of variation $\breve{\sigma}_a^2$, where we define $\breve{\lambda}^s := \breve{\lambda}^a/\breve{n}$. These correspond to $\lambda^{a,\breve{n}}$, $\sigma_a^{2,\breve{n}} = \sigma_s^{2,\breve{n}}$, and $\lambda^{s,\breve{n}}$ for n given by $\breve{n}$ in the notation of Section 3. Let $\mathfrak{o}^{\breve{n}}$ denote the number of servers that remain always in the "*on*" state and let $\breve{\mathfrak{f}}(i) := \mathfrak{f}^{\breve{n}}(i)$ denote the fraction of active servers when the state is $i \in E$. Let, in addition, $\bar{\Delta}^d$ denote the average service time and σ_d^2 denote the square coefficient of variation for the service time distribution for this physical system. For the controlled pure jump process, representing the state of the reserve machines, let $\breve{\lambda}_{ij}(\alpha)$ denote the actual transitions rates for this system when going from state i to j, for $i \neq j$, when the control is $\alpha \in \mathcal{U}$ and let $\breve{\lambda}_{ii}(\alpha) := -\sum_{j \in E \setminus \{i\}} \breve{\lambda}_{ij}(\alpha)$.

In order to use the switching diffusion limit for the model I, given by Theorem 5.5, let $\breve{b} : E \to \mathbb{R}$ be given by $\breve{b} := b^{\breve{n}}$, where b^n is defined in (3.2). We approximate the workload in the physical system at time t, $X(t)$, by the process $\breve{X}(t) := \sqrt{\breve{n}}\breve{x}(t)$, where $\breve{x}$ is given by:

$$\breve{x}(t) = \breve{x}_0 + \int_0^t \breve{b}(\breve{\theta}(s)) ds + \sqrt{\breve{\lambda}^s (\bar{\Delta}^d)^2 (\breve{\sigma}_a^2 + \sigma_d^2)}\breve{w}(t) + \breve{r}(t), \tag{5.25}$$
$$= \breve{x}_0 + \frac{\bar{\Delta}^d}{\sqrt{\breve{n}}} \int_0^t \breve{\lambda}^a - \lambda^d \breve{n} \breve{\mathfrak{f}}(\breve{\theta}(s)) ds + \frac{\bar{\Delta}^d}{\sqrt{\breve{n}}} \sqrt{\breve{\lambda}^a (\breve{\sigma}_a^2 + \sigma_d^2)}\breve{w}(t) + \breve{r}(t), \quad t \geq 0,$$

where $\breve{x}_0 := X(0)/\breve{n}$, $X(0)$ is the initial workload in the system, $\breve{w}$ denotes a standard Brownian motion, $\breve{r}$ the reflection process and $\breve{\theta}$ the controlled pure jump process,

which satisfies:

$$\mathbb{P}\Big(\breve{\theta}(t+\delta)=j|\breve{\theta}(t)=i,\breve{\mathcal{F}}_t\Big)=\int_{\mathcal{U}}\breve{\lambda}_{ij}(\alpha)v(\breve{x}(t),i)(d\alpha)\delta+o(\delta),\quad t\geq 0, \tag{5.26}$$

for $\delta>0$ and relaxed control v, where $\breve{\mathcal{F}}_t:=\sigma\{\breve{x}(s),\breve{\theta}(s);s\leq t\}$ for each $t\geq 0$.

An approximation for model II can be constructed as follows. Let $\breve{\lambda}^a$, $\breve{\sigma}_a^2$, and $\breve{b}$ be the parameters of the physical system as described in the previous paragraph. In addition, let $\breve{q}$ be a stochastic process that is a solution to

$$\breve{q}(t)=\breve{q}_0+\int_0^t\lambda^d\left(\breve{b}(\breve{\theta}(s))\vee-\breve{q}(s)\right)ds+\sqrt{\breve{\lambda}^s\,(\breve{\sigma}_a^2+1)}\breve{w}(t),\ \ t\geq 0, \tag{5.27}$$

where $\breve{q}_0:=\breve{n}^{-1/2}(Q(0)-\breve{n}\breve{\rho})$, $\breve{\rho}:=\breve{\lambda}^a\bar{\Delta}^d/\breve{n}$, $Q(0)$ is the initial number of customers in the system, $\breve{w}$ denotes a standard Brownian motion and $\breve{\theta}$ denotes the controlled pure jump process, which satisfies expression (5.26) with $\breve{q}$ in place of $\breve{x}$ and $\breve{\mathcal{F}}'_t:=\sigma\{\breve{q}(s),\breve{\theta}(s);\,s\leq t\}$ in place of $\breve{\mathcal{F}}_t$. Notice that $(\breve{q},\breve{\theta})$ is the limit switching diffusion of Theorem 5.7 with the physical system's data.

We approximate the number of customers in the system at time t by the process $\breve{Q}:=\sqrt{\breve{n}}\breve{\breve{q}}$, with $\breve{\breve{q}}$ given by $\breve{\breve{q}}(t)=\breve{q}(t)+\sqrt{\breve{n}}\breve{\rho}$, $t\geq 0$. The expression (5.27) can be simplified and written in terms of $(\breve{\breve{q}},\breve{\theta})$, as follows:

$$\begin{aligned}\breve{\breve{q}}(t)=\breve{\breve{q}}_0+\frac{1}{\sqrt{\breve{n}}}\int_0^t\breve{\lambda}^a-\lambda^d\left[\breve{n}\breve{\mathbb{f}}(\breve{\theta}(s))\wedge\sqrt{\breve{n}}\breve{\breve{q}}(s)\right]ds\\ +\frac{1}{\sqrt{\breve{n}}}\sqrt{\breve{\lambda}^a\,(\breve{\sigma}_a^2+1)}\breve{w}(t)+\breve{r}(t),\ \ t\geq 0\end{aligned} \tag{5.28}$$

where $\breve{\breve{q}}_0:=\breve{n}^{-1/2}Q(0)$ and we introduced here a reflection term $\breve{r}(t)$ in order to prevent the approximation $\breve{Q}$ from taking negative values. The reflection process $\breve{r}(t)$ satisfies $\breve{r}(0)=0$, it is non-decreasing and increases only at time t such that $\breve{\breve{q}}(t)=0$.

An interesting feature to be pointed out here is the similarity between the scaled approximation for the two models, given by (5.25) and (5.28). Since the service process for model I is parallelized, in the sense that the processing stations work together to complete the pending jobs, the term $\lambda^d\breve{n}\breve{\mathbb{f}}(\breve{\theta}(t))$ comes out in its drift function. This term can be understood as the system processing rate at time t. For model II, the analogous term is given by $\lambda^d(\breve{n}\breve{\mathbb{f}}(\breve{\theta}(t))\wedge\sqrt{\breve{n}}\breve{\breve{q}}(t))$, which depends on the number of customers $\sqrt{\breve{n}}\breve{\breve{q}}(t)$, since the servers are not parallelized and do not work together on pending jobs. In addition, since $\breve{x}$ represents the scaled workload, its expression is multiplied by the average service time $\bar{\Delta}^d$. This is not present in the expression of $\breve{\breve{q}}$, since it represents the scaled number of clients.

6. The Control Problem and a Numerical Method

In this section, we consider the control problem and present the numerical approach used to solve it. For simplicity, let (z,θ) represent either the approximation $(\breve{x},\breve{\theta})$ for model I, given by (5.25), or the approximation $(\breve{\breve{q}},\breve{\theta})$ for model II, given by (5.28). Under control v, we assume that (z,θ) satisfies the following regime-switching stochastic

differential equation with reflection:

$$\begin{cases} dz(t) = c(z(t),\theta(t))dt + \sigma dw(t) + dr(t), \\ \mathbb{P}\left(\theta(t+\delta) = j \,|\, \theta(t) = i, \mathcal{G}_t\right) = \int_{\mathcal{U}} \breve{\lambda}_{ij}(\alpha) v(z(t),i)(d\alpha)\delta + o(\delta) \end{cases} \tag{6.1}$$

where c (resp., σ) represents either the drift term (resp., diffusion coefficient) in (5.25) or in (5.28) and $\mathcal{G}_t$ represents either $\breve{\mathcal{F}}_t = \sigma\{\breve{x}(s), \breve{\theta}(s); s \le t\}$ or $\breve{\breve{\mathcal{F}}}'_t = \sigma\{\breve{\breve{q}}(s), \breve{\theta}(s); s \le t\}$.

Let us define the running cost function $K : \mathbb{R}_+ \times E \to \mathbb{R}$. For example, this function can represent the energy consumption rate, which depends on the number of machines turned on; it can represent a measure of performance, such as the number of pending tasks; or it can represent a combination of both. Assume that K is either continuous in its first argument or bounded. We consider the following ergodic control problem: determine a control policy v, which minimizes the cost:

$$\gamma(z_0, i_0, v) := \limsup_T \frac{1}{T} \int_0^t \mathbb{E}^v_{(z_0,i_0)} \left[K(z(s),\theta(s))\right] ds, \tag{6.2}$$

where (z_0, i_0) is the initial condition of the system and the superscript v over the expectation was added in order to emphasize that (z,θ) is controlled by v.

In general, closed-form analytical solutions for such control problems are not known. In addition, as far as the authors are aware, there is not any available theory which allows to get an analytical solution, specially in this case where the process is controlled by the pure jump process. Therefore, our approach here is to find a solution numerically. Since there are no numerical methods currently available for this specific class of control problems, we propose an approach based on the Markov chain approximation method (MCAM), where the original problem is set as a Markov decision process (MDP) after a proper discretization. This MDP will be constructed using finite difference approximations on the differential operators of the dynamic programming equation (DPE) for the original control problem.

In order to present the DPE for this control problem, let us denote by $\mathcal{D}_+$ the set of functions f of $C^2(\mathbb{R}_+ \times E)$ such that $f_x(0,i) \ge 0$ for each $i \in E$. For each $\alpha \in \mathcal{U}$, let us define $\mathcal{L}^\alpha$ to be the following operator on $\mathcal{D}_+$:

$$\mathcal{L}^\alpha f(\xi,i) = c(\xi,i) f_x(\xi,i) + \frac{\sigma^2}{2} f_{xx}(\xi,i) + \sum_{j\in E} \breve{\lambda}_{ij}(\alpha) f(\xi,i).$$

The DPE associated with the ergodic control problem with cost (6.2) is given by:

$$\begin{cases} \inf_{\alpha\in\mathcal{U}} \left\{\mathcal{L}^\alpha V(\xi,i) - \gamma + K(\xi,i)\right\} = 0, \quad \text{for } \xi > 0 \\ V_x(0,i) = 0 \end{cases} \tag{6.3}$$

for $i \in E$, $V \in \mathcal{D}_+$, and constant γ (see for instance [25, p. 65]).

We will not attempt to prove convergence of the numerical method proposed here, however the theorem below provides a motivation for the approach. It states that if a randomized Markovian control v^ϵ satisfies (6.3) approximately for some bounded function $V^\epsilon \in \mathcal{D}_+$ and constant γ^ϵ (i.e., with errors within an interval of length ϵ), then v^ϵ is an ϵ-optimum control. Therefore, if a solution is found for a discrete version

of this dynamic programming equation and this discrete solution can be smoothly interpolated, then it will be a good approximation to the continuous problem.

Theorem 6.1. *Let $\epsilon > 0$. Suppose that there are $V^\epsilon \in \mathcal{D}_+$ bounded and scalar γ^ϵ satisfying* (6.3) *"approximately," in the sense that:*

$$d^\epsilon(\xi, i) := \inf_{\alpha \in \mathcal{U}} \{\mathcal{L}^\alpha V^\epsilon(\xi, i) - \gamma^\epsilon + K(\xi, i)\} \tag{6.4}$$

satisfies $|d^\epsilon(\xi, i)| < \epsilon/2$, for all $\xi > 0$ and $i \in E$, and $V^\epsilon_x(0, i) = 0$ for all $i \in E$. Let v^ϵ be a randomized Markovian control such that

$$\int_{\mathcal{U}} \mathcal{L}^\alpha V^\epsilon(\xi, i) v^\epsilon(\xi, i)(d\alpha) - \gamma^\epsilon + K(\xi, i) = d^\epsilon(\xi, i), \tag{6.5}$$

for all $\xi > 0$ and $i \in E$. Then v^ϵ is an ϵ-optimum solution for the ergodic control problem with cost given by (6.2).

Proof. See the Appendix A. □

In order to construct the approximating Markov chain, the state space for the continuous component of the switching diffusion is discretized. For $h > 0$, let us denote this discretized state space by $S_h := \{0, h, 2h, 3h, \ldots\}$. Now, difference approximations are used for the differential operators in $\mathcal{L}^\alpha$. Let $V \in \mathcal{D}_+$, the forward and backward finite difference approximations for V_x are given by

$$D_h^+ V(\xi, i) := \frac{V(\xi + h, i) - V(\xi, i)}{h}, \qquad D_h^- V(\xi, i) := \frac{V(\xi, i) - V(\xi - h, i)}{h},$$

respectively, for $(\xi, i) \in S_h \times E$, where we use $V(\xi, i) \equiv V(0, i)$ for $\xi < 0$. It is well known that $V_x(\xi, i) = D_h^\pm V(\xi, i) + O(h)$ for all ξ and i, where $O(h)$ is a term such that $O(h) \to 0$ as $h \downarrow 0$. Using the central difference approximation for V_{xx}, we define

$$D_h^2 V(\xi, i) := \frac{V(\xi + h, i) - 2V(\xi, i) + V(\xi - h, i)}{h^2} \quad \text{for} \quad (\xi, i) \in S_h \times E,$$

which satisfies $V_{xx}(\xi, i) = D_h^2 V(\xi, i) + O(h)$. Again, we set $V(\xi, i) \equiv V(0, i)$ for $\xi < 0$. Therefore, for $\alpha \in \mathcal{U}$, we define $L_h^\alpha V(\xi, i)$ to be the finite difference approximation for $\mathcal{L}^\alpha V(\xi, i)$ as follows: $\mathcal{L}^\alpha V(\xi, i) = L_h^\alpha V(\xi, i) + O(h)$, where

$$L_h^\alpha V(\xi, i) := c^+(\xi, i) D_h^+ V(\xi, i) + c^-(\xi, i) D_h^- V(\xi, i) + \frac{\sigma^2}{2} D_h^2 V(\xi, i) + \sum_{j \in E} \breve{\lambda}_{ij}(\alpha) V(\xi, j),$$

$c^+(\xi, i) := c(\xi, i) \vee 0$ and $c^-(\xi, i) := c(\xi, i) \wedge 0$. This combination of the forward and backward difference approximation for V_x is used in order to have positive transition probabilities for the approximating Markov chain (see [25, chapter 5] for more details).

With this, we can define the following discrete DPE, for a function $V_h : S_h \times E \to \mathbb{R}$ and scalar γ_h:

$$\begin{cases} \inf_{\alpha \in \mathcal{U}} \{L_h^\alpha V_h(\xi, i) - \gamma_h + K(\xi, i)\} = 0 \quad \text{for} \quad (\xi, i) \in S_h \times E \\ D_h^- V_h(0, i) = 0. \end{cases} \tag{6.6}$$

By grouping the terms involving $V_h(\xi,i)$, $V_h(\xi+h,i)$ and $V_h(\xi-h,i)$ together for the first equation in (6.6), we have

$$V_h(\xi,i) = \inf_{\alpha\in\mathcal{U}} \Bigg\{ \Delta t^h K(\xi,i) + V_h(\xi,i)\Bigg[1 - \frac{M^h(\xi,i,\alpha)}{\bar{M}^h}\Bigg] + V_h(\xi+h,i)\Bigg[\frac{\sigma^2/2 + hc^+(\xi,i)}{\bar{M}^h}\Bigg]$$

$$+ V_h(\xi-h,i)\left[\frac{\sigma^2/2 - hc^-(\xi,i)}{\bar{M}^h}\right] + \sum_{j\in E\setminus\{i\}} \breve{\lambda}_{ij}(\alpha)\Delta t^h V_h(\xi,j) - \bar{\gamma}_h \Bigg\}, \tag{6.7}$$

where $\bar{\gamma}_h := \Delta t^h \gamma_h$, $M^h(\xi,i,\alpha) := |c(\xi,i)|h + \sigma^2 - \breve{\lambda}_{ii}(\alpha)h^2$, $\bar{M}^h := \max_{\xi,i,\alpha} M^h(\xi,i,\alpha)$, and $\Delta t^h := h^2/\bar{M}^h$. Notice that the condition $D_h^- V^h(0,i) = 0$ of (6.6) implies that $V_h(-h,i) = V_h(0,i)$ and, hence, we can replace $V_h(-h,i)$ by $V_h(0,i)$ in (6.7) when $\xi = 0$.

Comparing (6.7) with the dynamic programming equation for a Markov decision process with average reward criterion (e.g., [34, p. 443]) suggests the following transition probabilities for $\alpha \in \mathcal{U}$:

$$\begin{aligned}
&p^h((\xi,i),(\xi,i)|\alpha) := 1 - \frac{M^h(\xi,i,\alpha)}{\bar{M}^h} \quad \text{for} \quad \xi > 0,\ i \in E \\
&p^h((\xi,i),(\xi\pm h,i)|\alpha) := \frac{\sigma^2/2 \pm hc^{\pm}(\xi,i)}{\bar{M}^h}, \quad \text{for} \quad \xi > 0,\ i \in E \\
&p^h((\xi,i),(\xi,j)|\alpha) := \breve{\lambda}_{ij}(\alpha)\Delta t^h, \quad \text{for} \quad \xi \geq 0,\ i,j \in E,\ i \neq j.
\end{aligned} \tag{6.8}$$

For $\xi = 0$ and $i = j$, we have the following transition probability:

$$p^h((0,i),(0,i)|\alpha) := 1 - \frac{M^h(0,i,\alpha)}{\bar{M}^h} + \frac{\sigma^2/2 - hc^-(0,i)}{\bar{M}^h} \quad \text{for} \quad \xi = 0, i \in E. \tag{6.9}$$

The transition probability from (ξ,i) to any other state that is not covered by either (6.8) or (6.9) is defined to be zero. With these transition probabilities, we can find a numerical solution to (6.7) by using, for instance, the value iteration procedure or a linear programming formulation.

Remark 6.2 (Finite State Space for Numerical Solution). Since numerical methods require finite state space, an upper bound for the set S_h must be introduced. The idea is to let B (a multiple of h) be the largest value that x can take. Since we desire to work with ergodic queues, we can choose B to be large enough not to interfere with the process. Let $S_h^B := \{0,h,2h,3h,\ldots,B\}$ and redefine the transition probability from state (B,i) to (B,i) in a similar fashion to what is done in equation (6.9),

$$p^h((B,i),(B,i)|\alpha) := 1 - \frac{M^h(B,i,\alpha)}{\bar{M}^h} + \frac{\sigma^2/2 + hc^+(B,i)}{\bar{M}^h}, \tag{6.10}$$

for $\xi = B, i \in E$. Notice that this is analogous to having a reflecting boundary at $\xi = B$. That is, this probability will appear if we add the condition that $D_h^+ V_h(B,i) = 0$ to (6.6). Clearly, we also need to redefine the transition probabilities involving states (ξ,i) with $\xi > B$ to be zero.

Let us denote by $\zeta^h := \{\zeta_k^h\}_{k=0}^{\infty}$ a discrete time Markov chain taking values on $\{0, h, \ldots, B\} \times E$ and whose transition matrices are given by $p^h((x,i),(y,j)|\alpha)$, which are defined by (6.8), (6.9), and (6.10). It is desirable that this Markov chain is locally consistent with the reflected switching diffusion, which is solution of (6.1). Although, we are unaware of any work which deals specifically with the switching diffusion considered here, which is reflected and has switching rates depending on the control, it is known (see [25]) that for similar jump diffusions and under mild additional assumptions, local consistency is a sufficient condition for the weak convergence of appropriately constructed continuous time interpolation of these approximating Markov chain to the jump diffusions. Loosely speaking, this condition implies that the local mean and covariance of the Markov chain approximation matches that of the jump diffusion. We will verify the local consistency conditions presented by [35, 42] for switching diffusions (although we had to adapt slightly the condition presented there in order to cover the switching rate dependency on the control, which is present in the system considered here) in addition to the boundary local consistency, for reflected diffusions, which is given by (7.3a)-(7.3c) of [25, p. 137].

In order to present the theorem below, let us introduce the following notation. Let $\mathbb{E}_{(x,i)}^{\alpha,h}$ and $\mathbb{P}_{(x,i)}^{\alpha,h}$ denote the conditional expectation and probability distribution, respectively, given that $\zeta_k^h = (x,i)$ and the control is set to $\alpha \in \mathcal{U}$ at time k. Notice that we omit k from the notation of the expectation and the probability since ζ^h is time-homogeneous. In addition, let us define $\Delta\zeta_{1,k}^h := (\zeta_{k+1}^h)_1 - (\zeta_k^h)_1$, where $(\zeta_{k+1}^h)_l$ denotes l-th component of ζ_k^h, $l \in \{1,2\}$.

Theorem 6.3. *The discrete time Markov chain ζ^h satisfies the following local consistency conditions:*

$$\sup_{k,\omega} |\Delta\zeta_{1,k}^h| \to 0 \quad as \quad h \to 0; \tag{6.11}$$

for $0 < x < B$, $i, j \in E$, and $i \neq j$,

$$\mathbb{E}_{(x,i)}^{\alpha,h}\left[\Delta\zeta_{1,k}^h\right] = c(x,i)\Delta t^h + o(\Delta t^h), \tag{6.12}$$

$$\mathbb{E}_{(x,i)}^{\alpha,h}\left[\left(\Delta\zeta_{1,k}^h - \mathbb{E}_{(x,i)}^{\alpha,h}\left[\Delta\zeta_{1,k}^h\right]\right)^2\right] = \sigma^2\Delta t^h + o(\Delta t^h), \tag{6.13}$$

$$\mathbb{P}_{(x,i)}^{\alpha,h}\left\{(\zeta_{k+1}^h)_2 = j\right\} = \Delta t^h \breve{\lambda}_{ij}(\alpha) + o(\Delta t^h), \tag{6.14}$$

$$\mathbb{P}_{(x,i)}^{\alpha,h}\left\{(\zeta_{k+1}^h)_2 = i\right\} = 1 + \breve{\lambda}_{ii}(\alpha)\Delta t^h + o(\Delta t^h); \tag{6.15}$$

for $x \in \{0, B\}$, (6.14) *and* (6.15) *are satisfied and*

$$\mathbb{E}_{(0,i)}^{\alpha,h}\left[\Delta\zeta_{1,k}^h\right] = c_1 h + o(h), \tag{6.16}$$

$$\mathbb{E}_{(B,i)}^{\alpha,h}\left[\Delta\zeta_{1,k}^h\right] = -c_1 h + o(h), \tag{6.17}$$

$$\mathbb{E}_{(x,i)}^{\alpha,h}\left[\left(\Delta\zeta_{1,k}^h - \mathbb{E}_{(x,i)}^{\alpha,h}\left[\Delta\zeta_{1,k}^h\right]\right)^2\right] = o(h), \tag{6.18}$$

where $c_1 > 0$ and $c_2 > 0$ are constants; and there are $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ such that $p^h((0,i),(h,i)|\alpha) \geq \varepsilon_1$ and $p^h((B,i),(B-h,i)|\alpha) \geq \varepsilon_2$.

Proof. See the Appendix A. □

6.1. **A Numerical Approach to a Control Problem with Restriction.** In addition to the control problem presented previously, we will also consider the problem of minimizing (6.2) while satisfying a cost constraint. Formally, the idea is to find a randomized Markovian control $v : \mathbb{R}_+ \times E \to \mathcal{P}$ such that the following average cost is minimized:

$$\gamma(z_0, i_0, v) := \limsup_T \frac{1}{T} \int_0^T \mathbb{E}^v_{(z_0,i_0)} \left[K(z(s), \theta(s))\right] ds, \tag{6.19}$$

where (z_0, i_0) is the initial condition, while satisfying, for some constant W, the constraint:

$$\limsup_T \frac{1}{T} \int_0^T \mathbb{E}^v_{(z_0,i_0)} \left[R(z(s), \theta(s))\right] ds \leq W, \tag{6.20}$$

where $R : \mathbb{R}_+ \times E \to \mathbb{R}$ is another running cost function. This control problem is motivated by the idea that we would like to minimize power consumption while maintaining a reasonable quality of service. For instance, R might be used to calculate a performance measure of the system, such as the waiting time.

The approach taken here for this problem is numerical. Having defined the approximating Markov chain in the previous section, we now define a constrained Markov decision process as follows. Let $\zeta^h := \{\zeta^h_n\}_{n=1}^\infty$ be a controlled Markov chain taking values in $\{0, h, \ldots, B\} \times E$ with transition probabilities given by $p^h((x,i),(y,j)|\alpha)$, which is defined by (6.8), (6.9) and (6.10). Then the original control problem is approximated by the following problem: find a control policy v^h which minimizes the cost:

$$\begin{aligned} \gamma(\zeta_0, v^h) = \quad & \limsup_N \tfrac{1}{N} \mathbb{E}^{v^h}_{\zeta_0} \left[\textstyle\sum_{n=0}^N K(\zeta^h_n)\right] \\ \text{s.t.:} \quad & \limsup_N \tfrac{1}{N} \mathbb{E}^{v^h}_{\zeta_0} \left[\textstyle\sum_{n=0}^N R(\zeta^h_n)\right] \leq W, \end{aligned}$$

where ζ_0 is the initial condition of the system. This discretized problem can be solved via a linear programming formulation [1, 17].

7. Numerical Experiments

In this section, we consider some numerical experiments that illustrate the switching diffusion approximations proposed here. These experiments are divided in two parts. First we consider an experiment involving the queuing model II in Section 7.1. The approach in this case is to minimize a combined cost, which penalizes energy consumption (by a linear function of the number of machines turned on) and total number of pending tasks in the system. The setup is the same as the one proposed in [30] and the resulting optimal control obtained by the switching diffusion approximation and numerical methods proposed here is compared with the strategy developed in [30]. We then consider an experiment involving queuing model I in Section 7.2. This time, we consider an optimal control problem with restriction, where the objective is to minimize energy consumption, while satisfying a maximal average workload requirement.

In order to validate the results, event-driven simulations of these parallel processing system were implemented. In these simulations, the inter-arrival time distributions of the arriving jobs can be either hyper-exponential or exponential. For queuing model I, which considers systems where pending jobs are shared among the processing stations, we implemented the following job-splitting scheme: each job that enters the system is split into $\breve{\mathfrak{o}}$ tasks, where $\breve{\mathfrak{o}}$ is the number of machines that are always turned on for the system in consideration. These split tasks wait in queue and can be served by any available server. Each task has an exponential service time distribution with mean $\bar{\Delta}^d/\breve{\mathfrak{o}}$. This implies that the amount of work brought into the system by the l-th job, which is given by Δ_l^d in our queuing model, is Erlang distributed with mean $\bar{\Delta}^d$ and squared coefficient of variation given by $\sigma_d^2 = 1/\breve{\mathfrak{o}}$. This type of service parallelization is found in index servers of large web search engines [5]. Since the system modeled by queuing model II does not implement service parallelization, the simulation for this system considers that jobs are not split and are served in order of arrival by the first server that becomes available. For the simulation of model II, the processing times of the arriving jobs are exponentially distributed.

In the simulation of both systems (for model I and II), when a signal to shut down is sent by the controller, the system shuts down the first $\mathfrak{r}$ stations that become available. The stations take an exponential time to turn on and off with rates λ^{on} and λ^{off}, respectively.

7.1. **Combining Conflicting Objectives.** In this section, we consider a control problem involving the queuing model II. In order to have a basis of comparison for the models and numerical methods proposed here, we consider the same scenario presented by Mitrani in [30] and we compare the controls obtained via the numerical methods of the previous section with the strategy proposed by Mitrani in [30]. In [30], the control is found by a carefully designed heuristics for a controlled $M/M/n$ queue (which only considers systems with inter-arrival times that are exponentially distributed). Following the scenario presented by Mitrani, we consider that the state of the reserve machines are represented by the state space $E = \{0, 1, 3\}$, where the state 0 represents that the reserve machines are off, 1 represents the state where the reserve machines are turning on, and 2 represents the state where they are active. There is no state representing that the reserve machines are *turning* off, they are assumed to shutdown immediately when a signal is sent by the controller.

The control problem considered here, which is the same considered by Mitrani [30], is that of finding an optimal policy that minimizes a cost that penalizes a weighted average of the number of stations consuming power and the number of pending tasks of the system in equilibrium. Using the notation and the switching diffusion approximation presented here, the ergodic cost is given by

$$\gamma(x_0, i_0, v) = \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}^v_{(x_0,i_0)}\left[\int_0^T c_1\breve{Q}(s) + c_2\mathfrak{N}(\breve{\theta}(s))ds\right],$$

for a system with initial condition $(x_0, i_0) \in [0, \infty) \times E$, under the control v, and with $\breve{n}$ total processing stations, where $\breve{Q}(t) := \sqrt{\breve{n}}\breve{\breve{q}}(t)$, with $\breve{\breve{q}}$ satisfying (5.28), represents the number of customers in the system by time t, and $\mathfrak{N}(\breve{\theta}(t))$ represents the number

of servers consuming power at time t, where $\mathfrak{N}$ is given by $\mathfrak{N}(i) = \mathfrak{o}^{\breve{n}}$ for $i = 0$ and $\mathfrak{N}(i) = \breve{n}$ for $i \in \{1, 2\}$.

In order to represent the control, we define $\mathcal{U}$ to be $\{0, 1\}$, where a control set to $\alpha = 0$ represents that the system should remain as it is and $\alpha = 1$ indicates that the system should change state. This idea is implemented in the transition rates of the controlled pure jump process $\breve{\theta}$ as follows:

$$\breve{\lambda}_{0\,1}(\alpha) = \begin{cases} \lambda^c & \text{if } \alpha = 1 \\ 0 & \text{otherwise,} \end{cases} \qquad \breve{\lambda}_{1\,2}(\alpha) = \lambda^{on}, \qquad \breve{\lambda}_{2\,0}(\alpha) = \begin{cases} \lambda^c & \text{if } \alpha = 1 \\ 0 & \text{otherwise,} \end{cases}$$

where λ^{on} is the rate in which the machines turn on and λ^c is interpreted as the rate in which the control is implemented. Notice that $\breve{\lambda}_{1\,2}(\alpha)$ does not change with α, that is, the processing stations cannot be interrupted if they have initiated the start up process.

The control problem defined above was solved numerically using the method presented in Section 6, where the approximating Markov chain ζ_k^h for $(\breve{\bar{q}}, \breve{\theta})$ was defined with transition probabilities given by (6.8), (6.9) and (6.10). The ergodic control problem for this approximating Markov chain was solved using the value iteration procedure (e.g., [34, p. 472]). The optimal control solutions determined by the numerical method for this problem in every scenario considered below operate in the following fashion: an upper U and a lower D threshold bounds were found; if the queue length is above U and the reserves are off, then the reserves are turned on; and if the queue length is below D and the reserves are active, then they are turned off. In [30], the author assumed the same form for the control, but the threshold values U and D were different from the ones found here.

We consider two scenarios for the numerical experiments. For each one of these scenarios, we set $c_1 = 1$, $c_2 = 2$, and the service time to be exponentially distributed with rate $\lambda^d = 1$. For the first case, the inter-arrival times are exponentially distributed with rate $\breve{\lambda}^a$ (with squared coefficient of variation given by $\breve{\sigma}_a^2 = 1$). Table 1 displays the resulting upper U and lower D threshold found using the heavy traffic approach proposed here the upper and lower thresholds found by the heuristic proposed in [30] and the associated ergodic cost value. This ergodic cost value was computed by the closed form expression presented by [30] for a control based on these upper and lower thresholds. The table also shows the value for the rate $\breve{\lambda}^a$ used as well as the number of processing stations $\breve{n}$ and the number of reserve stations $\breve{\mathfrak{r}}$ that were used for each example considered. Notice that the ergodic cost for the control using the approach proposed here is lower than the determined by the heuristic approach for every case.

For the second case, we suppose that the inter-arrival times are hyper-exponentially distributed. The results for this case are displayed by Table 2 for different values of arrival rate $\breve{\lambda}^a$, squared coefficient of variation $\breve{\sigma}_a^2$, number of processing stations $\breve{n}$ and number of reserve stations $\breve{\mathfrak{r}}$. The resulting upper and lower thresholds values for both approaches are also displayed. This time, the ergodic cost displayed in the table was calculated by the event-driven queuing simulation of this system implementing the control with the given threshold values, since no closed form for such a system is

TABLE 1. Table shows the results for the ergodic cost for the first scenario, where the inter-arrival times are exponentially distributed, for the control found by the approach proposed here (heavy-traffic) and the approach proposed by [30] (heuristics). Different values for the inter-arrival rates $\breve{\lambda}^a$, number of processing stations $\breve{n}$ and number of reserve stations $\breve{\mathtt{r}}$ were used. U and D represent the upper and lower thresholds found for the controls by the different approaches.

Method	U	D	$\breve{\sigma}_a^2$	$\breve{\lambda}^a$	$\breve{n}$	$\breve{\mathtt{r}}$	Ergodic Cost
heavy-Traffic	10	9	1	4	10	5	14.123313
heuristics	9	4	1	4	10	5	17.309647
heavy-Traffic	19	17	1	10	20	9	33.066862
heuristics	19	10	1	10	20	9	37.356842

TABLE 2. Table shows the results for the ergodic cost for the second scenario, where the inter-arrival times are hiper-exponentially distributed, for the control found by the approach proposed here (heavy-traffic) and the approach proposed by [30] (heuristics). Different values for the inter-arrival rates $\breve{\lambda}^a$, squared coefficient of variation of the inter-arrival times $\breve{\sigma}_a^2$, number of processing stations $\breve{n}$ and number of reserve stations $\breve{\mathtt{r}}$ were used. U and D represent the upper and lower thresholds found for the controls by the different approaches. The ergodic cost was calculated by an event-driven queuing simulation that implements the controls. The numbers following the symbol $\pm$ near the computed averages for the simulation are the 95% t confidence bounds.

Method	U	D	$\breve{\sigma}_a^2$	$\breve{\lambda}^a$	$\breve{n}$	$\breve{\mathtt{r}}$	Ergodic Cost (sim)
heavy-traffic	12	9	10	4	10	5	16.521 $\pm$ 0.004
heuristics	9	4	10	4	10	5	17.491 $\pm$ 0.004
heavy-traffic	21	17	10	10	20	9	37.230 $\pm$ 0.003
heuristics	19	10	10	10	20	9	39.020 $\pm$ 0.004

known. Again, notice that the ergodic cost for the control derived via the heavy traffic approach proposed here is lower that that for the heuristics.

7.2. **A Control Problem with Restriction.** We now consider an optimal control problem with restriction for the queuing model II, whose objective is to minimize the number of stations turned on while attending a maximal workload requirement. Similarly to the control problem of the previous section, we consider the following set of control actions $\alpha \in \mathcal{U} = \{0, 1\}$ where $\alpha = 1$ indicates that the reserve stations should change the state, and $\alpha = 0$ indicates that it should remain in the current state. However, this time we choose the following set of states $E = \{0, 1, 2, 3\}$ for the pure jump process $\breve{\theta}$, which models the state of the reserve machines. The state 0 represents that the reserve machines are "off," 1 represents the state where the reserve machines

are "turning on," 2 represents the state where the reserve machines are "powering off," and 3 means that the reserve machines are "on." We use the following transition rates:

$$\breve{\lambda}_{0\,1}(\alpha) = \begin{cases} \lambda^c & \text{if } \alpha = 1 \\ 0 & \text{otherwise,} \end{cases} \qquad \breve{\lambda}_{3\,2}(\alpha) = \begin{cases} \lambda^c & \text{if } \alpha = 1 \\ 0 & \text{otherwise,} \end{cases}$$

$$\breve{\lambda}_{1\,3}(\alpha) = \lambda^{on}, \qquad \breve{\lambda}_{2\,0}(\alpha) = \lambda^{off},$$

where $\breve{\lambda}_{1\,3}(\alpha)$ and $\breve{\lambda}_{2\,0}(\alpha)$ do not change with α, that is, the machines cannot be interrupted in their process of turning on or off. The rate λ^c is interpreted as the rate in which the control is implemented. For every test performed here, we set $\lambda^{on} = \lambda^{off}$.

Consider a system with $\breve{n}$ processing stations of which $\breve{\mathfrak{o}}$ are always turned on. The control problem considered here is the one discussed in Section 6.1 for the switching diffusion approximation $(\breve{x}, \breve{\theta})$ satisfying (5.25) with cost rates K and R, used in (6.19) and (6.20), given by $K(\xi, i) = \breve{n}$ if $i \in \{1, 2, 3\}$ and $\breve{\mathfrak{o}}$ if $i = 0$; and $R(\xi, i) \equiv \xi$. Notice that K measures the number of processing stations consuming power at a given time and R measures the scaled workload in the system. Let $W := \breve{n}^{-1/2}\tilde{W}$ denote the constant associated with the restriction (6.20), where $\tilde{W}$ denotes the *unscaled* maximum mean workload constraint constant.

The following data were used in numerical experiments: $\breve{n} = 150$, $\breve{\mathfrak{o}} = 120$, $\lambda^d = 1$, $\sigma_d^2 = 1/120$, $\breve{\lambda}^a = 110$, $\breve{\sigma}_a^2$ varying in the set $\{1, 10, 20\}$, $\lambda^c = \breve{\lambda}^a$, and $\lambda^{on} = \lambda^{off}$ varying in the set $\{0.1, 1.0, 10\}$. Where the inter-arrival times are assumed to be hyper-exponentially distributed and the service times are Erlang distributed. The values for $\tilde{W}$, the *unscaled* maximum mean workload constraint constant, are given by 3.0, 20, and 50 for the systems with $\breve{\sigma}_a^2 = 1$, 10, and 20, respectively. Given the appropriate data for the system, the transition probabilities for the approximating Markov chain, given by (6.8), (6.9) and (6.10), were computed and the associated linear programming problem was constructed. The linear program was solved using IBM ILOG CPLEX Optimizer, and the values of B and h, used in the approximation, were chosen empirically in order to improve the numerical results. The resulting control was implemented in the computer simulation of the system.

Two types of resulting controls were observed. For the first kind, named here "type-1," an upper U and a lower D threshold were found. Similarly to what was discussed in the previous section, the control operates in the following fashion: if the workload is above U and the reserves are off, then the reserves are turned on. In addition, if the workload is below D and the reserves are active, then they are turned off. The second kind of control, named here "type-2", was observed when the reserve machines took longer times to turn on and off. In this case, an upper threshold U was found, but the lower threshold D was zero. In this case, the reserve machines were turned off with some (usually small) probability p.

The results of the simulated system with the optimal controls found by the numerical method are given in Table 3, which reports the mean workload (MWL) obtained in the queuing system simulation (Sim) and the mean number of servers consuming power (MS). For the mean number of servers consuming power (MS), we display the value which was obtained by the simulation and also by the linear programming solution of the approximating Markov decision process. The mean workload associated with

TABLE 3. The table shows the mean workload of the system (MWL) and the average number of machines consuming power (MS). In this latter case, it is shown the values measured by the simulation (Sim) and the values obtained by the heavy traffic approximation (HT) (through the linear programming solution). It is also shown the type of control (CT) obtained as solution of the optimal control problem. The numbers following the symbol ± near the computed averages for the simulation are the 95% t confidence bounds.

$\breve{\sigma}_a^2$	λ^{on}	CT	MWL (Sim)	$\tilde{W}$	MS (Sim)	MS (HT)
1.0	0.1	2	2.793 ±0.152	3.0	142.4 ±1.050	138.1
	1.0	2	2.526 ±0.045		143.8 ±0.281	136.1
	10	2	2.772 ±0.015		134.8 ±0.121	126.5
10	0.1	2	27.62 ±1.460	20	136.4 ±1.260	144.8
	1.0	2	23.21 ±0.211		132.7 ±0.152	141.0
	10	1	17.12 ±0.146		129.8 ±0.155	130.1
20	0.1	2	63.74 ±0.379	50	130.0 ±0.140	138.2
	1.0	2	45.53 ±0.348		127.4 ±0.118	128.7
	10	1	45.21 ±0.446		124.7 ±0.106	125.7

the linear programming solution is always at the upper bound value W, since it is the limiting constraint for the control problem. From the results, we observe that the values obtained via the linear programming solution are close to the ones observed in the simulation. In addition, notice that the "type-1" controller was only obtained when the reserve machines change rapidly from the active to inactive states (and vice-versa) and the inter-arrival coefficient of variation is high. This leads us to believe that the type-2 controller may be compensating periods of long queues by maintaining the system operating with the reserves machines turned on for an unnecessarily longer period of time. This behavior is observed when the variance in the arrival process is small and the speed in which the servers change state is slow.

In order to verify whether there is gain in performing this task of turning machines on and off for the different settings considered here, the following test is proposed. The mean number of servers consuming power (MS) is obtained from the system operating the optimal control calculated by the simulation. This number is then rounded to the nearest integer and applied to a system with no control. That is, a system that always uses the same number of active servers. We refer to this system as "no control with optimal mean number of servers (NCO)." Table 4 contains the results of the simulation of NCO and that of the system operating the optimal control (OC). We are interested in comparing the mean and variance of the workload in these simulations. In order to facilitate the reference to the results, some information from Table 3 is repeated in Table 4.

Notice that, when the inter-arrival coefficient of variation $\breve{\sigma}_a^2$ is small and the reserve machines take a longer time to change state, it is better to operate the system with the fixed optimal mean number of servers. That is, the mean workload and its standard

TABLE 4. The table shows the data obtained with the simulation of a system with no control using the optimal mean number of servers (NCO) and the system operating the optimal control (OC). The values shown are: the mean workload in the system (MWL), the standard deviation of the workload (Std.), and the number of machines consuming power (MS) for the NCO system. It is also shown the type of control (CT obtained) as solution of the optimal control problem. The numbers following the symbol ± near the computed averages for the simulation are the 95% t confidence bounds.

$\breve{\sigma}_a^2$	λ^{on}	CT	MWL (OC)	Std. (OC)	MWL (NCO)	Std. (NCO)	MS
1.0	0.1	2	2.793 ±0.152	3.490 ±0.200	2.024 ±0.017	2.089 ±0.026	142
	1.0	2	2.526 ±0.045	3.229 ±0.083	1.912 ±0.023	1.945 ±0.025	144
	10	2	2.772 ±0.015	2.612 ±0.022	2.500 ±0.028	2.547 ±0.034	135
10	0.1	2	27.62 ±1.460	38.86 ±0.874	16.75 ±0.090	19.81 ±0.200	136
	1.0	2	23.21 ±0.211	25.39 ±0.254	19.53 ±0.294	22.84 ±0.488	133
	10	1	17.12 ±0.146	16.63 ±0.138	23.03 ±0.237	26.07 ±0.346	130
20	0.1	2	63.74 ±0.379	74.79 ±0.533	45.11 ±0.261	51.94 ±0.378	130
	1.0	2	45.53 ±0.348	44.41 ±0.342	52.38 ±0.766	59.75 ±0.980	127
	10	1	45.21 ±0.446	41.97 ±0.420	62.91 ±1.170	69.63 ±1.640	125

deviation are lower for the uncontrolled system operating with the optimal mean number of machines. Notice also that those cases coincide with the cases where the type-2 controller was found to be optimal, except for the case where $\breve{\sigma}_a^2 = 20$ and $\lambda^{on} = 1$. However, the situation is different for the cases where the coefficient of variation is high and the rates to turn machines on and off are fast enough (when the type-1 controller is observed). In these cases, the mean workload and the standard deviation are lowest for the system operating the optimal control.

8. Conclusion

In this paper, the problem of managing power consumption in large parallel systems by turning some of the machines on and off was considered. The approach taken here was to derive a switching diffusion approximation for this system and pose the decision problem as a continuous time optimal stochastic control problem. Two models were proposed, the first assumed general service and arrival time distributions under an assumption of task parallelization and the second considered general arrival times and exponentially distributed service times in a first-come first-served service regime with multiple servers. We showed that these controlled system can be approximated by controlled switching diffusions with jumping rates depending on the state of the continuous component of the process (through its dependency on the control function). In addition, a numerical scheme, based on the Markov chain approximation method, was proposed in order to solve the optimal control problem. In order to validate the approach, some numerical experiments where performed and compared with a computer simulation.

Appendix A. Proof of Some Auxiliary Results

In this appendix, we present the proof of some auxiliary results.

Proof of Lemma 3.3. Tightness of $m^{a,n}$, $m^{d,n}$ and $\tilde{m}^{d,n}$ follows from the criterion given by Theorem 2.7(b) of [20, p. 10]. Asymptotic continuity follows from the fact that the maximum of the jumps of m^n on any finite time interval converges to zero in probability by the uniform integrability assumptions (e.g. [9, p. 148]). Now let us consider the second statement of the lemma. Define $\mathcal{T}^n$ to be:

$$\mathcal{T}^n(t) := n^{-1}\sum_{l=1}^{\lfloor nt\rfloor}\Delta_l^{s,n} = -\bar{\Delta}^{s,n}n^{-1/2}m^{a,n}(t) + \bar{\Delta}^{s,n}n^{-1}\lfloor nt\rfloor$$

for each $t \geq 0$. The process $-\bar{\Delta}^{s,n}n^{-1/2}m^{a,n}$ converges in probability to the zero process, by tightness of $\{m^{a,n}\}$. This implies that $\mathcal{T}^n$ converges in probability to the process taking values $\bar{\Delta}^s t$. Let J^n be given by:

$$J^n(t) := \inf\{u \geq 0 : \mathcal{T}^n(u) > t\} = \inf\left\{u \geq 0 : \sum_{l=1}^{\lfloor nu\rfloor}\Delta_l^{s,n} > nt\right\},$$

using Theorem 7.2 of [39, p. 82], we have that this process converges weakly to the process taking values $\lambda^s t$. Now, since we have that $\sup_{t\leq T}|J^n(t) - a^n(t)| \leq n^{-1}$, for any $T > 0$, the sequence $\{a^n\}$ converges weakly to the process taking values $\lambda^s t$, by Theorem 3.1 of [6, p.27]. Since the limit is not random, the weak convergence implies convergence in probability. □

Proof of Lemma 5.4. Let us first consider that G_d contains only one point of discontinuity denoted by d. For $\epsilon > 0$, let $g_n^\epsilon \in C^\infty(\mathbb{R})$ be a non-decreasing real function with the following properties as $n \to \infty$: (i) $g_n^\epsilon, g_n^{\epsilon\prime}$ converge uniformly in $\mathbb{R}$ to continuous functions $g^\epsilon, g^{\epsilon\prime}$, respectively; and (ii) $g_n^{\epsilon\prime\prime}$ converges almost everywhere to a function $g^{\epsilon\prime\prime}$ which takes value $\frac{1}{2\epsilon}$ on $N_\epsilon(d) := \{x \in \mathbb{R} \ : \ d-\epsilon \leq x \leq d+\epsilon\}$ and 0 everywhere else (see the construction of such a function g_n^ϵ in [8, p. 144]). For $B > 0$, let $\varphi_B \in C_0^\infty(\mathbb{R})$ be a real function with compact support whose value is 1 for $|\xi| < B$ and 0 for $|\xi| > B+1$. Let $\tilde{g}_{Bn}^\epsilon : \mathbb{R}_+ \to \mathbb{R}$ be given by $\tilde{g}_{Bn}^\epsilon(\xi, i) = g_n^\epsilon(\xi)\varphi_B(\xi)$ for each $i \in E$ and $\xi \in \mathbb{R}_+$. Since $\tilde{g}_{Bn}^\epsilon \in \mathcal{D}_+$ and (x, θ) solves the submartingale problem for $\mathcal{L}^\mu$, we have that:

$$\mathbb{E}\left[\tilde{g}_{Bn}^\epsilon(x(t),\theta(t)) - \tilde{g}_{Bn}^\epsilon(x(0),\theta(0)) - \int_0^t(\mathcal{L}^\mu\tilde{g}_{Bn}^\epsilon)(s,x(s),\theta(s))ds\right] \geq 0.$$

By letting $B \to \infty$ and then $n \to \infty$, we have

$$\mathbb{E}\left[g^\epsilon(x(t)) - g^\epsilon(x(0)) - \int_0^t b(\theta(s))g^{\epsilon\prime}(x(s)) - \frac{\sigma^2}{4\epsilon}\mathbb{I}_{[d-\epsilon,d+\epsilon]}(x(s)) \right.$$
$$\left. - \sum_{j\in E}\int_{\mathcal{U}} g^\epsilon(x(s))\lambda_{\theta(s)j}(\alpha)\mu_s(d\alpha)ds\right] \geq 0.$$

Let $\delta > 0$, the inequality above implies that

$$\frac{4\epsilon}{\delta\sigma^2}\mathbb{E}\left[g^\epsilon(x(t)) - g^\epsilon(x(0)) - \int_0^t b(\theta(s))g^{\epsilon\prime}(x(s)) - \sum_{j\in E}\int_{\mathcal{U}} g^\epsilon(x(s))\lambda_{\theta(s)j}(\alpha)\mu_s(d\alpha)ds\right]$$
$$\geq \frac{1}{\delta}\mathbb{E}\left[\int_0^t \mathbb{I}_{[d-\epsilon,d+\epsilon]}(x(s))ds\right] \geq \mathbb{P}\left(\int_0^t \mathbb{I}_{[d-\epsilon,d+\epsilon]}(x(s))ds \geq \delta\right),$$

where we used Markov's inequality in the last line. Letting $\epsilon \to 0$ gives the result, since $g^\epsilon(\xi), g^{\epsilon\prime}(\xi)$ are bounded in ϵ. Clearly, we can consider the case where the set G_d has a finite number of discontinuity points by defining a function g_n^ϵ for each point d in G_d. □

Proof of Theorem 6.1. Let (z, θ) be the solution to (6.1) with control v^ϵ. Then, using Itô's formula (e.g., Theorem 4.57 [18, p. 57]), the fact that the diffusion spends zero amount of time at $\xi = 0$ (w.r.t Lebesgue's measure) and the fact that $V_x^\epsilon(0, i) = 0$ together with the property (5.8) of the reflection term, we have that:

$$\text{(A.1)}\qquad \mathbb{E}^{v^\epsilon}_{(z_0,i_0)}\left[V^\epsilon(z(t),\theta(t))\right] - V^\epsilon(z_0,i_0)$$
$$= \mathbb{E}^{v^\epsilon}_{(z_0,i_0)}\left[\int_0^t\int_{\mathcal{U}} \mathcal{L}^\alpha V^\epsilon(z(s),\theta(s))v^\epsilon(z(s),\theta(s))(d\alpha)ds\right]$$
$$\leq \mathbb{E}^{v^\epsilon}_{(z_0,i_0)}\left[\int_0^t \frac{\epsilon}{2} + \gamma^\epsilon - K(z(s),\theta(s))ds\right],$$

where we used (6.5) and the fact that $d^\epsilon(z(s), \theta(s)) < \epsilon/2$. Dividing both sides by t and taking the limit, we have:

$$\text{(A.2)}\qquad \gamma^\epsilon + \frac{\epsilon}{2} \geq \limsup_t \frac{1}{t}\int_0^t \mathbb{E}^{v^\epsilon}_{(z_0,i_0)}\left[K(z(s),\theta(s))\right]ds =: \gamma(z_0, i_0, v^\epsilon).$$

Now let $\tilde{v}$ denote a randomized Markovian control with associated switching diffusion given by $(\tilde{z}, \tilde{\theta})$, satisfying (6.1) for the control $\tilde{v}$. Then, by the minimization in (6.4) and the fact that $d(\xi, i) \geq -\epsilon/2$, we have that

$$\int_{\mathcal{U}} \mathcal{L}^\alpha V^\epsilon(\tilde{z}(s),\tilde{\theta}(s))\tilde{v}(\tilde{z}(s),\tilde{\theta}(s))(d\alpha) - \gamma^\epsilon + K(\tilde{z}(s),\tilde{\theta}(s)) \geq -\frac{\epsilon}{2}$$

at times $s \geq 0$ such that $\tilde{z}(s) > 0$. Then, repeating the steps used to derive the first equality in (A.1), we have that

$$\mathbb{E}^{\tilde{v}}_{(z_0,i_0)}\Big[V^\epsilon(\tilde{z}(t),\tilde{\theta}(t))\Big] - V^\epsilon(z_0,i_0) = \mathbb{E}^{\tilde{v}}_{(z_0,i_0)}\Big[\int_0^t\int_{\mathcal{U}} \mathcal{L}^\alpha V^\epsilon(\tilde{z}(s),\tilde{\theta}(s))v^\epsilon(\tilde{z}(s),\tilde{\theta}(s))(d\alpha)ds\Big]$$
$$\geq \mathbb{E}^{\tilde{v}}_{(z_0,i_0)}\Big[\int_0^t -\frac{\epsilon}{2} + \gamma^\epsilon - K(\tilde{z}(s),\tilde{\theta}(s))ds\Big].$$

Dividing both sides by t and taking the limit, we have

$$\gamma^\epsilon - \frac{\epsilon}{2} \leq \limsup_t \frac{1}{t}\int_0^t \mathbb{E}^{\tilde{v}}_{(z_0,i_0)}\Big[K(\tilde{z}(s),\tilde{\theta}(s))\Big]ds =: \gamma(z_0, i_0, \tilde{v}).$$

Therefore, the above inequality implies that $\gamma(z_0, i_0, v^\epsilon) - \epsilon \leq \gamma(z_0, i_0, \tilde{v})$, which concludes the proof. □

Proof of Theorem 6.3. First notice that $|\Delta\zeta^h_{1,k}| \leq h$ by definition of the Markov chain and, therefore, (6.11) is satisfied. Let us now consider $0 < x < B$. It is straight forward to verify condition (6.12). In fact, we have that

$$\mathbb{E}^{\alpha,h}_{(x,i)}\left[\Delta\zeta^h_{1,k}\right] = \frac{h^2(c^+(x,i) + c^-(x,i))}{\bar{M}^h} = c(x,i)\Delta t^h,$$

where we used the fact that $c^+(x,i) + c^-(x,i) = c(x,i)$ and that $\Delta t^h = h^2/\bar{M}^h$. In addition, by the fact that $c^+(x,i) - c^-(x,i) = |c(x,i)|$, we can verify (6.13) by noticing that

$$\mathbb{E}^{\alpha,h}_{(x,i)}\left[(\Delta\zeta^h_{1,k})^2\right] = \frac{h^2\sigma^2}{\bar{M}^h} + \frac{h^3|c(x,i)|}{\bar{M}^h} = \sigma^2\Delta t^h + o(\Delta t^h),$$

and that $\left(\mathbb{E}^{\alpha,h}_{(x,i)}\left[\Delta\zeta^h_{1,k}\right]\right)^2 = o(\Delta t^h)$. Now, for conditions (6.14) and (6.15), we have

$$\mathbb{P}^{\alpha,h}_{(x,i)}\left\{(\zeta^h_{k+1})_2 = j\right\} = p^h((x,i),(x,j)|\alpha) = \breve{\lambda}_{ij}(\alpha)\Delta t^h \quad \text{and}$$
$$\mathbb{P}^{\alpha,h}_{(x,i)}\left\{(\zeta^h_{k+1})_2 = i\right\} = \sum_{y\in S^h(x)} p^h((x,i),(y,i)|\alpha) = 1 + \breve{\lambda}_{ii}(\alpha)\Delta t^h.$$

with $S^h(x) := \{x - h, x, x + h\}$. The same holds for $x \in \{0, B\}$ by (6.9) and (6.10).

Now, let $x = 0$, then we can verify (6.16) by noticing that

$$\mathbb{E}^{\alpha,h}_{(0,i)}\left[\Delta\zeta^h_{1,k}\right] = p^h((0,i),(h,i)|\alpha)h = h\frac{1}{2} + h\left(\frac{\sigma^2/2 + hc^+(0,i)}{\bar{M}^h} - \frac{1}{2}\right). \tag{A.3}$$

and by the fact that the term $(\sigma^2/2 + hc^+(0,i))/\bar{M}^h$ converges to $1/2$ as $h \to 0$, by the definition of $\bar{M}^h$. Now $\mathbb{E}^{\alpha,h}_{(0,i)}\left[\Delta\zeta^h_{1,k}\right] = p^h((0,i),(h,i)|\alpha)h^2$ and it will satisfy (6.18) for $x = 0$ by (A.3). By analogous arguments, we can verify (6.17) and (6.18) for $x = B$. Also, by the preceeding discussion, it is straight forward to verify that there are $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ such that $p^h((0,i),(h,i)|\alpha) \geq \varepsilon_1$ and $p^h((B,i),(B-h,i)|\alpha) \geq \varepsilon_2$. □

References

[1] E. Altman and F. Spieksma. The linear program approach in multi-chain markov decision processes revisited. *Mathematical Methods of Operations Research*, 42:169–188, 1995.

[2] D. Applebaum. *Lévy Processes and Stochastic Calculus*. Cambridge University Press, Cambridge; New York, 2nd edition edition, 2009.

[3] J. R. Artalejo, A. Economou, and M. J. Lopez-Herrero. Analysis of a multiserver queue with setup times. *Queueing Syst. Theory Appl.*, 51(1-2):53–76, 2005.

[4] L.A. Barroso, J. Dean, and U. Holzle. Web search for a planet: The google cluster architecture. *IEEE Micro*, 23(2):22–28, 2003.

[5] L.A. Barroso and U. Hölzle. The case for energy-proportional computing. *Computer*, 40(12):33–37, 2007.

[6] P. Billingsley. *Convergence of Probability Measures (2nd Edition)*. John Wiley & Sons, New York, 1999.

[7] V. S. Borkar. Controlled diffusion processes. *Probability Surveys*, 2:213–244, 2005.
[8] K. L. Chung and R. J. Williams. *Introduction to Stochastic Integration*. Birkhäuser, New York, 2nd ed. edition, 2013.
[9] S. N. Ethier and T. G. Kurtz. *Markov Process Characterization and Convergence*. John Wiley & Sons, Hoboken, NJ, 1986.
[10] M. D. Fragoso and J. Baczynski. Optimal control for continuous-time linear quadratic problems with infinite Markov jump parameters. *SIAM Journal on Control and Optimization*, 40:270–297, 2001.
[11] M. D. Fragoso and E. M. Hemerly. Optimal control for a class of noisy linear systems with markovian jumping parameters and quadratic cost. *International Journal of Systems Science*, 22(12):2553–2561, 1991.
[12] A. Gandhi, V. Gupta, and M.A. Kozuch M. Harchol-Balter. Optimality analysis of energy-performance trade-off for server farm management. *Perform. Eval.*, 67(11):1155–1171, 2010.
[13] M. Ghosh, A. Arapostathis, and S. Marcus. Optimal Control of Switching Diffusions with Application to Flexible Manufacturing Systems. *SIAM Journal on Control and Optimization*, 31(5):1183–1204, 1993.
[14] A. Greenberg, J. Hamilton, D.A. Maltz, and P. Patel. The cost of a cloud: research problems in data center networks. *SIGCOMM Comput. Commun. Rev.*, 39(1):68–73, 2008.
[15] X. Guo and O. Hernández-Lerma. *Continuous-Time Markov Decision Processes*. Stochastic Modelling and Applied Probability. Springer, Berlin, Heidelberg, 2009.
[16] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, 1981.
[17] A. Hordijk and L. C. M. Kallenberg. Constrained undiscounted stochastic dynamic programming. *Mathematics of Operations Research*, 9(2):276–289, 1984.
[18] J. Jacod and A.N. Shiryaev. *Limit theorems for stochastic processes*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, New York, 1987.
[19] H. Kaspi and K. Ramanan. SPDE limits of many-server queues. *The Annals of Applied Probability*, 23(1):145–229, 2013.
[20] T. G. Kurtz. *Approximation of Population Processes*. SIAM, Philadelphia, Pa, 1981.
[21] T. G. Kurtz. Martingale problems for constrained Markov processes. In *Recent Advances in Stochastic Calculus, J. S. Baras and V. Mirelli, Eds.*, pages 151–168. Springer-Verlag, New York, 1990.
[22] H. J. Kushner. *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*. Systems & Control: Foundations & Applications. Birkhäuser, Boston, 1990.
[23] H.J. Kushner. *Heavy traffic Analysis of Controlled Queueing and Communication Networks*. Springer-Verlag, New York, 2001.
[24] H.J. Kushner and Y. N. Chen. Optimal control of assignment of jobs to processors under heavy traffic. *Stochastics and Stochastic Reports*, 68:22–8, 1999.
[25] H.J. Kushner and P.G. Dupuis. *Numerical Methods for Stochastic Control Problems in Continuous Time*. Springer–Verlag, New York, 1992.

[26] N. Lee and V. G. Kulkarni. Optimal arrival rate and service rate control of multi-server queues. *Queueing Systems*, 76:37–50, 2014.

[27] A. Mandelbaum, W. A. Massey, and M. I. Reiman. Strong approximations for Markovian service networks. *Queueing Systems*, 30(1-2):149–201, 1998.

[28] M. Mazzucco and D. Dyachuk. Balancing electricity bill and performance in server farms with setup costs. *Future Generation Computer Systems*, 28(2):415 – 426, 2012.

[29] M. Mazzucco and D. Dyachuk. Optimizing cloud providers revenues via energy efficient server allocation. *Sustainable Computing: Informatics and Systems*, 2(1):1 – 12, 2012.

[30] I. Mitrani. Managing performance and power consumption in a server farm. *Annals of Operations Research*, 202(1):121–134, 2013.

[31] D. Nguyen and G. Yin. Modeling and Analysis of Switching Diffusion Systems: Past-Dependent Switching with a Countable State Space. *SIAM Journal on Control and Optimization*, 54(5):2450–2477, 2016.

[32] D. Niyato, S. Chaisiri, and L.B. Sung. Optimal power management for server farm to support green computing. In *Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid*, CCGRID '09, pages 84–91, Washington, DC, USA, 2009. IEEE Computer Society.

[33] A. A. Puhalskii and M. I. Reiman. The multiclass GI/PH/N queue in the Halfin-Whitt regime. *Advances in Applied Probability*, 32(2):564–595, 2000.

[34] M. L. Puterman. *Markov decision processes : discrete stochastic dynamic programming*. John Wiley and Sons, New York, 1994.

[35] Q. S. Song, G. Yin, and Z. Zhang. Numerical methods for controlled regime-switching diffusions and regime-switching jump diffusions. *Automatica*, 42(7):1147–1157, 2006.

[36] E. Le Sueur and G. Heiser. Dynamic voltage and frequency scaling: the laws of diminishing returns. In *Proceedings of the 2010 international conference on Power aware computing and systems*, HotPower'10, pages 1–8, Berkeley, CA, USA, 2010. USENIX Association.

[37] H. Tanaka. Stochastic differential equations with reflecting boundary condition in convex regions. *Hiroshima Mathematical Journal*, 9(1):163–177, 1979.

[38] W. Whitt. Heavy traffic limit theorems for queues: A survey. In M. Beckmann and H.P. Kunzi, editors, *Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, New York, 1974.

[39] W. Whitt. Some useful functions for functional limit theorems. *Mathematics of Operations Research*, 5(1):67–85, 1980.

[40] W. Whitt. *Stochastic-Process Limits*. Springer-Verlag, New York, 2002.

[41] F. Xi and C. Zhu. On the martingale problem and Feller and strong Feller properties for weakly coupled Lévy type operators. *Stochastic Processes and their Applications*, 128(12):4277–4308, 2018.

[42] G. Yin, C. Zhang, and L. Y. Wang. Numerical Methods for Controlled Switching Diffusions. In Ivan Lirkov and Svetozar Margenov, editors, *Large-Scale Scientific*

Computing, Lecture Notes in Computer Science, pages 33–44. Springer International Publishing, 2018.