Asymptotic optimality of switched control policies in a simple parallel server system under an extended heavy traffic condition

Rami Atar* Eyal Castiel* Marty Reiman†

November 23, 2023

Abstract

This paper studies a 2-class, 2-server parallel server system under the recently introduced extended heavy traffic condition [1], which states that the underlying 'static allocation' linear program (LP) is critical, but does not require that it has a unique solution. The main result is the construction of policies that asymptotically achieve a lower bound, proved in [1], on an expected discounted linear combination of diffusion-scaled queue lengths, and are therefore asymptotically optimal (AO). Each extreme point solution to the LP determines a control mode, i.e., a set of activities (class–server pairs) that are operational. When there are multiple solutions, these modes can be selected dynamically. It is shown that the number of modes required for AO is either one or two. In the latter case there is a switching point in the (normalized) workload domain, characterized in terms of a free boundary problem. Our policies are defined by identifying pairs of elementary policies and switching between them at this switching point. They provide the first example in the heavy traffic literature where weak limits under an AO policy are given by a diffusion process where both the drift and diffusion coefficients are discontinuous.

MSC 2020 classification: 60K25 ; 68M20 ; 93E20 ; 60F17 ; 90B36

Keywords: parallel server systems; decomposable service rates; extended heavy traffic condition; dynamic graph of basic activities; switched control systems; diffusion with discontinuous coefficients.

1 Introduction

1.1 Background

Parallel server systems (PSS) are queueing control problems in which a number of servers offer service to customers of different classes, and choices as to which customer class each server is dedicated to are made dynamically. Since its introduction in [9], its study in heavy traffic has attracted much attention due to its simple structure, its practical significance, and the theoretical challenges it poses. The problem formulation in [9] includes a key assumption, referred to as the *heavy traffic condition* (HTC), which states that an underlying 'static allocation' linear program

*Viterbi Faculty of Electrical and Computer Engineering, Technion, Haifa, Israel

†Department of Industrial Engineering and Operations Research, Columbia University, New York City, NY

(LP) satisfies a critical load condition and that it has a unique solution. Whereas critical load is universally considered a defining condition of any notion of heavy traffic, uniqueness of solutions has been assumed mainly because it simplifies the mathematical treatment. The *extended HTC* (EHTC), which merely states that the LP is at criticality but does not require uniqueness, has recently been introduced in [1] in order to address a considerably broader notion of heavy traffic. The main result of [1] is a lower bound on the asymptotically achievable cost in a general PSS under the EHTC. This paper focuses on the 2-class, 2-server PSS referred to in this introduction as the 2×2 PSS, which is the simplest case in which the EHTC is strictly broader than the HTC. The goal is to complement the results of [1] in this case by constructing policies that asymptotically achieve the lower bound, which hence are asymptotically optimal (AO) in heavy traffic.

The structure of the 2×2 PSS is as follows. Each of the two servers is capable of serving each of the two classes. The classes (respectively, servers) are usually indexed using the symbol i (respectively, k), and activities, namely class-server pairs, by $j = (i, k)$. Arriving customers await service in class-based queues, and upon receiving a single service, leave the system. The control decisions consist of routing (determining which server serves each customer) and sequencing (determining the order in which they are served). The rates of arrivals of customers of the two classes are denoted by λ_i^n, $i = 1, 2$, and the rates of service at each of the four activities are denoted by μ_{ik}^n, where n denotes the usual heavy traffic parameter. These rates are assumed to be asymptotic to $\lambda_i n + \hat{\lambda}_i n^{1/2}$ and $\mu_{ik} n + \hat{\mu}_{ik} n^{1/2}$, for some given λ_i, $\hat{\lambda}_i$, μ_{ik}, $\hat{\mu}_{ik}$. The cost consists of an expected infinite horizon discounted linear combination of the two queue lengths, and is rescaled at the diffusion scale.

Whereas the cost, and consequently the notion of AO, are set up at the diffusion scale, the underlying LP alluded to above addresses the behavior of the PSS at the fluid, or law-of-large-numbers (LLN) scale. Posed in terms of the first order parameters, λ_i, μ_{ik}, it is concerned with the mean fraction of time devoted by each server to each class. When the LP has a unique solution, at least one activity is *non-basic*, in the sense that the fraction allocated to it is zero. The so called *graph of basic activities* (GBA), formed by the activities with positive allocation fraction, is static. In this case, the critical load condition dictates that any policy not adhering to this solution, in the sense of effort allocation, causes the total queue length to blow up, and in particular cannot be AO. Under policies that adhere to this solution, the LLN assures that the aforementioned fractions of effort converge to those given by the LP solution (a necessary, but certainly not sufficient condition for AO). When there are multiple LP solutions, a result from [1] states that for the 2×2 PSS, the space of solutions, denoted by $\mathcal{S}_{\rm LP}$, forms a line segment ${\rm ch}(\xi^{*,1}, \xi^{*,2})$ in the space of 2×2 matrices (where ch denotes the convex hull). In each of the two extremal solutions, $\xi^{*,1}, \xi^{*,2}$ there is again at least one non-basic activity. We refer to these two extreme points as *control modes*, or simply *modes*. For similar reasons, any policy that does not lead to an unbounded cost should keep the system critically loaded at all times, and thus, asymptotically, the fractions of effort will vary dynamically within $\mathcal{S}_{\rm LP}$. In the control literature, a control process that takes values only at the vertices of the action space is called a bang-bang control. The analogue of this notion in our setting is a policy for which the limiting fractions of effort take values only in $\{\xi^{*,1}, \xi^{*,2}\}$, switching between the two extremal solutions. Some of the policies introduced in this paper are designed to act that way.

Contrary to the setting where the HTC holds, it is impossible to construct an AO policy based only on the first order data under the EHTC. A second order approximation of the PSS, which is often referred to as a *Brownian control problem* (BCP), is required. The BCP represents a diffusion limit of the PSS, in which Brownian motion (BM) replaces stochastic fluctuations associated with cumulative arrival and service processes. Closely related to the BCP is another diffusion control problem, called a *workload control problem* (WCP). Obtained by a certain projection of the BCP, it is a control problem in which the process is one-dimensional, representing the total workload asymptotics. The structure of the WCP obtained is quite simple to describe. The state process is a reflected diffusion on $\mathbb{R}_+$, with controlled drift and diffusion coefficients, $b = b(\xi)$, $\sigma = \sigma(\xi)$, where the control process, $\xi = \xi_t$ takes values in $\mathcal{S}_{\rm LP}$ and $\xi \mapsto (b(\xi), \sigma(\xi)^2)$ is an affine map. The cost is given as an expected discounted version of the state process itself. By a standard argument based on the Hamilton-Jacobi-Bellman (HJB) equation, there exists an optimal bang-bang control for the WCP. There can therefore be two possibilities for the WCP solution: the single mode case, where one of the modes is always used, and the dual mode case, where both modes are used by the optimal control in different parts of the state space. Note that in this case the GBA can be changed dynamically. The HJB equation also reveals the structure of the feedback function from state to control. This particular HJB equation was solved in [16]. It was shown that in the dual mode case there is a switching point $z^* \in (0, \infty)$, such that one of the modes is used when the state is below z^* and the other otherwise. The HJB equation can be viewed, in this case, as an equation involving a free boundary, in which the solution is a pair, where one component is the value function and the other is z^*. The results of [16] also characterize z^* as the unique solution to an explicit equation, as well as a solution to the HJB equation.

Our policies are obtained by 'translating' the WCP solution. In the case of a single mode, the prescribed policy corresponds either to a threshold policy of the form that first appeared in [3] (see below) or a simple priority policy, depending on the mode used and the cost. In the dual mode case, pairs of elementary policies are identified, which are combined together to form switched control policies, so that one is active when the normalized workload process is below the switching point and the other above it. In each case, the policy is designed to meet the target allocation efforts determined by the corresponding mode, and the set of operational activities is restricted by the corresponding GBA.

The paper closest to ours is the aforementioned [3], that studies a 2-server, 2-class PSS with 3 activities. This PSS is known as an 'N' network, because upon relabeling, the activities are given by $(1, 1)$, $(1, 2)$ and $(2, 2)$, forming the symbol N. (See Figure 1 (a).) In this network the number of solutions to the LP cannot exceed 1, and thus the requirement of a unique solution does not pose a restriction. In an earlier work, [8], it had been observed that when the larger '$c\mu$' value is in class 1, the BCP solution suggests that the queue length of class 1 customers and the idleness process at server 1 should both converge to zero at the diffusion scale, and that a simple priority policy does not achieve this. In [3] this was addressed by putting a threshold on class 1 queue length, that when exceeded, server 2 prioritizes class 1, and otherwise it prioritizes class 2. The size of the threshold must converge to zero at the diffusion scale so as to achieve the first goal. To achieve AO of a threshold policy with logarithmic (in n) size threshold, as used in [3], the interarrival and service times are assumed there to possess exponential moments. (More on the history of the problem and the works that contributed to its development can be read in [1].)

As already mentioned, one of the policies we implement is a threshold policy similar to the one used by [3]. However, our assumptions are positioned differently with respect to the threshold–moment tradeoff, assuming a larger (still $o(\sqrt{n})$), polynomial size threshold, but requiring only a polynomial moment assumption. We assume $2+\varepsilon$ moment assumptions for all of our policies except the single-mode threshold policy and the dual-mode policies that employ the threshold policy when the workload is above z^*. For these, a finite $\mathbf{m}_0$-th moment is assumed, where the number $4 < \mathbf{m}_0 < 5$ is indicated explicitly. Another difference between our results and those of [3] is that our policies do not use preemption. Although the policy introduced in [3] uses preemption, it is plausible that an analogous non-preemptive policy is also AO under similar conditions; see Corollary 2.15 and Remark 2.16 below. In this paper, our choice not to use preemption leads to non-trivial issues in the dual mode case. Instead of a simple switching between elementary policies when the workload crosses z^*, it is sometimes the case that one must wait for a particular server to become available before switching. This is described in §5.1.3.

It is also worth mentioning that we have argued in [1] that the AO of the threshold policy from [3] extends beyond the HTC to the case of multiple solutions and a single mode (under some assumptions which include the existence of exponential moments).

Beside the objective to break the uniqueness barrier, an additional source of motivation for this work stems from the relation between non-uniqueness and service rate decomposability. As stated in Lemma 2.1, for the 2×2 PSS, the LP exhibits multiple solutions if and only if the service rates decompose as $\mu_{ik} = \alpha_i \beta_k$. Service rates decompose this way when the mean size of a job is characteristic to the class (and then α_i is the reciprocal mean), and each server has its own processing speed (here given by β_k). As the HTC does not hold under decomposability, this important class of service rates has been left out by earlier work.

1.2 Results

The description of the policies given above is only a sketch. There are nontrivial issues that arise regarding the need to 'patch' 2 policy types, requiring us to slightly modify the policies, where the details differ from one pairing to another.

The main result states that, under the prescribed policies, the rescaled workload process converges in law to the diffusion process that solves the WCP, and these policies are AO. As far as convergence is concerned, in addition to the 'standard' issues involved in proving state-space collapse, we need to deal with issues related to switching control modes at z^*. Moreover, to obtain AO from weak convergence, uniform integrability needs to be established, and it is here where the $2+\varepsilon$ and $\mathbf{m}_0$ moment assumptions are used.

An approach to proving convergence to a diffusion with discontinuous coefficients, addressing especially the technicalities involved with the discontinuity of the diffusion coefficient, was developed in [11], going beyond the general framework for convergence of semimartingales such as that from [13]. Whereas the tools from [11] are not directly applicable in our setting, an argument which, as in [11], shows that the time spent near the discontinuity set is negligible, is also at the basis of our proof. The paper [11] also gives an example of a queueing model whose scaling limit yields a diffusion process with discontinuities in both drift and diffusion coefficients. Our dual mode case provides what seems to be the first example where this occurs under an AO policy of a queueing control problem (for AO in heavy traffic leading to discontinuity in the drift only, see [2]).

1.3 Organization of the paper

In §2.1 we describe our model and the control problem associated with it in more detail. The LP and the extended heavy traffic condition are introduced in §2.2 and preliminary results about the LP from [1] are stated. In §2.3, the WCP and the associated HJB equation are introduced, and in Proposition 2.6, it is stated that there exists a unique classical solution to the HJB equation. This proposition also provides a condition which determines whether an optimal solution to the WCP must employ a single mode or two modes (not to be confused with the number of modes in the space of LP solutions, which is always two under multiplicity), and asserts that in the dual mode case there exists a single switching point z^* in workload space. We also present in this section the lower bound from [1] stated in Theorem 2.4. The main result is stated in §2.4. The definitions of the proposed policies appear first, and then, in Theorem 2.13, the weak convergence and AO results are stated. Numerical results are presented in §2.5.

In §3 we state and prove some results related to the static allocation LP, providing, in particular, explicit expressions for the extreme points of the set of optimal solutions. Development of the WCP is carried out in §4. Preliminary results proved in [1] in a general case are included in this section. This section also contains proofs of results related to the HJB equation, some of which rely on [16].

The proof of our main result, Theorem 2.13, is the subject of §5. In §5.1, we present the general scheme for proving Theorem 2.13: the weak convergence result is stated in Theorem 5.1. We then present four propositions that are used for the proof. Each proposition corresponds to a specific section and step of the proof. Proposition 5.3, in §5.2.2, proves uniform integrability; state space collapse is proved in Proposition 5.4 in §5.2.3; a key non idling property is proved in Proposition 5.5 in §5.2.4; and a 'fast switching' property, showing the aforementioned property that the process spends asymptotically negligible time near the discontinuity, is proved in Propostion 5.6 in §5.2.6. Finally, the appendix contains proofs of several lemmas stated earlier.

1.4 Notation

$\mathbb{N}$, $\mathbb{R}$ and $\mathbb{R}_+$ are the sets of natural, real and, respectively, nonnegative real numbers. For $a, b \in \mathbb{R}$, $a \vee b$ and $a \wedge b$ denote the maximum and minimum of a and b, respectively, and $a^+ = a \vee 0$. For a set A, $\mathbb{1}_A$ denotes its indicator function. For $f : \mathbb{R}_+ \to \mathbb{R}$ and $t, \delta > 0$, $\|f\|_t = \sup_{s\in[0,t]} |f(s)|$ and

$$w_t(f, \delta) = \sup\{|f(s_1) - f(s_2)| : 0 \le s_1 \le s_2 \le (s_1 + \delta) \wedge t\}.$$

For $0 \le s \le t$, the notation $f[s, t]$ stands for $f(t) - f(s)$. For real-valued functions and processes, the notation $X(t)$ is used interchangeably with X_t. Given a Polish space E, $C_E[0, \infty)$ and $D_E[0, \infty)$ denote the spaces of E-valued, continuous and, respectively, càdlàg functions on $[0, \infty)$, equipped with the topology of convergence u.o.c. and, respectively, the J_1 topology. Denote by $C^+_{\mathbb{R}}[0, \infty)$ and $D^+_{\mathbb{R}}[0, \infty)$ the subset of $C_{\mathbb{R}}[0, \infty)$ and, respectively, $D_{\mathbb{R}}[0, \infty)$, of non-negative, non-decreasing functions, and by $C^{0,+}_{\mathbb{R}}[0, \infty)$ the subset of $C^+_{\mathbb{R}}[0, \infty)$ of functions that are null at zero. Write $X_n \Rightarrow X$ for convergence in law. A sequence of processes with sample paths in $D_E[0, \infty)$ is said to be C-tight if it is tight and the limit of every weakly convergent subsequence has sample paths in $C_E[0, \infty)$ a.s. The letter c denotes a deterministic constant whose value may change from one appearance to another.

2 Model and main results

The setup and results presented in this section have quite a few ingredients, as already mentioned in the introduction: the LP and modes, the diffusion scaling, the WCP and HJB equation, the switching point z^*, threshold and switching policies. Before going into the details we provide a roadmap. The 2-class, 2-server system, introduced in §2.1, is indexed by n and is observed at the diffusion scale. The assignment of jobs from the different classes to different servers is regarded as a control policy. A cost functional is considered, given by the expected discounted weighted sum of queue lengths, also rescaled at the diffusion scale; see (2.6). The weights of the queue lengths are given by a vector (h_1, h_2).

In order to formulate a policy-independent condition for heavy traffic, an LP (2.7) is introduced, involving first order arrival and service rates, where the variable to be determined is a 2×2 allocation matrix describing the (first order) fraction of effort each server dedicates to each class. It is assumed to be at criticality, in the sense that servers are fully occupied but queue lengths stay balanced. In that regard we adopt the heavy traffic notion of [9] and many other papers that followed, with one important distinction: We do not assume that there is a unique allocation matrix that maintains criticality, i.e. there may be multiple LP solutions. When there are multiple solutions, they are all given as convex combinations of two allocation matrices (Lemma 2.1), which we call modes. Because earlier work has addressed the unique solution case, we only treat the case of multiple solutions (Assumption 2.2). A result from [1] states that under this assumption the first order rates μ_{ik} are necessarily decomposable as $\alpha_i\beta_k$. Each of the modes induces a so-called graph of basic activities, as in Figure 1. The parameters α_i and the structure of the graphs influence the type of control policy to be proposed.

The WCP (2.16)–(2.17) is a control problem for a diffusion process in dimension 1, in which both the drift and the diffusion coefficient are controlled by a control process that takes values in the space of allocation matrices. The values of these coefficients, evaluated at the two modes, are denoted b_m and, respectively, σ_m, $m = 1, 2$. The WCP formally describes the control problem associated with the PSS at the diffusion limit, and its control process represents the dynamic selection of the mode at which the PSS operates. The significance of the WCP has two aspects. First, as was shown in [1], its solution gives a lower bound on the PSS cost asymptotics under any sequence of control policies (Theorem 2.4). Accordingly, any policy that achieves this bound is AO. Second, it suggests how AO policies should be structured. In particular: (a) State space collapse should hold, which in our setting involves two properties that should hold up to a level negligible at diffusion scale: (i) Both servers should be busy whenever there is any work in the system, and (ii) all queue length should be kept in the class i that minimizes $h_i\alpha_i$. Roughly speaking, (ii) implies that the class that maximizes $h_i\alpha_i$ should be prioritized. This type of sequencing policy is known as a $c\mu$ rule. (But, as shown by [8], and [3], something more than a simple priority policy may be needed here in order to accomplish (i).) (b) When, for either $(m, m') = (1, 2)$ or $(m, m') = (2, 1)$, one has $b_m \le b_{m'}$ and $\sigma_m \le \sigma_{m'}$, only mode m should be used. Otherwise both modes should be used, by dynamically selecting them at different parts of the state space. The partition of the state space is defined via a switching point z^*: When the diffusion-scaled workload is below z^*, the mode m with $b_m \ge b_{m'}$ should be used, otherwise m'. Determining this switching point is done by solving an HJB equation, (2.18)–(2.19).

The construction of an AO policy must take into account several considerations: Whether to operate in one or two modes, and in the latter case, their ordering in workload space; which class has high priority; and the structure of the graph of basic activities for each of the modes. Different types of policies apply in different cases (Definitions 2.9, 2.10, 2.11, 2.12). The main result states that these policies are AO and that the normalized workload converges weakly to the diffusion process given by the WCP state process under an optimal control (Theorem 2.13).

2.1 Queueing model, scaling and queueing control problem

The model under consideration is as in [1], specialized to the case of two job classes, two servers and four activities. We will refer to it as the 2×2 PSS when there is need to distinguish it from the general PSS treated in [1]. The symbol $i \in \{1, 2\}$ is used as a generic index to a class, and $k \in \{1, 2\}$ to a server. For a general PSS, an *activity* is a class-server pair (i, k) where server k is capable of serving class i. In this paper it is assumed that each server is capable of serving each class, hence there are four activities. They are labeled by (i, k) or sometimes by $j = (i, k)$.

The model consists of a sequence of systems, indexed by $n \in \mathbb{N}$, that are all defined on one probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For the nth system, one considers the following processes. The processes denoted $A^n = (A^n_i)$ and $S^n = (S^n_{ik})$ represent arrival and potential service counting processes. That is, $A^n_i(t)$ is the number of arrivals of class i jobs until time t, $i = 1, 2$, and $S^n_{ik}(t)$ is the number of service completions of class i jobs by server k, by the time server k has devoted t units of time to class i, $i = 1, 2$, $k = 1, 2$. Next, $X^n = (X^n_i)$, $I^n = (I^n_k)$, $D^n = (D^n_{ik})$ and $T^n = (T^n_{ik})$ denote queue length, cumulative idleness, departure, and cumulative busyness processes. In other words, $X^n_i(t)$ is the number of class i customers in the system at time t, $I^n_k(t)$ is the cumulative time server k has been idle by time t, $D^n_{ik}(t)$ is the number of class i departures from server k, and $T^n_{ik}(t)$ is the cumulative time devoted by server k to class i. The process T^n_{ik} takes the form $T_{ik}(t) = \int_0^t \Xi^n_{ik}(s)ds$, where $\Xi^n_{ik}(t)$ is the fraction of effort devoted by server k to class-i jobs at t. In particular, $\sum_i \Xi^n_{ik}(t) \le 1$ for every k. Thus Ξ^n is referred to as the allocation process.

The aforementioned arrival and potential service processes are constructed as follows. Arrival rates λ^n_i and service rates μ^n_{ik} are given, satisfying, for some constants $\lambda_i \in (0, \infty)$, $\mu_{ik} \in (0, \infty)$, $\hat{\lambda}_i \in \mathbb{R}$, $\hat{\mu}_{ik} \in \mathbb{R}$,

$$\hat{\lambda}^n_i := n^{-1/2}(\lambda^n_i - n\lambda_i) \to \hat{\lambda}_i,$$
$$\hat{\mu}^n_{ik} := n^{-1/2}(\mu^n_{ik} - n\mu_{ik}) \to \hat{\mu}_{ik},$$

as $n \to \infty$. For each i a renewal process $\check{A}_i$ is given, with interarrival distribution that has mean 1 and squared coefficient of variation $0 < C^2_{A_i} < \infty$. Similarly, for each (i, k), a renewal process $\check{S}_{ik}$ is given with mean 1 interarrival and squared coefficient of variation $0 < C^2_{S_{ik}} < \infty$. It is assumed that A^n and S^n are given by

$$A^n_i(t) = \check{A}_i(\lambda^n_i t), \qquad S^n_{ik}(t) = \check{S}_{ik}(\mu^n_{ik} t).$$

It is assumed moreover that the six processes $\check{A}_i$, $\check{S}_{ik}$ are mutually independent, have strictly positive inter-arrival distributions and right-continuous sample paths. The (IID) interarrivals of $\check{A}_i$ and $\check{S}_{ik}$ are denoted by $\check{a}_i(l)$ and $\check{u}_{ik}(l)$, $l \in \mathbb{N}$, respectively, and those of the accelerated processes A^n_i and S^n_{ik} are given by

$$a^n_i(l) = \frac{1}{\lambda^n_i}\check{a}_i(l), \qquad u^n_{ik}(l) = \frac{1}{\mu^n_{ik}}\check{u}_{ik}(l). \tag{2.1}$$

The system is assumed to start empty, that is, $X^n(0) = 0$ for all n. Simple relations between the processes are

$$D^n_{ik}(t) = S^n_{ik}(T^n_{ik}(t)), \tag{2.2}$$

$$X^n_i(t) = A^n_i(t) - \sum_k D^n_{ik}(t), \tag{2.3}$$

$$I^n_k(t) = t - \sum_i T^n_{ik}(t), \tag{2.4}$$

$$\text{the sample paths of } X^n_i \text{ are nonnegative, and those of } I^n_k \text{ are in } C^{0,+}_{\mathbb{R}}[0,\infty). \tag{2.5}$$

The tuple $(\check{A}, \check{S})$ is referred to as the *stochastic primitives*. In our formulation we will consider T^n as the control process (equivalently, the allocation process $\varXi^n$ may be regarded the control). In view of equations (2.2), (2.3), (2.4), given the stochastic primitives, the control uniquely determines the processes D^n, X^n, I^n. Let an additional process be defined on the probability space denoted by $\varUpsilon = (\varUpsilon(l), l \in \mathbb{N})$, taking values in a Polish space $\mathcal{S}_{\rm rand}$ and assumed to be independent of the stochastic primitives, for each n (there is no need to let $\varUpsilon$ vary with n, as the primitives are all defined on the same probability space). It is included in the model in order to allow the construction of randomized controls; for more details about its potential use see [1, Remark 2.1.ii].

The process T^n is said to be an *admissible control for the queueing control problem (QCP) for the n-th system* if for each (i,k), T^n_{ik} has sample paths in $C^{0,+}_{\mathbb{R}}[0,\infty)$ that are 1-Lipschitz, and the associated processes D^n, X^n and I^n given by (2.2), (2.3) and (2.4) satisfy (2.5); furthermore, T^n is adapted to the filtration $\{\mathcal{F}^n_t\}$ defined by $\mathcal{F}^n_t = \sigma\{(A^n(s), D^n(s), s \in [0,t]), \varUpsilon\}$. Denote by $\mathcal{A}^n$ the collection of all admissible controls for the QCP for the n-th system. As argued in [1, Remark 2.1.i], this definition allows for the control to depend on the history of all processes involved in the model (in addition to the auxiliary randomness $\varUpsilon$).

The queue length process normalized at the diffusion scale is defined by $\hat{X}^n_i(t) = n^{-1/2}X^n_i(t)$. The cost of interest for the n-th system is given by

$$\hat{J}^n(T^n) = \mathbb{E}\int_0^\infty e^{-\gamma t} h(\hat{X}^n(t))dt, \qquad T^n \in \mathcal{A}^n, \tag{2.6}$$

where $\gamma > 0$ and $h(x) = h_1x_1 + h_2x_2$, with constants $h_1, h_2 > 0$, and $\hat{X}^n$ is the rescaled queue length process associated with the admissible control T^n. The value for the n-th system is defined by

$$\hat{V}^n = \inf\{\hat{J}^n(T^n) : T^n \in \mathcal{A}^n\}.$$

This completes the description of the queueing models and QCP. The complete set of problem data consists of the stochastic primitives mentioned above and the collection of parameters

$$(\lambda_i), (\mu_{ik}), (\hat{\lambda}_i), (\hat{\mu}_{ik}), (C_{A_i}), (C_{S_{ik}}), \gamma, (h_i).$$

We sometimes refer to (λ_i), (μ_{ik}) as the first order data and to $(\hat{\lambda}_i)$, $(\hat{\mu}_{ik})$, (C_{A_i}), $(C_{S_{ik}})$ as the second order data.

2.2 The linear program and extended heavy traffic condition

Given the first order data $\lambda_i > 0$, $\mu_{ik} > 0$, $i, k = 1, 2$, consider the following linear program (LP) for the unknowns $(\xi_{ik}) \in \mathbb{R}_+^{2\times 2}$ and $\rho \in \mathbb{R}$.

Linear Program. Minimize ρ subject to

$$\begin{cases} \displaystyle\sum_{k=1}^{2} \xi_{ik}\mu_{ik} = \lambda_i & i = 1, 2, \\ \displaystyle\sum_{i=1}^{2} \xi_{ik} \leqslant \rho & k = 1, 2, \\ \xi_{ik} \geqslant 0 & i, k = 1, 2. \end{cases} \tag{2.7}$$

Denote the optimal objective value of (2.7) by ρ^*.

Extended heavy traffic condition. $\rho^* = 1$.

The extended heavy traffic condition (EHTC) is broader than the *heavy traffic condition* that has been extensively used in the literature, which requires, in addition to $\rho^* = 1$, that there be a unique corresponding ξ.

Under the EHTC, any solution is of the form $(\xi, 1)$. Let $\mathcal{S}_{\rm LP}$ denote the subset of $\mathbb{R}^{2\times 2}$ for which the set of all solutions is given by $\mathcal{S}_{\rm LP} \times \{1\}$. We say that the *EHTC with multiplicity* (EHTCM) holds if the EHTC holds and the LP has multiple solutions (that is, there exist two distinct pairs $(\xi^{(1)}, 1)$ and $(\xi^{(2)}, 1)$ satisfying (2.7)). Following [1], we say that the service rates μ_{ik} are *decomposable* if $\mu_{ik} = \alpha_i\beta_k$ for all i, k, for some constants α_i and β_k.

A matrix $\xi \in \mathbb{R}_+^{2\times 2}$ is called *column-stochastic* if $\sum_i \xi_{ik} = 1$ for both $k = 1, 2$. A column-stochastic matrix is called a *mode* if (at least) one of its columns is either $(0, 1)^T$ or $(1, 0)^T$. A mode is said to be *degenerate* if it has more than one zero entry; otherwise it is said to be *nondegenerate*. A pair of nondegenerate modes is said to be a *class-switched* (*server-switched*) pair of modes if the zero entries in the two modes are in distinct rows but the same column (respectively, distinct columns but the same row). For example, the graphs in Figure 1(a) and (b) correspond to a class-switched pair of modes, whereas those in Figure 1(a) and (c) correspond to a server-switched pair.

The following condition will be referred to as the *nondegeneracy condition*, namely

$$\lambda_i \neq \mu_{ik} \text{ for all } (i, k) \in \{1, 2\}^2. \tag{2.8}$$

The following is proved in §3.

Lemma 2.1. *Let the EHTC hold.*

1. *For any solution $(\xi, 1)$, ξ is column-stochastic.*
2. *The LP (2.7) has multiple solutions if and only if (μ_{ik}) are decomposable.*
3. *If the LP has multiple solutions then there exists a pair of modes $(\xi^{*,1}, \xi^{*,2})$ such that*

$$\mathcal{S}_{\rm LP} = {\rm ch}(\{\xi^{*,1}, \xi^{*,2}\}). \tag{2.9}$$

4. *If the LP has multiple solutions and the nondegeneracy condition* (2.8) *holds then both* $\xi^{*,1}$ *and* $\xi^{*,2}$ *of* (2.9) *are nondegenerate. Moreover, they form either a class-switched or a server-switched pair.*

The main result will be proved under the following.

Assumption 2.2. 1. *The EHTCM holds.*

2. *The nondegeneracy condition* (2.8) *holds.*

Note that the case where the EHTC holds but EHTCM does not hold is already covered in the work [3], although, as mentioned in the introduction, under different assumptions on preemption and moment conditions. (See Corollary 2.15 and Remark 2.16 for implications of our results to the case where uniqueness holds, and more on the relation to [3] in that case.)

In view of Lemma 2.1(2), under Assumption 2.2, the rates (μ_{ik}) are decomposable. Thus $\mu_{ik} = \alpha_i\beta_k$, and clearly there is a degree of freedom in choosing (α_i) and (β_k). In this paper we will always assume that they are chosen so that $\sum_k \beta_k = 1$, and it is easy to see that, given (μ_{ik}), this normalization uniquely determines these parameters.

It is also guaranteed by the lemma that, under Assumption 2.2, the extreme points of $\mathcal{S}_{\rm LP}$ are two nondegenerate modes $\xi^{*,1}, \xi^{*,2}$ forming a class- or a server-switched pair. Once a labeling of these modes has been fixed, we will sometimes slightly abuse the terminology by referring to them as modes 1 and 2 rather than modes $\xi^{*,1}$ and $\xi^{*,2}$.

In earlier work on PSS, under the assumption that the LP has a unique solution $(\xi^*, 1)$, activities are categorized as *basic* or *nonbasic* according to the positivity of the fraction allocated to them by ξ^*, that is, an activity (i, k) is *basic* if $\xi^*_{ik} > 0$ and *nonbasic* if $\xi^*_{ik} = 0$. We extend this terminology to the case of multiple solutions as follows. For $m \in \{1, 2\}$, an activity (i, k) is said to be *basic in mode* m if the allocation associated to it by this mode does not vanish, namely $\xi^{*,m}_{ik} > 0$. If $\xi^{*,m}_{ik} = 0$ (respectively, $\xi^{*,m}_{ik} = 1$) it is said to be *non-basic (respectively, full) in mode* m.

Figure 1 demonstrates the second part of Lemma 2.1(4), namely that a pair of modes can be class-switched, as in Figure 1(a) and (b), or server-switched, as in Figure 1(a) and (c), but the LP does not give rise to a pair such as Figure 1(b) and (c). The terms class-switched and server-switched will sometimes be abbreviated as **CS** and **SS**.

A mode is said to be in *canonical form* if its first column is $(1, 0)^T$. It is clear by the definition of a mode that it is always possible to relabel the classes and the servers so that a given mode is in canonical form, and that if the mode is nondegenerate there is only one such relabeling. The graph of a mode in canonical form is shown in Figure 1(a). Because of its resemblance to the symbol N, this form is sometimes called an *N-system.*

If the EHTCM holds but (2.8) does not, that is, there exist i, k such that $\lambda_i = \mu_{ik}$, the situation is different: at least one of the modes will be degenerate, i.e., have two non-basic activities (see Lemma 3.1 below). In the terminology of linear programming this corresponds to a case where one of the basic solutions of the LP (2.7) is degenerate (cf. [7, Definition 3.1]). The degenerate case is not covered in this paper. The key difficulty in the degenerate case is that, in at least one of the modes, the servers do not communicate in the graph of basic activities, so the pooling required to reach the cost lower bound is not possible. (See [1, §2.4] for a more detailed discussion of this issue.)

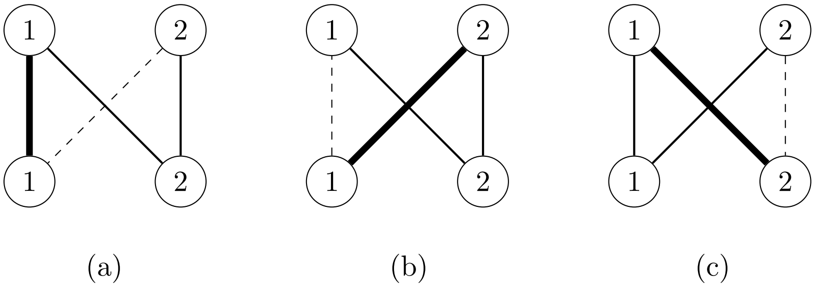

Figure 1: *The full and non-basic activities are shown in thick and dashed lines, respectively. Graph (a) corresponds to a mode in canonical form. Graphs (a,b) correspond to a pair of modes where the non-basic activity switches a class, whereas in (a,c) it switches a server. The pair (b,c), in which the non-basic activity switches both a class and a server is neither class- nor server-switched.*

Under Assumption 2.2, both modes have a single non-basic activity. Then, given a mode, the graph of basic activities has exactly three edges, and one can speak of the single-activity class (the one associated with only one nonbasic activity), the dual-activity class, and similarly, the single- and dual-activity server. These terms allow us to refer to the roles of classes and servers in the graph without considering a particular labeling. It is also useful to accompany these terms with matching notation. For a mode ξ, let the single- (respectively, dual-) activity class be denoted by $i_1(\xi)$ (respectively, $i_2(\xi)$), and similarly, let the single- (respectively, dual-) activity server be denoted by $k_1(\xi)$ (respectively, $k_2(\xi)$).

The following example, which corresponds to Examples (A) and (B) in [1], should help to make the above ideas more concrete.

Example 2.3. *Let (μ_{ik}) be given by*

$$\mu_{11} = 3, \qquad \mu_{12} = 4, \qquad \mu_{21} = 6, \qquad \mu_{22} = 8.$$

Then μ_{ik} are decomposable as $\alpha_i\beta_k$, where $\alpha_1 = 7, \alpha_2 = 14, \beta_1 = 3/7$, and $\beta_2 = 4/7$. For (λ_i), consider two cases, namely

$$\text{(A)} \quad \lambda_1 = 5,\ \lambda_2 = 4, \qquad \text{(B)} \quad \lambda_1 = 3.5,\ \lambda_2 = 7.$$

The linear program takes the form: Minimize ρ subject to

$$\begin{cases} 3\xi_{11} + 4\xi_{12} = \lambda_1, \\ 6\xi_{21} + 8\xi_{22} = \lambda_2, \\ \xi_{11} + \xi_{21} \le \rho, \\ \xi_{12} + \xi_{22} \le \rho, \\ \min \xi_{ik} \ge 0. \end{cases} \tag{2.10}$$

The solutions to the LP were calculated in [1], and it was found that in both cases $\rho^* = 1$, and the two modes are given by

$$\xi^{*,1} = (1, \tfrac{1}{2}, 0, \tfrac{1}{2})^T, \qquad \xi^{*,2} = (\tfrac{1}{3}, 1, \tfrac{2}{3}, 0)^T$$

in Case (A) and

$$\xi^{*,1} = (1, \tfrac{1}{8}, 0, \tfrac{7}{8})^T, \qquad \xi^{*,2} = (0, \tfrac{7}{8}, 1, \tfrac{1}{8})^T$$

in Case (B). In particular, the system is critically loaded and there are multiple ways to allocate the effort so as to meet the demand. That is, the EHTCM holds. Figures 2 and 3 depict the graphs of basic activities corresponding to these modes. The examples differ in the way the two modes are paired. In Case (A) (resp., (B)), the non-basic activity switches a server (a class). This distinction is of crucial significance to the way AO policies are designed in this paper.

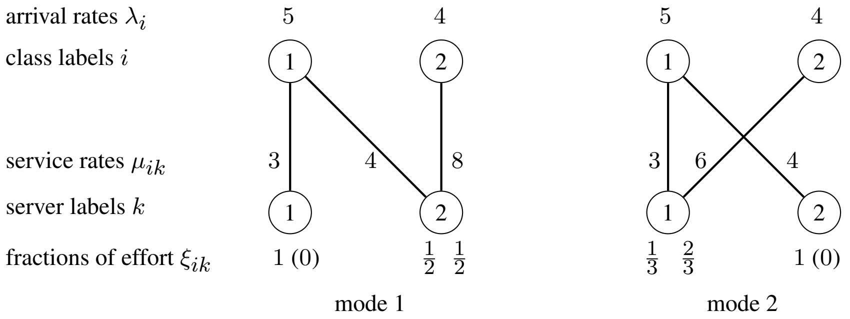
arrival rates λ_i
class labels i
service rates μ_{ik}
server labels k
fractions of effort ξ_{ik}
5
4
1
2
3
4
8
1
2
1 (0)
$\frac{1}{2}$ $\frac{1}{2}$
mode 1
5
4
1
2
3
6
4
1
2
$\frac{1}{3}$ $\frac{2}{3}$
1 (0)
mode 2

Figure 2: *The two modes in Case (A).*

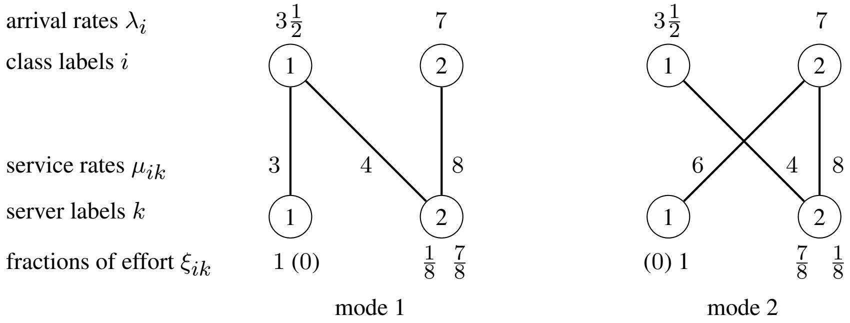

Figure 3: *The two modes in Case (B).*

2.3 Workload control problem

The WCP was derived and studied in [1] under the EHTC. We describe this problem in the special case needed here, namely under the setting of a 2×2 PSS and assuming that the EHTCM holds. In particular, as mentioned above, the parameters (α_i), (β_k) are uniquely determined by the problem data. Define the workload process and its scaled version as

$$W^n(t) = \sum_i \frac{X_i^n(t)}{\alpha_i}, \qquad \hat{W}^n(t) = \sum_i \frac{\hat{X}_i^n(t)}{\alpha_i}. \tag{2.11}$$

(It follows from [1, Lemma 2.4(2)] that (2.11) above agrees with the definition of W^n and $\hat{W}^n$ given in [1]). Let the process that appears in the definition of the cost (2.6) be denoted by

$$\hat{H}_t^n = h(\hat{X}^n(t)) = h_1 \hat{X}_1^n(t) + h_2 \hat{X}_2^n(t). \tag{2.12}$$

Throughout, denote by $p, q \in \{1,2\}$ the two distinct indices for which

$$h_p\alpha_p \geq h_q\alpha_q, \tag{2.13}$$

where in the special case $h_1\alpha_1 = h_2\alpha_2$, set $p = 1$ and $q = 2$. The policies constructed in this paper aim at keeping $\hat{X}^n_p$ close to zero. Hence we call p the *high priority class* (HPC) and q the *low priority class* (LPC). Next, let $\sigma_{A,i} = \lambda_i^{1/2}C_{A_i}$, $\sigma_{S,ik} = \mu_{ik}^{1/2}C_{S_{ik}}$, and

$$b(\xi) = \sum_i \frac{\hat{\lambda}_i - \sum_k \hat{\mu}_{ik}\xi_{ik}}{\alpha_i}, \qquad \sigma(\xi)^2 = \sum_i \frac{\sigma^2_{A,i} + \sum_k \sigma^2_{S,ik}\xi_{ik}}{\alpha_i^2}, \qquad \xi = (\xi_{ik}) \in \mathcal{S}_{\mathrm{LP}}. \tag{2.14}$$

Let also

$$b_m = b(\xi^{*,m}), \qquad \sigma_m = \sigma(\xi^{*,m}), \qquad m = 1, 2. \tag{2.15}$$

It was shown in [1] that the asymptotics of the pair $(\hat{W}^n, \varXi^n)$ are governed by a state-control pair of processes $(Z, \varXi)$, where Z is a one-dimensional controlled diffusion given by

$$Z_t = z + \int_0^t b(\varXi_s)ds + \int_0^t \sigma(\varXi_s)dB_s + L_t, \tag{2.16}$$

$\varXi$ is a control process, B is a standard BM (SBM), L is a reflection term at zero, and $z \geqslant 0$. A precise definition is as follows.

A tuple $\mathfrak{S} = (\varOmega', \mathcal{F}', (\mathcal{F}'_t), \mathbb{P}', B, \varXi, Z, L)$ is said to be an *admissible control system for the WCP with initial condition* z if $(\varOmega', \mathcal{F}', (\mathcal{F}'_t), \mathbb{P}')$ is a filtered probability space, B, $\varXi$, Z and L are processes defined on it, B is a SBM and an $(\mathcal{F}'_t)$-martingale, $\varXi$ is $(\mathcal{F}'_t)$-progressively measurable taking values in $\mathbb{R}^{2\times 2}$ and satisfying $\mathbb{P}'$(for a.e. t, $\varXi_t \in \mathcal{S}_{\mathrm{LP}}) = 1$, Z is continuous nonnegative and $(\mathcal{F}'_t)$-adapted, L has sample paths in $C^{0,+}_{\mathbb{R}}[0,\infty)$ and is $(\mathcal{F}'_t)$-adapted, and equation (2.16) and the identity $\int_{[0,\infty)} Z_t dL_t = 0$ are satisfied $\mathbb{P}'$-a.s.

Denoting by $\mathcal{A}_{\mathrm{WCP}}(z)$ the collection of all control systems for the WCP with initial condition z, the cost and value are defined as

$$J_{\mathrm{WCP}}(z, \mathfrak{S}) = \mathbb{E}_{\mathfrak{S}} \int_0^\infty e^{-\gamma t} Z_t dt, \qquad V_{\mathrm{WCP}}(z) = \inf\{J_{\mathrm{WCP}}(z, \mathfrak{S}) : \mathfrak{S} \in \mathcal{A}_{\mathrm{WCP}}(z)\}. \tag{2.17}$$

The significance of the WCP lies in the fact that it provides a lower bound on the large n asymptotics of the n-th system value. More precisely, the main result of [1], when specialized to the 2 × 2 PSS, states the following.

Theorem 2.4 ([1]). *Let Assumption 2.2 (multiplicity and nondegeneracy) hold. Then*

$$\liminf_n \hat{V}^n \geqslant V_0 := h_q\alpha_q V_{\mathrm{WCP}}(0).$$

To prove this result one only needs to verify that the assumptions from [1] hold under Assumption 2.2 of this paper. This is done in §3.

Remark 2.5. *The general result from [1] does not assume multiplicity and, moreover, uses different notation based on the formulation of a dual to the LP. In the 2 × 2 setting with multiplicity, the lower bound from [1] reduces to the above expression given in terms of the parameters* (α_i) *in place of the dual.*

In view of Theorem 2.4, a sequence $T^n \in \mathcal{A}^n$ of admissible controls for the QCP, also referred to as a *sequence of policies*, is said to be *asymptotically optimal* (AO) if

$$\limsup_{n\to\infty} \hat{J}^n(T^n) \leqslant V_0.$$

A heuristic discussion of the WCP and Sheng's toroise–hare problem. The WCP is important not only because it provides a lower bound on performance, but also because one can learn from it how to construct a policy for the queueing model that performs near optimality. This problem has been solved completely. Before presenting its solution we discuss its nature informally. Assume first that the pairs (b_1, σ_1) and (b_2, σ_2) are such that $b_1 < b_2$ while $\sigma_1 = \sigma_2$. Then it is intuitively clear and can be shown by a simple coupling that $J_{\rm WCP}$ is minimized by always using mode 1, because it has a smaller drift. Next consider the case where $b_1 = b_2$ but $\sigma_1 < \sigma_2$. Mode 2 has greater variance and thus one expects that the constraining mechanism, causing the diffusion to bounce back from the boundary, will be more active under mode 2 than under mode 1. Hence again using mode 1 at all times is optimal. More generally, mode 1 is optimal when $b_1 \le b_2$ and $\sigma_1 \le \sigma_2$.

A more interesting case is when $b_1 < b_2$ but $\sigma_1 > \sigma_2$, referred to in [16] as the *tortoise–hare problem.* It seems reasonable to use the mode with smaller drift when the diffusion process is far from the origin, and switch to the mode with smaller variance when it is close to the origin, where the aforementioned boundary effect is more prominent.

These heuristic arguments were validated rigorously in [16]. In particular, in the case $b_1 < b_2$, $\sigma_1 > \sigma_2$, it was shown that there exist a point $z^* \in (0, \infty)$ such that it is optimal to select mode 2 (resp., 1) when Z_t is in $[0, z^*)$ (resp., $[z^*, \infty)$). The identification of z^* and the proof of the result were based on an HJB equation. The precise details are as follows.

The HJB equation. The value function can be characterized in terms of an HJB equation (see [6] for an introduction to the subject). To present this equation, for $(v_1, v_2, \xi) \in \mathbb{R}^2 \times \mathcal{S}_{\rm LP}$, let

$$\bar{\mathbb{H}}(v_1, v_2, \xi) = b(\xi)v_1 + \frac{\sigma(\xi)^2}{2}v_2, \qquad \mathbb{H}(v_1, v_2) = \inf_{\xi\in\mathcal{S}_{\rm LP}} \bar{\mathbb{H}}(v_1, v_2, \xi) = \min_{\xi\in\{\xi^{*,1},\xi^{*,2}\}} \bar{\mathbb{H}}(v_1, v_2, \xi),$$

where the identity on the RHS follows from the fact that both b and σ^2 are affine as a function of ξ, and $\mathcal{S}_{\rm LP}$ is the convex hull of $\{\xi^{*,1}, \xi^{*,2}\}$. A classical solution to the HJB equation is a $C^2(\mathbb{R}_+ : \mathbb{R})$ function u satisfying

$$\mathbb{H}(u'(z), u''(z)) + z - \gamma u(z) = 0, \qquad z \in (0, \infty), \tag{2.18}$$

and the boundary conditions at 0 and ∞,

$$u'(0) = 0, \qquad u(z) < c(1 + z), z \in \mathbb{R}_+, \text{ for some constant } c. \tag{2.19}$$

Given a C^2 function u, denote $\mathbb{H}^u_m(z) = \bar{\mathbb{H}}(u'(z), u''(z), \xi^{*,m})$, $m = 1, 2$, and $\mathbb{H}^u(z) = \min_m \mathbb{H}^u_m(z)$.

The following two conditions play an important role in what follows. They correspond to the two cases discussed above, and will be referred to as the *single mode* case and, respectively, the *dual mode* case (not to be confused with uniqueness and multiplicity of the LP solution):

$$\text{there exist distinct } m, m' \in \{1, 2\} \text{ such that } b_m \leqslant b_{m'} \text{ and } \sigma_m \leqslant \sigma_{m'}, \tag{2.20}$$

$$\text{there exist distinct } m, m' \in \{1,2\} \text{ such that } b_m < b_{m'} \text{ and } \sigma_m > \sigma_{m'}. \tag{2.21}$$

The C^2 smoothness of the value function is tied to the question of existence of a classical solution to the HJB equation. Owing to the uniform ellipticity ($\sigma(\xi)^2 > 0$ at both $\xi = \xi^{*,1}$ and $\xi^{*,2}$), one can show that a classical solution uniquely exists [1, Proposition 2.5], a type of result that, in the general context of optimal switching of a diffusion process, has been called *the principle of smooth fit.* The results of [16] alluded to above shed more light on the specific problem at hand, providing structural properties (parts 2 and 3 of the result below) that are harder to obtain via the general approach.

Proposition 2.6. 1. *There exists a unique classical solution u to* (2.18)–(2.19). *Moreover, $u = V_{\mathrm{WCP}}$.*

2. *If* (2.20) *holds then, with m as in* (2.20),

$$\mathbb{H}^u(z) = \mathbb{H}^u_m(z) \quad z \in \mathbb{R}_+. \tag{2.22}$$

3. *Alternatively, if* (2.21) *holds then there exists $z^* \in (0,\infty)$ such that, with the pair (m,m') of* (2.21),

$$\mathbb{H}^u(z) = \begin{cases} \mathbb{H}^u_{m'}(z) & z < z^*, \\ \mathbb{H}^u_m(z) & z > z^*. \end{cases} \tag{2.23}$$

The proof of this result appears in §4. Some details on the construction from [16] (as corrected in [17]), by which parts 2 and 3 above were proved, appear in Appendix B.

Further terminology is as follows. In the single mode case, the mode $\xi^{*,m}$ for which m satisfies (2.20) will be referred to as the *active mode* and denoted ξ^A, because the above result indicates that it is optimal to always select $\varXi_t = \xi^{*,m}$. In the dual mode case, the modes $\xi^{*,m'}$ and $\xi^{*,m}$ for which m' and m satisfy (2.21) will be referred to as ξ^L, the *lower* and, respectively, ξ^H the *higher workload mode*, and z^* as the *switching point.* These terms refer to the fact that the result suggests that it is optimal to select $\varXi_t = \xi^L$ (respectively, $\varXi_t = \xi^H$) when $Z_t < z^*$ (respectively, $Z_t > z^*$).

Example 2.7. *We go back to Example 2.3, focusing on Case (A), adding now information on the second order data. Assume that the squared coefficients of variation $C^2_{A_i}$ and $C^2_{S_{ik}}$ are all* 1 *except $C^2_{S_{11}} = 4$. Set both $\hat\lambda_i$ to* 0. *As for $\hat\mu_{ik}$, consider two cases. In Case (A1), all $\hat\mu_{ik}$ are set to* 0. *In Case (A2), they are all* 0 *except $\hat\mu_{11} = 1$. For each of the modes, computing the drift and squared diffusion coefficients via* (2.14)–(2.15) *gives, in Case (A1),*

$$b_1 = 0, \quad \sigma_1^2 = \frac{3}{7}, \quad b_2 = 0, \quad \sigma_2^2 = \frac{15}{49}. \tag{2.24}$$

In Case (A2),

$$b_1 = -\frac{1}{7}, \quad \sigma_1^2 = \frac{3}{7}, \quad b_2 = -\frac{1}{21}, \quad \sigma_2^2 = \frac{15}{49}. \tag{2.25}$$

Note that (A1) is a single mode case whereas (A2) is a dual mode case. Roughly speaking, we may infer that, in the former case, in order to perform near optimality, it is necessary to keep the proportions of the work allocated to the different activities close to the fractions given by ξ^A. That is, the policies should be designed to achieve $\varXi^n \Rightarrow \xi^A$. As for Case (A2), recall that Z

approximates $\hat{W}^n$. Hence in this case it is necessary for the work allocation to vary over time in such a way that the aforementioned proportions are close to ξ^L when $\hat{W}^n(t) < z^*$ and to ξ^H when $\hat{W}^n(t) > z^*$. This is again only a rough statement; a precise formulation of this behavior appears next.

The controlled diffusion. Controlling (2.16) according to the above description results in two different diffusion processes. In the single mode case, the optimally controlled process Z is given by

$$Z_t^{(1)} = z + b^A t + \sigma^A B_t + L_t^{(1)}, \tag{2.26}$$

with $b^A = b(\xi^A)$ and $\sigma^A = \sigma(\xi^A)$, which is nothing but a reflecting BM with drift b^A and diffusivity σ^A. In the dual mode case, consider the SDE

$$Z_t^{(2)} = z + \int_0^t b^*(Z_s^{(2)})ds + \int_0^t \sigma^*(Z_s^{(2)})dB_s + L_t^{(2)}, \tag{2.27}$$

where, throughout, we denote

$$b^* = b \circ \varphi^*, \qquad \sigma^* = \sigma \circ \varphi^*, \qquad \varphi^*(z) = \xi^L \mathbb{1}_{[0,z^*]}(z) + \xi^H \mathbb{1}_{(z^*,\infty)}(z), \qquad z \in \mathbb{R}_+. \tag{2.28}$$

For this equation, weak existence and uniqueness of solutions hold, as we shall argue in Lemma 4.1. As a result, there exists a control system for the WCP that behaves exactly as described above, with $\mathit{\Xi}_t = \xi^L$ (respectively, $\mathit{\Xi}^H$) when $Z_t^{(2)} \le z^*$ ($> z^*$), and moreover, this description uniquely determines the law of the process $Z^{(2)}$.

As for the asymptotics of the QCP, the preceding discussion, and the fact that the system starts empty, suggest that in order to achieve the lower bound, the convergence

$$(\hat{X}_p^n, \hat{W}^n) \Rightarrow (0, Z^{(1)}), \qquad \text{with } z = 0, \tag{2.29}$$

should hold in the single mode case, and

$$(\hat{X}_p^n, \hat{W}^n) \Rightarrow (0, Z^{(2)}), \qquad \text{with } z = 0, \tag{2.30}$$

in the dual mode case, where $Z^{(1)}$ is given by (2.26), $(Z^{(2)}, \mathit{\Xi}, L, B)$ is a weak solution to (2.27), and $z = 0$.

2.4 Asymptotic optimality results

This section is devoted to the description of several policies that are shown to be AO under different conditions. We have already assumed that the interarrival times of the primitive processes possess finite second moments. Our main results require a stronger assumption.

Assumption 2.8. *There exists* $\mathbf{m} > 2$ *such that*

$$\max_{i,k} \mathbb{E}[\check{a}_i(1)^{\mathbf{m}}] \vee \mathbb{E}[\check{u}_{ik}(1)^{\mathbf{m}}] < \infty.$$

Whereas the assumption $\mathbf{m} > 2$ is required for all our results, some of them will require yet a stronger moment assumption, namely $\mathbf{m} > \mathbf{m}_0$, where, throughout, we denote

$$\mathbf{m}_0 = \frac{1}{2}(5 + \sqrt{17}).$$

Different policies are proposed in different cases. The distinction between the various cases is based on whether the single-mode condition (2.20) or the dual-mode condition (2.21) holds, and further, for each of the relevant modes (ξ^A in the former case and both ξ^L and ξ^H in the latter), whether the HPC is the single- or dual-activity class.

As a rule, all policies we describe are non-preemptive, that is, the processing of a job is not interrupted once started. A job is said to be *in the queue* if it is waiting to be served, whereas it is *in the system* if it is either in the queue or being processed. (In what is a bit of an abuse of terminology we use the term *queue length* to refer to the number in the system.) A server is said to be *available* at a time t if either it has just completed a job or has already been idle at that time.

Some of the policies to be described are defined in terms of a sequence of thresholds, Θ^n, put on the queue length at one of the two buffers. Under Assumption 2.8, $\mathbf{m} > 2$. Fix $\bar{a}$ satisfying

$$\frac{1}{2} - \bar{\zeta}(\mathbf{m}) < \bar{a} < \frac{1}{2} \qquad \text{where} \qquad \bar{\zeta}(\mathbf{m}) = \begin{cases} \frac{\mathbf{m}-2}{4\mathbf{m}}, & \mathbf{m} \in (2, \mathbf{m}_0], \\ \frac{\mathbf{m}-2}{4\mathbf{m}} \wedge \frac{\mathbf{m}^2-5\mathbf{m}+2}{2\mathbf{m}(3\mathbf{m}-2)} = \frac{\mathbf{m}^2-5\mathbf{m}+2}{2\mathbf{m}(3\mathbf{m}-2)}, & \mathbf{m} \in (\mathbf{m}_0, \infty). \end{cases} \tag{2.31}$$

Set the sequence of threshold levels Θ^n and their normalized version $\hat{\Theta}^n$ to

$$\Theta^n = \lceil n^{\bar{a}} \rceil, \qquad \hat{\Theta}^n = n^{-1/2}\Theta^n. \tag{2.32}$$

Definition 2.9. *(Server dedicated to / prioritizes a class).*

1. *A server is said to be* dedicated *to class i at a given time if it acts as follows: if available at that time, it admits a job from class i provided there is one in the queue, or there is a new class-i arrival, but does not admit a job from the other class.*

2. *A server is said to* prioritize *class i at a given time if it acts as follows: if available at that time, it admits a job from class i provided there is one in the queue; otherwise it admits a job from the other class provided there is one in the queue. If the server is idle at that time and there is a new arrival of any class, it admits this arrival unless the other server is dedicated to that class and is free at that time (in which case the other server admits it).*

Definition 2.10. *(*$\mathbf{P}$*,* $\mathbf{T}_1$ *and* $\mathbf{T}_2$ *rules). Let a mode* ξ *be given. At any moment in time the single-activity server is dedicated to the dual-activity class.*

1. *The servers are said to obey the* priority rule*, abbreviated* $\mathbf{P}$ *rule, at a given time if the dual-activity server prioritizes the single-activity class at that time.*

2. *The servers are said to obey the* single-activity class threshold rule*, abbreviated* $\mathbf{T}_1$ *rule, at a given time if in the n-th system, the dual-activity server prioritizes the single-activity class when the queue length of the single-activity class equals or exceeds* Θ^n *at that time, and otherwise prioritizes the dual-activity class.*

3. *The servers are said to obey the* dual-activity class threshold rule*, abbreviated* $\mathbf{T}_2$ *rule, at a given time if in the* n*-th system, the dual-activity server prioritizes the dual-activity class when the queue length of the dual-activity class equals or exceeds* Θ^n *at that time, and otherwise prioritizes the single-activity class.*

Recall the notation ξ^A, ξ^L, ξ^H and p from §2.3. In the case of a single mode, the following two policies are proposed. (In Definitions 2.11 and 2.12 below, the text in square brackets is not a part of the definition, but serves to indicate when each policy is to be applied).

Definition 2.11. *(Single mode policies* $\mathbf{P}$ *and* $\mathbf{T}_2$*). Let the single mode condition* (2.20) *hold.*

1. *The* $\mathbf{P}$ *policy [to be applied when* $i_1(\xi^A) = p$*] is as follows: The servers obey the* $\mathbf{P}$ *rule corresponding to* ξ^A *at all times.*

2. *The* $\mathbf{T}_2$ *policy [to be applied when* $i_2(\xi^A) = p$*] is as follows: The servers obey the* $\mathbf{T}_2$ *rule corresponding to* ξ^A *at all times.*

In the dual mode case, servers switch between two rules, depending, roughly speaking, on whether the rescaled workload is below or above the switching point z^*. This makes the structure of the policies more complicated. In particular, there is potential loss of capacity in every switching, especially since our policies are nonpreemptive. Moreover, the number of switchings grows without bound, as the rescaled workload converges to a diffusion. Therefore one has to be careful about how to assure that the rule obeyed is updated soon enough after the workload level has crossed the switching point so as not to compromise optimality. The considerations differ in the different cases, and give rise to the use of the $\mathbf{T}_1$ rule as well as the sampling rules in Definition 2.12 below. Although it is possible that AO is not too sensitive to these fine details, proving that might be quite hard.

The precise definition requires the use of a state variable called *current mode*, that determines which rule is applicable. This variable is not updated continuously in time about whether $W^n > n^{1/2}z^*$ but only at certain sampling times, as detailed below. The four policies used in the dual mode case are as follows.

Definition 2.12. *(Dual mode policies* $\mathbf{PP}$*,* $\mathbf{T}_2\mathbf{T}_2$*,* $\mathbf{T}_1\mathbf{T}_2$ *and* $\mathbf{T}_2\mathbf{T}_1$*). Let the dual-mode condition* (2.21) *hold.*

1. *The* $\mathbf{PP}$ *policy [applied when* $i_1(\xi^L) = i_1(\xi^H) = p$*] is as follows.*

   (a) *The workload is sampled at each service completion of the single activity server. If the workload is below* $n^{1/2}z^*$*, the current mode is set to* $\xi = \xi^L$*; otherwise it is set to* $\xi = \xi^H$*.*

   (b) *The servers always obey the* $\mathbf{P}$ *rule w.r.t. the current mode* ξ*.*

2. *The* $\mathbf{T}_2\mathbf{T}_2$ *policy [applied when* $i_2(\xi^L) = i_2(\xi^H) = p$*] is as follows.*

   (a) *The workload is sampled at each service completion of the single activity server. If the workload is below* $n^{1/2}z^*$*, the current mode is set to* $\xi = \xi^L$*; otherwise it is set to* $\xi = \xi^H$*.*

   (b) *The servers always obey the* $\mathbf{T}_2$ *rule w.r.t. the current mode* ξ*.*

3. *The* $\mathbf{T}_1\mathbf{T}_2$ *policy [applied when* $i_1(\xi^L) = i_2(\xi^H) = p$*] is as follows.*

(a) *The workload is sampled at each arrival and service completion. If the workload is below $n^{1/2}z^*$, the current mode is set to $\xi = \xi^L$; otherwise it is set to $\xi = \xi^H$.*

(b) *Whenever $\xi = \xi^L$, the servers obey the $\mathbf{T}_1$ rule w.r.t. ξ; whenever $\xi = \xi^H$, they obey the $\mathbf{T}_2$ rule w.r.t. ξ.*

4. *The $\mathbf{T}_2\mathbf{T}_1$ policy [applied when $i_2(\xi^L) = i_1(\xi^H) = p$] is as $\mathbf{T}_1\mathbf{T}_2$, except that the roles of $\mathbf{T}_1$ and $\mathbf{T}_2$ are interchanged.*

Our main result states conditions under which each of the six policies just introduced are AO.

Theorem 2.13. *Let Assumptions 2.2 and 2.8 hold. In parts 1(b), 2(b) and 2(c) below, assume moreover that $\mathbf{m} > \mathbf{m}_0$.*

1. *Assume that the single-mode condition* (2.20) *holds.*

   (a) *If $i_1(\xi^A) = p$ then under the $\mathbf{P}$ policy* (2.29) *holds and this policy is AO.*

   (b) *If $i_2(\xi^A) = p$ then under the $\mathbf{T}_2$ policy* (2.29) *holds and this policy is AO.*

2. *Assume that the dual-mode condition* (2.21) *holds.*

   (a) *If $i_1(\xi^L) = i_1(\xi^H) = p$ then under the $\mathbf{PP}$ policy* (2.30) *holds and this policy is AO.*

   (b) *If $i_2(\xi^L) = i_2(\xi^H) = p$ then under the $\mathbf{T}_2\mathbf{T}_2$ policy* (2.30) *holds and this policy is AO.*

   (c) *If $i_1(\xi^L) = i_2(\xi^H) = p$ then under the $\mathbf{T}_1\mathbf{T}_2$ policy* (2.30) *holds and this policy is AO.*

   (d) *If $i_2(\xi^L) = i_1(\xi^H) = p$ then under the $\mathbf{T}_2\mathbf{T}_1$ policy* (2.30) *holds and this policy is AO.*

The proof of this result is in §5. Table 1 summarizes how to determine which of the above six case applies, based on the problem data. We discuss some of the fine details of our construction in the dual mode case in §5.1.3.

Remark 2.14. *In §5.1, Theorem 5.1 provides more detailed information than* (2.29)–(2.30) *on the weak limit of the processes involved.*

Although the main goal of this paper is to address the case where the LP has multiple solutions, the following result about the case where it has a unique solution is a consequence of our treatment.

Corollary 2.15. *Let the 'standard' HTC hold; that is, the EHTC with a unique LP solution $(\xi^*, 1)$. Let also Assumption 2.8 hold. In part (b) below, assume moreover that $\mathbf{m} > \mathbf{m}_0$.*

(a) *If $i_1(\xi^*) = p$ then under the $\mathbf{P}$ policy* (2.29) *holds and this policy is AO.*

(b) *If $i_2(\xi^*) = p$ then under the $\mathbf{T}_2$ policy* (2.29) *holds and this policy is AO.*

Remark 2.16. *This result is closely related to the main result of [3], which proved AO of a threshold policy under the standard HTC. In a remark in [3, page 622] it was conjectured that the threshold policy without preemption has the same behavior in the heavy traffic limit as the one with preemption. Corollary 2.15 confirms that AO can indeed be achieved by a non-preemptive threshold policy, although, strictly speaking, it does not prove the precise conjecture from [3] as our threshold sizes differ from those of [3].*

Identifying an AO policy assuming LP solution multiplicity

Input: the problem data, $(\lambda_i), (\mu_{ik}), (\hat{\lambda}_i), (\hat{\mu}_{ik}), (\sigma_i), (\sigma_{ik}), \gamma, (h_i)$. Output: an AO policy.

1. Find the two modes (i.e., extreme solutions) of the LP (2.7), $\xi^{*,m}$, $m = 1, 2$. In particular, since the LP is assumed to have multiple solutions, μ_{ik} are given as $\alpha_i \beta_k$, and the formulas in Lemma 3.1 express $\xi^{*,m}$, $m = 1, 2$ in terms of α_i and β_k.

2. Compute the drift–diffusion pairs (b_m, σ_m), $m = 1, 2$. For this, use formula (2.15).

3. Decide *single* or *dual* mode case, and find the active/lower/higher workload modes, as follows. For $(m, m') = (1, 2)$ or $(2, 1)$, if $b_m \le b_{m'}$ and $\sigma_m \le \sigma_{m'}$ then this is the single mode case, with m the active mode: $\xi^A = \xi^{*,m}$.
Otherwise, this is the dual mode case, and if $\sigma_m < \sigma_{m'}$ then m (resp., m') is the lower (higher) workload mode: $\xi^L = \xi^{*,m}$, $\xi^H = \xi^{*,m'}$.

4. In the dual mode case, compute the switching point z^*. This is done by numerically solving the HJB equation (2.18)–(2.19), or alternatively, equation (B.3).

5. Determine high priority class: $p = \arg\max_i h_i \alpha_i$.

6. For $\xi = \xi^A$ (single mode) or $\xi = \xi^L$ and ξ^H (dual mode), denote by $i_1(\xi)$ (resp., $i_2(\xi)$) the class that is assigned one (resp., two) server(s) under mode ξ.

7. Given the number of modes, the modes ξ^A or ξ^L and ξ^H, the indices p and $i_1(\xi)$, $i_2(\xi)$ for the relevant modes ξ, determine which of the six cases of Theorem 2.13 applies.

Table 1

Example 2.17. *We continue Cases (A1) and (A2) from Example 2.7, specifying now which part of the main result applies in each case. Recall from Example 2.3 that $\alpha_1 = 7$ and $\alpha_2 = 14$. Recall from Example 2.7 that Figure 2 depicts the two modes and the corresponding graphs, and that the drift-diffusion pairs are given by* (2.24) *and* (2.25)*. Case (A1) is a single mode because $b_1 = b_2$, and the active mode is $\xi^A = \xi^{*,2}$ because $\sigma_2 < \sigma_1$. As can be seen in Figure 2 (right), this mode has $i_1(\xi^A) = 2$. Now assume that the costs are $(h_1, h_2) = (1, 1)$. Then the class that maximizes $h_i\alpha_i$ is $p = 2$. Thus $i_1(\xi^A) = p$ and Theorem 2.13(1a) holds. If instead $(h_1, h_2) = (3, 1)$ then $p = 1$ and therefore Theorem 2.13(1b) holds.*

In Case (A2) we have $b_1 < b_2$ but $\sigma_2 < \sigma_1$, hence this is a dual mode case with $\xi^L = \xi^{*,2}$ and $\xi^H = \xi^{*,1}$. By Figure 2, $i_1(\xi^L) = 2$ and $i_1(\xi^H) = 2$. Again, if $(h_1, h_2) = (1, 1)$ then $p = 2$ and we see that Theorem 2.13(2a) applies, but if $(h_1, h_2) = (3, 1)$, $p = 1$ and Theorem 2.13(2b) applies.

2.5 Some numerical results

The dual mode policies that we introduce are admittedly complicated, and the proof of their asymptotic optimality is quite involved. A natural question thus arise here: Is all of this complexity worthwhile? What is the gain from it?

From a purely theoretical/mathematical context the answer is simple: If the dual mode policy has a cost that is strictly below that of both single mode controls then a non-switching policy cannot be asymptotically optimal. From a practical viewpoint this answer falls short. Implementing the dual mode policy is clearly more complicated than implementing a single mode policy. A key question thus arises: Is the gain from using a dual mode policy sufficient to overcome the effort involved in implementing it? The answer to this question is context dependent, and clearly depends on the the cost of the effort required to implement the dual mode control, which we do not include as part of our model. Thus we do not quite answer this question.

We do, however, partially answer the question of how much larger is the cost of single mode policies over the optimal switching policy. We do this in two numerical examples. All of the numerical results that we present here are obtained by finding the unique root of equation B.3, and then using B.1 and B.2. Note that to use B.3, B.1, and B.2, the identity of the 2 modes needs to be flipped around because in case (A2) we have $b_1 < b_2$, and the analysis in Appendix B requires $b_1 \geqslant b_2$. The mode identities in our presentation remains as in (A2).

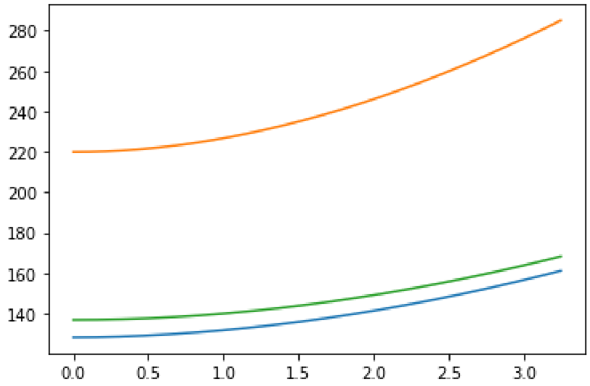

Figure 4: *Value functions in Case (A2).*

The first numerical results are for the case (A2) introduced in Example 2.7, with $\gamma = 0.01$. In particular, in Figure 4 we plot 3 curves: $V_1(z), V_2(z)$ and $V_{WCP}(z)$. Here V_i is the cost function for using mode $i, 1 = 1, 2$. (The orange/top curve corresponds to V_1, and the green/middle curve corresponds to V_2.) It is clear from Figure 4 that there is a gap between $V_{WCP}(z)$ and both $V_1(z)$ and $V_2(z)$. We define the 'gain' from using the optimal switching policy as

$$G := \frac{V_1(0) \wedge V_2(0)}{V_{WCP}(0)}.$$

Thus $G \geq 1$. For case (A2), $G = 1.07$. If the cost related to implementing the more complicated switching policy is great enough it may indeed be decided that $G = 1.07$ is not sufficient to overcome this cost.

This raises a more general question: Can we place an a priori upper bound on G? It is beyond the scope of this paper to answer this question in any precise mathematical manner, but we provide a set of numerical results *suggesting* that the answer may be no: Given any $M < \infty$, it is possible to find a set of parameters $\{(b_i, \sigma_i^2), i = 1, 2)\}$ such that $G > M$. We examine a set of parameters loosely related to case (A2). In particular, we fix $b_2 = -1/21$ and $\sigma_2^2 = 15/49$, which are the parameters in case (A2), throughout. We also use $\gamma = 0.01$. We then take 6 different values for b_1, and set σ_1^2 so that

$$\frac{\sigma_1^2}{|b_1|} = \frac{\sigma_2^2}{|b_2|}.$$

In particular, we take the b_1 values to be $\{-3/21, -6/21, -12/21, -24/21, -48/21, -96/21\}$. Table 2 contains the results of this numerical experiment. Note that $G = 1.26$ when $b_1 = -3/21$, and grows to $G = 5.01$ when $b_1 = -96/21$.

b_1	σ_1^2	z^*	G	V_{WCP}
$-3/21$	$45/49$	1.91	1.26	174
$-6/21$	$90/49$	1.49	1.56	141
$-12/21$	$180/49$	1.13	2.01	109
$-24/21$	$360/49$	0.84	2.67	82
$-48/21$	$720/49$	0.61	3.63	61
$-96/21$	$1440/49$	0.44	5.01	44

Table 2: Gain from using switching policy

In a very real sense the situation presented in Figure 4 is a 'lucky' one for a system controller who does not want to go through the trouble of implementing the switching policy, since the loss is only 7%. The results of Table 2 stand in contrast to that. To see this in graphical form, in Figure 5 we present the same plot as in Figure 4, but corresponding to the parameters in the 3^{rd} row of Table 2 . (In this case V_1 and V_2 have switched places: The green/top curve corresponds to V_2, and the orange/middle curve corresponds to V_1.)

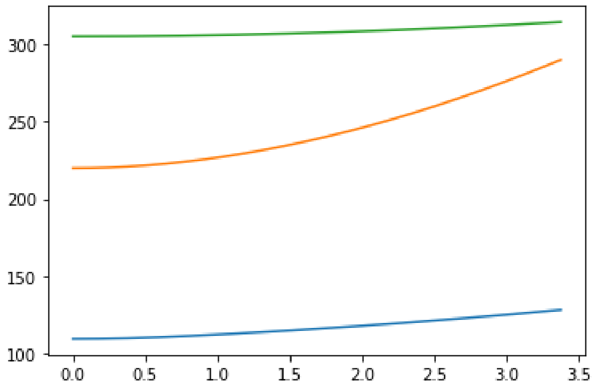

Figure 5: *Value functions for row 3 of Table 2.*

3 The LP under the EHTC

In this section we prove Lemma 2.1. In addition, in a sequence of three lemmas, we provide an explicit solution to the LP and a criterion for determining whether the two modes are class-switched or server-switched. Finally, Theorem 2.4 is proved. The section is structured as follows. In §3.1 we first prove Lemma 2.1(1–3) based mostly on results from [1]. Then we state Lemmas 3.1 and 3.2, which provide the LP solution, and Lemma 3.3 which is concerned with how the modes are paired. Lemma 2.1(4) is then proved based on Lemma 3.3. In §3.2 we prove Lemmas 3.1–3.3 and Theorem 2.4.

3.1 LP-related lemmas

Proof of Parts 1–3 of Lemma 2.1.

1. Let $\xi \in \mathcal{S}_{\rm LP}$. It is impossible that $\sum_i \xi_{ik} < 1$ for both $k = 1$ and 2 as this contradicts the EHTC $\rho^* = 1$. Assume then that, say, $\sum_i \xi_{i1} < 1$. Then $\xi_{11} < 1$. Define

$$\tilde{\xi} = \xi + \begin{pmatrix} \varepsilon & -c\varepsilon \\ 0 & 0 \end{pmatrix},$$

where $c = \mu_{12}^{-1}\mu_{11}$. Then for $\varepsilon > 0$ small, $\tilde{\xi}$ satisfies (2.7) with $\rho < 1$, which again contradicts the EHTC. This proves Part 1.

2. This follows from [1, Lemma 2.4(4)]. Note that uniqueness of the dual problem, which is a standing assumption in [1], is not used in the proof of this statement.

3. This statement follows from [1, Lemma 2.3(1)] and [1, Lemma 2.4(4)], where again the uniqueness of the dual is not used. □

The following lemma computes the two modes.

Lemma 3.1. *Let the EHTCM hold. Then*

$$\sum_i \frac{\lambda_i}{\alpha_i} = \sum_k \beta_k = 1, \tag{3.1}$$

where the last equality merely expresses the normalization convention mentioned earlier in §2.2. Moreover, any $\xi \in \mathcal{S}_{\rm LP}$ is determined by its entry ξ_{11} via

$$\xi = \begin{pmatrix} \xi_{11} & \dfrac{\lambda_1}{\alpha_1\beta_2} - \dfrac{\beta_1}{\beta_2}\xi_{11} \\ 1-\xi_{11} & 1 - \dfrac{\lambda_1}{\alpha_1\beta_2} + \dfrac{\beta_1}{\beta_2}\xi_{11} \end{pmatrix}. \tag{3.2}$$

The two modes $\xi^{*,1}$ and $\xi^{*,2}$ can be expressed by (3.2) *with ξ_{11} given by*

$$\xi^{*,1}_{11} = \max\Big(0, \frac{\lambda_1}{\alpha_1\beta_1} - \frac{\beta_2}{\beta_1}\Big), \qquad \xi^{*,2}_{11} = \min\Big(\frac{\lambda_1}{\alpha_1\beta_1}, 1\Big). \tag{3.3}$$

Recall that under the nondegeneracy condition, for any mode there is a unique relabelling of classes and servers which transforms it to a canonical form. The following result shows that both modes, once put in canonical form, are given by the same formula.

Lemma 3.2. *Let Assumption 2.2 (EHTCM and nondegeneracy) hold. Fix $m \in \{1,2\}$. Relabel classes and servers so that $\xi^{*,m}$ is in canonical form. Then $\lambda_1 > \alpha_1\beta_1$ and*

$$\xi^{*,m} = \begin{pmatrix} 1 & \dfrac{\lambda_1}{\alpha_1\beta_2} - \dfrac{\beta_1}{\beta_2} \\ 0 & \dfrac{\lambda_2}{\alpha_2\beta_2} \end{pmatrix}. \tag{3.4}$$

In particular, if $\xi, \xi' \in \mathcal{S}_{\rm LP}$ and there are i,k such that $\xi_{ik} = \xi'_{ik} = 0$ then $\xi = \xi'$.

Lemma 2.1(4), which is yet to be proved, states that under the nondegeneracy condition, the two modes must be either class- or server-switched. The following lemma contains this result, and in addition provides a criterion for distinguishing between these cases. We will say that the *class-switching condition* holds if

$$\max_i \frac{\lambda_i}{\alpha_i} < \max_k \beta_k, \tag{3.5}$$

and the *server-switching condition* holds if

$$\max_i \frac{\lambda_i}{\alpha_i} > \max_k \beta_k. \tag{3.6}$$

Lemma 3.3. *Let Assumption 2.2 hold. Then both modes are nondegenerate. Moreover, under the class-switching condition* (3.5)*, the modes are class-switched (as, for example, in Figure 1(a) and (b)), and under the server-switching condition* (3.6)*, they are server-switched (as, for example, in Figure 1(a) and (c)).*

Proof of Part 4 of Lemma 2.1. The statement is contained in Lemma 3.3. □

Remark 3.4. *Note that cases 2(c) and 2(d) of Theorem 2.13 correspond to class-switched modes (for example, under case 2(c) one has $i_1(\xi^L) = i_2(\xi^H) = p$ hence the non-basic activity must have switched from class p to class q when moving from ξ^L to ξ^H). By Lemma 3.3, this occurs under* (3.5)*. Also note that in both cases, the proposed policies apply a different rule for the lower and upper workload modes. On the other hand, cases 2(a) and 2(b) of Theorem 2.13 correspond to server-switching, and hold under* (3.6)*, and our policies are such that the same rule is used for the lower and upper workload modes.*

3.2 Proof of Lemmas 3.1–3.3 and Theorem 2.4

Proof of Lemma 3.1. In view of Lemma 2.1(1), every solution ξ is column-stochastic, and, recalling $\mu_{ik} = \alpha_i\beta_k$, must satisfy

$$\begin{aligned}
&\xi_{11}\beta_1 + \xi_{12}\beta_2 = \frac{\lambda_1}{\alpha_1}, \\
&\xi_{21}\beta_1 + \xi_{22}\beta_2 = \frac{\lambda_2}{\alpha_2}, \\
&\xi_{11} + \xi_{21} = 1, \\
&\xi_{12} + \xi_{22} = 1, \\
&\xi_{i,k} \geqslant 0, \ \ i, k \in \{1, 2\}.
\end{aligned} \tag{3.7}$$

Identity (3.1) follows.

Next, because the expression (3.2) is also column-stochastic, proving that any solution ξ is determined by ξ_{11} as in (3.2) amounts to proving that ξ_{12} is given as in (3.2). This follows from the first line in (3.7).

It remains to prove (3.3). By the expression just obtained for ξ_{12} it follows that as long as

$$\xi_{11} \geqslant \frac{\lambda_1}{\alpha_1\beta_1} - \frac{\beta_2}{\beta_1},$$

we obtain $\xi_{12} \leqslant 1$. Clearly, in addition, $\xi_{11} \geqslant 0$ must hold. Similarly, as long as

$$\xi_{11} \leqslant \frac{\lambda_1}{\alpha_1\beta_1},$$

we obtain $\xi_{12} \geqslant 0$, and in addition, $\xi_{11} \leqslant 1$ must hold. As a result, it is necessary that

$$\xi_{11} \in \Big[\max\Big(0, \frac{\lambda_1}{\alpha_1\beta_1} - \frac{\beta_2}{\beta_1}\Big), \min\Big(\frac{\lambda_1}{\alpha_1\beta_1}, 1\Big)\Big]. \tag{3.8}$$

Moreover, setting ξ_{11} to each of the two endpoints of the interval indicated in (3.8) and letting ξ be the corresponding expression from (3.2) gives rise to a solution satisfying all of (3.7), as can be checked directly. Because by (3.2) a solution ξ is an affine function of its entry ξ_{11}, these two endpoints correspond to the two extreme points of $\mathcal{S}_{\rm LP}$, that is, to the two modes $\xi^{*,1}$, $\xi^{*,2}$. This proves the lemma. □

Proof of Lemma 3.2. Note that relations (3.7) are invariant to relabeling of classes and servers. Hence so is relation (3.2), which was derived solely from (3.7). Let m be given and assume a relabeling has been performed to put $\xi^{*,m}$ in canonical form. Then $\xi^{*,m}$ satisfies (3.2) with its first column given by $(1, 0)^T$. Consequently $\xi^{*,m}_{11} = 1$. Substituting $\xi^{*,m}_{11} = 1$ into (3.2) proves (3.4). Because under the nondegeneracy assumption there can be only one zero entry, in (3.4) we have $\xi^{*,m}_{12} > 0$. Hence $\lambda_1 > \alpha_1\beta_1$. The final assertion follows from (3.4) using again the fact that there can be at most one zero entry. □

The four possible graphs and their relabelings are described in Figure 6. Namely, if (i', k) is the non-basic activity in $\xi^{*,m}$, then defining

$$\begin{aligned}\widetilde{\xi}^{*,m}_{11} &= \xi^{*,m}_{ik} = 1,\\ \widetilde{\xi}^{*,m}_{22} &= \xi^{*,m}_{i'k'},\\ \widetilde{\xi}^{*,m}_{21} &= \xi^{*,m}_{i'k} = 0, \text{ and}\\ \widetilde{\xi}^{*,m}_{12} &= \xi^{*,m}_{ik'},\end{aligned}$$

$\widetilde{\xi}^{*,m}$ is obtained from $\xi^{*,m}$ upon relabeling in the form of an "N".

Proof of Lemma 3.3. The nondegeneracy of both modes follows from Lemma 3.2.

Next, let the class switching condition (3.5) hold. Because of (3.1),

$$\max_k \beta_k > \max_i \frac{\lambda_i}{\alpha_i} \geqslant \min_i \frac{\lambda_i}{\alpha_i} > \min_k \beta_k. \tag{3.9}$$

Recall from the proof of Lemma 3.1 that the two endpoints of the interval defined in (3.8) correspond to the two modes. Consider the right endpoint. If the minimum in expression in (3.8) is 1 then by (3.2), the non-basic activity in that mode is $(2,1)$, and moreover, $\frac{\lambda_1}{\alpha_1} > \beta_1$. By (3.1), this gives $\frac{\lambda_2}{\alpha_2} < \beta_2$. In view of (3.9) this gives

$$\beta_2 > \max_i \frac{\lambda_i}{\alpha_i}.$$

Hence the maximum in (3.8) is 0. By (3.2), this shows that the non-basic activity in the other mode is $(1,1)$. If, on the other hand, the minimum in (3.8) is $\frac{\lambda_1}{\alpha_1\beta_1}$ then $\frac{\lambda_1}{\alpha_1} < \beta_1$ and the non-basic activity in the corresponding mode is $(1,2)$. Similarly, by (3.9)

$$\min_i \frac{\lambda_i}{\alpha_i} > \beta_2.$$

This means the maximum in (3.8) is not 0. The non-basic activity in the other mode is then $(2,2)$. In both cases, the two modes form a class-switched pair as claimed.

Consider now the server switching condition (3.6). Because of (3.1),

$$\max_i \frac{\lambda_i}{\alpha_i} > \max_k \beta_k \geqslant \min_k \beta_k > \min_i \frac{\lambda_i}{\alpha_i}. \tag{3.10}$$

If the minimum in (3.8) is 1 then $\frac{\lambda_1}{\alpha_1} > \beta_1$ and the non-basic activity in one mode is $(2,1)$. By (3.10), $\frac{\lambda_1}{\alpha_1} > \beta_2$. Hence the maximum in (3.8) is not zero and the non-basic activity in the other mode is $(2,2)$. Finally, if the minimum in (3.8) is not 1 then $\frac{\lambda_1}{\alpha_1} < \beta_1$ and the non-basic activity in one mode is $(1,2)$. By (3.10), $\frac{\lambda_1}{\alpha_1} < \beta_2$. Hence the maximum in (3.8) is zero and the non-basic activity in the other mode is $(1,1)$. In both cases, the two of modes forms a server-switched pair. □

Proof of Theorem 2.4. This lower bound is precisely the one stated in [1, Theorem 2.6], when specialized to the 2 × 2 PSS. To prove that it is valid we must verify that the standing assumption of [1], namely [1, Assumption 2.2], holds.

First, [1, Assumption 2.2.1], which states that the EHTC holds, is valid because of our Assumption 2.2.1. Next, [1, Assumption 2.2.2], when translated to the notation of this paper, states that

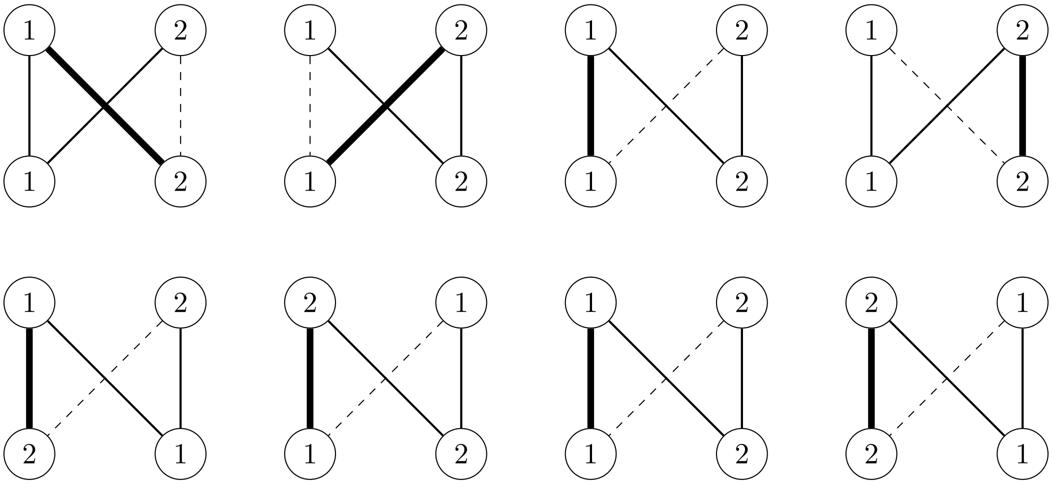

Figure 6: Top: Four possible graphs of basic activities. Bottom: Corresponding relabelings for canonical form.

every $\xi \in \mathcal{S}_{\rm LP}$ is column-stochastic. This holds by our Lemma 2.1.1. It remains to show that [1, Assumption 2.2.3], which states that the dual of (2.7) has a unique solution, is satisfied.

To this end we shall adopt in the remainder of this proof some notation and terminology from [1]. By Lemma 3.2, the non basic activity is different in both modes $\xi^{*,1}$, $\xi^{*,2}$, for else one would have $\xi^{*,1} = \xi^{*,2}$, contradicting the EHTCM. As a result, with the terminology introduced in [1, Section 2], all the activities are potentially basic. Using strict complementary slackness ((36) in [15, Chapter 7]) in the same way as in [1, Lemma 2.3.4], any (y, z) solution of the dual satisfies

$$y_i = \mu_j z_k, \qquad i, k \in \{1, 2\},\ j = (i, k).$$

It follows that $y_1 y_2 = \mu_{11}\mu_{21} z_1^2 = \mu_{12}\mu_{22} z_2^2$. By positivity of both z_k, this gives $z_1 = c z_2$ for some constant $c > 0$. Moreover, one of the constraints of the dual problem [1, eq. (2.8)] is $z_1 + z_2 = 1$. Therefore (z_k) are uniquely determined. As a consequence so are (y_i), which shows that the dual problem has at most 1 solution. The existence of a dual solution follows from the EHTC, as shown in [1, Lemma 2.4.2]. This completes the verification of [1, Assumption 2.2]. □

4 The WCP and HJB equation

In this section, Proposition 2.6 is proved. Lemma 4.1, which is used to prove it, contains two additional results: An identification of optimal control systems for the WCP, and weak uniqueness of solutions to the SDE (2.27), both needed for the weak convergence proofs in §5.

Proof of Part 1 of Proposition 2.6. This is a special case of [1, Proposition 2.5]. We comment that the fact that $V_{\rm WCP}$ is a classical solution to (2.18)–(2.19) has been established already in [16]. However, uniqueness of solutions is not covered there. □

In what follows, u always denotes the unique solution to (2.18)–(2.19).

In the following lemma, Parts 1 and 2 are largely based on results from [16]. For completeness, we have included details on the construction from [16] (as corrected in [17]) in Appendix B.

Lemma 4.1. *1. (Optimality in the single mode case). Assume* (2.20) *and recall that in this case $\xi^A = \xi^{*,m}$ for m as in* (2.20)*. Then, with u as above, equation* (2.22) *of Proposition 2.6 holds. Moreover, $J_{\rm WCP}(z, \mathfrak{S}^{(1)}) = V_{\rm WCP}(z)$ for the admissible control system $\mathfrak{S}^{(1)} = (\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P}, B, \Xi, Z^{(1)}, L^{(1)})$ where $(Z^{(1)}, L^{(1)}, B)$ is the RBM from* (2.26) *(assumed to be constructed on the original probability space), $\mathcal{F}_t = \sigma\{B_s : s \in [0,t]\}$ and $\Xi_t = \xi^A$.*

2. (Optimality in the dual mode case). Assume (2.21) *and recall that in this case $(\xi^L, \xi^H) = (\xi^{*,m'}, \xi^{*,m})$. Then there exists a switching point $z^* \in (0,\infty)$ such that* (2.23) *of Proposition 2.6 holds. Moreover, SDE* (2.27) *possesses a weak solution $(\Omega', \mathcal{F}', \{\mathcal{F}'_t\}, \mathbb{P}', B, Z^{(2)}, L^{(2)})$. Furthermore, one has $J_{\rm WCP}(z, \mathfrak{S}^{(2)}) = V_{\rm WCP}(z)$ for the admissible control system defined by $\mathfrak{S}^{(2)} = (\Omega', \mathcal{F}', \{\mathcal{F}'_t\}, \mathbb{P}', B, \Xi, Z^{(2)}, L^{(2)})$, where $\Xi_t = \varphi^*(Z^{(2)}_t)$.*

3. Weak uniqueness holds for solutions to SDE (2.27)*.*

Proof of Parts 2 and 3 of Proposition 2.6. These results are contained in Parts 1 and 2 of Lemma 4.1. □

For Markov control problems, a map from the state space to the control action space is often called a *stationary (feedback) control policy*, or a *policy* for short. For our WCP, a policy is thus a measurable map $\bar{\xi} : \mathbb{R}_+ \to \mathcal{S}_{\rm LP}$. This term is used in [16] and we adopt it in the next proof. Parts 1 and 2 take full advantage of several results from [16], where a classical solution to equation (2.18)–(2.19) is constructed, and a description of an optimal policy is provided.

Proof of Parts 1 and 2 of Lemma 4.1. The single mode condition (2.20) corresponds to eq. [41]–[42] on p. 105 of [16, Section 5.3]. The dual mode condition (2.21) corresponds to the complementary case. It is stated in [16, Theorem 1, Chapter 4] that a policy is optimal if and only if the cost associated to it is C^2 and satisfies [16, eq. (14), Chapter 4], which is the HJB equation (2.18)–(2.19) in our notation. However, the equation studied there is more general, and in order to reduce it to our (2.18)–(2.19) one must take the switching costs to vanish (by setting the expression $K = 0$), and the reflection-absorption parameter to correspond to reflection only (by setting $\lambda = 1$).

Under the single mode condition, the policy constructed has the simple form

$$\bar{\xi}(z) = \xi^{*,m}, \qquad z \in \mathbb{R}_+,$$

with the same m as in (2.20). It is shown that it is an optimal policy by computing the cost associated it and showing that it solves the HJB equation with its boundary conditions. The fact that (2.22) holds in this case is a direct consequence of the fact that the cost associated to the single mode policy indeed solves the HJB equation. This completes the proof of part 1.

As for part 2, the claim that the SDE (2.27) possesses a weak solution is proved in [1, Lemma 4.1]. Under the dual mode case, the policy constructed in [16] is

$$\bar{\xi}_{z^*}(z) = \xi^{*,m'} \mathbb{1}_{z \le z^*} + \xi^{*,m} \mathbb{1}_{z > z^*}, \qquad z \in \mathbb{R}_+, \tag{4.1}$$

with the same m and m' as in (2.21). By [16, Theorem 4, Chapter 3], any policy of this form gives rise to a cost function that is C^1 in all of $(0,\infty)$, and C^2 in $(0,\infty) \setminus \{z^*\}$. The value of z^* which leads to an optimal policy is found by the principle of smooth fit, namely by requiring that the second derivative is continuous also at z^*. The equation that states this smoothness condition, [16, (61), Chapter 5] (as corrected in [17, (3.21)]), turns out to have a unique solution $z^* \in (0,\infty)$. The solution to the HJB equation is then the cost associated to $\bar{\xi}_{z^*}$ with that value of z^*. The statement of part 2, by which the corresponding control system is optimal for the WCP, therefore follows from [16, Theorem 1, Chapter 4]. Once again, the very fact that this function solves the HJB equation implies that equation (2.23) holds in this case. □

Proof of Part 3 of Lemma 4.1. We slightly simplify the notation by writing (2.27)–(2.28) as

$$Z_t = z + \int_0^t b^*(Z_s)ds + \int_0^t \sigma^*(Z_s)dB_s + L_t,$$

where, with $a^*(x) = \sigma^*(x)^2$, there exist constants $b_0, b_1 \in \mathbb{R}$ and $a_0 > 0$, $a_1 > 0$, such that

$$b^*(x) = \begin{cases} b_0, & x \le z^*, \\ b_1, & x > z^*, \end{cases} \qquad a^*(x) = \begin{cases} a_0, & x \le z^* \\ a_1, & x > z^* \end{cases} .$$

In the remainder of this proof, a^* and b^* are written as a and b. Let $\mathcal{A}$ be the operator

$$\mathcal{A}f(x) := b(x)f'(x) + \frac{1}{2}a(x)f''(x),$$

on the domain $\mathcal{D}(\mathcal{A}) := \{f \in C_0^\infty[0,\infty) : f'(0) = 0\}$, where $C_0^\infty[0,\infty)$ denotes the set of compactly supported members of $C^\infty[0,\infty)$. If $f \in \mathcal{D}(\mathcal{A})$ and (Z, L, B) is a weak solution to the SDE then by Ito's formula, the boundary property of L and the boundary condition $f'(0) = 0$, the process $f(Z_t) - \int_0^t \mathcal{A}f(Z_s)ds$ is a martingale. Therefore it follows from the existence result in Part 2 of the lemma that, for every probability distribution ν on $[0,\infty)$, there exists a solution of the $D_{[0,\infty)}[0,\infty)$ martingale problem for $(\mathcal{A}, \nu)$. We will use the definition of the stopped martingale problem of Ethier and Kurtz [5], Chapter 4, Section 6. Then, for every probability distribution ν on $[0,\infty)$, there exists a solution of the stopped martingale problem for $\big(\mathcal{A}, \nu, [0, \frac{2}{3} z^*)\big)$ and of the stopped martingale problem for $\big(\mathcal{A}, \nu, (\frac{1}{3} z^*, \infty)\big)$.

Define the operators $\mathcal{A}_0$ and $\mathcal{A}_1$ in the following way:

$$\mathcal{A}_0 f(x) := b_0 f'(x) + \frac{1}{2}a_0 f''(x), \qquad \text{on the domain } \mathcal{D}(\mathcal{A}_0) := \mathcal{D}(\mathcal{A})$$

$$\mathcal{A}_1 f(x) := b(x)f'(x) + \frac{1}{2}a(x)f''(x), \qquad \text{on the domain } \mathcal{D}(\mathcal{A}_1) := C_0^\infty(\mathbb{R}).$$

The $D_{[0,\infty)}[0,\infty)$ martingale problem for $\mathcal{A}_0$ is well posed (for instance by Corollary 8.1.2, Theorem 4.4.1 and Proposition 4.3.1 of [5]). Therefore, for every probability distribution ν on $[0,\infty)$, there exists a unique solution of the stopped martingale problem for $\big(\mathcal{A}_0, \nu, [0, \frac{2}{3} z^*)\big)$ by Theorem 4.6.1 of [5]. Since, for every $f \in \mathcal{D}(\mathcal{A}_0) = \mathcal{D}(\mathcal{A})$,

$$\mathcal{A}_0 f\big|_{[0,\frac{2}{3} z^*]} = \mathcal{A}f\big|_{[0,\frac{2}{3} z^*]},$$

every solution of the stopped martingale problem for $\big(\mathcal{A}, \nu, [0, \frac{2}{3} z^*)\big)$ is also a solution of the stopped martingale problem for $\big(\mathcal{A}_0, \nu, [0, \frac{2}{3} z^*)\big)$, therefore the solution of the stopped martingale problem for $\big(\mathcal{A}, \nu, [0, \frac{2}{3} z^*)\big)$ is unique.

Next, the $D_{\mathbb{R}}[0,\infty)$ martingale problem for $\mathcal{A}_1$ is well posed by Exercise 7.3.3 of [18] (it is shown there that the $C_{\mathbb{R}}[0,\infty)$ martingale problem for $\mathcal{A}_1$ is well posed, but every solution of the $D_{\mathbb{R}}[0,\infty)$ martingale problem for $\mathcal{A}_1$ has paths in $C_{\mathbb{R}}[0,\infty)$ almost surely). Therefore, for every probability distribution ν on $[0,\infty)$, there exists a unique solution of the stopped martingale problem for $\big(\mathcal{A}_1,\nu,(\frac{1}{3}\,z^*,\infty)\big)$ by Theorem 4.6.1 of [5]. Since for every $f_1 = \mathcal{D}(\mathcal{A}_1)$ there exists $f \in \mathcal{D}(\mathcal{A})$ such that

$$f\big|_{[\frac{1}{3}\,z^*,\infty)} = f_1\big|_{[\frac{1}{3}\,z^*,\infty)}, \qquad \mathcal{A}f\big|_{[\frac{1}{3}\,z^*,\infty)} = A_1 f_1\big|_{[\frac{1}{3}\,z^*,\infty)},$$

every solution of the stopped martingale problem for $\big(\mathcal{A},\nu,(\frac{1}{3}\,z^*,\infty)\big)$ is also a solution of the stopped martingale problem for $\big(\mathcal{A}_1,\nu,(\frac{1}{3}\,z^*,\infty)\big)$, therefore the solution of the stopped martingale problem for $\big(\mathcal{A},\nu,(\frac{1}{3}\,z^*,\infty)\big)$ is unique.

Now one can apply Theorem 4.6.2 in [5] with $U_1 := [0,\frac{2}{3}\,z^*)$, $U_k := (\frac{1}{3}\,z^*,\infty)$ for $k \geq 2$, to conclude that, for every probability distribution ν on $[0,\infty)$, the solution of the $D_{[0,\infty)}[0,\infty)$ martingale problem for $(\mathcal{A},\nu)$ is unique.

Finally, because, as already mentioned, every solution to the SDE is a solution to the martingale problem, the weak uniqueness of solutions to the SDE follows. □

Remark 4.2. *(Symmetry conditions). As Lemma 4.1 states,* (2.20) *is a necessary and sufficient condition for the optimal control of the WCP to not switch between modes. It is natural to ask whether it can be determined if the single or the dual-mode condition holds under various symmetry conditions. Specifically, consider*

$$\frac{\hat{\mu}_{1k}}{\alpha_1} = \frac{\hat{\mu}_{2k}}{\alpha_2}, \qquad k = 1,2, \tag{4.2}$$

$$\frac{\hat{\mu}_{i1}}{\beta_1} = \frac{\hat{\mu}_{i2}}{\beta_2}, \qquad i = 1,2, \tag{4.3}$$

$$C_{S_{i1}} = C_{S_{i2}}, \qquad i = 1,2. \tag{4.4}$$

As it turns out, each of the above three conditions is sufficient for (2.20)*. This is proved in Lemma C.1 in the appendix. As a result, each of these conditions is sufficient for non-switching. These conditions can arise naturally, and occur in certain cases in the literature.*

5 Asymptotic optimality

In this section we prove Theorem 2.13. Toward this, an important intermediate goal is to establish a weak convergence result, stated in Theorem 5.1. The proofs of both theorems rely on four main steps stated in Propositions 5.3–5.6.

5.1 Weak convergence

5.1.1 Statement of weak convergence result

We will adopt the following convention regarding the six cases listed in Theorem 2.13. When a statement is said to hold in a certain case of Theorem 2.13, it is meant that the assumptions as well as the policy specified in this case are in force. For example, saying that a certain statement holds in case 1(a) of Theorem 2.13 means that it holds when the single-mode condition (2.20) and the condition $i_1(\xi^A) = p$ hold, and the $\mathbf{P}$ policy is applied, and moreover, the weaker moment assumption $\mathbf{m} > 2$ is assumed (but $\mathbf{m} > \mathbf{m}_0$ in case 1(b), for example). When a claim is stated without specifying a case, it is meant that it holds in each one of the six.

In addition to the rescaled processes already defined, the weak convergence results will be concerned with additional rescaled processes, namely

$$\hat{A}_i^n(t) = n^{-1/2}(A_i^n(t) - \lambda_i^n t), \qquad \hat{S}_{ik}^n(t) = n^{-1/2}(S_{ik}^n(t) - \mu_{ik}^n t), \tag{5.1}$$

$$\hat{I}_k^n(t) = n^{1/2} I_k^n(t), \qquad \hat{L}^n(t) = \beta_1 \hat{I}_1^n(t) + \beta_2 \hat{I}_2^n(t). \tag{5.2}$$

With this notation, the balance equation (2.3) for X^n translates under scaling to

$$\hat{X}_i^n(t) = \hat{A}_i^n(t) + n^{-1/2}\lambda_i^n t - \sum_k \hat{S}_{ik}^n(T_{ik}^n(t)) - \sum_k n^{-1/2}\mu_{ik}^n T_{ik}^n(t). \tag{5.3}$$

$$= \hat{A}_i^n(t) - \sum_k \hat{S}_{ik}^n(T_{ik}^n(t)) + (n^{1/2}\lambda_i + \hat{\lambda}_i^n)t - \sum_k T_{ik}^n(t)(n^{1/2}\alpha_i\beta_k + \hat{\mu}_{ik}^n). \tag{5.4}$$

Thus if we denote

$$\hat{F}_t^n = \sum_i \frac{\hat{A}_i^n(t)}{\alpha_i} - \sum_{ik} \frac{\hat{S}_{ik}^n(T_{ik}^n(t))}{\alpha_i} + \sum_i \frac{\hat{\lambda}_i^n t}{\alpha_i} - \sum_{ik} \frac{\hat{\mu}_{ik}^n}{\alpha_i} T_{ik}^n(t), \tag{5.5}$$

we obtain the following representation for the workload:

$$\hat{W}_t^n = \hat{F}^n(t) + \hat{L}^n(t). \tag{5.6}$$

The convergence of the rescaled primitives is a direct consequence of the central limit theorem for renewal processes [4, §17]. Namely, the tuple $(\hat{A}_i^n, \hat{S}_{ik}^n)$ converges to what we denote by (A_i, S_{ik}), comprising 6 mutually independent BMs with zero drift and diffusivity given by the constants $\lambda_i^{1/2} C_{A_i}$ and $\mu_{ik}^{1/2} C_{S_{ik}}$, which in §2.3 we have denoted by $\sigma_{A,i}$ and $\sigma_{S,ik}$, respectively.

Theorem 5.1. *Let the assumptions of Theorem 2.13 hold. Then* $(T^n, \hat{W}^n, \hat{L}^n, \hat{X}^n) \Rightarrow (T, W, L, X)$, *where the latter is defined as follows.*
1. In cases 1(a)–(b) of Theorem 2.13, (W, L) *is the RBM and boundary term given by* (2.26) *and initial condition* $z = 0$, *and* $T_t = \xi^A t$ *for all* t.
2. In cases 2(a)–(d) of Theorem 2.13, (W, L) *is the (unique in law) weak solution to the SDE* (2.27) *with initial condition* $z = 0$, *and letting* $\Xi_t = \varphi^*(W_t)$, *one has* $T_t = \int_0^t \Xi_s ds$ *for* $t \geq 0$.
3. In all cases, $X_p = 0$ *and* $X_q = \alpha_q W$.

The significance of this result is that it implies that, under each of the proposed policies, limits of the processes exist and form admissible control systems for the WCP, which are, in view of Lemma 4.1.1–2, optimal.

We next present a lemma from [1] concerning limits of $(\hat{A}^n, \hat{S}^n, T^n, \hat{F}^n)$ under general sequences of policies. The *Skorohod map*, $\Gamma : D_{\mathbb{R}}[0,\infty) \to D_{\mathbb{R}_+}[0,\infty) \times D_{\mathbb{R}}^{+}[0,\infty)$, takes a function ψ to a pair (φ, η), where

$$\varphi(t) = \psi(t) + \eta(t), \qquad \eta(t) = \sup_{0 \leqslant s \leqslant t} \psi(s)^-, \qquad t \geqslant 0. \tag{5.7}$$

The corresponding maps $\psi \mapsto \varphi$ and $\psi \mapsto \eta$ are denoted by Γ_1 and Γ_2, respectively.

Lemma 5.2. *Let $\{T^n\}$, $T^n \in \mathcal{A}^n$ be any sequence of admissible controls for the QCP for which $\limsup_n \hat{J}^n(T^n) < \infty$. Then the following conclusions hold. The sequence $(\hat{A}^n, \hat{S}^n, T^n)$ is C-tight. Along any convergent subsequence where $(\hat{A}^n, \hat{S}^n, T^n) \Rightarrow (A, S, T)$, one has*

$$(\hat{A}^n, \hat{S}^n, T^n, \hat{F}^n) \Rightarrow (A, S, T, F),$$

where F is defined below in (5.8). There exist on $(\Omega, \mathcal{F})$ processes (B, Ξ, Z', L') and a filtration $\{\mathcal{H}_t\}$ such that the tuple $\mathfrak{S} = (\Omega, \mathcal{F}, \{\mathcal{H}_t\}, \mathbb{P}, B, \Xi, Z', L')$ forms an admissible control system for the WCP with initial condition 0. These processes satisfy the relations

$$\begin{aligned} T &= \int_0^\cdot \Xi_s ds, \qquad (Z', L') = \Gamma[F], \\ F_t &= \sum_i \frac{A_i(t) + \hat{\lambda}_i t}{\alpha_i} - \sum_{i,k} \frac{S_{ik}(T_{ik}(t)) + \hat{\mu}_{ik} T_{ik}(t)}{\alpha_i} \\ &= \int_0^t b(\Xi_s) ds + \int_0^t \sigma(\Xi_s) dB_s. \end{aligned} \tag{5.8}$$

Proof. This is the content of [1, Lemmas 5.1 and 5.5]. □

This lemma is our starting point for proving Theorem 5.1. Whereas it relates limits of processes associated with the QCP to an admissible control system for the WCP, note that it does not make any claim regarding $\hat{X}^n$ or $\hat{W}^n$, hence by itself is not sufficient to relate the prelimit cost (defined in terms of $\hat{X}^n$) to the WCP cost. In particular, the pair (Z', L') need not be the weak limit of $(\hat{W}^n, \hat{L}^n)$. To proceed one must show that under the proposed policies, along the sequence specified in Lemma 5.2, one has

$$(\hat{A}^n, \hat{S}^n, T^n, \hat{F}^n, \hat{W}^n, \hat{L}^n) \Rightarrow (A, S, T, F, W, L) \tag{5.9}$$

where $(W, L) = (Z', L')$. Once this is achieved, the lemma guarantees that the weak limit (W, L) satisfies

$$W_t = \int_0^t b(\Xi_s) ds + \int_0^t \sigma(\Xi_s) dB_s + L_t,$$

with $\int_{[0,\infty)} W_t dL_t = 0$, assuring that the limit tuple indeed forms an admissible system for the WCP. To go from here to the statements made in Theorem 5.1, one further needs to show that (W, L) are as given in (2.26) or (2.27), and that $X_p = 0$.

The Propositions presented below address these issues as follows. Propositions 5.3 provides uniform integrability required to eventually deduce Theorem 2.13 from Theorem 5.1, and in addition ensures that the prelimit cost remains bounded so Lemma 5.2 may be applied. Proposition 5.4 shows that $\hat{X}^n_p \to 0$ in probability. Proposition 5.5 states precisely what is needed to attain (5.9). Finally, Proposition 5.6 implies that (2.27) holds in the dual mode case.

5.1.2 Main steps toward weak convergence

Throughout what follows, the assumptions of Theorem 2.13 are in force. The four main steps required to achieve weak convergence, and later, Theorem 2.13, are as follows.

Proposition 5.3. *There exists $\varepsilon_0 > 0$ such that*

$$\limsup_n \mathbb{E} \int_0^\infty e^{-\gamma t} (\hat{H}_t^n)^{1+\varepsilon_0} dt < \infty.$$

As already mentioned, this uniform integrability result will allow us to deduce convergence of the costs, as stated in Theorem 2.13, from the convergence stated in Theorem 5.1. Moreover, because it also implies boundedness of the cost under the proposed policies, it enables us to use Lemma 5.2. The proof is given in §5.2.2.

Proposition 5.4. *For every $t_0 > 0$, as $n \to \infty$,*

$$\mathbb{P}\left(\|\hat{X}_p^n\|_{t_0} \geqslant 2\hat{\Theta}^n \right) \to 0.$$

The above type of result is often referred to as state-space collapse (SSC), as it asserts that asymptotically all workload is kept in one buffer, a property crucially used in establishing the one-dimensional state space description of the limiting dynamics, as well as asymptotic optimality, since all workload is held in the 'less expensive' class. It is proved in §5.2.3.

Proposition 5.5. *Consider a subsequence as in Lemma 5.2, where $(\hat{A}^n, \hat{S}^n, T^n) \Rightarrow (A, S, T)$. Then along this sequence one has $(\hat{A}^n, \hat{S}^n, T^n, \hat{F}^n, \hat{W}^n, \hat{L}^n) \Rightarrow (A, S, T, F, W, L)$, where F is given by (5.8), and $(W, L) = \Gamma(F)$. In particular, $(\hat{W}^n, \hat{L}^n)$ is a C-tight sequence, and the conclusions of Lemma 5.2 hold with $(W, L) = (Z', L')$.*

The proof appears in §§5.2.4–5.2.5.

Finally, the policies $\mathbf{P}$ and $\mathbf{T}_2$ that are proposed under the single mode condition do not use the non-basic activity of the active mode ξ^A. For the four policies employed under the dual-mode condition, we need the following control over the use of the non-basic activities corresponding to ξ^L and ξ^H.

Proposition 5.6. *Consider the same subsequence as in Proposition 5.5 and cases 2(a)–(d) of Theorem 2.13. If (i^l, k^l) (resp. (i^h, k^h)) denotes the non-basic activity in ξ^L (resp. ξ^H), then for any $t_0 > 0$ and $\varepsilon > 0$,*

$$\int_0^{t_0} \mathbb{1}_{\hat{W}_t^n \leqslant z^* - \varepsilon} dT^n_{i^l k^l}(t) \to 0, \qquad \int_0^{t_0} \mathbb{1}_{\hat{W}_t^n \geqslant z^* + \varepsilon} dT^n_{i^h k^h}(t) \to 0, \tag{5.10}$$

in probability, as $n \to \infty$.

To explain the role of this proposition, recall that if $\xi \in \mathcal{S}_{\rm LP}$ then the condition $\xi_{ik} = 0$ for some activity (i, k) not only implies that ξ is one is the modes $\xi^{*,1}$ or $\xi^{*,2}$ but also identifies which one by Lemma 3.2. Since any limit T of T^n is given as $\int_0^\cdot \Xi_s ds$, $\Xi_t \in \mathcal{S}_{\rm LP}$, this proposition implies that when workload is either below or above the switching point, the resource allocation asymptotically follows the respective mode of operation ξ^L or ξ^H. This is very close to stating that the pair (W, L) follows the SDE (2.27), and indeed is the basis for proving this fact. The proof of this proposition appears in §5.2.6.

5.1.3 Considerations for the construction of dual-mode policies

As noted above, two of the key results that we need to prove Theorem 2.13 are state-space collapse (Proposition 5.4), and a boundary property stating that there is asymptotically no idleness of either server when there is work in the system (Proposition 5.5). The proofs of these results differ by case/policy, and, for dual-mode policies, rely on the rules used as well as the way that workload is sampled. Here we provide a brief description of the reasons underlying the policy definitions that we have used.

- A difficulty arises in the proof of Proposition 5.4 in case 2(a), where the policy switches between two **P** rules. Only the dual activity server in the current mode processes the HPC and the identity of the dual activity server changes when switching modes. At each switching time, if the server processing the HPC in the new mode is busy with low priority jobs and the service of the high priority job at the other server ends, no server will process high priority jobs for a time $O(n^{-1})$. In principle, this time could accumulate to let the number of high priority jobs increase to a non-negligible value. In order to prevent this, we designed switching between **P** rules to only occur at the time of service completion at the single activity server. When switching, this server becomes dual activity and gives priority to the HPC (which is the single activity class in **P**).

  In case 2(a), there is always at least one server processing the HPC, regardless of switching between modes.

- Similarly, in order to prove Proposition 5.5 in case 2(b), we need to show that the number of HPC is not zero when there are low priority jobs in the system. To make sure that the high priority class does not receive too much service, switching between two $\mathbf{T}_2$ rules only occurs at the time of service completion at the single activity server. When switching, the single activity server becomes dual activity and now gives priority to the low priority jobs if the number of HPC jobs is low.

  In case 2(b), as long as the number of HPC jobs is below the threshold there is at most one server working on HPC jobs, regardless of switching between modes.

- In case 2(c) it should be noted that in the lower mode the system looks similar to case 1(a), so one could consider using a **P** rule. Doing so, however, would cause difficulty in proving Proposition 5.5 for this case. Using a **P** rule keeps the number of HPC jobs $O(1)$. At the moment of switching into the $\mathbf{T}_2$ rule this can lead to the new single activity server, which is dedicated to HPC, incurring idle time when there is work in the system. To avoid this situation the $\mathbf{T}_1$ rule was used instead of **P**, guaranteeing that, at a switching time, there will not be too few HPC jobs in the system. A similar argument applies to case 2(d).

5.1.4 Proof of weak convergence

Here we prove Theorem 5.1 based on the four propositions.

Proof of Theorem 5.1. First, by Proposition 5.3 one has that $\limsup_n \hat{J}^n(T^n) < \infty$ for each one of the relevant policies T^n. As a result, the assumptions of Lemma 5.2 and Proposition

5.5 are valid. To summarize the conclusions from these results, fix a subsequence along which $(\hat{A}^n, \hat{S}^n, T^n) \Rightarrow (A, S, T)$. Then there exists a tuple $(\{\mathcal{H}_t\}, F, W, L, \Xi, B)$ such that, along this sequence,

$$(\hat{A}^n, \hat{S}^n, \hat{F}^n, T^n, \hat{W}^n, \hat{L}^n) \Rightarrow (A, S, F, T, W, L),$$

where F is given in terms of A, S, T by (5.8), and W and L are given by $(W, L) = \Gamma(F)$. Moreover, $\mathfrak{S} = (\Omega, \mathcal{F}, \{\mathcal{H}_t\}, \mathbb{P}, B, \Xi, W, L)$ forms an admissible control for initial condition 0, and $T = \int_0^\cdot \Xi_s ds$. In particular,

$$W = F + L = \int_0^\cdot b(\Xi_s)ds + \int_0^\cdot \sigma(\Xi_s)dB_s + L_t. \tag{5.11}$$

Next, by Proposition 5.4, $\hat{X}_p^n \to 0$ in probability. Also, by (2.11), $\hat{X}_q^n = \alpha_q \hat{W}^n - \alpha_p \hat{X}_p^n$, and therefore we now have, along the subsequence, $(\hat{A}^n, \hat{S}^n, \hat{F}^n, T^n, \hat{W}^n, \hat{L}^n, \hat{X}^n) \Rightarrow (A, S, F, T, W, L, X)$, where we denote

$$X_p = 0, \qquad X_q = \alpha_q W. \tag{5.12}$$

Note that the control system $\mathfrak{S}$ thus constructed may depend on the subsequence. However, consider the following.

Claim. In cases 1(a)–(b) of Theorem 2.13, (W, L) *is the RBM and its boundary term given by* (2.26)*, and* $T_t = \xi^A t$ *for all* t*. In cases 2(a)–(d) of Theorem 2.13,* (W, L) *is a weak solution to the SDE* (2.27)*, and moreover,* $\Xi_t = \varphi^*(W_t)$ *for a.e.* t*.*

Suppose the above claim holds true. Then the law of (W, L) is uniquely determined: in case 1 as a RBM; in case 2 as a weak solution to (2.27), for which weak uniqueness holds by Lemma 4.1.3. In particular, this law does not depend on the subsequence. Moreover, since by this claim and (5.12), the pair of processes (T, X) is uniquely determined by W (away from a $\mathbb{P}$-null set), it follows that the law of (T, W, L, X) does not depend on the subsequence. This yields the convergence $(T^n, \hat{W}^n, \hat{L}^n, \hat{X}^n) \Rightarrow (T, W, L, X)$ along the full sequence and completes the proof of the result. In what follows, the claim is proved.

Consider first the single mode case, namely cases 1(a)–(b) of Theorem 2.13. The policies employed are $\mathbf{P}$ and $\mathbf{T}_2$, and both do not use the non-basic activity of the active mode ξ^A. In other words, if we denote this activity by (i^a, k^a) then under these policies, $T^n_{i^a k^a} = 0$ for all n. As a consequence, the limit process T must satisfy $T_{i^a k^a}(t) = 0$ a.s., hence $\Xi_{i^a k^a}(t) = 0$ for a.e. t, a.s. By the uniqueness statement made in Lemma 3.2, whenever $\tilde{\xi} \in \mathcal{S}_{\rm LP}$ and $\tilde{\xi}_{i^a k^a} = 0$, one must have $\tilde{\xi} = \xi^A$. It follows that $\Xi_t = \xi^A$ for a.e. t, and hence $T_t = \xi^A t$. As a consequence, (5.11) holds as $W_t = b(\xi^A)t + \sigma(\xi^A)B_t + L_t$. That is, the pair (W, L) satisfies (2.26). This proves the first part of the claim.

Next consider the dual mode, namely, cases 2(a)–(d) of Theorem 2.13. First, we will show based on Proposition 5.6 that, for every $\varepsilon > 0$ and $t_0 > 0$,

$$\int_0^{t_0} \mathbb{1}_{\{W_t < z^* - \varepsilon\}} dT_{i^l k^l}(t) = 0, \qquad \int_0^{t_0} \mathbb{1}_{\{W_t > z^* + \varepsilon\}} dT_{i^h k^h}(t) = 0 \tag{5.13}$$

holds a.s. Let g be a continuous function such that

$$\mathbb{1}_{w < z^* - \varepsilon} \leqslant g(w) \leqslant \mathbb{1}_{w < z^* - \frac{\varepsilon}{2}}, \qquad w \in \mathbb{R}_+. \tag{5.14}$$

By the continuous mapping theorem, we have along the subsequence $(g(\hat{W}^n), T^n) \Rightarrow (g(W), T)$. In addition, T^n is continuous with finite variation over compacts. By Theorem 2.2 of [12],

$$\int_0^{\cdot} g(\hat{W}^n_t) dT^n_{i^l k^l}(t) \Rightarrow \int_0^{\cdot} g(W_t) dT_{i^l k^l}(t). \tag{5.15}$$

By Proposition 5.6, $\int_0^{t_0} \mathbb{1}_{\{\hat{W}^n_t < z^* - \frac{\varepsilon}{2}\}} dT^n_{i^l k^l}(t) \to 0$ in probability. Hence the LHS in (5.15) converges to zero in probability. Thus the RHS in (5.15) equals zero a.s., and therefore by the first inequality in (5.14), the first part of (5.13) is proved. The second part of (5.13) is proved analogously.

Next, clearly $\varXi_{i^l k^l}(t) \mathbb{1}_{\{\varXi_t = \xi^L\}} = 0$. Hence by (5.13), $\int_0^{t_0} \mathbb{1}_{\{W_t < z^* - \varepsilon\}} \mathbb{1}_{\{\varXi_t \neq \xi^L\}} \varXi_{i^l k^l}(t) dt = 0$. Arguing again by the uniqueness statement in Lemma 3.2, one has $\varXi_{i^l k^l}(t) > 0$ whenever $\varXi_t \neq \xi^L$. It follows that, a.s.,

$$\int_0^{t_0} \mathbb{1}_{\{W_t < z^* - \varepsilon\}} \mathbb{1}_{\{\varXi_t \neq \xi^L\}} dt = 0, \qquad \int_0^{t_0} \mathbb{1}_{\{W_t > z^* + \varepsilon\}} \mathbb{1}_{\{\varXi_t \neq \xi^H\}} dt = 0, \tag{5.16}$$

where the second equality is proved analogously to the first one. Going back to (5.11), note that by (5.16) one has both $\int_0^{t_0} b(\varXi_s) \mathbb{1}_{\{W_t < z^* - \varepsilon\}} \mathbb{1}_{\{\varXi_t \neq \xi^L\}} ds = 0$ and $\int_0^{t_0} \sigma(\varXi_s) \mathbb{1}_{\{W_t < z^* - \varepsilon\}} \mathbb{1}_{\{\varXi_t \neq \xi^L\}} dB_s = 0$. A similar statement hold for $\{W_t > z^* + \varepsilon\}$ and ξ^H. Hence by (5.11) and the definition (2.28) of the functions b^* and σ^*, it follows that

$$W_t = \int_0^t \mathbb{1}_{|W_s - z^*| \geqslant \varepsilon} b^*(W_s) ds + \int_0^t \mathbb{1}_{|W_s - z^*| \geqslant \varepsilon} \sigma^*(W_s) dB_s + \int_0^t \mathbb{1}_{|W_s - z^*| < \varepsilon} b(\varXi_s) ds + \int_0^t \mathbb{1}_{|W_s - z^*| < \varepsilon} \sigma(\varXi_s) dB_s + L_t,$$

for $t \leq t_0$. Because t_0 is arbitrary, this is true for all t. Hence, denoting $\delta^\varepsilon_t = \mathbb{1}_{|W_t - z^*| < \varepsilon}$ and using the boundedness of $b(\cdot)$,

$$U_t := \left| W_t - \int_0^t b^*(W_s) ds - \int_0^t \sigma^*(W_s) dB_s - L_t \right| \leq \gamma^\varepsilon_t + |\tilde{\gamma}^\varepsilon_t| + |\check{\gamma}^\varepsilon_t|,$$

where

$$\gamma^\varepsilon_t = c \int_0^t \delta^\varepsilon_s ds, \qquad \tilde{\gamma}^\varepsilon_t = \int_0^t \delta^\varepsilon_s \sigma(\varXi_s) dB_s, \qquad \check{\gamma}^\varepsilon_t = \int_0^t \delta^\varepsilon_s \sigma^*(W_s) dB_s.$$

Notice that U_t does not depend on ε, so if we manage to prove that the RHS converges to zero in probability as $\varepsilon \downarrow 0$, it follows that, for every t, $U_t = 0$ a.s. Hence by continuity of this process, $U_t = 0$ for all t, a.s. That is, the processes (W, L, B) satisfy (2.27) a.s.

To this end, apply the occupation times formula, [14, Corollary VI.1.6], by which for any continuous semimartingale Y, one has, a.s.,

$$\int_0^t \mathbb{1}_{\{Y_s = 0\}} d\langle Y, Y \rangle_s = \int_{-\infty}^{\infty} \mathbb{1}_{y=0} L^y_t(Y) dy,$$

with $L^y(Y)$ the local time of Y at y. Consider the above with $Y_t = W_t - z^*$. Clearly the RHS in the above display is zero. Moreover, $\langle Y, Y \rangle_t = \int_0^t \sigma(\varXi_s)^2 ds$, and since σ is bounded away from zero, we obtain that $\int_0^t \mathbb{1}_{\{W_s=z^*\}}ds = 0$ a.s. For fixed ω, one has for all t, $\mathbb{1}_{\{|W_t-z^*|<\varepsilon\}} \to \mathbb{1}_{\{W_t=z^*\}}$ as $\varepsilon \downarrow 0$. Hence, by dominated convergence, $\gamma_t^\varepsilon \to 0$ a.s. as $\varepsilon \downarrow 0$.

Next, for the stochastic integral $\tilde{\gamma}^\varepsilon$ we have

$$\mathbb{E}(\tilde{\gamma}^\varepsilon)^2 = \mathbb{E}\int_0^t (\delta_s^\varepsilon \sigma(\varXi_s))^2 ds \le c\mathbb{E}\int_0^t \delta_s^\varepsilon ds = c\mathbb{E}(\gamma_t^\varepsilon).$$

By local boundedness of γ^ε and its a.s. convergence to 0 as $\varepsilon \downarrow 0$, it follows that $\tilde{\gamma}_t^\varepsilon \to 0$ in probability. A similar argument holds for $\check{\gamma}_t^\varepsilon$. We conclude that (W, L, B) satisfies (2.27). Finally, by (5.16) and the fact $\int_0^t \mathbb{1}_{\{W_s=z^*\}}ds = 0$ a.s. just proved, one has $\varXi_t = \varphi^*(W_t)$ for a.e. t, a.s., which completes the proof of the claim, and also of the result. □

5.1.5 Proof of Theorem 2.13 and Corollary 2.15

As a consequence of Theorem 5.1 and Proposition 5.3 we have the following.

Proof of Theorem 2.13. The weak convergence statements asserted in the theorem are already established in Theorem 5.1. For the AO results we need to show that in each of the six cases $\hat{J}^n(T^n) \to V_0 = h_q\alpha_q V_{\rm WCP}(0)$. Combining Theorem 5.1 with the identification of an optimal control for the WCP given in Lemma 4.1 shows that $\hat{X}^n \Rightarrow X$ where $X_p = 0$ and $X_q = \alpha_q W$, W is given by (2.26) or (2.27) in the respective cases, and moreover

$$V_0 = h_q\alpha_q\mathbb{E}\int_0^\infty e^{-\gamma t}W_t dt.$$

The convergence stated above implies

$$\hat{H}_t^n = h \cdot \hat{X}_t^n \Rightarrow h_q\alpha_q W_t. \tag{5.17}$$

By (2.6), $\hat{J}^n(T^n) = \mathbb{E}\int_0^\infty e^{-\gamma t}\hat{H}_t^n dt$. Hence the convergence $\hat{J}^n(T^n) \to V_0$ will follow from (5.17) once uniform integrability is established. Arguing along the lines of [3] (pp. 640–643), introduce the measure $dm = \gamma e^{-\gamma t}dt$ on $(\mathbb{R}_+, \mathcal{R}_+)$ and invoke the Skorohod representation theorem to obtain from (5.17) that $\hat{H}^n \to h_q\alpha_q W$ $(m \times \mathbb{P})$-a.e. Accordingly, uniform integrability of H^n w.r.t. $m \times \mathbb{P}$ suffices to obtain $\hat{J}^n(T^n) \to V_0$. However, this is ensured by Proposition 5.3, for

$$\limsup_n \int_{[0,\infty)\times\varOmega} (\hat{H}^n)^{1+\varepsilon_0} d(m \times \mathbb{P}) = \gamma \limsup_n \mathbb{E}\int_0^\infty e^{-\gamma t}(\hat{H}_t^n)^{1+\varepsilon_0}dt < \infty,$$

and the result is proved. □

Proof of Corollary 2.15. As far as the proof of the weak convergence (2.29) and AO are concerned, there is no difference between the setting of Theorem 2.13(1), where the LP has multiple solutions but the single mode condition holds, and the case where it has a single solution. Therefore this result is a corollary of Theorem 2.13(1). □

5.2 Proof of Propositions 5.3–5.6

In §5.2.1, we provide several useful estimates. Then, Propositions 5.3, 5.4, 5.5 and 5.6 are proved in §5.2.2, §5.2.3, §§5.2.4–5.2.5, and §5.2.6, respectively.

5.2.1 Auxiliary lemmas

Two estimates on the rescaled primitives, $\hat{A}^n_i$ and $\hat{S}^n_{ik}$, are provided in (5.18) and Lemma 5.7, and a certain estimate on the maximum service duration is given in Lemma 5.9.

Because the assumptions on A^n_i and S^n_{ik} are similar, the estimates are stated for $\hat{A}^n_i$ but apply also for $\hat{S}^n_{ik}$. Recall that $\check{a}_i(l)$ are interarrival times of A_i and thus $a^n_i(l) := (\lambda^n_i)^{-1}\check{a}_i(l)$ are the interarrival times of A^n_i. Recall $E[\check{a}_i(1)^{\mathbf{m}}] < \infty$ for a constant $\mathbf{m} > 2$. The first estimate is [10, Theorem 4] which states that for any $2 \le \kappa \le \mathbf{m}$,

$$E[\|\hat{A}^n_i\|^\kappa_t] \le c(1+t)^{\kappa/2} \tag{5.18}$$

for a constant c that does not depend on n or t. The next useful estimate is as follows.

Lemma 5.7. *Let $\nu_1, \nu_2 \in (0,1)$ be such that $\nu_1 \leqslant \nu_2 + \frac{1}{2}$ and assume that $h_0 := (\frac{\mathbf{m}}{2} - 1)\nu_1 - \mathbf{m}\nu_2 > 0$ (note, in particular, that for every $\nu_1 \in (0, \frac{1}{2}]$ there exists $\nu_2 > 0$ satisfying these conditions). Fix $c_1 > 0$. Then for any $h < h_0$,*

$$P(w_{t_0}(\hat{A}^n_i, n^{-\nu_1}) \geqslant c_1 n^{-\nu_2}) \leqslant c n^{-h}(1+t_0)^{\mathbf{m}/2}, \qquad n \in \mathbb{N},\ t_0 \ge 1,$$

where $c = c(c_1, \nu_1, \nu_2, \mathbf{m})$ does not depend on n or t_0.

The proof appears in Appendix A.

Remark 5.8. *We sometimes use the balance equation in the following form, which follows from* (5.4)*, namely*

$$\hat{X}^n_i(t) = \hat{f}^n_i(t) + n^{1/2}\int_0^t \Big(\lambda_i - \sum_k \mu_{ik}\Xi^n_{ik}(s)\Big)ds \tag{5.19}$$

where

$$\hat{f}^n_i(t) := \hat{A}^n_i(t) - \sum_k \hat{S}^n_{ik}(T^n_{ik}(t)) + t\hat{\lambda}^n_i - \sum_k T^n_{ik}(t)\hat{\mu}^n_{ik}.$$

Note that $\hat{F}^n$ of (5.5) *is given by $\hat{F}^n(t) = \sum_i \alpha_i^{-1}\hat{f}^n_i(t)$. Then using the* 1*-Lipschitz property of the trajectories of T^n_{ik}, it is easy to see that both estimates above imply some estimates for $\hat{f}^n_i$. In particular, by* (5.18)*, $\mathbb{E}\|\hat{f}^n_i\|^\kappa_t \le c(1+t)^\kappa$. Moreover, under the assumptions of Lemma 5.7 and the additional assumption that $\nu_2 < \nu_1$, the conclusion of the lemma holds for $\hat{f}^n_i$ and $\hat{F}^n$.*

Next, we give an estimate on the maximal service duration and interarrival time up to a given time. The *time in service by t* of a given job is defined as the time that the job has spent in service up to time t. Let $\mathrm{TIS}(n,i,k,l,t)$ denote the time in service by t of the lth job in activity (i,k). If service to job l has completed by time t then clearly $\mathrm{TIS}(n,i,k,l,t) = u^n_{ik}(l)$, but if it is still in service, $\mathrm{TIS}(n,i,k,l,t) < u^n_{ik}(l)$. Of course, $\mathrm{TIS}(n,i,k,l,t) = 0$ for jobs for which service has not started by t. For $t > 0$ and a real-valued path φ, denote

$$\varLambda(\varphi,t) = \sup\{t_2 - t_1 :\ 0 \le t_1 \le t_2 \le t,\ \varphi_{t_2} = \varphi_{t_1}\}.$$

Then, for activity (i,k), the maximal time in service by time t, namely $\sup_l \mathrm{TIS}(n,i,k,l)$, is bounded above by $\varLambda(S^n_{ik},t)$. We will need an upper bound on the service time as well as the interarrival times, and to this end define

$$e^n_{\max}(t) = \max_{i,k}\varLambda(S^n_{ik},t) \vee \max_i \varLambda(A^n_i,t). \tag{5.20}$$

This process also bounds from above all service durations completed by time t.

Lemma 5.9. *One has $\mathbb{E}[(e^n_{\max})^2] \le cn^{-1}(1+t)$ for a constant c that does not depend on n or t. Moreover, for any $t < \infty$ and $c_1 > 0$, $\mathbb{P}\left(e^n_{\max}(t) \geqslant c_1 n^{\bar{a}-1}\right) \to 0$, as $n \to +\infty$.*

The proof appears in Appendix A.

5.2.2 Uniform integrability

Here we prove Proposition 5.3. We sometimes need to refer to the variable keeping track of the current mode, which in the various policies is defined in slightly different ways according to different sampling times. Denote

$$\mathrm{MODE}^n(t) = \begin{cases} \xi^L & \text{if the current mode is lower workload,} \\ \xi^H & \text{if the current mode is higher workload.} \end{cases}$$

Proof of Proposition 5.3. The proof has three parts; Part 1 appears below, while Parts 2 and 3 are defferred to Appendix A. Fix any one of the sequences of policies $T^n \in \mathcal{A}^n$ for which we attempt to prove AO.

Part 1. This part is concerned with the case where, whenever the workload in the system is sufficiently large, policy $\mathbf{P}$ is active. This covers case 1(a) of Theorem 2.13, where the system has a single mode and the fixed priority policy $\mathbf{P}$ is applied, as well as case 2(a) where the dual mode policy $\mathbf{PP}$ is applied. In this case we will prove that the statement of the lemma holds with $\varepsilon_0 = 1$, and specifically that, under the given sequence, $\mathbb{E}[(\hat{H}^n_t)^2]$ is bounded by a polynomial in t for all n. By (2.11), this is equivalent to the same property holding for $\mathbb{E}[(\hat{W}^n_t)^2]$.

To prove the result in this case, assume that in the single (resp., dual) mode case, the active mode ξ^A (resp., the high workload mode ξ^H) is in canonical form. Thus, provided that the workload in the system exceeds z^* (when applicable), class 1 (resp., 2) is the dual (single) activity class, and server 1 (resp., 2) is the single (dual) activity server. In addition, $p = 2$: class 2 is the HPC. First we provide a bound on the second moment of $\hat{X}^n_2$. In the dual mode case, fix K large enough so that $\hat{X}^n_2(t) \ge K$ implies $\hat{W}^n_t \ge z^* + 1$; in the single mode case let $K = 1$. Given $t > 0$, consider the event $\hat{X}^n_2(t) > K$. Let $\tau = \tau_n(t)$ be defined by

$$\tau = \sup\{s \in [0,t] : \hat{X}^n_2(s) \le K\}.$$

Because the system starts empty, $0 \le \tau \le t$, and because the jumps of the normalized queue length are of size $n^{-1/2}$, $\hat{X}^n_2(\tau) \le K + 1$. Thus

$$\hat{X}^n_2(t) \le \hat{X}^n_2(t) - \hat{X}^n_2(\tau) + K + 1.$$

By (5.3), denoting $C^n(t) = \sum_i \|\hat{A}^n_i\|_t + \sum_{ik} \|\hat{S}^n_{ik}\|_t$,

$$\hat{X}^n_2(t) - \hat{X}^n_2(\tau) \le 2C^n(t) + n^{-1/2}\lambda^n_2(t-\tau) - n^{-1/2}\mu^n_{22}(T^n_{22}(t) - T^n_{22}(\tau)).$$

Because $\hat{X}^n_2 \ge K$ in $[\tau, t]$, the priority policy corresponding to a mode given in canonical form (either ξ^A or ξ^H) is in force after an initial time before the current mode updates and there are class 2 jobs to serve. We can bound the time until server 2 prioritizes class 2 and starts serving it at full rate in $[\tau, t]$ by $e^n_{\max}(t)$. If there currently is a class 2 job being served by server 2, server 2 only serves class 2 jobs for the whole period and

$$T^n_{22}(t) - T^n_{22}(\tau) = t - \tau.$$

If there currently is a class 1 job being served by server 2, because service is non-preemptive, server 2 has to finish this job. If at that time the current mode is ξ^H (resp. ξ^A in the single mode case), server 2 prioritizes class 2 from that point on. If at that time the current mode is ξ^L the workload is then sampled because a service finished at the single activity server. The current mode changes to ξ^H and stays until t. In both cases we get

$$T^n_{22}(t) - T^n_{22}(\tau) \geq t - \tau - e^n_{\max}.$$

Also, in the case under consideration one has $\mu_{22} > \lambda_2$. Hence for all sufficiently large n, $\mu^n_{22} > \lambda^n_{22}$. Moreover, $c_1 := \sup_n n^{-1}\mu^n_{22} < \infty$. Hence

$$\hat{X}^n_2(t) \leq 2C^n(t) + c_1 n^{1/2} e^n_{\max}(t) + K + 1.$$

The above inequality holds also on the complementary event, namely when $\hat{X}^n_2(t) \leq K$. Thus using (5.18) and Lemma 5.9 we obtain

$$\mathbb{E}[\|\hat{X}^n_2\|^2_t] \leq c(1+t). \tag{5.21}$$

In the next step we bound $\hat{W}^n_t$. Let $\tilde{K}$ be a constant that is sufficiently large to ensure that $\hat{X}^n_1(t) \geq \tilde{K}$ implies $\hat{W}^n(t) \geq z^* + 1$ (where again, $\tilde{K} = 1$ in the single mode case). Given $t > 0$, consider the event $\hat{X}^n_1(t) > \tilde{K}$ and let $\sigma = \sigma_n(t)$ be

$$\sigma = \sup\{s \in [0,t] : \hat{X}^n_1(s) \leq \tilde{K}\}.$$

Then clearly $\sigma \in [0,t]$. Because during $[\sigma, t]$ there are at least two jobs of class 1 in the system, both servers are never idle during $[\sigma, t]$. Since one has $\hat{X}^n_1(\sigma) \leq \tilde{K} + n^{-1/2}$, it follows that $\hat{W}^n_\sigma \leq c(1 + \tilde{K} + \hat{X}^n_2(\sigma))$. Hence

$$\hat{W}^n_t \leq \hat{W}^n_t - \hat{W}^n_\sigma + c(1 + \tilde{K} + \|\hat{X}^n_2\|_t).$$

By (5.6) and the nonidling of both servers during $[\sigma, t]$, by which $\hat{L}^n$ remains flat during this interval, we have $\hat{W}^n_t - \hat{W}^n_\sigma = \hat{F}^n_t - \hat{F}^n_\sigma$. Thus by (5.5), for a constant c (that may depend on $\tilde{K}$),

$$\hat{W}^n_t \leq c(C^n(t) + t + 1 + \|\hat{X}^n_2\|_t).$$

Clearly, the above bound is valid also in the complementary event, $\hat{X}^n_1(t) \leq \tilde{K}$. We can therefore apply (5.18) and (5.21), to obtain $\mathbb{E}[(\hat{W}^n_t)^2] \leq c(1+t)$ for some constant c, for all n and t. Consequently the same holds for the second moment of $\hat{H}^n_t$, and the result follows. This completes Part 1 of the proof. The remaining parts appear in Appendix A. □

5.2.3 State space collapse

We now prove Proposition 5.4. Assume that either the active mode ξ^A or the lower workload mode ξ^L (whichever is applicable) is in canonical form. Let

$$\tau = \tau^n_c = \inf\left\{t \geqslant 0 : \hat{X}^n_p(t) \geqslant 2\hat{\Theta}^n\right\}.$$

This random time is used outside this proof with the notation τ_c^n, but in this proof the shorter notation τ is used. Then

$$\mathbb{P}\left(\sup_{t\leqslant t_0}\hat{X}_p^n(t)\geqslant 2\hat{\Theta}^n\right)\leq \mathbb{P}\left(\tau\leqslant t_0\right).$$

We will prove the lemma by showing that the RHS above converges to zero as $n\to\infty$. On the event $\{\tau\leqslant t_0\}$, define

$$\sigma=\sigma^n=\sup\left\{t\leqslant\tau:\,\hat{X}_p^n(t)\leqslant\frac{3\hat{\Theta}^n}{2}\right\}.$$

The proof relies on the fact that, under our policies, when the number of HPC jobs is above Θ^n, it is served at a rate that is enough to deplete the queue in all cases. The first step toward this goal is the following.

Lemma 5.10. *There exist constants $c_1, c_2 > 0$ such that, on $\{\tau\leqslant t_0\}$,*

$$\int_\sigma^\tau\big(\lambda_p-\sum_k\mu_{pk}\Xi_{pk}^n(t)\big)dt\leqslant -c_1(\tau-\sigma)+c_2e_{\max}^n, \tag{5.22}$$

with $e_{\max}^n$ defined in (5.20).

Proof. First, by definition of σ, and τ, and the fact that jumps of $\hat{X}^n$ are of size $n^{-1/2}<\hat{\Theta}^n/2$ for large n, we obtain

$$\inf_{t\in[\sigma,\tau]}\hat{X}_p^n(t)\geqslant\hat{\Theta}^n.$$

We address here one case only; the proof under the remaining cases is defferred to the appendix.

- **Case 1(a)**: In this case, we use the **P** policy, which is single mode. Because the active mode is in canonical form and $i_1(\xi^A)=p$, $p=2$ and server 2 prioritizes class 2. It is possible that server 2 is busy with the "wrong" class of job at time σ but as soon as that job finishes, server 2 will only serve the class 2 jobs in the system:

  $$\int_\sigma^\tau\Xi_{22}^n(t)dt\geqslant\tau-\sigma-e_{\max}^n.$$

  Thus,

  $$\int_\sigma^\tau\big(\lambda_2-\sum_k\mu_{2k}\Xi_{2k}^n(t)\big)dt\leqslant(\tau-\sigma)(\lambda_2-\mu_{22})+\mu_{22}e_{\max}^n.$$

  To show that $\lambda_2-\mu_{22}<0$, note that since the active mode is in canonical form, $\frac{\lambda_1}{\alpha_1}>\beta_1$, which implies $\frac{\lambda_2}{\alpha_2}<\beta_2$ by (3.1).

The remaining cases appear in Appendix A. □

Now that Lemma 5.10 has been proved in all cases, the proof of Proposition 5.4 does not need to differentiate between them.

Proof of Proposition 5.4. Let $\nu_2=1/2-\bar{a}\leqslant 1/4$ and $\nu_1\in(\nu_2,1/2)$. Recall that $\hat{\Theta}^n=n^{-1/2}\lceil n^{\bar{a}}\rceil$, as defined in (2.32), so that $\hat{\Theta}^n=n^{-1/2}\lceil n^{1/2-\nu_2}\rceil\geqslant n^{-\nu_2}$. Notice that, as required in Lemma 5.7, $\nu_1\leqslant\nu_2+\frac{1}{2}$. Let us introduce the event

$$\Omega_1=\left\{\tau\leqslant t_0,\,\hat{X}_p^n(\tau)-\hat{X}_p^n(\sigma)\geqslant\frac{\hat{\Theta}^n}{4},\inf_{t\in[\sigma,\tau]}\hat{X}_p^n(t)\geqslant\hat{\Theta}^n\right\}.$$

For n large enough one has $\hat{X}^n_p(\sigma) < \frac{7}{4}\hat{\Theta}^n$. Consequently $\{\tau \le t_0\} = \Omega_1$. We obtain

$$\mathbb{P}(\tau \leqslant t_0) = \mathbb{P}(\Omega_1) = \mathbb{P}\left(\Omega_1 \cap \{\tau - \sigma \leqslant n^{-\nu_1}\}\right) + \mathbb{P}\left(\Omega_1 \cap \{\tau - \sigma > n^{-\nu_1}\}\right). \tag{5.23}$$

Picking up on (5.19), by Lemma 5.10

$$\begin{aligned}\hat{X}^n_p(\tau) - \hat{X}^n_p(\sigma) &= \hat{f}^n_p(\tau) - \hat{f}^n_p(\sigma) + \sqrt{n}\int_\sigma^\tau \big(\lambda_p - \sum_k \mu_{pk}\Xi^n_{pk}(t)\big)dt \\ &\leqslant \hat{f}^n_p(\tau) - \hat{f}^n_p(\sigma) - c\sqrt{n}(\tau - \sigma) + c\sqrt{n}e^n_{\max}.\end{aligned}$$

Notice that on the event $\{\tau - \sigma \leqslant n^{-\nu_1},\, \tau \leqslant t_0\}$,

$$\hat{X}^n_p(\tau) - \hat{X}^n_p(\sigma) \leqslant w_{t_0}(\hat{f}^n_p, n^{-\nu_1}) + c\sqrt{n}e^n_{\max}.$$

Thus

$$\begin{aligned}\mathbb{P}\left(\Omega_1 \cap \{\tau - \sigma \leqslant n^{-\nu_1}\}\right) &\leqslant \mathbb{P}\left(w_{t_0}(\hat{f}^n_p, n^{-\nu_1}) + e^n_{\max}\sqrt{n} \geqslant \frac{n^{-\nu_2}}{2}\right) \\ &\leqslant \mathbb{P}\left(w_{t_0}(\hat{f}^n_p, n^{-\nu_1}) \geqslant \frac{n^{-\nu_2}}{4}\right) + \mathbb{P}\left(ce^n_{\max}\sqrt{n} \geqslant \frac{n^{-\nu_2}}{4}\right).\end{aligned}$$

Recall that $\nu_1 \in (\nu_2, 1/2)$, so $n^{-\nu_1} < n^{-\nu_2}$. By Lemma 5.7 and Remark 5.8, the first term converges to zero. By lemma 5.9 and $\nu_2 = 1/2 - \bar{a}$,

$$\mathbb{P}\left(e^n_{\max} \geqslant \frac{c}{4}n^{-\nu_2 - 1/2}\right) = \mathbb{P}\left(e^n_{\max} \geqslant \frac{c}{4}n^{\bar{a}-1}\right) \to 0.$$

For the second term in (5.23), notice that on $\{\tau - \sigma > n^{-\nu_1},\, \tau \leqslant t_0\}$,

$$\hat{X}^n_p(\tau) - \hat{X}^n_p(\sigma) \leqslant 2\|\hat{f}^n_p(t)\|_{t_0} + e^n_{\max}\sqrt{n} - c\sqrt{n}n^{-\nu_1},$$

and

$$\hat{X}^n_p(\tau) - \hat{X}^n_p(\sigma) \geqslant \frac{\hat{\Theta}^n}{2} - n^{-1/2}.$$

Hence

$$\mathbb{P}\left(\Omega_1 \cap \{\tau - \sigma > n^{-\nu_1}\}\right) \leqslant \mathbb{P}\left(2\|\hat{f}^n_p\|_{t_0} + e^n_{\max}\sqrt{n} \geqslant c\sqrt{n}n^{-\nu_1} + \frac{\hat{\Theta}^n}{2} - n^{-1/2}\right).$$

By Lemma 5.9, $e^n_{\max}$ is smaller than $n^{\bar{a}-1} = o(n^{-1/2})$. By Remark 5.8, $2\|\hat{f}^n_p\|_{t_0}$ is a tight sequence of RVs (for t_0 fixed). Because $\nu_1 < 1/2$, $\sqrt{n}n^{-\nu_1} \to +\infty$. The claim follows. □

5.2.4 Boundary behavior

The goal of this section and the following one is to prove Proposition 5.5. The key is the follows lemma, which states, roughly speaking, that $\hat{L}^n$ is approximately the boundary term corresponding to $\hat{F}^n$.

Let $c_3 = \frac{3}{\alpha_1 \wedge \alpha_2}$.

Lemma 5.11. *Fix $t_0 > 0$. With*

$$\bar{R}^n_t = \int_0^t \mathbb{1}_{\hat{W}^n_s \geqslant c_3 \hat{\Theta}^n} d\hat{L}^n_s, \tag{5.24}$$

$\bar{R}^n_{t_0} \to 0$ *in probability as $n \to \infty$.*

This lemma is proved in the next subsection. Let us show how Proposition 5.5 now follows.

Proof of Proposition 5.5. In addition to $\bar{R}^n$ just introduced, we shall need the following definitions:

$$\begin{aligned} Z^n_t &= \max(\hat{W}^n_t - c_3\hat{\Theta}^n,\, 0), \\ \widetilde{R}^n_t &= \int_0^t \mathbb{1}_{\hat{W}^n_s < c_3\hat{\Theta}^n} d\hat{L}^n_s, \\ \Delta^n_t &= Z^n_t - \hat{W}^n_t. \end{aligned}$$

By (5.6) and definition of the new processes, one has

$$Z^n_t = \hat{F}^n_t + \Delta^n_t + \bar{R}^n_t + \widetilde{R}^n_t.$$

Consider the pair $(Z^n, \widetilde{R}^n)$. The first component is nonnegative. The second is nonnegative, nondecreasing, and moreover,

$$\int_{[0,\infty)} Z^n_t d\widetilde{R}^n_t = \int_{[0,\infty)} \max(\hat{W}^n_t - c_3\hat{\Theta}^n,\, 0)\mathbb{1}_{\hat{W}^n_t < c_3\hat{\Theta}^n} d\hat{L}^n_t = 0.$$

Hence by Skorohod's Lemma, $(Z^n, \widetilde{R}^n) = \Gamma(\hat{F}^n + \Delta^n + \bar{R}^n)$. Hence

$$\hat{W}^n = \Gamma_1(\hat{F}^n + \Delta^n + \bar{R}^n) - \Delta^n, \qquad \hat{L}^n = \Gamma_2(\hat{F}^n + \Delta^n + \bar{R}^n) + \bar{R}^n. \tag{5.25}$$

Recall that we are considering a subsequence along which (according to the assumptions and to Lemma 5.2), $(\hat{A}^n, \hat{S}^n, T^n, \hat{F}^n) \Rightarrow (A, S, T, F)$, with F as in (5.8). Moreover, by the definition of Z^n and Δ^n, we have $0 \le -\Delta^n_t \le c_3\hat{\Theta}^n$, whereas by Lemma 5.11, $\bar{R}^n \to 0$ in probability. By the continuity of the map Γ we therefore conclude from (5.25) that, on the same subsequence,

$$(\hat{A}^n, \hat{S}^n, T^n, \hat{F}^n, \hat{W}^n, \hat{L}^n) \Rightarrow (A, S, T, F, W, L),$$

where $(W, L) = \Gamma(F)$. This completes the proof. □

5.2.5 Proof of Lemma 5.11

We first explain how the proof changes between the cases.

- Under the **P** policy, both servers can process the low priority jobs. This means that idling can only occur if the low priority class has few jobs and in that case the total number of jobs is also low by Proposition 5.4.

- Under the $\mathbf{T}_2$ rule, no server can incur idleness when the high priority class is above the threshold. In addition, the high priority class takes a small time to reach above Θ^n and does not empty below two jobs after that time with high probability unless the total workload is close to 0 (Lemma 5.13). This means that neither server will idle except when there are almost no jobs in the system.
- Under the **PP** policy, Proposition 5.4 is still enough to prove Lemma 5.11 in the same way as in the single mode case **P**.
- Under the $\mathbf{T}_2\mathbf{T}_2$ policy, Lemma 5.12 and 5.13 hold. Because of that, Lemma 5.11 holds for the same reason as in the non switching case $\mathbf{T}_2$.
- When using a **P** rule, the high priority class could become zero with a lot of LPC jobs in the system, and switching to the $\mathbf{T}_2$ rule could lead to idleness even though there are a lot of low priority jobs (approximately $\alpha_q z^*$). Thus we introduce the $\mathbf{T}_1$ rule in place of the **P** rule in this case. This ensures that the number of high priority jobs does not decrease too much during the corresponding period.

In some cases, some idleness can occur when the high priority class starts with too few jobs but in those cases, it takes a small time to leave such states.

In cases 1(a) and 2(a), Lemma 5.11 is a direct consequence of the state space collapse. Let us introduce two random times and a lemma that we will use in cases 1(b), 2(b)–(d). Fix $t_0 > 0$. Let

$$\rho^n = \inf\left\{t \geqslant 0 :\, \hat{X}_p^n(t) \geqslant \hat{\Theta}^n\right\} \wedge t_0,$$

$$\tau_r^n = \inf\left\{t > \rho^n :\, \hat{X}_p^n(t) = 2n^{-1/2}, \text{ and } \hat{X}_q^n(t) \geqslant \hat{\Theta}^n\right\}.$$

Lemma 5.12. *If any of the following conditions hold, we have*

$$\bar{R}_{\rho^n}^n \to 0 \textit{ in probability.} \tag{5.26}$$

- *Case 1(b),* (2.20) *and* $i_2(\xi^A) = p$.
- *Case 2(b),* (2.21) *and* $i_2(\xi^L) = i_2(\xi^H) = p$.
- *Cases 2(c) and 2(d),* (2.21) *and* $i_1(\xi^L) = i_2(\xi^H)$.

Lemma 5.13. *Under the same assumptions as the previous lemma,*

$$\mathbb{P}\left(\tau_r^n \leqslant t_0\right) \to 0. \tag{5.27}$$

The proofs of Lemmas 5.12 and 5.13 appear in Appendix A.

Remark 5.14. *The above two lemmas do not address cases 1(a) and 2(a). The reason for this is that in these cases we can directly prove Lemma 5.11 when using a* **P** *or* **PP** *policy.*

Proof of Lemma 5.11. Here we only treat one case, deferring the remaining cases to the appendix.

- **Case 1(a), P policy:**

In this case $p = 2$, and no server can be idle when $X_1 \geqslant 2$. Thus

$$\int_0^{t_0} \mathbb{1}_{\hat{X}_1^n(t) \geqslant 2n^{-1/2}} d\hat{L}_t^n = 0. \tag{5.28}$$

For any $\delta > 0$,

$$\begin{aligned}
\mathbb{P}\left(\bar{R}_{t_0}^n \geqslant \delta\right) &\leqslant \mathbb{P}\left(\int_0^{t_0} (\mathbb{1}_{\hat{X}_2^n(t) \geqslant 2\hat{\Theta}^n} + \mathbb{1}_{\hat{X}_1^n(t) \geqslant 2n^{-1/2}}) d\hat{L}_t^n \geqslant \delta\right) \\
&= \mathbb{P}\left(\int_0^{t_0} \mathbb{1}_{\hat{X}_2^n(t) \geqslant 2\hat{\Theta}^n} d\hat{L}_t^n \geqslant \delta\right) \\
&\leqslant \mathbb{P}\left(\mathbb{1}_{\sup_{t \leqslant t_0} \hat{X}_2^n(t) \geqslant 2\hat{\Theta}^n} \hat{L}^n(t_0) \geqslant \delta\right) \\
&\leqslant \mathbb{P}\left(\sup_{t \leqslant t_0} \hat{X}_2^n(t) \geqslant 2\hat{\Theta}^n\right) \\
&= \mathbb{P}\left(\tau_c^n \leqslant t_0\right).
\end{aligned}$$

By Proposition 5.4,

$$\mathbb{P}\left(\tau_c^n \leqslant t_0\right) \to 0. \tag{5.29}$$

The treatment of the remaining cases appears in Appendix A. □

5.2.6 Fast switching

Proof of Proposition 5.6. Fix t_0 and $\varepsilon > 0$. Assume that ξ^L is in canonical form. Then $(i^l, k^l) = (2, 1)$. Let

$$\begin{aligned}
\tau_f &= \inf\left\{t \geqslant 0 : \int_0^t \mathbb{1}_{\hat{W}_t^n \leqslant z^* - \varepsilon} dT_{21}^n(t) > 0\right\}, \\
t_{\min} &= \sup\left\{t \leqslant \tau_f : \hat{W}_t^n \geqslant z^*\right\}, \\
t_{\max} &= \inf\left\{t \geqslant t_{\min} : \hat{W}_t^n \leqslant z^* - \varepsilon\right\}, \\
\tau_1 &= \inf\left\{t \geqslant t_{\min} : \mathrm{MODE}^n(t) = \xi^L\right\},
\end{aligned}$$

where the dependence on n is suppressed. The first statement of the lemma will be proved once we show $\mathbb{P}(\tau_f \leqslant t_0) \to 0$. We omit the proof of the second statement, which is similar. To this end, let

$$\kappa^n = \inf\{s \geqslant 0 \text{ s.t. } \exists t \leqslant t_0,\, \hat{W}_{t-s}^n \geqslant z^*,\, \hat{W}_t^n \leqslant z^* - \varepsilon\}.$$

If $\sup_{t \leqslant t_0} \hat{W}_t^n \leqslant z^*$, the current mode never changes so the single activity server is dedicated to only one class for the whole period and the non-basic activity is never used. Thus,

$$\{\tau_f \leqslant t_0\} \subset \{t_{\max} \leqslant t_0\}.$$

Note that we have used the fact that, for all of our policies, jobs are never routed to a non basic activity of the current mode.

The remainder of the argument is based on the following fact, which must be argued separately for each case. This is concerned with the difference $\tau_1 - t_{\min}$, which while case dependent, can in all cases be shown to satisfy

$$\lim_{n\to+\infty} \mathbb{P}\left(\tau_1 - t_{\min} > e^n_{\max}\right) = 0. \tag{5.30}$$

By Proposition 5.5, $\hat{W}^n$ is C-tight. Hence for any constant $c > 0$, $\mathbb{P}\left(c\kappa^n \leqslant n^{\bar{a}-1}\right) \to 0$. On the other hand, by Lemma 5.9, $\mathbb{P}\left(e^n_{\max} \geqslant n^{\bar{a}-1}\right) \to 0$. Thus

$$\mathbb{P}\left(e^n_{\max} \geqslant c\kappa^n\right) \to 0. \tag{5.31}$$

In order for $\int_0^{t_0} \mathbb{1}_{W^n_t \leqslant z^*-\varepsilon} dT^n_{21}(t)$ to become positive, $t_{\max}$ must be smaller than t_0 and

$$t_{\max} - \tau_1 < e^n_{\max}.$$

It is not possible for $t_{\max} - \tau_1 \geqslant e^n_{\max}$ to occur on $\{\tau_f \leqslant t_0\}$. If that were the case, server 1 would necessarily finish service of the job it was in the process of serving at time τ_1 before the workload has time to reach $z^* - \varepsilon$. When $\text{MODE}^n(t) = \xi^L$ in canonical form, server 1 can only take new class 1 jobs regardless of the rule. Even if class 1 has no job in the queue, the non basic activity is not used after the possible residual job that was in service at time τ_1. In addition, by definition of $t_{\min}$, this is the last time the mode switches from upper to lower workload before τ_f. This would prevent $\int_0^{\tau_f} \mathbb{1}_{\hat{W}^n_t \leqslant z^*-\varepsilon} dT^n_{21}(t)$ from becoming positive and is also the reason why $t_{\max}$ needs to be smaller than t_0 for τ_f to be smaller than t_0.

By definition of $t_{\max}$ and κ^n,

$$\mathbb{P}\left(t_{\max} \leqslant t_0,\, t_{\max} - t_{\min} < \kappa^n\right) = 0.$$

We next show that

$$\lim_{n\to+\infty} \mathbb{P}\left(t_{\max} \leqslant t_0,\, \tau_1 \geqslant t_{\max}\right) = 0. \tag{5.32}$$

We have

$$\begin{aligned}
\mathbb{P}\left(t_{\max} \leqslant t_0,\, \tau_1 \geqslant t_{\max}\right) &= \mathbb{P}\left(t_{\max} \leqslant t_0,\, \tau_1 \geqslant t_{\max}, t_{\max} - t_{\min} \geqslant \kappa^n\right)\\
&\leqslant \mathbb{P}\left(\tau_1 \geqslant t_{\min} + \kappa^n\right)\\
&\leqslant \mathbb{P}\left(\tau_1 \geqslant t_{\min} + \kappa^n, e^n_{\max} \leqslant \kappa^n\right) + \mathbb{P}\left(\tau_1 \geqslant t_{\min} + \kappa^n, e^n_{\max} > \kappa^n\right)\\
&\leqslant \mathbb{P}\left(\tau_1 \geqslant t_{\min} + e^n_{\max}\right) + \mathbb{P}\left(e^n_{\max} > \kappa^n\right).
\end{aligned}$$

Both terms converge to zero, the first by (5.30), and the second by (5.31).

We can now prove the lemma based on (5.32), (5.30) and (5.31). We have

$$\begin{aligned}
\mathbb{P}\left(\tau_f \leqslant t_0\right) &= \mathbb{P}\left(\tau_f \leqslant t_0,\, t_{\max} \leqslant t_0\right)\\
&= \mathbb{P}\left(\tau_f \leqslant t_0,\, t_{\max} \leqslant t_0,\, \tau_1 \geqslant t_{\max}\right) + \mathbb{P}\left(\tau_f \leqslant t_0,\, t_{\max} \leqslant t_0,\, \tau_1 \leqslant t_{\max},\, t_{\max} - \tau_1 < e^n_{\max}\right)\\
&\leqslant \mathbb{P}\left(t_{\max} \leqslant t_0, \tau_1 \geqslant t_{\max}\right) + \mathbb{P}\left(t_{\max} \leqslant t_0,\, t_{\max} - t_{\min} \leqslant 2e^n_{\max}\right) + \mathbb{P}\left(\tau_1 - t_{\min} > e^n_{\max}\right)\\
&\leqslant \mathbb{P}\left(t_{\max} \leqslant t_0, \tau_1 \geqslant t_{\max}\right) + \mathbb{P}\left(\kappa^n \leqslant 2e^n_{\max}\right) + \mathbb{P}\left(t_{\max} \leqslant t_0,\, t_{\max} - t_{\min} < \kappa^n\right)\\
&\quad + \mathbb{P}\left(\tau_1 - t_{\min} > e^n_{\max}\right).
\end{aligned}$$

The first term goes to zero by (5.32), the second by (5.31), the third is zero, and the fourth goes to zero by (5.30). This completes the proof.

It remains to prove (5.30). As noted above, the proof differs by case.

Case 2(a): The current mode changes to ξ^L after the first service completion of a job at the single activity server for ξ^H (which is server 2) at or after $t_{\min}$. Note that $t_{\min}$ must correspond to a service completion. If $t_{\min}$ corresponds to a service completion at server 2, the $\tau_1 = t_{\min}$. If $t_{\min}$ corresponds to a service completion at server 1, then the mode will not cange, and τ_1 will correspond to the next service completion at server 2. This will occur before $t_{\min} + e^n_{\max}$ if server 2 is busy (serving class 1) at $t_{\min}$. Note that, on $\{\tau^n_c \geqslant t_0\}$, $\hat{X}^n_2(t_{\min}) < 2\hat{\Theta}^n$, so that

$$X^n_1(t_{\min}) \geqslant \alpha_1(W^n_{t_{\min}} - 2\alpha_2^{-1}\Theta^n) \geqslant \alpha_1(n^{1/2}z^* - 1 - 2\alpha_2^{-1}\Theta^n) \geqslant 2,$$

so server 2 is busy at $t_{\min}$. Thus

$$\mathbb{P}\left(\tau_1 - t_{\min} > e^n_{\max}\right) \leqslant \mathbb{P}\left(\tau^n_c \leqslant t_0\right),$$

which converges to zero by Proposition 5.4. This proves (5.30).

Case 2(b): The current mode changes after the first service completion of a job at the single activity server (which is server 2) at or after $t_{\min}$. Server 2 is dedicated to HPC jobs. Under $\{\tau^n_r \geqslant t_0, \rho^n \leqslant t_{\min}\}$, there are at least 2 HPC jobs in the system at time $t_{\min}$ so the single activity server cannot be idling at that time. Thus

$$\mathbb{P}\left(\tau_1 - t_{\min} > e^n_{\max}\right) \leqslant \mathbb{P}\left(\tau^n_r \leqslant t_0\right) + \mathbb{P}\left(\rho^n > t_{\min}\right).$$

By Lemma 5.13, $\mathbb{P}\left(\tau^n_r \leqslant t_0\right)$ converges to zero. In addition,

$$\begin{aligned}\mathbb{P}\left(\rho^n > t_{\min}\right) &= \mathbb{P}\left(\rho^n > t_{\min}, t_{\min} > \tilde{\tau}\right) + \mathbb{P}\left(\rho^n > t_{\min}, t_{\min} \leqslant \tilde{\tau}\right)\\ &\leqslant \mathbb{P}\left(\rho^n > \tilde{\tau}\right) + \mathbb{P}\left(t_{\min} \leqslant \tilde{\tau}\right)\\ &\leqslant \mathbb{P}\left(\rho^n > \tilde{\tau}, \tau^n_c \geqslant t_0\right) + \mathbb{P}\left(\rho^n > \tilde{\tau}, \tau^n_c < t_0\right) + \mathbb{P}\left(\hat{X}^n_2(t_{\min}) < \alpha_2 z^*/2\right)\\ &\leqslant \mathbb{P}\left(\rho^n > \tilde{\tau}, \tau^n_c \geqslant t_0\right) + 2\mathbb{P}\left(\tau^n_c < t_0\right).\end{aligned}$$

where $\tilde{\tau}$ is defined in the proof of Lemma 5.12. Both terms converge to zero, the first by Lemma 5.13, and the second by Proposition 5.4. This proves (5.30).

Cases 2(c) and 2(d): The current mode changes after the first service completion or arrival of a low or high priority job after $t_{\min}$ so under $\{t_{\max} \leqslant t_0\}$,

$$\tau_1 - t_{\min} \leqslant e^n_{\max}.$$

Thus (5.30) follows from Lemma 5.9. □

A Proofs of lemmas

Proof of Lemma 5.7. In this proof, $\mathbf{m}$ is written as m. Note first that it suffices to prove

$$\mathbb{P}(w_{t_0}(\hat{A}^n, n^{-\nu_1}) \geqslant n^{-\nu_2}) \leqslant cn^{-h_0}(1+t_0)^{m/2}, \qquad n \in \mathbb{N}, \tag{A.1}$$

where $c = c(\nu_1, \nu_2, m)$ does not depend on n or t_0. Indeed, if $c_1 \geqslant 1$, the result follows directly from (A.1). If $c_1 \in (0,1)$, let $\bar{\nu}_2 > \nu_2$ be such that it satisfies all hypotheses of the lemma. Namely $\bar{\nu}_2 \in (0,1)$, $\nu_1 \leq \bar{\nu}_2 + \frac{1}{2}$, $\bar{h}_0 := (\frac{m}{2} - 1)\nu_1 - m\bar{\nu}_2 > 0$. Then by (A.1), for n such that $c_1 n^{-\nu_2} \geqslant n^{-\bar{\nu}_2}$,

$$\mathbb{P}(w_{t_0}(\hat{A}^n, n^{-\nu_1}) \geqslant c_1 n^{-\nu_2}) \leqslant c(1+t_0)^{\frac{m}{2}} n^{-\bar{h}_0}.$$

Moreover, $\bar{h}_0$ can be made arbitrarily close to h_0 by choosing $\bar{\nu}_2$ close to ν_2. This shows that the desired inequality holds for all large n, and by making $c = c(c_1, \nu_1, \nu_2, m)$ larger, for all $n \in \mathbb{N}$.

We now prove (A.1). As before, c denotes a positive constant whose value may change from line to line; here it may depend on (ν_1, ν_2, m) but not on n, t_0. By (5.1),

$$\hat{A}^n(t) - \hat{A}^n(s) = n^{-1/2}(A^n(t) - A^n(s)) - n^{-1/2}\lambda^n(t-s).$$

Thus

$$\mathbb{P}(w_{t_0}(\hat{A}^n, n^{-\nu_1}) \geqslant n^{-\nu_2}) \leqslant \mathbb{P}(A^n(t_0) > 2\lambda^n t_0) + \mathbb{P}(\Omega^n_+) + \mathbb{P}(\Omega^n_-), \tag{A.2}$$

where

$$\Omega^n_+ = \{A^n(t_0) \leqslant 2\lambda^n t_0, \exists s, t \in [0, t_0], 0 \leqslant t - s \leqslant n^{-\nu_1}, A^n(t) - A^n(s) \geqslant n^{\frac{1}{2}-\nu_2} + \lambda^n(t-s)\},$$

$$\Omega^n_- = \{A^n(t_0) \leqslant 2\lambda^n t_0, \exists s, t \in [0, t_0], 0 \leqslant t - s \leqslant n^{-\nu_1}, A^n(t) - A^n(s) \leqslant -n^{\frac{1}{2}-\nu_2} + \lambda^n(t-s)\}.$$

Consider the event Ω^n_+. Because $a^n(l)$ are the interarrival times of A^n, on this event there must exist $l^0 \leqslant 2\lambda^n t_0$ and $R \leqslant n^{-\nu_1}$ such that

$$\sum_{l=l^0}^{l^0 + n^{\frac{1}{2}-\nu_2} + \lambda^n R} a^n(l) \leqslant R.$$

Recall that $\mathbb{E}\check{a}(l) = 1$. Letting $\bar{a}^n(l) = a^n(l) - (\lambda^n)^{-1}$, we have $\mathbb{E}\bar{a}^n(l) = 0$. Then taking $r = \lambda^n R$, using $\lambda^n \leqslant c_1 n$ where $c_1 = \sup \frac{\lambda^n}{n} < \infty$, we have

$$\begin{aligned}
\mathbb{P}(\Omega^n_+) &\leqslant \mathbb{P}(\exists j \leqslant 2\lambda^n t_0, \exists r \leqslant n^{-\nu_1}\lambda^n, \sum_{l=l^0+1}^{l^0 + n^{\frac{1}{2}-\nu_2} + r} a^n(l) \leqslant (\lambda^n)^{-1} r) \\
&\leqslant \mathbb{P}(\exists j \leqslant 2c_1 n t_0, \exists r \leqslant c_1 n^{1-\nu_1}, \sum_{l=l^0+1}^{l^0 + n^{\frac{1}{2}-\nu_2} + r} \bar{a}^n(l) \leqslant -(\lambda^n)^{-1} n^{\frac{1}{2}-\nu_2}).
\end{aligned}$$

Let $M^n_{l^1} = \sum_{l=1}^{l^1} \bar{a}^n(l)$, $l^1 = 0, 1, 2, \ldots$, and note that it is a martingale. For a real-valued function X on $\mathbb{Z}_+$ let

$$\operatorname{osc}(X, l_1, l_2) = \max\{|X(l_3) - X(l_4)| : l_3, l_4 \in [l_1, l_2]\}, \qquad 0 \leq l_1 \leq l_2.$$

Then using $\frac{1}{2} - \nu_2 \leqslant 1 - \nu_1$ and denoting $\rho = [c_1 n^{1-\nu_1}]$,

$$\begin{aligned}
\mathbb{P}(\Omega^n_+) &\leqslant \mathbb{P}(\exists j \leqslant 2c_1 n t_0, \operatorname{osc}(M^n, l^0, l^0 + 2c_1 n^{1-\nu_1}) \geqslant c_1^{-1} n^{-\frac{1}{2}-\nu_2}) \\
&\leqslant \mathbb{P}(\exists j \in [0, 2c_1 n t_0] \cap \{0, \rho, 2\rho, \ldots\},\ \operatorname{osc}(M^n, l^0, l^0 + \rho) \geqslant c_1^{-1} n^{-\frac{1}{2}-\nu_2}/3) \\
&\leqslant 1 - (1 - \mathbb{P}(\|M^n\|_\rho \geqslant c_1^{-1} n^{-\frac{1}{2}-\nu_2}/6))^{c t_0 n^{\nu_1}}.
\end{aligned}$$

Because $n^{-1}\lambda^n \to \lambda > 0$, we have the lower bound $\lambda^n > cn$ for some $c > 0$ and all large n. Hence Burkholder's inequality shows that

$$\mathbb{E}(\|M^n\|_\rho)^m \leqslant c\mathbb{E}(\rho|\bar{a}^n(1)|^2)^{\frac{m}{2}} \leqslant c n^{(1-\nu_1)\frac{m}{2}} (\lambda^n)^{-m} \leqslant c n^{-(1+\nu_1)\frac{m}{2}}.$$

Hence for any $a > 0$, $\mathbb{P}(\|M^n\|_\rho > a) \leqslant ca^{-m}n^{-(1+\nu_1)\frac{m}{2}}$, and therefore

$$\mathbb{P}(\Omega_+^n) \leqslant 1 - (1 - cn^{-(\nu_1 - 2\nu_2)\frac{m}{2}})^{ct_0 n^{\nu_1}}.$$

Note that $\nu_1 > 2\nu_2$ by the assumption $h_0 > 0$. Now, if $\varepsilon \in (0, \frac{1}{2})$ and $a > 0$ then, with $c_2 = 2\log 2$, $1 - (1-\varepsilon)^a \leqslant 1 - e^{-c_2 a\varepsilon} \leqslant c_2 a\varepsilon$. This gives

$$\mathbb{P}(\Omega_+^n) \leqslant cn^{-(\nu_1 - 2\nu_2)\frac{m}{2}} t_0 n^{\nu_1} = ct_0 n^{-h_0}.$$

By a similar argument, the same estimate holds for $\mathbb{P}(\Omega_-^n)$. An application of (5.18) gives

$$\mathbb{P}(A^n(t_0) > 2\lambda^n t_0) \leqslant \mathbb{P}(\|\hat{A}^n\|_{t_0} > cn^{1/2}) \leqslant c\frac{\mathbb{E}(\|\hat{A}^n\|_{t_0}^m)}{n^{m/2}} \leqslant cn^{-\frac{m}{2}}(1+t_0)^{\frac{m}{2}}.$$

Hence by (A.2),

$$\mathbb{P}(w_{t_0}(\hat{A}^n, n^{-\nu_1}) \geqslant n^{-\nu_2}) \leqslant ct_0 n^{-h_0} + c(1+t_0)^{\frac{m}{2}} n^{-\frac{m}{2}}.$$

It follows from $\nu_1, \nu_2 \in (0,1)$ that $h_0 < \frac{m}{2}$. The result follows. □

Proof of Lemma 5.9. Fix i, k. Denote $Y_t^n = \Lambda(S_{ik}^n, t)$. For any $u > 0$, if $Y_t^n \geq u$ then there must exist $0 \leq t_1 < t_2 \leq t$ with $t_2 - t_1 = u$ and $S_{ik}^n(t_2-) = S_{ik}^n(t_1)$, hence by the definition (5.1) of $\hat{S}_{ik}^n$,

$$\hat{S}_{ik}^n(t_1) - \hat{S}_{ik}^n(t_2-) = n^{-1/2}\mu_{ik}^n(t_2 - t_1) \geq c_2 n^{1/2} u,$$

for some constant c_2 that depends only on the sequence $\{\mu_{ik}^n\}$. Hence $Y_t^n \leq 2c_2^{-1} n^{-1/2}\|\hat{S}_{ik}^n\|_t$. The first result now follows from (5.18) with $\kappa = 2$.

To prove the second statement we will proceed in two steps. First, we will show that the number of interarrival times involved in the maximum is at most cn with probability going to 1, then show that the maximum over cn variables has the right order of magnitude.

To simplify the notation, fix (i,k) and remove them from the notation of S_{ik}^n, u_{ik}^n, $\hat{\mu}_{ik}^n$, etc. The claim will be proved for $e_{\max}^n(t)$ defined as in (5.20) but without maximizing over (i,k); clearly, this is sufficient. Note that

$$S^n(t) = \sup\left\{s \geqslant 0 : \sum_{l=1}^{s} u^n(l) \leqslant t\right\}.$$

Let

$$K^n = \inf\left\{s \geqslant 0 : \sum_{l=1}^{s} u^n(l) \geqslant t_0\right\}.$$

Then

$$e_{\max}^n(t_0) \leqslant \sup\left\{u^n(s) : \sum_{l=1}^{s-1} u^n(l) \leqslant t_0\right\} \leqslant \sup\left\{u^n(s) :\ s \leqslant K^n\right\}.$$

For any $c_2 \geqslant 0$,

$$\begin{aligned}\mathbb{P}\left(K^n \geqslant c_2 n\right) &\leqslant \mathbb{P}\left(\sum_{p=1}^{c_2 n} u^n(p) \leqslant t_0\right)\\ &= \mathbb{P}\left(\sum_{p=1}^{c_2 n} \check{u}(p) \leqslant \mu^n t_0\right)\\ &= \mathbb{P}\left(\frac{1}{c_2 n}\sum_{p=1}^{c_2 n} \check{u}(p) \leqslant \frac{\mu t_0}{c_2} + \frac{\hat{\mu}^n t_0}{c_2} n^{-1/2}\right).\end{aligned}$$

If $c_2 > \mu t_0$, by the law of large numbers, the RHS converges to 0 as $n \to \infty$. Next, by independence of service times, for any $c > 0$,

$$U_n := \mathbb{P}\left(\max_{l \leqslant c_2 n} u^n(l) \geqslant c n^{\bar{a}-1}\right) = 1 - \mathbb{P}\left(u^n(1) \leqslant c n^{\bar{a}-1}\right)^{c_2 n}.$$

Using $1-(1-x)^n \leqslant c_3 n x$, letting $c_4 = c_2 c_3$ and denoting $\bar{\varepsilon} = \mathbf{m} - 2 > 0$, we obtain

$$\begin{aligned}U_n &\leqslant c_4 n \mathbb{P}\left(u^n(1) \geqslant c n^{\bar{a}-1}\right)\\ &\leqslant c_4 n \mathbb{P}\left(\check{u}(1) \geqslant c\mu^n n^{\bar{a}-1}\right)\\ &= c_4 n \mathbb{P}\left((\check{u}(1))^{2+\bar{\varepsilon}} \geqslant \left(c\mu^n n^{\bar{a}-1}\right)^{2+\bar{\varepsilon}}\right)\\ &\leqslant c_4 n \frac{\mathbb{E}\left[(\check{u}(1))^{2+\bar{\varepsilon}}\right]}{\left(c\mu^n n^{\bar{a}-1}\right)^{2+\bar{\varepsilon}}},\end{aligned}$$

where the last inequality uses Assumption 2.8. Next, by the definition in (2.31) of $\bar{a}$ we have that $\bar{a} > \frac{1}{2} - \frac{\bar{\varepsilon}}{4(\bar{\varepsilon}+2)}$, hence

$$\begin{aligned}(2+\bar{\varepsilon})(1+\bar{a}-1) &= \bar{a}(2+\bar{\varepsilon})\\ &\geqslant (2+\bar{\varepsilon})(\frac{1}{4} + \frac{1}{4+2\bar{\varepsilon}})\\ &= \frac{1}{2} + \frac{\bar{\varepsilon}}{4} + \frac{2+\bar{\varepsilon}}{4+2\bar{\varepsilon}} = 1 + \frac{\bar{\varepsilon}}{4}.\end{aligned}$$

Thus

$$n\frac{\mathbb{E}\left[(\check{u}(1))^{2+\bar{\varepsilon}}\right]}{\left(c\mu^n n^{\bar{a}-1}\right)^{2+\bar{\varepsilon}}} = O(n^{-\frac{\bar{\varepsilon}}{4}}).$$

Hence for any $c > 0$ there exists c_2 such that

$$\begin{aligned}\mathbb{P}\left(e^n_{\max} \geqslant c n^{\bar{a}-1}\right) &\leqslant \mathbb{P}\left(K^n \geqslant c_2 n\right) + \mathbb{P}\left(e^n_{\max} \geqslant c n^{\bar{a}-1},\, K^n \leqslant c_2 n\right)\\ &\leqslant \mathbb{P}\left(K^n \geqslant c_2 n\right) + \mathbb{P}\left(\max_{l \leqslant c_2 n} u^n(l) \geqslant c n^{\bar{a}-1}\right),\end{aligned}$$

and both terms have been shown to converge to zero. □

Proof of Proposition 5.3 (continued from §5.2.2). Part 1 of the proof of this proposition appears in §5.2.2; here we provide the remaining parts.

Part 2. Consider the case where $\mathbf{T}_1$ is applicable when the workload is sufficiently large. This covers case 2(d) of Theorem 2.13, in which the policy $\mathbf{T}_2\mathbf{T}_1$ applies (none of our proposed policies implement $\mathbf{T}_1$ as a single mode policy). The proof given in Part 1 is applicable, for the following reasons. In the first step, during the analyzed time interval $[\tau, t]$, one has $\hat{X}^2 \geq K > \hat{\Theta}^n$, and therefore there is no difference between how $\mathbf{P}$ and $\mathbf{T}_1$ behave during this interval. In addition, the workload is sampled at each arrival/service, which means that

$$T^n_{22}(t) - T^n_{22}(\tau) \geq t - \tau - e^n_{\max}.$$

In the second step, the argument given for nonidling of both servers during $[\sigma, t]$ again holds here similarly to Part 1, upon noticing that, since σ corresponds to an arrival, it is a sampling time, and so the current mode is either already the high mode or switches to the high mode at that time. (It is possible that, if σ is a mode switching time, then server 1 was idle just before σ. But, if so, it starts to serve class 1 at σ.) The remaining details need no adaptation.

Part 3. Finally, consider the policies which employ $\mathbf{T}_2$ for high workload levels, namely the policies $\mathbf{T}_2$, $\mathbf{T}_2\mathbf{T}_2$ and $\mathbf{T}_1\mathbf{T}_2$, covering all remaining cases of Theorem 2.13. In these cases the stronger moment assumption is in force, and the goal is to prove that there exists $\varepsilon_0 > 0$ such that $\mathbb{E}[(\hat{W}^n_t)^{1+\varepsilon_0}] \leq \text{pol}(t)$ for all n and t, for some polynomial pol. As before, let ξ^A or ξ^H be in canonical form, in the single and, respectively, dual mode case. Fix a constant $K > z^*$ in the dual mode case and $K = 1$ in the single mode case. Given t consider the event $\hat{W}^n_t > K$. Let

$$\tau_1 = \tau^n_1 = \sup\{s \in [0, t] : \hat{W}^n_s \leq K\}.$$

Then $\hat{W}^n_{\tau_1} \leq K + 1$, and moreover $\hat{W}^n \geq K$ during $[\tau_1, t]$. Although this lower bound on the workload is sufficiently large to guarantee that $\hat{W}^n$ corresponds to the higher workload mode, it is possible that at time τ_1 the current mode variable still equals the lower workload mode. We argue that the time it takes to switch to the upper workload mode is bounded by $2e^n_{\max}$ in both $\mathbf{T}_1\mathbf{T}_2$ and $\mathbf{T}_2\mathbf{T}_2$. In the former case, the current mode switches as soon as a there is a new arrival or departure. In the latter case, one possibly has to complete the service of a job at server 2, which is server $k_1(\xi^L)$. It is possible that there are no jobs allowed to be routed to server 2 at time τ_1. Wait for an arrival of class 1 job, and service completion of this job at server 2. At this time it is guaranteed that mode has switched to ξ^H if it was not ξ^H earlier. Thus if we let

$$\tau_2 = \inf\{s \geq \tau_1 : \text{MODE}^n(s) = \xi^H\} \wedge t,$$

we have $\tau_2 - \tau_1 \leq 2e^n_{\max} \wedge t$.

According to the rules of $\mathbf{T}_2$, server 2 must be busy throughout the interval $[\tau_2, t]$. Thus $\hat{I}^n_2(t) = \hat{I}^n_2(\tau_2)$. Hence by (5.6), and using $\hat{I}^n_k[a, b] \leq n^{1/2}(b - a)$ (by (5.2)), we have

$$\begin{aligned}
\hat{W}^n_t &= \hat{W}^n_{\tau_1} + \hat{F}^n[\tau_1, t] + \hat{L}^n[\tau_1, t] \\
&= \hat{W}^n_{\tau_1} + \hat{F}^n[\tau_1, t] + \beta_1\hat{I}^n_1[\tau_1, t] + \beta_2\hat{I}^n_2[\tau_1, t] \\
&\leq K + 1 + 2\|\hat{F}^n\|_t + cn^{1/2}e^n_{\max} + \beta_1\hat{I}^n_1[\tau_2, t]
\end{aligned}$$

holds on the event $\hat{W}^n_t > K$. On the complementary event, $\hat{W}^n_t \leq K$. Use (5.5) and (5.18) to obtain the bound $\mathbb{E}[\|\hat{F}^n\|^{1+\varepsilon_0}_t] \leq \{\mathbb{E}[\|\hat{F}^n\|^2_t]\}^{(1+\varepsilon_0)/2} \leq c(1 + t)^{(1+\varepsilon_0)/2}$. Use Lemma 5.9 to bound the second moment of $n^{1/2}e^n_{\max}$ by $c(1+t)$. By Minkowski's inequality and the inequality $(a+b)^{1+\varepsilon_0} \le 4a^{1+\varepsilon_0} + 4b^{1+\varepsilon_0}$ which holds for $a, b \ge 0$, $\varepsilon_0 \in (0,1)$, this yields

$$\mathbb{E}[(\hat{W}^n_t)^{1+\varepsilon_0}] \le c(1+t)^{(1+\varepsilon_0)/2} + c\mathbb{E}[\mathbb{1}_{\{\hat{W}^n_t > K\}} \hat{I}^n_1[\tau_2, t]^{1+\varepsilon_0}].$$

Hence it suffices to bound the last term above by a polynomial in t.

To this end, let $\tau_3 = \inf\{s \ge \tau_2 : \hat{X}^n_1(s) \ge \hat{\Theta}^n\} \wedge t$. Then

$$\begin{aligned} \mathbb{E}[\mathbb{1}_{\{\hat{W}^n_t > K\}} \hat{I}^n_1[\tau_2, t]^{1+\varepsilon_0}] &\le 4\Delta^n_1 + 4\Delta^n_2, \quad \text{where} \\ \Delta^n_1 &= \mathbb{E}[\mathbb{1}_{\{\hat{W}^n_t > K\}} \hat{I}^n_1[\tau_2, \tau_3]^{1+\varepsilon_0}] \\ \Delta^n_2 &= \mathbb{E}[\mathbb{1}_{\{\hat{W}^n_t > K, \tau_3 < t\}} \hat{I}^n_1[\tau_3, t]^{1+\varepsilon_0}]. \end{aligned}$$

To bound Δ^n_1, consider the event $\{\hat{W}^n_t > K, \tau_3 > \tau_2\}$. On it, during the interval $[\tau_2, \tau_3]$, one has $\hat{X}^n_1 < \hat{\Theta}^n$, hence by the rules of $\mathbf{T}_2$, server 2 prioritizes class 2, except possibly it completes a service that started when $\hat{X}^n_1 \geqslant \hat{\Theta}^n$. Moreover, $\hat{W}^n \ge K$ during the same interval, by which we know that there are multiple class-2 jobs in the system, and thus server 2 gives no service to class 1, with the only exception of service to a job that started when $\hat{X}^n_1 \geqslant \hat{\Theta}^n$. Hence the departure process associated with activity $(1,2)$ increases by at most 1 during this interval, that is, $0 \le D^n_{12}[\tau_2, \tau_3] \le 1$. Recalling the definition of $\hat{S}^n$ in (5.1), this can be expressed as $0 \le e^n_1[\tau_2, s] \le n^{-1/2}$, $s \in [\tau_2, \tau_3]$, where

$$e^n_1(s) := n^{-1/2} D^n_{12}(s) = \hat{S}^n_{12}(T^n_{12}(s)) + n^{-1/2}\mu^n_{12} T^n_{12}(s).$$

Using this in (5.4) gives, for $s \in [\tau_2, \tau_3]$,

$$\hat{X}^n_1(s) = \hat{X}^n_1(\tau_2) + \hat{f}^n_1[\tau_2, s] - e^n_1[\tau_2, s] + n^{1/2}\lambda_1(s - \tau_2) - n^{1/2}\mu_{11} T^n_{11}[\tau_2, s],$$

where

$$\hat{f}^n_1(s) = \hat{A}^n_1(s) - \hat{S}^n_{11}(T^n_{11}(s)) + (\hat{\lambda}^n_1 s - \hat{\mu}^n_{11} T^n_{11}(s)).$$

Next, by (2.4),

$$T^n_{11}[\tau_2, s] = (s - \tau_2) - I^n_1[\tau_2, s] - T^n_{21}[\tau_2, s].$$

However, activity $(2,1)$ is not in use by $\mathbf{T}_2$. Denoting $c_1 = \lambda_1 - \mu_{11}$ and $g(s) = c_1 s$, this yields

$$\hat{X}^n_1(s) = \hat{X}^n_1(\tau_2) + \hat{f}^n_1[\tau_2, s] - e^n_1[\tau_2, s] + n^{1/2} g[\tau_2, s] + \mu_{11}\hat{I}^n_1[\tau_2, s], \qquad s \in [\tau_2, \tau_3]. \tag{A.3}$$

Because $\mathbf{T}_2$ is used after the update of modes, it must be true that $c_1 = \lambda_1 - \mu_{11} > 0$.

We use the following property of the Skorohod map. Let $\eta = \Gamma_2[\psi]$. Then by (5.7), $\eta(s) = \sup_{0 \le \theta \le s}[\psi(\theta)^-]$. Assume that $\psi = \psi_1 + \psi_2$ where $\psi_2 \ge 0$. Then

$$\begin{aligned} \eta(s) = \Gamma_2[\psi_1 + \psi_2](s) &= \sup_{\theta \in [0,s]} [(\psi_1(\theta) + \psi_2(\theta))^-] \\ &\le \sup_{\theta \in [0,s]} [\psi_1(\theta)^-] \\ &\le \|\psi_1\|_s. \end{aligned}$$

This property is used as follows. On the interval $[\tau_2, \tau_3]$, $\hat{I}^n_1$ can increase only at times when $\hat{X}^n_1 = 0$; and the latter process is nonnegative. This shows that $\mu_{11}\hat{I}^n_1$ serves as the Skorohod term in (A.3) on that interval, and thus by the non-negativity of $\hat{X}_1^n(\tau_2)$ and c_1, and the bound $|e_1^n[\tau_2, s]| \le n^{-1/2}$, this shows that

$$\mu_{11}\hat{I}_1^n[\tau_2, \tau_3] \le \sup_{\theta \in [\tau_2, \tau_3]} |\hat{f}_1^n[\tau_2, \theta]| + \sup_{\theta \in [\tau_2, \tau_3]} |e_1^n[\tau_2, \theta]| \le 2\|\hat{f}_1^n\|_t + n^{-1/2}.$$

Hence $\varDelta_1^n \le c\{\mathbb{E}[\|\hat{f}_1^n\|_t^2]\}^{(1+\varepsilon_0)/2} + 1 \le c(1+t)$.

Next we bound $\varDelta_2^n$. To this end, consider the event $\varOmega^n := \{\hat{W}_t^n > K, \tau_3 < t, \hat{I}_1^n(t) > \hat{I}_1^n(\tau_3)\}$. Because by its definition, $\hat{I}_1^n(t) \le n^{1/2}t$, we have

$$\varDelta_2^n \le \mathbb{P}(\varOmega^n)(n^{1/2}t)^{1+\varepsilon_0}.$$

On the event $\varOmega^n$ let

$$\tau_5 = \inf\{s > \tau_3 : \hat{X}_1^n(s) = 0\}, \qquad \tau_4 = \sup\{s < \tau_5 : \hat{X}_1^n(s) \ge \hat{\varTheta}^n\}.$$

Then on $\varOmega^n$ it must hold that $\tau_3 \le \tau_4 < \tau_5 \le t$. The arguments which lead to (A.3) are valid for the time interval $[\tau_4, \tau_5]$. As a result,

$$\hat{X}_1^n[\tau_4, \tau_5] = \hat{f}_1^n[\tau_4, \tau_5] - e_1^n[\tau_4, \tau_5] + n^{1/2}g[\tau_4, \tau_5] + \mu_{11}\hat{I}_1^n[\tau_4, \tau_5],$$

with $|e_1^n[\tau_4, s]| \le n^{-1/2}$ for all $s \in [\tau_4, \tau_5]$. Now, $\hat{I}_1^n$ remains flat on the interval $[\tau_4, \tau_5]$, and moreover, $\hat{X}_1^n(\tau_4) = \hat{\varTheta}^n - n^{-1/2}$ and $\hat{X}_1^n(\tau_5) = 0$. This gives

$$\hat{f}_1^n[\tau_4, \tau_5] + n^{1/2}g[\tau_4, \tau_5] \le -\hat{\varTheta}^n + 2n^{-1/2}.$$

Given any $\delta > 0$, using the nonnegativity of the second term on the LHS, in the case that $\tau_5 - \tau_4 \le n^{-\delta}$ one must have $\hat{f}_1^n[\tau_4, \tau_5] \le -\hat{\varTheta}^n/2$. On the other hand, in the case $\tau_5 - \tau_4 > n^{-\delta}$ one must have $\hat{f}_1^n[\tau_4, \tau_5] + c_1 n^{1/2} n^{-\delta} < 0$. As a result,

$$\mathbb{P}(\varOmega^n) \le p_1^n + p_2^n := \mathbb{P}(w_t(\hat{f}_1^n, n^{-\delta}) \ge \hat{\varTheta}^n/2) + \mathbb{P}(2\|\hat{f}_1^n\|_t > c_1 n^{\frac{1}{2}-\delta}).$$

We have by (2.32) $\hat{\varTheta}^n = n^{-\frac{1}{2}}\lceil n^{\bar{a}}\rceil$, where we recall that $\bar{a} < \frac{1}{2}$. Thus by Lemma 5.7, with $\nu_1 = \delta$, $\nu_2 = \frac{1}{2} - \bar{a}$, one has, for any $\varepsilon_1 > 0$,

$$p_1^n \le c(1+t)^{\frac{\mathbf{m}}{2}} n^{-h_0+\varepsilon_1}$$

where

$$h_0 = (\frac{\mathbf{m}}{2} - 1)\delta - \mathbf{m}(\frac{1}{2} - \bar{a}), \qquad \delta \le 1 - \bar{a}.$$

Next, by (5.18) and Chebychev's inequality,

$$p_2^n \le c(1+t)^{\frac{\mathbf{m}}{2}} n^{-\frac{\mathbf{m}}{2}+\delta\mathbf{m}}.$$

As a result, we have

$$\varDelta_2^n \le c(n^{\zeta_1(\delta,\varepsilon_0,\varepsilon_1)} + n^{\zeta_2(\delta,\varepsilon_0)})(1+t)^{\frac{\mathbf{m}}{2}+1+\varepsilon_0},$$

where

$$\zeta_1(\delta, \varepsilon_0, \varepsilon_1) = -\Big(\frac{\mathbf{m}}{2} - 1\Big)\delta + \mathbf{m}\Big(\frac{1}{2} - \bar{a}\Big) + \frac{1}{2} + \frac{\varepsilon_0}{2} + \varepsilon_1, \qquad \zeta_2(\delta, \varepsilon_0) = -\frac{\mathbf{m}}{2} + \delta\mathbf{m} + \frac{1}{2} + \frac{\varepsilon_0}{2}.$$

By our assumptions, we have $\mathbf{m} > \mathbf{m}_0 = \frac{1}{2}(5+\sqrt{17})$ and $\bar{a} > \frac{1}{2} - \frac{\mathbf{m}^2-5\mathbf{m}+2}{2\mathbf{m}(3\mathbf{m}-2)}$. Using this, a calculation shows that with the choice $\delta = \mathbf{m}/(3\mathbf{m}-2) \in (1/3, 1/2)$, one has $\zeta_1(\delta,0,0) \vee \zeta_2(\delta,0) < 0$. It follows that there exist $\varepsilon_0 > 0$ and $\varepsilon_1 > 0$ for which $\zeta_1(\delta,\varepsilon_0,\varepsilon_1) \vee \zeta_2(\delta,\varepsilon_0) < 0$. Therefore $\varDelta_2^n \le c(1+t)^{\frac{\mathbf{m}}{2}+1+\varepsilon_0}$ and the proof is complete. □

Proof of Lemma 5.10 (continued from §5.2.3). The proof of the lemma in Case 1(a) appears in §5.2.3; here we cover the remaining cases.

- **Case 1(b)**: In this case, we use the $\mathbf{T}_2$ policy which is single mode. Between σ and τ, $\hat{X}_p^n(t) \geqslant \hat{\Theta}^n$. Thus, both servers give priority to the HPC at all time in $[\sigma, \tau]$. It is possible that the dual activity server is occupied with the low priority class at σ but as soon as the current job is served, the high priority class gets priority on one server and dedication by the other server. By definition of $e^n_{\max}$, we have for any $k \in \{1,2\}$,

  $$\int_\sigma^\tau \varXi^n_{pk}(t)dt \geqslant \tau - \sigma - e^n_{\max}\mathbb{1}_{k=k_2(\xi^A)}.$$

  Thus

  $$\int_\sigma^\tau \big(\lambda_p - \sum_k \mu_{pk}\varXi^n_{pk}(t)\big)dt \leqslant (\tau-\sigma)(\lambda_p - \sum_k \mu_{pk}) + \mu_{pk_2(\xi^A)}e^n_{\max}.$$

  Finally, $\lambda_p - \sum_k \mu_{pk} < 0$ by (3.1) and $\min_i \lambda_i > 0$.

- **Case 2(a)**: In this case, we use the **PP** policy, which only changes mode at the completion of a service at the single activity server. This is precisely the server that gives priority to the high priority class after switching mode because we are in a **SS** case. Since there are always HPC jobs to serve between σ and τ, we claim that excluding an initial period, at any given time the HPC is being served by at least one server, as discussed in Section 5.1.3. We now discuss how long this initial period can be.

  The only way some service is lost is if there were no high priority jobs at some point in the system, both servers become busy with low priority jobs and σ occurs before the service completion of those jobs. After completion of those two jobs, the HPC keeps priority on at least one server regardless of switching of the current mode. For the initial jobs, with respect to $\text{MODE}^n(\sigma)$ either the service ends first at the dual activity server or the service ends first at the single activity server.

  In the first case, a service of the HPC job starts at the dual activity server because it has priority there until either τ is reached or a mode switch occurs. In the second case the available server is currently dedicated to the LPC. Since this corresponds to a service completion at the single activity server, the workload is sampled and there could also be a switching of modes. If the mode changes then this server becomes dual activity and the HPC has priority there. If there is no mode switch, the service of another LPC job starts. LPC jobs are served at this server until the mode switches.

  Thus, if the service ends at the dual activity server before the mode switches then an HPC job will start there. If the mode switches before the service ends at the current dual activity server, then the current single activity server becomes dual activity and an HPC job begins service there.

In either case the the HPC is served by at least one server after that time because it has priority on the dual activity server, and the dual activity server is always available when switching modes. In addition, the time it takes for the HPC to begin service is smaller than the service of the job present at the dual activity server at time σ, which is smaller than $e^n_{\max}$. Hence

$$\int_\sigma^\tau \big(\mu_{p1}\varXi^n_{p1}(t) + \mu_{p2}\varXi^n_{p2}(t)\big)\,dt \geqslant \min_k \mu_{pk}(\tau - \sigma) - \max_k \mu_{pk} e^n_{\max}.$$

This yields

$$\int_\sigma^\tau \big(\lambda_p - \sum_k \mu_{pk}\varXi^n_{pk}(t)\big)dt \leqslant (\tau-\sigma)(\lambda_p - \min_k \mu_{pk}) + \max_k \mu_{pk} e^n_{\max}.$$

In addition, $\lambda_p - \min_k \mu_{pk} < 0$. Putting ξ^L in canonical form forces in this case $\frac{\lambda_1}{\alpha_1} > \beta_1$ and $p = 2$ because $i_1(\xi^L) = p$. In addition, by (3.6), either $\frac{\lambda_1}{\alpha_1} > \beta_1 \vee \beta_2$ or $\frac{\lambda_2}{\alpha_2} > \beta_1 \vee \beta_2$. Thus $\frac{\lambda_1}{\alpha_1} > \max_k \beta_k$, which implies $\frac{\lambda_2}{\alpha_2} < \min_k \beta_k$ by (3.1).

- **Case 2(b)**: In this case, we use the $\mathbf{T}_2\mathbf{T}_2$ policy. Since the number of HPC jobs stays above $\varTheta^n$ throughout the period $[\sigma, \tau]$, both servers only take new jobs from the HPC regardless of the mode. Similarly to 1(b), there is at most one job served at either activity before the HPC gets served. Hence, for any $k \in \{1, 2\}$,

  $$\int_\sigma^\tau \varXi^n_{pk}(t)dt \geqslant \tau - \sigma - e^n_{\max}.$$

  Thus

  $$\int_\sigma^\tau \big(\lambda_p - \sum_k \mu_{pk}\varXi^n_{pk}(t)\big)dt \leqslant (\tau-\sigma)(\lambda_p - \sum_k \mu_{pk}) + \sum_k \mu_{pk} e^n_{\max}.$$

  We obtain the result similarly as before because $\lambda_p - \sum_k \mu_{pk} < 0$.

- **Cases 2(c) and 2(d)**: the HPC is single activity in one mode and dual activity in the other. Recall that this case is **CS**. This means the dual activity server stays the same regardless of switching modes. When a $\mathbf{T}_1$ rule is applied, the dual activity server prioritizes the single activity class as long as there are more than $\varTheta^n$ jobs of this class. The HPC is single activity when we apply the $\mathbf{T}_1$ rule. Similarly, under a $\mathbf{T}_2$ rule, the dual activity server prioritizes the dual activity class as long as there are more than $\varTheta^n$ jobs of this class and the HPC is dual activity when we apply the $\mathbf{T}_2$ rule. Once again put ξ^L in canonical form, server 2 is the dual activity server in both modes. Regardless of which rule is used, as long as the number of HPC jobs is above $\varTheta^n$, server 2 only takes new jobs from the HPC. As before, we have to exclude a residual service of a low priority job that started before σ. Hence

  $$\int_\sigma^\tau \varXi^n_{p2}(t)dt \geqslant \tau - \sigma - e^n_{\max}.$$

  It remains to show that the activity processing HPC jobs throughout the period is enough to deplete them. Because of the canonical form of ξ^L, $\frac{\lambda_1}{\alpha_1} > \beta_1$ and thus by (3.5), $\frac{\lambda_1}{\alpha_1} < \beta_2$. Moreover, $\frac{\lambda_2}{\alpha_2} < \beta_2$ by (3.1). Whichever the HPC is, server 2 has enough capacity to deplete it. In other words, $\lambda_p - \mu_{p2} < 0$ and

  $$\int_\sigma^\tau \big(\lambda_p - \sum_k \mu_{pk}\varXi^n_{pk}(t)\big)dt \leqslant (\tau-\sigma)(\lambda_p - \mu_{p2}) + \mu_{p2} e^n_{\max}.$$

□

Proof of Lemma 5.12. The proof follows reasoning similar to the proof of Proposition 5.4. Fix δ. Let us introduce

$$\sigma_r^n = \sup\left\{ t \leqslant \tau_r^n : \hat{X}_p^n(t) \geqslant \hat{\Theta}^n, \text{ or } \hat{X}_q^n(t) \leqslant \frac{\hat{\Theta}^n}{2} \right\}.$$

In cases 2(b), (c) and (d), introduce

$$\widetilde{\tau}^n = \inf\left\{ t \geqslant 0 : \hat{X}_q^n(t) \geqslant \frac{z^* \alpha_q}{2} \right\},$$

$$\widetilde{\sigma}^n = \sup\left\{ t \leqslant \widetilde{\tau}^n : \hat{X}_q^n(t) \leqslant \frac{z^* \alpha_q}{4} \right\}.$$

To simplify, we write throughout this proof

$$\rho^n = \rho, \, \tau_r^n = \tau, \, \widetilde{\tau}^n = \widetilde{\tau}, \, \widetilde{\sigma}^n = \widetilde{\sigma} \text{ and } \sigma_r^n = \sigma.$$

We briefly describe the interaction between the times we just introduced. Under the state space collapse, $\tilde{\tau}$ has to occur before the first time the current mode switches. The times σ and $\tilde{\sigma}$ allow us to have some knowledge about the state of queue lengths, uniformly over an interval. The first step of the proof of (5.26) is establishing that

$$\mathbb{P}\left(\tau_c^n \geqslant t_0, \, \widetilde{\tau} \leqslant \rho\right) \to 0 \tag{A.4}$$

holds in cases 2(b), (c) and (d). In those cases, under the event $\{\tau_c^n \geqslant t_0\} \cap \{\widetilde{\tau} \geqslant \rho\}$, for any $t \leqslant \rho$,

$$\text{MODE}^n(t) = \text{MODE}^n(0) = \xi^L. \tag{A.5}$$

In addition, in case 1(b), for any $t \leqslant \rho$

$$\text{MODE}^n(t) = \xi^A, \tag{A.6}$$

which means (A.4) is not needed in this case.

We now proceed with the proof of (A.4). On $\{\tau_c^n \geqslant t_0, \, \widetilde{\tau} \leqslant \rho\}$, the following hold:

1. The number of HPC job is below $\hat{\Theta}^n$ and there are LPC jobs in the system during $[\widetilde{\sigma}, \widetilde{\tau}]$:

$$\sup_{t \in [\widetilde{\sigma}, \widetilde{\tau}]} \hat{X}_p^n(t) < \hat{\Theta}^n \text{ and } \inf_{t \in [\widetilde{\sigma}, \widetilde{\tau}]} \hat{X}_q^n(t) > 0.$$

2. The current mode is the same as in the initial state: for any $t \in [\widetilde{\sigma}, \widetilde{\tau}]$,

$$\text{MODE}^n(t) = \text{MODE}^n(0) = \xi^L.$$

3. In addition, the number of LPC jobs must have grown during that time:

$$\hat{X}_q^n(\widetilde{\tau}) - \hat{X}_q^n(\widetilde{\sigma}) \geqslant \frac{z^* \alpha_q}{4}.$$

1. comes from the definition of ρ, $\tilde{\sigma}$ and $\tilde{\tau}$. 2. comes from the fact that the number of HPC jobs is bounded by $2\Theta^n$ under $\{\tau_c^n \geqslant t_0\}$ and the number of LPC jobs is bounded by $\frac{z^*\alpha_q}{2}$ before $\tilde{\tau}$ and the workload cannot cross above z^*. 3. comes from the definition of $\tilde{\tau}$ and $\tilde{\sigma}$. Because of 2., only the rule in the lower workload mode is in use. When using a $\mathbf{T}_1$ rule, both servers take new jobs from the LPC because of 1. and $\lambda_q - \sum_k \mu_{qk} < 0$. When using a $\mathbf{T}_2$ rule, again because of 1., the dual activity server takes new jobs from the LPC. Since ξ^L is in canonical form $\frac{\lambda_1}{\alpha_1} > \beta_1$, $k_2(\xi^L) = 2$ and $q = 2$ because of the $\mathbf{T}_2$ rule. This means that $\frac{\lambda_2}{\alpha_2} < \beta_2$ and thus $\lambda_q - \mu_{qk_2(\xi^L)} < 0$. Hence, with (5.19), regardless of the case there exists $c > 0$ such that

$$\hat{X}_q^n(\widetilde{\tau}) \leqslant \hat{X}_q^n(\widetilde{\sigma}) + \hat{f}_q^n[\widetilde{\sigma}, \widetilde{\tau}] - c\sqrt{n}(\widetilde{\tau} - \widetilde{\sigma}) + \sum_k \mu_{qk} e_{\max}^n.$$

From this, we obtain two bounds:

$$\hat{X}_q^n(\widetilde{\tau}) - \hat{X}_q^n(\widetilde{\sigma}) \leqslant w_{t_0}(\hat{f}_q^n, \widetilde{\tau} - \widetilde{\sigma}) - c\sqrt{n}(\widetilde{\tau} - \widetilde{\sigma}) + \sum_k \mu_{qk} e_{\max}^n$$
$$\hat{X}_q^n(\widetilde{\tau}) - \hat{X}_q^n(\widetilde{\sigma}) \leqslant 2\|\hat{f}_q^n\|_{t_0} - c\sqrt{n}(\widetilde{\tau} - \widetilde{\sigma}) + \sum_k \mu_{qk} e_{\max}^n$$

We now prove (A.4): let $r_n = \frac{\log(n+1)}{\sqrt{n}}$. Because of 3.,

$$\begin{aligned}
\mathbb{P}\left(\tau_c^n \geqslant t_0,\, \widetilde{\tau} \leqslant \rho\right) &\leqslant \mathbb{P}\left(\widetilde{\tau} \leqslant \rho,\, \hat{f}_q^n[\widetilde{\sigma}, \widetilde{\tau}] - c\sqrt{n}(\widetilde{\tau} - \widetilde{\sigma}) + \sum_k \mu_{qk} e_{\max}^n \geqslant \frac{z^*\alpha_q}{4}\right) \\
&\leqslant \mathbb{P}\left(w_{t_0}(\hat{f}_q^n, \widetilde{\tau} - \widetilde{\sigma}) - c\sqrt{n}(\widetilde{\tau} - \widetilde{\sigma}) + \sum_k \mu_{qk} e_{\max}^n \geqslant \frac{z^*\alpha_q}{4}, \widetilde{\tau} - \widetilde{\sigma} \leqslant r^n\right) \\
&\quad + \mathbb{P}\left(2\|\hat{f}_q^n\|_{t_0} - c\sqrt{n}(\widetilde{\tau} - \widetilde{\sigma}) + \sum_k \mu_{qk} e_{\max}^n \geqslant \frac{z^*\alpha_q}{4}, \widetilde{\tau} - \widetilde{\sigma} > r^n\right) \\
&\leqslant \mathbb{P}\left(w_{t_0}(\hat{f}_q^n, r_n) + \sum_k \mu_{qk} e_{\max}^n \geqslant \frac{z^*\alpha_q}{4}\right) + \mathbb{P}\left(2\|\hat{f}_q^n\|_{t_0} + \sum_k \mu_{qk} e_{\max}^n \geqslant c\log(n+1)\right).
\end{aligned}$$

By Lemma 5.9, $e_{\max}^n$ is smaller than $n^{\bar{a}-1} = o(1)$. By Remark 5.8, $\hat{f}_q^n$ are C-tight and $\|\hat{f}_q^n\|_{t_0}$ are tight RVs. Both terms must then converge to zero. This concludes the proof of (A.4) in all relevant cases for this lemma.

Case 2(c), $\mathbf{T}_1$ rule:
By splitting the integral that defines $\bar{R}^n$, and using (5.24) and (A.5),

$$\begin{aligned}
\bar{R}_\rho^n &\leqslant \int_0^\rho \mathbb{1}_{\hat{W}_t^n \geqslant c_3\hat{\Theta}^n,\, \tau_c^n \geqslant t_0,\, \widetilde{\tau} \geqslant \rho} d\hat{L}_t^n + \int_0^\rho \mathbb{1}_{\tau_c^n < t_0} + \mathbb{1}_{\widetilde{\tau} \leqslant \rho,\, \tau_c^n \geqslant t_0} d\hat{L}_t^n \\
&\leqslant \int_0^\rho \mathbb{1}_{\hat{X}_q^n(t) \geqslant \hat{\Theta}^n, \mathrm{MODE}^n(t) = \xi^L} d\hat{L}_t^n + \int_0^\rho \mathbb{1}_{\tau_c^n < t_0} + \mathbb{1}_{\widetilde{\tau} \leqslant \rho,\, \tau_c^n \geqslant t_0} d\hat{L}_t^n
\end{aligned}$$

The first term is zero by definition of the $\mathbf{T}_1$ rule because q is the dual activity class. The second term goes to zero by Proposition 5.4 and (A.4). This concludes the proof of (5.26) in case 2(c).

Case 1(b), 2(b) and 2(d), $\mathbf{T}_2$ rule:

We now deal with all cases that use a $\mathbf{T}_2$ rule in lower workload at the same time thanks to (A.5)/(A.6). Recall that $p = 1 = i_2(\xi^L)$ and $k_1(\xi^L) = 1$. By splitting the integral, we obtain

$$\bar{R}^n_\rho = \int_0^\rho \mathbb{1}_{\hat{W}^n_t \geqslant c_3\hat{\Theta}^n,\, \tau^n_c \geqslant t_0} d\hat{L}^n_t + \int_0^\rho \mathbb{1}_{\hat{W}^n_t \geqslant c_3\hat{\Theta}^n,\, \tau^n_c < t_0} d\hat{L}^n_t.$$

By (5.2), $\hat{L}^n = \sum_k \beta_k \hat{I}^n_k$. By definition of the $\mathbf{T}_2$ rule, as long as there are at least 2 customers in the system, the dual activity server cannot idle. With ξ^L in canonical form, server 2 is the dual activity server, so under the event $\{\tau^n_c \geqslant t_0\} \cap \{\widetilde{\tau} \geqslant \rho\}$,

$$\int_0^\rho \mathbb{1}_{\hat{W}^n_t \geqslant 2(\alpha_1 \wedge \alpha_2)^{-1} n^{-1/2},\, \tau^n_c \geqslant t_0,\, \widetilde{\tau} \geqslant \rho} d\hat{I}^n_2(t) = 0.$$

Since $\int_0^\rho \mathbb{1}_{\hat{W}^n_t \geqslant c_3\hat{\Theta}^n,\, \tau^n_c < t_0} d\hat{L}^n_t \Rightarrow 0$ by Proposition 5.4, we are left to deal with

$$\int_0^\rho \mathbb{1}_{\hat{W}^n_t \geqslant c_3\hat{\Theta}^n,\, \tau^n_c \geqslant t_0,\, \widetilde{\tau} \geqslant \rho} d\hat{L}^n_t = \beta_1 \int_0^\rho \mathbb{1}_{\hat{W}^n_t \geqslant c_3\hat{\Theta}^n,\, \tau^n_c \geqslant t_0,\, \widetilde{\tau} \geqslant \rho} d\hat{I}^n_1(t).$$

We introduce the following times:

$$\bar{\tau}^n = \inf\left\{ t \geqslant 0 : \int_0^t \mathbb{1}_{\hat{W}^n_t \geqslant c_3\hat{\Theta}^n} d\hat{L}^n_t \geqslant \frac{\delta}{2} \right\},$$

and

$$\bar{\sigma}^n = \sup\left\{ t \leqslant \bar{\tau}^n : \hat{X}^n_2(t) = 0 \right\}.$$

We want to show that

$$\mathbb{P}\left(\rho \geqslant \bar{\tau}\right) \to 0.$$

On the event $\{\rho \geqslant \bar{\tau}\}$, we have :

- First, by definition of $\bar{\sigma}$,
$$\inf_{t \in (\bar{\sigma}, \bar{\tau})} \hat{X}^n_2(t) > 0.$$
- Second, by definition of ρ,
$$\sup_{t \in [0, \bar{\tau}]} \hat{X}^n_1(t) < \hat{\Theta}^n.$$
- Finally, by definition of $\bar{\sigma}$,
$$\hat{X}^n_2(\bar{\sigma}) \leqslant n^{-1/2}.$$

In order for server 1 to be idle at time $\bar{\tau}$ (so that $d\hat{I}^n_1(\bar{\tau}) > 0$), it is necessary that $\hat{X}^n_1(\bar{\tau}) = 0$. This means that in order to also have $\hat{W}^n_{\bar{\tau}} \geqslant c_3\hat{\Theta}^n$, we need to have $\hat{X}^n_2(\bar{\tau}) \geqslant 2\hat{\Theta}^n$, or

$$\hat{X}^n_2(\bar{\tau}) - \hat{X}^n_2(\bar{\sigma}) \geqslant 2\hat{\Theta}^n - n^{-1/2} \geqslant \hat{\Theta}^n. \tag{A.7}$$

We can now give the balance equation for $\hat{X}_2^n$ between $\bar{\sigma}$ and $\bar{\tau}$ under the $\mathbf{T}_2$ rule. Under the $\mathbf{T}_2$ rule, server 2 prioritizes class 2 for the whole period because the number of HPC jobs remains below $\hat{\Theta}^n$. Since $\lambda_1 > \mu_{11}$ we obtain $\lambda_2 < \mu_{22}$ from (3.1). By the same reasoning as (A.3) there exists $c_2 > 0$ such that

$$\hat{X}_2^n[\bar{\sigma},\bar{\tau}] \leqslant \hat{f}_2^n[\bar{\sigma},\bar{\tau}] - c_2\sqrt{n}(\bar{\tau}-\bar{\sigma}) + \sqrt{n}e_{\max}^n \sum_k \mu_{2k}. \tag{A.8}$$

Combining (A.7) and (A.8) we obtain (on $\{\rho \geqslant \bar{\tau}\}$)

$$\hat{f}_2^n[\bar{\sigma},\bar{\tau}] - c_2\sqrt{n}(\bar{\tau}-\bar{\sigma}) + \sqrt{n}e_{\max}^n \sum_k \mu_{2k} \geqslant \hat{\Theta}^n.$$

As in the proof of Proposition 5.4, we distinguish between two cases: $\bar{\tau}-\bar{\sigma}$ smaller or larger than $n^{-\nu_1}$. On $\bar{\tau}-\bar{\sigma} \leqslant n^{-\nu_1}$, $\hat{f}_2^n[\bar{\sigma},\bar{\tau}] \leqslant w_{t_0}(\hat{f}_2^n, n^{-\nu_1})$, so that

$$\mathbb{P}\left(\hat{f}_2^n[\bar{\sigma},\bar{\tau}] - c_2\sqrt{n}(\bar{\tau}-\bar{\sigma}) + \sqrt{n}e_{\max}^n \sum_k \mu_{2k} \geqslant \hat{\Theta}^n, \bar{\tau}-\bar{\sigma} \leqslant n^{-\nu_1}\right)$$
$$\leqslant \mathbb{P}\left(w_{t_0}(\hat{f}_2^n, n^{-\nu_1}) \geqslant \hat{\Theta}^n/2\right) + \mathbb{P}\left(\sqrt{n}e_{\max}^n \sum_k \mu_{2k} \geqslant \hat{\Theta}^n/2\right).$$

On $\bar{\tau}-\bar{\sigma} > n^{-\nu_1}$, $\hat{f}_2^n[\bar{\sigma},\bar{\tau}] \leqslant 2\|\hat{f}_2^n\|_{t_0}$ and $c_2\sqrt{n}(\bar{\tau}-\bar{\sigma}) \geqslant c_2 n^{\frac{1}{2}-\nu_1}$, so that

$$\mathbb{P}\left(\hat{f}_2^n[\bar{\sigma},\bar{\tau}] - c_2\sqrt{n}(\bar{\tau}-\bar{\sigma}) + \sqrt{n}e_{\max}^n \sum_k \mu_{2k} \geqslant \hat{\Theta}^n, \bar{\tau}-\bar{\sigma} > n^{-\nu_1}\right) \leqslant$$
$$\mathbb{P}\left(2\|\hat{f}_2^n\|_{t_0} \geqslant c_2 n^{\frac{1}{2}-\nu_1}/2\right) + \mathbb{P}\left(\sqrt{n}e_{\max}^n \sum_k \mu_{2k} \geqslant c_2 n^{\frac{1}{2}-\nu_1}/2\right).$$

To summarize,

$$\begin{aligned}\mathbb{P}(\rho \geqslant \bar{\tau}) &\leqslant \mathbb{P}(\tau_c^n \leqslant t_0) + \mathbb{P}(\tau_c^n \geqslant t_0,\, \widetilde{\tau} < \rho) + \mathbb{P}\left(\tau_c^n \geqslant t_0,\, \widetilde{\tau} \geqslant \rho,\, \bar{\tau} \leqslant \rho,\, \bar{\tau}-\bar{\sigma} \leqslant n^{-\nu_1}\right) \\ &\quad + \mathbb{P}\left(\tau_c^n \geqslant t_0,\, \widetilde{\tau} \geqslant \rho,\, \bar{\tau} \leqslant \rho,\, \bar{\tau}-\bar{\sigma} > n^{-\nu_1}\right) \\ &\leqslant \mathbb{P}(\tau_c^n \leqslant t_0) + \mathbb{P}(\tau_c^n \geqslant t_0,\, \widetilde{\tau} < \rho) + \mathbb{P}\left(w_{t_0}(\hat{f}_2^n, n^{-\nu_1}) \geqslant \hat{\Theta}^n/2\right) + \mathbb{P}\left(\sqrt{n}e_{\max}^n \sum_k \mu_{2k} \geqslant \hat{\Theta}^n/2\right) \\ &\quad + \mathbb{P}\left(2\|\hat{f}_2^n\|_{t_0} \geqslant c_2 n^{\frac{1}{2}-\nu_1}/2\right) + \mathbb{P}\left(\sqrt{n}e_{\max}^n \sum_k \mu_{2k} \geqslant c_2 n^{\frac{1}{2}-\nu_1}/2\right).\end{aligned}$$

All terms converge to zero. The first two have already been treated in Proposition 5.4 and (A.4) respectively. The third term converges by Lemma 5.7, the fourth and sixth by Lemma 5.9, and the fifth by tightness of $\hat{f}_2^n$.

This completes the proof of (5.26) and Lemma 5.12 in all of the relevant cases. □

Proof of Lemma 5.13.

Case 1(b), $\mathbf{T}_2$ policy:

We begin the proof by a full analysis in the case where a $\mathbf{T}_2$ policy is used, and then explain how to deal with the other cases. In this case, with the active mode in canonical form, $p = 1$. Recall the definition of the random time τ:

$$\tau = \inf \left\{ t > \rho : \hat{X}_1^n(t) = 2n^{-1/2}, \text{ and } \hat{X}_2^n(t) \geqslant \hat{\Theta}^n \right\},$$

and

$$\sigma = \sup \left\{ t \leqslant \tau^n : \hat{X}_1^n(t) \geqslant \hat{\Theta}^n, \text{ or } \hat{X}_2^n(t) \leqslant \frac{\hat{\Theta}^n}{2} \right\}.$$

For any $t \in [\sigma, \tau]$,

- $\hat{X}_1^n(t) < \hat{\Theta}^n$,
- and $\hat{X}_2^n(t) > \frac{\hat{\Theta}^n}{2}$.

In this case, only server 1 processes class 1 jobs and does so at most at full rate. Indeed, only server 1 gives priority to class 1 jobs and the other server is busy with class 2 jobs present in the system. Similarly, only server 2 processes class 2 jobs and does so at full rate since there are always class 2 jobs and the number of HPC jobs is below Θ^n between σ and τ. Similarly to Lemma 5.10, $e^n_{\max}$ is such that for any $s, t \in [\sigma, \tau]$, $s \leqslant t$, including a service that could start before σ at the wrong activity (server 2 in this case),

$$\int_s^t \big(\lambda_1 - \sum_k \mu_{1k} \Xi_{1k}^n(s)\big) ds \geqslant (\lambda_1 - \mu_{11})(t - s) - \mu_{12} e^n_{\max},$$

and excluding a service of a HPC job by server 2 starting before σ,

$$\int_s^t \big(\lambda_2 - \sum_k \mu_{2k} \Xi_{2k}^n(s)\big) ds \leqslant (\lambda_2 - \mu_{22})(t - s) + \mu_{22} e^n_{\max}.$$

With the active mode in canonical form, we have $\lambda_1 > \mu_{11}$. By (3.1), this means $\lambda_2 < \mu_{22}$. Thus $\lambda_1 - \mu_{11} > 0$ and $\lambda_2 - \mu_{22} < 0$.

There are two possibilities for the state of the system at time σ: we have either

- $\hat{X}_1^n(\sigma) = \hat{\Theta}^n - n^{-1/2}$,
- or $\hat{X}_2^n(\sigma) \in [\frac{\hat{\Theta}^n}{2} + \frac{1}{2} n^{-1/2}, \frac{\hat{\Theta}^n}{2} + n^{-1/2}]$.

We will decompose the event $\Omega^n := \{\tau \leqslant t_0\}$ in two events:

$$\Omega^n = \Omega_1^n \cup \Omega_2^n,$$

with

$$\Omega_1^n := \Omega^n \cap \left\{ \hat{X}_1^n(\sigma) = \hat{\Theta}^n - n^{-1/2} \right\},$$

and

$$\Omega_2^n := \Omega^n \cap \left\{ \hat{X}_2^n(\sigma) \in [\frac{\hat{\Theta}^n}{2} + \frac{1}{2} n^{-1/2}, \frac{\hat{\Theta}^n}{2} + n^{-1/2}] \right\}.$$

To lighten notation, introduce also

$$\widetilde{\Omega}^n := \left\{ \tau \leqslant t_0, \sup_{t \in [\sigma,\tau]} \hat{X}_1^n(t) < \hat{\Theta}^n, \inf_{t \in [\sigma,\tau]} \hat{X}_2^n(t) > \frac{\hat{\Theta}^n}{2} \right\}.$$

Similarly to Proposition 5.4, let $\nu_2 = 1/2 - \bar{a} \leqslant 1/4$ and $\nu_1 \in (\nu_2, 1/2)$ so that $\hat{\Theta}^n = n^{-1/2} \lceil n^{1/2-\nu_2} \rceil \geqslant n^{-\nu_2}$. Then

$$\begin{aligned} \mathbb{P}\left(\Omega_1^n\right) &= \mathbb{P}\left(\hat{X}_1^n(\sigma) = \hat{\Theta}^n - n^{-1/2}, \hat{X}_1^n(\tau) \leqslant 2n^{-1/2}, \widetilde{\Omega}^n \right) \\ &\leqslant \mathbb{P}\left(\hat{X}_1^n(\tau) - \hat{X}_1^n(\sigma) \leqslant -\frac{\hat{\Theta}^n}{2}, \widetilde{\Omega}^n \right) \\ &= \mathbb{P}\left(\hat{X}_1^n(\tau) - \hat{X}_1^n(\sigma) \leqslant -\frac{\hat{\Theta}^n}{2}, \widetilde{\Omega}^n, \tau - \sigma \leqslant n^{-\nu_1} \right) \\ &\quad + \mathbb{P}\left(\hat{X}_1^n(\tau) - \hat{X}_1^n(\sigma) \leqslant -\frac{\hat{\Theta}^n}{2}, \widetilde{\Omega}^n, \tau - \sigma > n^{-\nu_1} \right). \end{aligned}$$

On $\widetilde{\Omega}^n$, we also have

$$\begin{aligned} \hat{X}_1^n(\tau) - \hat{X}_1^n(\sigma) &= \hat{f}_1^n(\tau) - \hat{f}_1^n(\sigma) + \sqrt{n} \int_\sigma^\tau \left(\lambda_1 - \sum_k \mu_{1k} \Xi_{1k}^n(t) \right) dt \\ &\geqslant \hat{f}_1^n(\tau) - \hat{f}_1^n(\sigma) + (\lambda_1 - \mu_{11}) \sqrt{n}(\tau - \sigma) - \mu_{12} e_{\max}^n \sqrt{n}. \end{aligned}$$

On the event $\left\{ \widetilde{\Omega}^n, \tau - \sigma > n^{-\nu_1} \right\}$, the reasoning is very similar to the proof of Proposition 5.4:

$$\begin{aligned} &\mathbb{P}\left(\hat{X}_1^n(\tau) - \hat{X}_1^n(\sigma) \leqslant -\frac{\hat{\Theta}^n}{2}, \widetilde{\Omega}^n, \tau - \sigma > n^{-\nu_1} \right) \\ &\qquad \leqslant \mathbb{P}\left(2 \sup_{t \leqslant t_0} \left| \hat{f}_1^n(t) \right| + \mu_{12} \sqrt{n} e_{\max}^n \geqslant (\lambda_1 - \mu_{11}) \sqrt{n} n^{-\nu_1}/2 \right). \end{aligned} \tag{A.9}$$

The last probability goes to zero by tightness of $\hat{f}_1^n$, $\nu_1 < 1/2$ and Lemma 5.9. When $\tau - \sigma \leqslant n^{-\nu_1}$ the situation is again the same as in the proof of Proposition 5.4:

$$\begin{aligned} &\mathbb{P}\left(\hat{X}_1^n(\tau) - \hat{X}_1^n(\sigma) \leqslant -\frac{\hat{\Theta}^n}{2}, \widetilde{\Omega}^n, \tau - \sigma \leqslant n^{-\nu_1} \right) \\ &\qquad \leqslant \mathbb{P}\left(w_{t_0}(\hat{f}_1^n, n^{-\nu_1}) + \mu_{12} e_{\max}^n \sqrt{n} \geqslant \frac{n^{-\nu_2}}{2} \right). \end{aligned} \tag{A.10}$$

The right hand side also converges to zero by Lemmas 5.7, Remark 5.8 and Lemma 5.9, similarly to the proof of Proposition 5.4.

The second event Ω_2^n is treated similarly:

$$\begin{aligned}
\mathbb{P}\left(\Omega_2^n\right) &\leqslant \mathbb{P}\left(\hat{X}_2^n(\sigma) \in [\frac{\hat{\Theta}^n}{2} + \frac{1}{2}n^{-1/2}, \frac{\hat{\Theta}^n}{2} + n^{-1/2}],\, \hat{X}_2^n(\tau) \geqslant \hat{\Theta}^n,\, \widetilde{\Omega}^n\right) \\
&\leqslant \mathbb{P}\left(\hat{X}_2^n(\tau) - \hat{X}_2^n(\sigma) \geqslant \frac{\hat{\Theta}^n}{3}, \widetilde{\Omega}^n\right) \\
&= \mathbb{P}\left(\hat{X}_2^n(\tau) - \hat{X}_2^n(\sigma) \geqslant \frac{\hat{\Theta}^n}{3}, \widetilde{\Omega}^n, \tau - \sigma \leqslant n^{-\nu_1}\right) \\
&\quad + \mathbb{P}\left(\hat{X}_2^n(\tau) - \hat{X}_2^n(\sigma) \geqslant \frac{\hat{\Theta}^n}{3}, \widetilde{\Omega}^n, \tau - \sigma > n^{-\nu_1}\right).
\end{aligned}$$

On this event, we also have

$$\begin{aligned}
\hat{X}_2^n(\tau) - \hat{X}_2^n(\sigma) &= \hat{f}_2^n(\tau) - \hat{f}_2^n(\sigma) + \sqrt{n}\int_\sigma^\tau \left(\lambda_2 - \sum_k \mu_{2k}\Xi_{2k}^n(t)\right) dt \\
&\leqslant \hat{f}_2^n(\tau) - \hat{f}_2^n(\sigma) + (\lambda_2 - \mu_{22})\sqrt{n}(\tau - \sigma) + \mu_{22}e_{\max}^n\sqrt{n}.
\end{aligned}$$

On the event $\{\tau - \sigma > n^{-\nu_1}\}$, similar to the treatment of Ω_1^n:

$$\begin{aligned}
&\mathbb{P}\left(\hat{X}_2^n(\tau) - \hat{X}_2^n(\sigma) \geqslant \frac{\hat{\Theta}^n}{3}, \widetilde{\Omega}^n, \tau - \sigma > n^{-\nu_1}\right) \\
&\qquad \leqslant \mathbb{P}\left(2\sup_{t\leqslant t_0}|\hat{f}_2^n(t)| + \mu_{22}\sqrt{n}e_{\max}^n \geqslant -(\lambda_2 - \mu_{22})\sqrt{n}n^{-\nu_1}/3\right). \qquad \text{(A.11)}
\end{aligned}$$

The last probability goes to zero by tightness of $\hat{f}_2^n$, $\nu_1 < 1/2$ and Lemma 5.9. When $\tau - \sigma \leqslant n^{-\nu_1}$ the situation is also the same as for Ω_1^n:

$$\begin{aligned}
&\mathbb{P}\left(\hat{X}_2^n(\tau) - \hat{X}_2^n(\sigma) \geqslant \frac{\hat{\Theta}^n}{3}, \widetilde{\Omega}^n, \tau - \sigma \leqslant n^{-\nu_1}\right) \\
&\qquad \leqslant \mathbb{P}\left(w_{t_0}(\hat{f}_2^n, n^{-\nu_1}) + \mu_{22}e_{\max}^n\sqrt{n} \geqslant \frac{n^{-\nu_2}}{3}\right). \qquad \text{(A.12)}
\end{aligned}$$

The right hand side also converges to zero by Lemma 5.7 similarly to the Ω_1^n case. This concludes the proof in case 1(b).

We now present the changes required to adapt the proof to the other cases described in the lemma:

Case 2(b), $\mathbf{T}_2\mathbf{T}_2$ policy:

This case involves switching between two $\mathbf{T}_2$ rules: $i_2(\xi^L) = i_2(\xi^H) = p$. Because ξ^L is in canonical form, $p = 1$. In this case, mode switching only occurs at a service completion at the single activity server that is dedicated to HPC jobs. There could be either type of job being served at either server immediately before time σ. We will see that after those jobs exit the system, there is at least one server processing the LPC and at most one processing the HPC. If the first job to finish is from the dual activity server, the service of a low priority job starts because there are fewer than Θ^n HPC jobs. This server will continue to take LPC jobs until the minimum between the next mode switching time and τ. If the first job to finish at or after σ is at the single activity server, either the current mode switches or the service of an HPC job starts. If there is no mode switch then this server will continue to take HPC jobs until the minimum between the next mode switching time and τ. When the dual activity server completes its job it will, as noted before, serve LPC jobs. If there is a mode switch then the formerly single activity server becomes dual activity, and (because there are fewer than Θ^n HPC jobs) begins service on an LPC job. When the formerly dual activity server completes its job, it becomes single activity and serves HPC. This continues until the minimum between the next mode switching time and τ.

After the two jobs present at σ, there cannot be a time where both servers are occupied with HPC jobs. This is because the HPC is only processed by the server dedicated to it and there can be no residual service of an HPC job at the single activity server whenever the current mode switches. Similarly, after both initial jobs have been processed, there is always at least one server occupied with LPC jobs. This is because one server gives priority to the LPC and the service of a LPC job starts whenever the current mode switches.

Keeping the definition of τ, σ, just as in the single mode case, for any $t, s \in [\sigma, \tau]$, including/excluding a service that could start before σ at the wrong activity, there is always at most one server processing HPC jobs and at least one server processing LPC jobs between σ and τ. Thus,

$$\int_s^t \big(\lambda_1 - \sum_k \mu_{1k} \varXi^n_{1k}(s)\big) ds \geqslant (\lambda_1 - \max_k \mu_{1k})(t-s) - \sum_k \mu_{1k} e^n_{\max},$$

and

$$\int_s^t \big(\lambda_2 - \sum_k \mu_{2k} \varXi^n_{2k}(s)\big) ds \leqslant (\lambda_2 - \min_k \mu_{2k})(t-s) + \sum_k \mu_{2k} e^n_{\max}.$$

With one mode in canonical form we have $\frac{\lambda_1}{\alpha_1} > \beta_1$. In addition, by Lemma 3.3, in case 2(b), (3.6) holds and thus $\frac{\lambda_1}{\alpha_1} > \max_k \beta_k$, which also means $\frac{\lambda_2}{\alpha_2} < \min_k \beta_k$ by (3.1).

The rest of the proof is the same as in the single mode case, obtaining (A.9), (A.10), (A.11) and (A.12). This concludes the proof in case 2(b).

Case 2(c), $\mathbf{T}_1\mathbf{T}_2$ policy:

With ξ^L in canonical form, we have $p = 2$, so τ and σ are defined as

$$\tau = \inf \left\{ t > \rho : \hat{X}^n_2(t) = 2n^{-1/2}, \text{ and } \hat{X}^n_1(t) \geqslant \hat{\varTheta}^n \right\},$$

and

$$\sigma = \sup \left\{ t \leqslant \tau^n : \hat{X}^n_2(t) \geqslant \hat{\varTheta}^n, \text{ or } \hat{X}^n_1(t) \leqslant \frac{\hat{\varTheta}^n}{2} \right\}.$$

In this case, in the upper workload mode only the single activity server is allowed to begin service of the HPC between σ and τ, while in the lower workload mode neither server can begin service of HPC between σ and τ. Between those times, there are always low priority class jobs to serve, and the number of HPC jobs stays below Θ^n. The most service the HPC can get between σ and τ occurs if there are no switches and the current mode is always upper workload. In this mode, the single activity server (server 1) is dedicated to service of the HPC. Hence, including a service that could start before σ at the wrong activity,

$$\int_s^t \big(\lambda_2 - \sum_k \mu_{2k} \Xi^n_{2k}(s)\big) ds \geqslant (\lambda_2 - \mu_{21})(t-s) - \mu_{21} e^n_{\max}.$$

Between σ and τ, server 2 gives priority to class 1 regardless of switches. Excluding a service of a HPC job that could have started before σ,

$$\int_s^t \big(\lambda_1 - \sum_k \mu_{1k} \Xi^n_{1k}(s)\big) ds \leqslant (\lambda_1 - \mu_{12})(t-s) - \mu_{12} e^n_{\max}.$$

Because we have one mode in canonical form, $\frac{\lambda_1}{\alpha_1} > \beta_1$. In addition, by Lemma 3.3 in case 2(c), (3.5) holds, which means that $\frac{\lambda_1}{\alpha_1} < \beta_2$. Finally $\frac{\lambda_2}{\alpha_2} > \beta_1$ by the previous observation and (3.1).

The rest of the proof is the same as in the single mode case, obtaining (A.9), (A.10), (A.11) and (A.12).

Case 2(d), T$_2$T$_1$policy:

This case is handled the same way as 2(c), by interchanging the roles of upper workload and lower workload mode. □

Proof of Lemma 5.11 (continued from §5.2.5).

- **Case 1(b), T$_2$ policy:**

In this case, we already know by Lemma 5.12 that $\mathbb{P}\left(\bar{R}^n_\rho \geqslant \frac{\delta}{2}\right) \to 0$. We need to show the same thing for the time after ρ:

$$\mathbb{P}\left(\int_\rho^{t_0} \mathbb{1}_{\hat{W}^n_t \geqslant c_3 \hat{\Theta}^n} d\hat{L}^n_t \geqslant \frac{\delta}{2}\right) \to 0.$$

First, by putting ξ^A in canonical form, $p=1$. When $X^n_1(t)$ is above the threshold, both servers can serve class 1 jobs (high priority class) so almost surely,

$$\int_\rho^{t_0} \mathbb{1}_{\hat{X}^n_1(t) \geqslant \hat{\Theta}^n} d\hat{L}^n_t = 0. \tag{A.13}$$

Similarly,

$$\int_\rho^{t_0} \mathbb{1}_{\hat{X}^n_2(t) \geqslant 2n^{-1/2}} d\hat{I}^n_2(t) = 0, \tag{A.14}$$

and

$$\int_\rho^{t_0} \mathbb{1}_{2n^{-1/2} \leqslant \hat{X}^n_1(t) \leqslant \hat{\Theta}^n} d\hat{I}^n_1(t) = 0. \tag{A.15}$$

Notice now that because of the three identities,

$$\begin{aligned}\mathbb{P}\left(\int_\rho^{t_0} \mathbb{1}_{W_t^n \geqslant c_3 \Theta^n} d\hat{L}_t^n \geqslant \frac{\delta}{2}\right) &\leqslant \mathbb{P}\left(\int_\rho^{t_0} \left(\mathbb{1}_{\hat{X}_1^n(t) \geqslant \hat{\Theta}^n} + \mathbb{1}_{\hat{X}_1^n(t) \leqslant \hat{\Theta}^n, \hat{X}_2^n \geqslant \hat{\Theta}^n}\right) d\hat{L}_t^n \geqslant \frac{\delta}{2}\right)\\ &= \mathbb{P}\left(\beta_1 \int_\rho^{t_0} \mathbb{1}_{\hat{X}_1^n(t) \leqslant \hat{\Theta}^n, \hat{X}_2^n \geqslant \hat{\Theta}^n} d\hat{I}_1^n(t) \geqslant \frac{\delta}{2}\right)\\ &= \mathbb{P}\left(\beta_1 \int_\rho^{t_0} \mathbb{1}_{\hat{X}_1^n(t) \leqslant 2n^{-1/2}, \hat{X}_2^n \geqslant \hat{\Theta}^n} d\hat{I}_1^n(t) \geqslant \frac{\delta}{2}\right)\end{aligned}$$

By Lemma 5.13,

$$\mathbb{P}\left(\int_\rho^{t_0} \mathbb{1}_{\hat{X}_1^n(t) \leqslant 2n^{-1/2}, \hat{X}_2^n \geqslant \hat{\Theta}^n} d\hat{I}_1^n(t) \geqslant \frac{\delta}{2}\right) \leqslant \mathbb{P}\left(\tau_r^n \leqslant t_0\right) \to 0. \tag{A.16}$$

- **Case 2(a), PP policy:**

In this case, $p = 2$. The reasoning in case 1(a) is still valid when switching between two P rules because (5.28) and (5.29) still hold.

- **Case 2(b), $\mathbf{T}_2\mathbf{T}_2$ policy:**

The result in this case has a proof very similar to case 1(b) because Lemmas 5.12 and 5.13 are still valid in this case. We already know by Lemma 5.12 that

$$\mathbb{P}\left(\int_0^{\rho} \mathbb{1}_{\hat{W}_t^n \geqslant c_3 \hat{\Theta}^n} d\hat{L}_t^n \geqslant \frac{\delta}{2}\right) \to 0.$$

We need to show the same thing for the time after ρ:

$$\mathbb{P}\left(\int_\rho^{t_0} \mathbb{1}_{\hat{W}_t^n \geqslant c_3 \hat{\Theta}^n} d\hat{L}_t^n \geqslant \frac{\delta}{2}\right) \to 0.$$

First, by putting ξ^L in canonical form $p = 1$.

The idea is to split the integrals between the times MODE$(t) = \xi^L$ and the times MODE$(t) = \xi^H$. In this case, we still have (A.13) but this time (A.14) and (A.15) only hold when MODE$(t) = \xi^L$. In the other case, we have

$$\int_\rho^{t_0} \mathbb{1}_{\hat{X}_2^n(t) \geqslant 2n^{-1/2}, \mathrm{MODE}(t) = \xi^H} d\hat{I}_1^n(t) = 0,$$

$$\int_\rho^{t_0} \mathbb{1}_{2n^{-1/2} \leqslant \hat{X}_1^n(t) \leqslant \hat{\Theta}^n, \mathrm{MODE}(t) = \xi^H} d\hat{I}_2^n(t) = 0.$$

We obtain

$$\int_\rho^{t_0} \mathbb{1}_{\mathrm{MODE}(t) = \xi^L, W_t^n \geqslant c_3 \Theta^n} d\hat{L}_t^n \leqslant \beta_1 \int_\rho^{t_0} \mathbb{1}_{\mathrm{MODE}(t) = \xi^L, \hat{X}_1^n(t) \leqslant 2n^{-1/2}, \hat{X}_2^n(t) \geqslant \hat{\Theta}^n} d\hat{I}_1^n(t),$$

and

$$\int_{\rho}^{t_0} \mathbb{1}_{\mathrm{MODE}(t)=\xi^H, W_t^n \geqslant c_3 \varTheta^n} d\hat{L}_t^n \leqslant \beta_2 \int_{\rho}^{t_0} \mathbb{1}_{\mathrm{MODE}(t)=\xi^H, \hat{X}_1^n(t) \leqslant 2n^{-1/2}, \hat{X}_2^n(t) \geqslant \hat{\varTheta}^n} d\hat{I}_2^n(t).$$

Finally, we obtain the result, since

$$\mathbb{P}\left(\int_{\rho}^{t_0} \mathbb{1}_{\hat{X}_1^n(t) \leqslant 2n^{-1/2}, \hat{X}_2^n(t) \geqslant \hat{\varTheta}^n} d\hat{L}_t^n \geqslant \frac{\delta}{2}\right) \leqslant \mathbb{P}\left(\tau_r^n \leqslant t_0\right) \to 0.$$

- <u>**Case 2(c)(d), T$_1$T$_2$/T$_2$T$_1$ policy:**</u>

We give the proof in case 2(d) but the reasoning is similar in case 2(c) by interchanging the role of ξ^L and ξ^H. In this case $p = 1$ with ξ^L in canonical form. The idea is similar to the 2(b) case. We already know by Lemma 5.12 that

$$\mathbb{P}\left(\int_{0}^{\rho} \mathbb{1}_{\hat{W}_t^n \geqslant c_3 \hat{\varTheta}^n} d\hat{L}_t^n \geqslant \frac{\delta}{2}\right) \to 0.$$

We need to show the same thing for the time after ρ:

$$\mathbb{P}\left(\int_{\rho}^{t_0} \mathbb{1}_{\hat{W}_t^n \geqslant c_3 \hat{\varTheta}^n} d\hat{L}_t^n \geqslant \frac{\delta}{2}\right) \to 0.$$

We will split the integral using $\mathbb{1}_{\mathrm{MODE}(t)=\xi^L} + \mathbb{1}_{\mathrm{MODE}(t)=\xi^H}$. We use a $\mathbf{T}_2$ rule when W^n is small and a $\mathbf{T}_1$ rule when W^n is large but that distinction is not important here. In terms of almost sure non idling properties, we have

$$\int_{\rho}^{t_0} \mathbb{1}_{\mathrm{MODE}(t)=\xi^L, \hat{X}_1^n(t) \geqslant \hat{\varTheta}^n} d\hat{L}_t^n = 0, \qquad \int_{\rho}^{t_0} \mathbb{1}_{\mathrm{MODE}(t)=\xi^L, \hat{X}_2^n(t) \geqslant 2n^{-1/2}} d\hat{I}_2^n(t)(t) = 0,$$

$$\int_{\rho}^{t_0} \mathbb{1}_{\mathrm{MODE}(t)=\xi^L, 2n^{-1/2} \leqslant \hat{X}_1^n(t) \leqslant \hat{\varTheta}^n} d\hat{I}_1^n(t) = 0, \qquad \int_{\rho}^{t_0} \mathbb{1}_{\mathrm{MODE}(t)=\xi^H, \hat{X}_2^n \geqslant 2n^{-1/2}} d\hat{L}_t^n = 0.$$

From these almost sure identities, we obtain

$$\begin{aligned}
\mathbb{P}\left(\int_{\rho}^{t_0} \mathbb{1}_{W_t^n \geqslant c_3 \varTheta^n} d\hat{L}_t^n \geqslant \frac{\delta}{2}\right) &\leqslant \mathbb{P}\left(\int_{\rho}^{t_0} \mathbb{1}_{\mathrm{MODE}(t)=\xi^L, W_t^n \geqslant c_3 \varTheta^n} d\hat{L}_t^n \geqslant \frac{\delta}{4}\right) \\
&\quad + \mathbb{P}\left(\int_{\rho}^{t_0} \mathbb{1}_{\mathrm{MODE}(t)=\xi^H, W_t^n \geqslant c_3 \varTheta^n} d\hat{L}_t^n \geqslant \frac{\delta}{4}\right) \\
&\leqslant \mathbb{P}\left(\beta_1 \int_{\rho}^{t_0} \mathbb{1}_{\mathrm{MODE}(t)=\xi^L, \hat{X}_1^n(t) \leqslant 2n^{-1/2}, \hat{X}_2^n(t) \geqslant \hat{\varTheta}^n} d\hat{I}_1^n(t) \geqslant \frac{\delta}{4}\right) \\
&\quad + \mathbb{P}\left(\int_{\rho}^{t_0} \mathbb{1}_{\mathrm{MODE}(t)=\xi^H, \hat{X}_2^n(t) \leqslant 2n^{-1/2}, \hat{X}_1^n(t) \geqslant 2\hat{\varTheta}^n} d\hat{L}_t^n \geqslant \frac{\delta}{4}\right) \\
&\leqslant \mathbb{P}\left(\tau_r^n \leqslant t_0\right) + \mathbb{P}\left(\tau_c^n \leqslant t_0\right)
\end{aligned}$$

Both probabilities go to zero (by Lemma 5.13 for the first and Proposition 5.4 for the second) so the result is proved in this case as well.

As mentionened, the proof is the same in case 2(c): this time, the policy uses a $\mathbf{T}_1$ rule when W^n is small and $\mathbf{T}_2$ rule when W^n is large. Keeping ξ^L in canonical form, $p = 2$. In addition, in terms of almost sure non idling properties, we have

$$\int_\rho^{t_0} \mathbb{1}_{\mathrm{MODE}(t)=\xi^H, \hat{X}_2^n(t) \geqslant \hat{\Theta}^n} d\hat{L}_t^n = 0, \qquad \int_\rho^{t_0} \mathbb{1}_{\mathrm{MODE}(t)=\xi^H, \hat{X}_1^n(t) \geqslant 2n^{-1/2}} d\hat{I}_2^n(t) = 0,$$

$$\int_\rho^{t_0} \mathbb{1}_{\mathrm{MODE}(t)=\xi^H, 2n^{-1/2} \leqslant \hat{X}_2^n(t) \leqslant \hat{\Theta}^n} d\hat{I}_1^n(t) = 0, \qquad \int_\rho^{t_0} \mathbb{1}_{\mathrm{MODE}(t)=\xi^L, \hat{X}_1^n \geqslant 2n^{-1/2}} d\hat{L}_t^n = 0.$$

We can obtain the result using the same decomposition. □

B Solution of the HJB equation and free boundary point

In this appendix we present the expression found in [17, Section 3, Case 1] for the solution to the HJB equation. (This solution is a slightly corrected version of the solution presented in [16, Section 5.3].) It includes an equation that uniquely characterizes the free boundary point (or switching point) z^* in the dual mode case. It is assumed in [17], without loss of generality, that $b_1 \geqslant b_2$. We assume further, for simplicity, (and, again, without loss of generality) that if $b_1 = b_2$ then $\sigma_1 \geqslant \sigma_2$. Note that, with these indexing assumptions, $m = 2$ in whichever of the complementary conditions (2.20) or (2.21) that holds. Throughout this section, denote the unique classical solution to (2.18)–(2.19) by $u(x)$, $x \in \mathbb{R}_+$, and let x serve as the initial condition for the WCP, which elsewhere in this paper is denoted by z. Let

$$\beta = \frac{b_1 + \sqrt{b_1^2 + 2\gamma\sigma_1^2}}{\sigma_1^2}, \qquad \rho = \frac{b_2 + \sqrt{b_2^2 + 2\gamma\sigma_2^2}}{\sigma_2^2}, \qquad \nu = \frac{b_1 - \sqrt{b_1^2 + 2\gamma\sigma_1^2}}{\sigma_1^2}.$$

Theorem B.1 ([17, Section 3, Case 1])**.** *Under* (2.20), $u(x) = \frac{x}{\gamma} + \frac{b_2}{\gamma^2} + \frac{1}{\gamma\rho} e^{-\rho x}$, $x \in \mathbb{R}_+$.

Next, consider condition (2.21). Because b_1 and b_2 are distinct in this case, we have $b_1 > b_2$. The policy (4.1) from §4 corresponding to switching at z is given in the present notation by $\bar{\xi}_z(x) = \xi^{*,1}\mathbb{1}_{x \leq z} + \xi^{*,2}\mathbb{1}_{x > z}$. Let $\mathfrak{S}_z^{(2)}$ be the admissible control system from Lemma 4.1.2, with a generic switching point z in place of the specific z^*. Let the corresponding expression $J_{\mathrm{WCP}}(x, \mathfrak{S}_z^{(2)})$, which is nothing but the cost associated with the switching policy $\bar{\xi}_z$, be denoted by $J(x, \bar{\xi}^z)$. Following is an expression for this cost. For $z > 0$, let

$$\begin{aligned}
A(z) &= \frac{\nu\beta(e^{(\nu-\beta)z} + e^{-(\nu-\beta)z} - 2)}{\nu\beta(e^{(\nu-\beta)z} + e^{-(\nu-\beta)z} - 2) + \rho(e^{-\nu z} - e^{-\beta z})(\nu e^{\nu z} - \beta e^{\beta z})}, \\
C(z) &= \frac{(\nu - \beta)(e^{-\nu z} - e^{-\beta z})}{\nu\beta(e^{(\nu-\beta)z} + e^{-(\nu-\beta)z} - 2) + \rho(e^{-\nu z} - e^{-\beta z})(\nu e^{\nu z} - \beta e^{\beta z})}, \\
D(z) &= \frac{(\beta - \nu)\rho(e^{\beta z} - e^{\nu z})}{\nu\beta(e^{(\nu-\beta)z} + e^{-(\nu-\beta)z} - 2) + \rho(e^{-\nu z} - e^{-\beta z})(\nu e^{\nu z} - \beta e^{\beta z})}, \\
F(z) &= \frac{e^{-\nu z} - e^{-\beta z} + (\beta - \nu)C(z)}{\nu e^{-\beta z} - \beta e^{-\nu z}}.
\end{aligned}$$

($F(\cdot)$ is not to be confused with the process F defined in the body of the paper). Then, for $x \leqslant z$,

$$\begin{aligned} J(x,\bar{\xi}^z) = \frac{x}{\gamma} + \frac{b_1\left[(e^{-\nu z}-e^{-\beta z}) + (A(z)-1)(e^{-\nu x}-e^{-\beta x}) + D(z)(e^{-\nu z-\beta x}-e^{-\beta z-\nu x})\right]}{\gamma^2(e^{-\nu z}-e^{-\beta z})} \\ + \frac{b_2\left[(1-A(z))(e^{-\nu x}-e^{-\beta x}) - D(z)(e^{-\nu z-\beta x}-e^{-\beta z-\nu x})\right]}{\gamma^2(e^{-\nu z}-e^{-\beta z})} \\ + \frac{C(z)(e^{-\nu x}-e^{-\beta x}) - F(z)(e^{-\nu z-\beta x}-e^{-\beta z-\nu x})}{\gamma(e^{-\nu z}-e^{-\beta z})}, \end{aligned} \tag{B.1}$$

and for $x \geqslant z$,

$$J(x,\bar{\xi}^z) = \frac{x}{\gamma} + \frac{b_1 A(z)}{\gamma^2 e^{\rho(x-z)}} + \frac{b_2}{\gamma^2}(1-e^{-\rho(x-z)}A(z)) + \frac{C(z)}{\gamma e^{\rho(x-z)}}. \tag{B.2}$$

It is here where the principle of smooth fit is applied. For the cost to be C^2 (in x), it must satisfy $J''(z-,\bar{\xi}^z) = J''(z+,\bar{\xi}^z)$. Using the expressions (B.1) and (B.2), this condition can be translated to the following equation

$$\begin{aligned} 0 = \left(\frac{b_1-b_2}{\gamma^2}\right)\left[\frac{(A(z)-1)(\nu^2 e^{-\nu z}-\beta^2 e^{-\beta z})}{e^{-\nu z}-e^{-\beta z}} - \rho^2 A(z) + \frac{(\beta^2-\nu^2)D(z)e^{-(\nu+\beta)z}}{e^{-\nu z}-e^{-\beta z}}\right] \\ + \frac{1}{\gamma}\left[\frac{C(z)((\nu^2 e^{-\nu z}-\beta^2 e^{-\beta z})}{e^{-\nu z}-e^{-\beta z}} - \rho^2 C(z) - \frac{(\beta^2-\nu^2)F(z)e^{-(\nu+\beta)z}}{e^{-\nu z}-e^{-\beta z}}\right]. \end{aligned} \tag{B.3}$$

Theorem B.2 ([17, Section 3, Case 1]). *Let* (2.21) *hold. Then* (B.3) *has a unique solution* $z^* \in (0,\infty)$. *Moreover,* $u(x) = J(x,\bar{\xi}^{z^*})$, $x \in \mathbb{R}_+$.

C Symmetry conditions

The following result is related to Remark 4.2.

Lemma C.1. *1. If either* (4.2) *or* (4.3) *holds then* $b_1 = b_2$. *In particular,* (2.20) *holds.*
2. If (4.4) *holds then* $\sigma_1 = \sigma_2$. *In particular,* (2.20) *holds.*

Proof. 1. Recall the expressions for b_1 and b_2,

$$b_m = b(\xi^{*,m}) = \sum_i \frac{\hat{\lambda}_i - \sum_k \hat{\mu}_{ik}\xi^{*,m}_{ik}}{\alpha_i}.$$

The difference between b_1 and b_2 is thus the difference between $\gamma_m := \sum_{i,k} \frac{\hat{\mu}_{ik}\xi^{*,m}_{ik}}{\alpha_i}$. For $\xi^{*,1}$, we distinguish these cases: either $\frac{\lambda_1}{\alpha_1} > \beta_2$ or $\frac{\lambda_1}{\alpha_1} < \beta_2$. For $\xi^{*,2}$, we distinguish these cases: either

$\frac{\lambda_1}{\alpha_1} < \beta_1$ or $\frac{\lambda_1}{\alpha_1} > \beta_1$. We will see that in each of the four cases $b_1 = b_2$.

$$\begin{aligned}
\frac{\lambda_1}{\alpha_1} > \beta_2 : &\quad \gamma_1 = \frac{1}{\alpha_1}\left[\hat\mu_{12} + \hat\mu_{11}(\frac{\lambda_1}{\alpha_1\beta_1} - \frac{\beta_2}{\beta_1})\right] + \frac{\hat\mu_{21}}{\alpha_2}(1 - \frac{\lambda_1}{\alpha_1\beta_1} + \frac{\beta_2}{\beta_1}),\\
\frac{\lambda_1}{\alpha_1} < \beta_2 : &\quad \gamma_1 = \frac{\hat\mu_{12}\lambda_1}{\alpha_1^2\beta_2} + \frac{1}{\alpha_2}\left[\hat\mu_{21} + \hat\mu_{22}(1 - \frac{\lambda_1}{\alpha_1\beta_2})\right],\\
\frac{\lambda_1}{\alpha_1} < \beta_1 : &\quad \gamma_2 = \frac{\hat\mu_{11}\lambda_1}{\alpha_1^2\beta_1} + \frac{1}{\alpha_2}\left[\hat\mu_{22} + \hat\mu_{21}(1 - \frac{\lambda_1}{\alpha_1\beta_1})\right],\\
\frac{\lambda_1}{\alpha_1} > \beta_1 : &\quad \gamma_2 = \frac{1}{\alpha_1}\left[\hat\mu_{11} + \hat\mu_{12}(\frac{\lambda_1}{\alpha_1\beta_2} - \frac{\beta_1}{\beta_2})\right] + \frac{\hat\mu_{22}}{\alpha_2}(1 - \frac{\lambda_1}{\alpha_1\beta_2} + \frac{\beta_1}{\beta_2}).
\end{aligned}$$

We now take the difference for each pair:

$$\begin{aligned}
\frac{\lambda_1}{\alpha_1} > \beta_2 \,\&\, \frac{\lambda_1}{\alpha_1} < \beta_1 : &\quad b_1 - b_2 = \frac{\hat\mu_{12} - \hat\mu_{11}\frac{\beta_2}{\beta_1}}{\alpha_1} - \frac{\hat\mu_{22} - \hat\mu_{21}\frac{\beta_2}{\beta_1}}{\alpha_2},\\
\frac{\lambda_1}{\alpha_1} > \beta_2 \,\&\, \frac{\lambda_1}{\alpha_1} > \beta_1 : &\quad b_1 - b_2 = (1 - \frac{\lambda_1}{\alpha_1\beta_1} + \frac{\beta_2}{\beta_1})\left[\frac{1}{\alpha_1}(\hat\mu_{12}\frac{\beta_1}{\beta_2} - \hat\mu_{11}) + \frac{1}{\alpha_2}(\hat\mu_{21} - \hat\mu_{22}\frac{\beta_1}{\beta_2})\right],\\
\frac{\lambda_1}{\alpha_1} < \beta_2 \,\&\, \frac{\lambda_1}{\alpha_1} < \beta_1 : &\quad b_1 - b_2 = \frac{\lambda_1}{\alpha_1^2}\left[\frac{\hat\mu_{12}}{\beta_2} - \frac{\hat\mu_{11}}{\beta_1}\right] + \frac{\lambda_1}{\alpha_1\alpha_2}\left[\frac{\hat\mu_{21}}{\beta_1} - \frac{\hat\mu_{22}}{\beta_2}\right],\\
\frac{\lambda_1}{\alpha_1} < \beta_2 \,\&\, \frac{\lambda_1}{\alpha_1} > \beta_1 : &\quad b_1 - b_2 = \frac{\hat\mu_{12}\frac{\beta_1}{\beta_2} - \hat\mu_{11}}{\alpha_1} + \frac{\hat\mu_{21} - \hat\mu_{22}\frac{\beta_1}{\beta_2}}{\alpha_2}.
\end{aligned}$$

Now, if $\frac{\hat\mu_{i1}}{\beta_1} = \frac{\hat\mu_{i,2}}{\beta_2}$, $i = 1, 2$, we get

$$\hat\mu_{11}\frac{\beta_2}{\beta_1} - \hat\mu_{12} = \hat\mu_{22} - \hat\mu_{21}\frac{\beta_2}{\beta_1} = 0,$$

and consequently $b_1 = b_2$ as claimed. If $\frac{\hat\mu_{1k}}{\alpha_1} = \frac{\hat\mu_{2k}}{\alpha_2}$, $k = 1, 2$, it is not hard to see that again the expressions can be rewritten with a different factorization to get $b_1 = b_2$.

2. Under (4.4), denote $C_i = C_{S_{i1}} = C_{S_{i2}}$. Then for $\xi \in \mathcal{S}_{\mathrm{LP}}$,

$$\begin{aligned}
\sigma(\xi)^2 &= \sum_i \frac{\sigma_{A,i}^2 + \sum_k \sigma_{S_{ik}}^2 \xi_{ik}}{\alpha_i^2}\\
&= \sum_i \frac{\sigma_{A,i}^2 + \sum_k C_i^2 \mu_{ik}\xi_{ik}}{\alpha_i^2}\\
&= \sum_i \frac{\sigma_{A,i}^2 + C_i^2\lambda_i}{\alpha_i^2},
\end{aligned}$$

where the last equality follows from (2.7). Hence $\sigma_1 = \sigma_2$ as claimed. □

Acknowledgement. We are very grateful to Cristina Costantini for providing us with the proof of Lemma 4.1 Part 3 and to two anonymous referees, as well as the anonymous associate editor, for their thoughtful suggestions that helped improve the exposition considerably. RA is supported by ISF grant 1035/20.

References

[1] R. Atar, E. Castiel, and M. Reiman. Parallel server systems under an extended heavy traffic condition: A lower bound. *arXiv:2201.07855*, *Ann. Appl. Probab.*, to appear.

[2] R. Atar and A. Lev-Ari. Workload-dependent dynamic priority for the multiclass queue with reneging. *Mathematics of Operations Research*, 43(2):494–515, 2018.

[3] S. L. Bell and R. J. Williams. Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: asymptotic optimality of a threshold policy. *Ann. Appl. Probab.*, 11(3):608–649, 2001.

[4] P. Billingsley. *Convergence of Probability Measures.* Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999. x+277 pp. A Wiley-Interscience Publication.

[5] S. N. Ethier and T. G. Kurtz. *Markov Processes, Characterization and Convergence.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1986. x+534 pp.

[6] W. H. Fleming and H. M. Soner. *Controlled Markov Processes and Viscosity Solutions*, volume 25. Springer Science & Business Media, 2006.

[7] D. Goldfarb and M. J. Todd. Chapter II linear programming. *Handbooks in Operations Research and Management Science*, 1:73–170, 1989.

[8] J. M. Harrison. Heavy traffic analysis of a system with parallel servers: asymptotic optimality of discrete-review policies. *Annals of Applied Probability*, pages 822–848, 1998.

[9] J. M. Harrison and M. J. López. Heavy traffic resource pooling in parallel-server systems. *Queueing Systems Theory Appl.*, 33(4):339–368, 1999. ISSN 0257-0130. URL `https://doi.org/10.1023/A:1019188531950`.

[10] E. V. Krichagina and M. I. Taksar. Diffusion approximation for gi/g/1 controlled queues. *Queuing Syst.*, 12: 333–368, 05 1992.

[11] N. V. Krylov and R. Liptser. On diffusion approximation with discontinuous coefficients. *Stochastic Processes and Their Applications*, 102(2):235–264, 2002.

[12] T. G. Kurtz and P. Protter. Weak limit theorems for stochastic integrals and stochastic differential equations. *Ann. Probab.*, 19(3):1035–1070, 1991.

[13] R. S. Liptser and A. N. Shiriaev. *Statistics of random processes: General theory*, volume 394. Springer, 1977.

[14] D. Revuz and M. Yor. *Continuous Martingales and Brownian Motion*, volume 293 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences].* Springer-Verlag, Berlin, third edition, 1999. xiv+602 pp.

[15] A. Schrijver. *Theory of Linear and Integer Programming.* Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, Ltd., Chichester, 1986. xii+471 pp.

[16] D. Sheng. *Some Problems in the Optimal Control of Diffusions.* PhD Thesis, Stanford University, California, 1978.

[17] D. Sheng. Two-mode control of reflecting brownian motion. *Unpublished manuscript*, 1981.

[18] D. W. Stroock and S. S. Varadhan. *Multidimensional Diffusion Processes*, volume 233. Springer Science & Business Media, 1997.