

# PARALLEL SERVER SYSTEMS UNDER AN EXTENDED HEAVY TRAFFIC CONDITION: A LOWER BOUND

BY RAMI ATAR<sup>1,\*</sup>, EYAL CASTIEL<sup>1,†</sup> AND MARTIN I. REIMAN<sup>2,‡</sup>

<sup>1</sup>Viterbi Faculty of Electrical and Computer Engineering, Technion, Haifa, Israel, \*[rami@technion.ac.il](mailto:rami@technion.ac.il);

†[eyal.ca@campus.technion.ac.il](mailto:eyal.ca@campus.technion.ac.il)

<sup>2</sup>Department of Industrial Engineering and Operations Research, Columbia University, New York City, NY,

‡[martyreiman@gmail.com](mailto:martyreiman@gmail.com)

The standard setting for studying parallel server systems (PSS) at the diffusion scale is based on the heavy traffic condition (HTC), which assumes that the underlying static allocation linear program (LP) is critical and has a unique solution. This solution determines the graph of basic activities, which identifies the set of activities (i.e., class–server pairs) that are operational. In this paper we explore the extended HTC, where the LP is merely assumed to be critical. Because multiple solutions are allowed, multiple sets of operational activities, referred to as modes, are available. Formally, the scaling limit for the control problem associated with the model is given by a so-called workload control problem (WCP) in which a cost associated with a diffusion process is to be minimized by dynamically switching between these modes. Our main result is that the WCP’s value constitutes an asymptotic lower bound on the cost associated with the PSS model.

## CONTENTS

1	Introduction . . . . .	1
2	Model and main results . . . . .	6
3	The LP under the EHTC . . . . .	17
4	The BCP, WCP and HJB equation . . . . .	20
5	An asymptotic lower bound . . . . .	31
A	Some useful linear programming results . . . . .	41
	Funding . . . . .	42
	References . . . . .	42

## 1. Introduction.

1.1. *Background and motivation.* Parallel server systems (PSSs) constitute a class of queueing control problems that have received a lot of attention since their introduction in [15]. The overwhelming majority of this effort has been on heavy traffic limits and associated asymptotic optimality (AO). Almost all of this work has been carried out under the assumption that a particular linear program (LP), called the ‘static allocation problem’ in [15] has a unique solution. (This LP was called the static planning problem in [14], which deals with a broader set of problems.) In particular, the notion of heavy traffic for PSS defined in [15] requires that the static allocation problem has a unique solution. But it is not difficult to construct a PSS that is intuitively in ‘heavy traffic’ and yet, because its associated static allocation problem does not have a unique solution, does not satisfy the the heavy traffic

---

*MSC2020 subject classifications:* Primary: 60K25; 90B22; 68M20; 49K45; Secondary: 60F17; 93E20; 49K15.

*Keywords and phrases:* parallel server systems, heavy traffic, diffusion limits, Brownian control problem, Hamilton-Jacobi-Bellman equation, strict complementary slackness.

definition of [15]. Thus what is termed heavy traffic in [15] is really, in a sense, *restricted* heavy traffic. In this paper we remove the requirement that the solution of the static planning problem is unique. Because the term heavy traffic in the context of PSS has already been defined in [15], we use the term extended heavy traffic.

A few further comments on the uniqueness assumption are in order. First, this assumption carries with it the benefit of substantial simplification of the problem, described further below. As a consequence, as also described further below, there has been great progress on this problem in [2],[3], [15], and [22] under the uniqueness assumption. On the other hand, there is no strong reason to believe that the uniqueness assumption holds in typical problems of practical interest. For example, it fails to hold under natural properties arising in the design of service systems, such as the decomposability condition described at the end of this subsection. Thus it seems reasonable to suspect that it often does not hold. While the great progress just cited justifies the insightful restriction of uniqueness as a temporary expediency, we feel that it is time to move beyond it. The purpose of this paper is to begin the exploration of what happens if the uniqueness assumption does not hold. This paper provides an asymptotic lower bound for a general PSS. A companion paper [1] provides asymptotic optimality results for the simplest non-trivial PSS where uniqueness fails to hold.

The specific system that we consider, described in more detail in §2, has  $I$  classes of customers and  $K$  servers. A class-server pair  $j = (i, k)$  is said to be an activity if server  $k$  is capable of serving class  $i$ . The number of activities is denoted by  $J$ . The service rate  $\mu_j$  depends on the activity, namely the class-server combination. As is the case in all PSS models, customers arrive at a class based queue where they await service, receive a single service, and leave the system. The controls available here are both routing (which server serves each customer) and sequencing (in what order they are served). Our objective is to minimize the expected infinite horizon total discounted cost, where the cost rate is a linear function of the queue length.

Despite the simple specification of this control problem, determining an optimal control is typically very challenging. This motivated Harrison and Lopez [15] to look for control policies that are AO in heavy traffic. When the control problem involves only sequencing - the order in which customers are served, the traffic intensity can be obtained immediately from the arrival and service rates. On the other hand, when the control problem involves routing, as it does in PSSs, the traffic intensity may depend on the control policy that is used. As a consequence, it is not immediately obvious how to define heavy traffic. This is the context in which Harrison [14] introduced the static planning problem, noting the connection to earlier work of [21], which introduced a multicommodity network flow problem to define heavy traffic for dynamic routing problems.

Harrison [12] pioneered the notion of using Brownian Control Problems (BCPs) to obtain AO control policies for queueing network control problems. As described in [15], the solution of such a problem can be broken into 4 steps: (1) Derive a BCP that plausibly approximates the original control problem; (2) Solve the BCP; (3) Translate the solution of the BCP into a control policy for the original problem; and (4) Prove that this policy is AO.

In general, the BCP that arises from queueing control problems, including PSSs, is multidimensional, and consequently difficult to solve. It was shown in [16] how state space collapse can lead to an ‘equivalent workload formulation’ of lower dimension. In the context of PSSs, a dramatically simplifying condition known as ‘complete resource pooling’ (CRP), under which, in heavy traffic, the servers can effectively be ‘pooled’ together into a single ‘super server’, was introduced in [15]. They showed that the CRP condition yields a BCP whose equivalent workload formulation is one dimensional.

A key assumption in both [14], [15] is that the static planning LP has a unique solution. This assumption, combined with CRP, gives rise to a substantial simplification: It was shown

in [15] that the equivalent workload formulation of the BCP has a simple pathwise solution, which minimizes the instantaneous cost rate at every point in time with probability one, completing steps (1) and (2) under the uniqueness assumption. Almost all subsequent work that we are aware of has involved this assumption. The exceptions (that we are aware of) are described below. It was understood and acknowledged that relaxing the uniqueness assumption would lead to additional difficulties. For example, in [15] Harrison and Lopez note that ‘multiple optima lead to Brownian approximations of a more complicated form than we are prepared to treat’.

Steps (3) and (4) above were first completed in the context of a simple special case known as an ‘N’ network. This has 2 customer classes, 2 servers, and 3 activities. (Server 1 can only serve class 1.) The ‘N’ structure assures that the static allocation LP can have at most 1 solution. An ‘N’ network with specific distributional assumptions and parameters was considered in [13], where a discrete review policy, obtained via the ‘BIGSTEP’ method, was introduced and shown to be AO. This same N network (but with more general distributions and parameter values) was considered by Bell and Williams in [2]. The parameter values are chosen to yield heavy traffic. Under the objective of minimizing the expected infinite horizon total discounted cost they prove AO of a threshold policy.

Bell and Williams [3] treat the general PSS under the assumptions that the static allocation LP has a unique solution, and that the CRP condition holds. Under those conditions AO of an intricately designed priority-threshold policy is proved.

PSSs were considered by Mandelbaum and Stolyar [22] under a cost function that is increasing and strictly convex. In this context they showed that a simple generalized  $c\mu$ -rule is AO. They also effectively assumed uniqueness of the solution to the static allocation LP, along with CRP. They acknowledge difficulty if uniqueness does not hold, and mention that, when uniqueness fails to hold due to having groups of identical servers, by viewing the groups as individual servers their results continue to hold.

Stolyar [28] introduces a condition that he calls First-Order-CRP (FO-CRP), which relaxes the requirement that the static allocation LP has a unique solution. However, the main heavy traffic AO results are proved under the more restrictive CRP condition, which requires uniqueness.

Of course, PSSs that do not satisfy CRP, as well as more complex stochastic networks remain of great interest. Some progress on PSSs that do not satisfy CRP was obtained by Petic and Williams [23], who prove some structural results along with an optimal control for the BCP under certain conditions. It is conjectured there that a particular continuous review threshold policy is AO under those conditions.

In the context of more complex stochastic networks, Budhiraja and Ghosh [5] show that the solution of the BCP provides an asymptotic lower bound for a class of controlled queueing networks, and in [6] they show that the value function converges to this lower bound. It is worth noting that [23], [5] and [6] all assume that the static planning LP has a unique solution.

As mentioned above, there is no strong reason to believe that the static allocation LP will satisfy the uniqueness assumption in typical problems of practical interest. In fact, there is good reason to believe that it will often not hold. An interesting and natural case is where the service rates satisfy a decomposition property. Let  $\mu_{ik}$  denote the rate at which class  $i$  jobs are processed by server  $k$ , and assume that every class-server pair is an activity. As we show, assuming that these rates (to first order) are decomposable as  $\mu_{ik} = \alpha_i \beta_k$  always gives rise to non-uniqueness. In the case  $I = K = 2$  we show that such a decomposition is also necessary for non-uniqueness. Service rates decompose this way when the mean size of a job is characteristic to the class (and then  $\alpha_i$  is the reciprocal mean), and each server has its own processing speed (here given by  $\beta_k$ ). Because all earlier results on the PSS in heavy traffic were obtained under the HTC, this class of service rates has been excluded.

This decomposition does not have to hold exactly, just to first order, allowing the second order data, which determine drift and variance, to be more general. In addition, this decomposition need not hold for all classes and servers. Non-uniqueness can arise when there are subsets of classes and/or servers of size at least two for which the decomposition holds, even if not all class-server pairs form activities. Examples (D) and (E) in §2.5 provides examples of this.

1.2. *Results.* A brief informal description of the role played by the uniqueness assumption in earlier work will help explain the main issues that must be dealt with in its absence. Whereas the aforementioned BCP describes the diffusion scale deviations of all relevant stochastic processes (such as queue lengths, cumulative busyness time on each activity, etc.) about a central model, the static allocation problem is concerned with the fluid scale behavior and determines the central model itself. Specifically, this LP describes the mean fraction of time devoted by each server to the different classes. When the LP has a unique solution, any sequence of policies along which the diffusion-scaled cost remains bounded as the scaling parameter grows must at the very least ‘adhere’ to this solution at the fluid scale. That is, such a sequence of policies should have the property that the diffusion scale deviations of the relevant processes about the central model are at most of order 1, so they can relate to the processes comprising the BCP. As a consequence, such a sequence must have the much weaker, fluid scale property, that the fractions of cumulative busyness for each activity converge to this unique solution at the scaling limit. The BCP is set up in a way that reflects behavior according to this unique solution.

The activities to which the LP solution allocates a strictly positive (respectively, zero) fraction of effort are said to be basic (respectively, non-basic). In the work [3] (which, like this paper, considers linear holding costs) the non-basic activities are not used by the prescribed policies even for diffusion-scale corrections. In this sense, the solution to the LP determines not only the fluid limit of the model but also which of the activities are operational.

When the LP has multiple solutions, many of the aspects described above break down. A BCP can still be written down, describing deviations about a fluid-scale allocation in the space of solutions, but the fluid-scale allocation now becomes a stochastic process in the space of solutions that is controlled by the decision maker. Thus the model has a non-deterministic fluid limit. The set of basic, hence operational, activities has to be selected dynamically. The introduction of this dynamics influences all stages of the analysis: The form of the BCP and the method of solving it, techniques for proving weak convergence required for establishing the lower bound (and, in the paper [1], the design of policies and the proof of their AO).

The class of PSS treated in this paper is quite general, except for the restriction to systems where the workload process is one-dimensional. In that we have roughly followed [15] and [3], which restrict their attention to systems that satisfy CRP. As mentioned previously, under CRP the servers can be pooled into a single super server, yielding a one-dimensional workload process. It was shown in Proposition 4 of [15] that under CRP (which assumed uniqueness of the LP solution) the equivalent workload formulation is one-dimensional. We do not prove an analog of Proposition 4 of [15] under non-uniqueness. On the other hand, we show that the condition for reduction to a one-dimensional problem associated with the workload process is the same as with a unique solution, namely that the dual of the static allocation LP has a unique solution. Because the definition of CRP has been tightly tied to uniqueness we do not use the term in relation to our results.

In what follows we summarize our results.

*Workload control problem (WCP) and Hamilton-Jacobi-Bellman (HJB) equation.* Under multiplicity, the set of LP solutions has  $2 \leq M < \infty$  extreme points that we call control modes, or simply modes. Each mode corresponds to a different allocation and a different

set of operational activities. As mentioned above, one of the dynamically controlled BCP ingredients is the fluid-scale limit of the allocation process. This process is in general random.

We derive the BCP and then introduce the WCP, which is the equivalent workload formulation of the BCP in our case. The WCP is formulated in terms of a one-dimensional stochastic differential equation (SDE) with controlled drift and diffusion coefficients, and reflection at zero. It is dramatically different from the WCP obtained under the uniqueness assumption, in which the coefficients are not controlled (and are constant), and the solution is a reflected Brownian motion (BM). It is a standard fact that the solution to a control problem of this kind can be characterized in terms of viscosity solutions of an HJB equation. The general theory allows one to express an optimal (or nearly optimal) control as a feedback from the diffusion process (where the feedback function can be specified in terms of the solution to the HJB equation). In our case we can prove, based on PDE theory on fully nonlinear uniformly elliptic equations, that the HJB equation has a unique classical solution. Moreover, the solution to the WCP is given by a diffusion with discontinuous drift and diffusion coefficients. Examples of queueing models and queueing control models in heavy traffic which lead to a diffusion process with a discontinuous diffusion coefficient have appeared in the literature [19], [24]. To the best of our knowledge this is the first time this occurs in the context of optimal/asymptotically optimal control of a queueing system.

*Asymptotic lower bound.* Our main result states that the solution of the WCP provides an asymptotic lower bound on achievable cost. The lack of a pathwise solution to the WCP, and the related representation of the limit process introduces technical difficulties. To prove the result one must show that scaling limits of the queueing model processes form an admissible control system for the WCP. The specific tuple that forms such a control system includes a filtration and a standard BM (SBM) on this filtration, which serves as the driving BM in the aforementioned SDE. The construction of this BM can be achieved via a transformation of certain processes associated with the scaling limits of the arrival processes and random time changes of potential service processes of the  $J$  activities, once it is established that these processes are all martingales on a common filtration. The random time changes appear for exactly the reason that the modes vary stochastically. Whereas in the single mode case the fluid limits of cumulative busyness processes are deterministic and the time changed processes are easily shown to be martingales, this conclusion is not immediate in the multiple mode case. Our approach to this difficulty is based on multi-parameter filtrations and a result due to Kurtz [20] on multi-parameter time change. One of the main steps toward using this result is to show that the prelimit cumulative busyness processes define multi-parameter stopping times on a filtration generated by the model's primitives.

Our results do not fit neatly into the 4 steps of [15]. This paper covers step (1), but also parts of steps (2) and (4). It partly covers step (2) by characterizing the solution of the WCP in terms of an HJB equation. It also covers part of step (4) by providing a target that, if achieved, yields AO. The remainders of steps (2) and (4), along with step (3), are carried out in [1] for the simplest PSS, consisting of 2 classes, 2 servers and 4 activities ( $I = K = 2, J = 4$ ).

1.3. *Organization of the paper.* In §2.1, we present the notation and model in more detail. We define the extended HTC in §2.2, and state lemmas regarding the static allocation LP and its dual. We introduce the WCP and state our main result in §2.3. A discussion on degeneracy of the LP solution appears in §2.4. Several examples of PSS with multiple LP solutions are provided in §2.5. In §3, we prove the lemmas related to the LP. §4 treats the BCP, WCP and HJB equation. In §4.1, the BCP and WCP are derived. §4.2 provides a candidate solution to the WCP based on the HJB equation. §4.3 proves that a classical solution to the HJB equation uniquely exists and establishes the optimality of the WCP's candidate solution. It also addresses a special case where the WCP's solution is to always use one particular mode.



Finally, we prove the main result in §5. The main lemmas on which this proof is based appear in §5.1, and are proved in §5.2. For ease of reference several results related to linear programming that we use are collected together in the Appendix.

1.4. *Notation and terminology.* The following notation will be used throughout.  $\mathbb{N}$ ,  $\mathbb{R}$  and  $\mathbb{R}_+$  are the sets of natural, real and, respectively, non-negative real numbers. The symbol  $\mathbf{i}$  denotes  $\sqrt{-1}$ . For  $a, b \in \mathbb{R}$ ,  $a \vee b$  denotes the maximum of  $a$  and  $b$ ,  $a \wedge b$  the minimum of  $a$  and  $b$ ,  $a^+ = a \vee 0$ , and  $a^- = -(a \wedge 0)$ . For any  $P \in \mathbb{N}$ ,  $0_P$  (resp.  $1_P$ ) denote column vectors in  $\mathbb{R}^P$  whose all components are 0 (resp. 1). For a set  $A$ ,  $\mathbb{1}_A$  denotes the indicator function of  $A$ . For  $a, b \in \mathbb{R}^P$ , write  $a \leq b$  (equivalently  $b \geq a$ ) for the componentwise inequalities. For a set  $A \subset \mathbb{R}^P$ ,  $\text{ch}(A)$  denotes the closed convex hull of  $A$ . For real-valued functions and processes, the notation  $X(t)$  is used interchangeably with  $X_t$ . Given a Polish space  $E$ , denote by  $C_E[0, \infty)$  and  $D_E[0, \infty)$  the spaces of  $E$ -valued, continuous and, respectively, càdlàg functions on  $[0, \infty)$ . Equip the former with the topology of convergence u.o.c. (uniformly over compacts) and the latter with the  $J_1$  topology. For an element of  $C_E[0, \infty)$  (or  $D_E[0, \infty)$ ), we use the notation  $\|f\|_t = \sup_{s \in [0, t]} |f_s|$ . Denote by  $C_{\mathbb{R}}^+[0, \infty)$  (respectively,  $D_{\mathbb{R}}^+[0, \infty)$ ) the subset of  $C_{\mathbb{R}}[0, \infty)$  (respectively,  $D_{\mathbb{R}}[0, \infty)$ ) of non-negative, non-decreasing functions, and by  $C_{\mathbb{R}}^{0,+}[0, \infty)$  the set of functions in  $C_{\mathbb{R}}^+[0, \infty)$  that are null at zero. Write  $X_n \Rightarrow X$  for convergence in law. A tight 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 positive deterministic constant whose value may change from one appearance to another.

In addition to abbreviations already introduced, we will use RV for random variable(s); IID for independent identically distributed; LHS (resp. RHS) for left (resp. right) hand side.

## 2. Model and main results.

2.1. *Queueing model, scaling and queueing control problem.* The queueing model is described in terms of a sequence of systems, indexed by  $n \in \mathbb{N}$ , all defined on one probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . Let us first introduce the structure of the system. Classes belong to the set  $\mathcal{I}$  of cardinality  $I$  and throughout,  $i$  will be used as a generic element of  $\mathcal{I}$ , indexing a class. Similarly, servers belong to the set  $\mathcal{K}$  of cardinality  $K$  with elements denoted  $k$ . Additionally, if a server  $k$  is able to serve a particular class  $i$ , the pair  $(i, k)$  is called an *activity*. The set of all activities is denoted by  $\mathcal{J} \subset \mathcal{I} \times \mathcal{K}$  and its cardinality by  $J$ . Elements of  $\mathcal{J}$  are denoted  $j = (i, k)$ . We will write  $\mathcal{J}_i$  (resp.  $\mathcal{J}^k$ ) for the set of all activities involving class  $i$  (resp., server  $k$ ).

For the  $n$ th system, the processes of interest are as follows. First, the processes  $A^n = (A_i^n)_{i \in \mathcal{I}}$ ,  $S^n = (S_j^n)_{j \in \mathcal{J}}$  represent arrival and potential service counting processes. That is,  $A_i^n(t)$  is the number of arrivals of class  $i$  jobs until time  $t$ ,  $i \in \mathcal{I}$ , and  $S_j^n(t)$  is the number of service completions on activity  $j$ , by the time the server has devoted  $t$  units of time to this activity,  $j \in \mathcal{J}$ . Next,  $X^n = (X_i^n)_{i \in \mathcal{I}}$ ,  $I^n = (I_k^n)_{k \in \mathcal{K}}$ ,  $D^n = (D_j^n)_{j \in \mathcal{J}}$  and  $T^n = (T_j^n)_{j \in \mathcal{J}}$  denote queue length, cumulative idleness, departure, and cumulative busyness processes. In other words,  $X_i^n(t)$  is the number of class  $i$  customers in the system at time  $t$ ,  $I_k^n(t)$  is the cumulative time server  $k$  has been idle by time  $t$ ,  $D_j^n(t)$  is the number departures due to activity  $j$ , and  $T_j^n(t)$  is the cumulative time devoted to activity  $j$ . The process  $T_j^n$  takes the form  $T_j^n(t) = \int_0^t \Xi_j^n(s) ds$ , where  $\Xi_j^n(t)$  is the fraction of effort devoted to activity  $j$ . In particular,  $\sum_{j \in \mathcal{J}^k} \Xi_j^n(t) \leq 1$  for every  $k$ . Thus  $\Xi^n$  is referred to as the *allocation process*.

The probabilistic model for the primitive processes is as follows. Arrival rates  $\lambda_i^n$  and service rates  $\mu_j^n$  are given, satisfying, for some constants  $\lambda_i \in (0, \infty)$ ,  $\hat{\lambda}_i \in \mathbb{R}$  for any  $i \in \mathcal{I}$ ,

and  $\mu_j \in (0, \infty)$ ,  $\hat{\mu}_j \in \mathbb{R}$  for any  $j \in \mathcal{J}$ ,

$$\begin{aligned}\hat{\lambda}_i^n &:= n^{-1/2}(\lambda_i^n - n\lambda_i) \rightarrow \hat{\lambda}_i, \\ \hat{\mu}_j^n &:= n^{-1/2}(\mu_j^n - n\mu_j) \rightarrow \hat{\mu}_j,\end{aligned}$$

as  $n \rightarrow \infty$ . For each  $i$  a renewal process  $\check{A}_i$  is given, with interarrival distribution that has mean 1 and squared coefficient of variation  $0 < C_{A_i}^2 < \infty$ . Similarly, for each  $j$ , a renewal process  $\check{S}_j$  is given with mean 1 interarrival and squared coefficient of variation  $C_{S_j}^2 < \infty$ . It is assumed that  $A^n$  and  $S^n$  are given by

$$A_i^n(t) = \check{A}_i(\lambda_i^n t), \quad S_j^n(t) = \check{S}_j(\mu_j^n t).$$

We call the parameters  $\{\lambda_i\}_{i \in \mathcal{I}}$ ,  $\{\mu_j\}_{j \in \mathcal{J}}$  the *first order data* and  $\{\hat{\lambda}_i\}_{i \in \mathcal{I}}$ ,  $\{\hat{\mu}_j\}_{j \in \mathcal{J}}$ ,  $\{C_{A_i}\}_{i \in \mathcal{I}}$  and  $\{C_{S_j}\}_{j \in \mathcal{J}}$  the *second order data*. It is assumed that the  $I + J$  processes  $\check{A}_i$ ,  $\check{S}_j$  are mutually independent, have strictly positive inter-arrival distributions and right-continuous sample paths. The (IID) interarrivals of  $\check{A}_i$  and  $\check{S}_j$  are denoted by  $\check{a}_i(p)$  and  $\check{u}_j(p)$ , respectively, and those of the accelerated processes  $A_i^n$  and  $S_j^n$  are given by

$$(2.1) \quad a_i^n(p) = \frac{1}{\lambda_i^n} \check{a}_i(p), \quad u_j^n(p) = \frac{1}{\mu_j^n} \check{u}_j(p).$$

The system is assumed to start empty, that is,  $X^n(0) = 0$  for all  $n$ .

The first order service rates are encoded in an  $I \times J$  matrix  $R$ . The element  $R_{ij} = \mu_j$  gives the rate at which activity  $j$  processes class  $i$  if  $j \in \mathcal{J}_i$  and one sets 0 otherwise. Additionally, we will need a  $K \times J$  matrix  $G$ . Elements of  $G$  are either 0 or 1, and are given by  $G_{k,j} = \mathbb{1}_{j \in \mathcal{J}^k}$ . The matrix  $G$  serves as interference constraint between the activities.

Simple relations between the processes introduced above are expressed as follows,

$$(2.2) \quad D_j^n(t) = S_j^n(T_j^n(t)),$$

$$(2.3) \quad X_i^n(t) = A_i^n(t) - \sum_{j \in \mathcal{J}_i} D_j^n(t),$$

$$(2.4) \quad I_k^n(t) = t - (GT^n(t))_k,$$

$$(2.5) \quad X_i^n \text{ is non-negative for all } i, \text{ and } I_k^n \text{ is non-negative non-decreasing for all } k.$$

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, which is obviously equivalent to having the allocation process  $\Xi^n$  as 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  $\mathcal{Y} = (\mathcal{Y}(k), k \in \mathbb{N})$ , taking values in a Polish space  $\mathcal{S}_{\text{rand}}$  and assumed to be independent of the stochastic primitives, for each  $n$  (there is no need to let  $\mathcal{Y}$  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.

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  $j \in \mathcal{J}$ ,  $T_j^n$  has sample paths in  $C_{\mathbb{R}}^{0,+}[0, \infty)$  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  $\{\mathcal{F}_t^n\}_{t \geq 0}$  defined by  $\mathcal{F}_t^n = \sigma\{(A^n(s), D^n(s), s \in [0, t]), \mathcal{Y}\}$ . Denote by  $\mathcal{A}^n$  the collection of all admissible controls for the QCP for the  $n$ -th system.

REMARK 2.1. i. It is desired to support models in which the controls depend on the history of the queue length and cumulative idleness processes in addition to that of the arrival and departures. Note that our definition allows such dependence because, in view of (2.3), (2.4), the sigma field  $\sigma\{(A^n(s), D^n(s), X^n(s), I^n(s), s \in [0, t]), \mathcal{Y}\}$  is equal to  $\mathcal{F}_t^n$ . Similarly, it is desired to allow control decisions to depend on past control decisions. This is allowed by our definition, where indeed  $T_t^n$  can be selected based on any information of  $T^n|_{[0, t]}$ , as follows from the fact the process is adapted, and so such information is contained in the filtration.

ii. The auxiliary randomness  $\mathcal{Y}$  can be used to create randomized controls. For example, a non-preemptive policy can decide which class is assigned to a server when it becomes available based on the value of one RV at a time, taken from a sequence of IID decision variables with prescribed distribution. In a preemptive-resume setting, one can randomize, for example, the time after a new arrival of a job of a high priority class, at which a job of lower priority class is preempted.

iii. In the problem formulation of [2, 3] the control process  $T^n$  may anticipate the future. This difference from our setting is due to the fact that, in the above references, the underlying workload control problem has the property that the optimal control with and without a nonanticipativity constraint are identical. However, this does not occur in our context.

The queue length process normalized at the diffusion scale is defined for each  $i \in \mathcal{I}$  by  $\hat{X}_i^n(t) = n^{-1/2}X_i^n(t)$ . As in most earlier treatments of the PSS, we consider the model with a discounted cost. For the  $n$ -th system it is given by

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

where  $h(x) = h \cdot x$  and  $\gamma > 0$  and  $h \in (0, \infty)^I$  are constants, 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 model and QCP. The complete set of problem data consists of the stochastic primitives mentioned above and the collection of parameters

$$\lambda, \mu, \hat{\lambda}, \hat{\mu}, C_A, C_S, \gamma, h, G, R.$$

2.2. *The linear program, dual problem and extended heavy traffic condition.* The notion of heavy traffic is tied to a critical load condition. For systems without control the arrival and service rates determine the traffic intensity, and hence the heavy traffic condition. In systems with routing control, such as PSS, the traffic intensity could depend on the control policy, and the heavy traffic condition involves an optimization problem. In particular, for the PSS, it was proposed in [15] to formulate the notion of heavy traffic via a linear program (LP) that they called the ‘static allocation problem’, involving only first order data, which expresses how to best allocate the servers’ effort. The condition then asserts that the LP solution yields a critical load at each server. This LP is as follows.

Given the first order data, the LP involves the unknowns  $\xi \in \mathbb{R}_+^J$  and  $\rho \in \mathbb{R}$ . The components  $\xi_j = \xi_{ik}$  correspond to the fraction of effort server  $k$  dedicates to jobs from class  $i$  on the fluid scale, whereas  $\rho$  is an upper bound on the load on each server.



*Linear Program.* Minimize  $\rho$  subject to

$$(2.7) \quad \begin{cases} R\xi = \lambda, \\ G\xi \leq \rho 1_K, \\ \xi \geq 0_J. \end{cases}$$

Denote the optimal objective value (minimum achievable  $\rho$ ) of (2.7) by  $\rho^*$ .

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

The extended heavy traffic condition (EHTC) is broader than the *heavy traffic condition* introduced in [15] that has been extensively used in the literature, in which it has also been assumed that there is only one  $\xi$  satisfying (2.7) with  $\rho = 1$ .

Under the EHTC, any solution is of the form  $(\xi, 1)$ . Let  $\mathcal{S}_{LP}$  denote the closed and convex subset of  $\mathbb{R}_+^J$  for which the set of all solutions is given by  $\mathcal{S}_{LP} \times \{1\}$ . With a slight abuse of terminology, we will sometimes refer to  $\mathcal{S}_{LP}$  as the set of solutions to the LP. We also introduce another linear program which is the dual of (2.7):

*Dual Problem.* Maximize  $y \cdot \lambda$  subject to

$$(2.8) \quad \begin{cases} \sum_k z_k = 1, \\ yR \leq zG, \\ z \geq 0_K. \end{cases}$$

(The standard terminology in linear programming is to refer to the LP in (2.7) as the *primal* LP. We slightly abuse this terminology by simply referring to our primal LP as ‘the LP’.) A solution to (2.8) is given by  $(y^*, z^*) \in \mathbb{R}^I \times \mathbb{R}_+^K$ . As we will argue later, the existence of a solution to (2.7) with  $\rho^* = 1$  implies the existence of at least one solution to (2.8). The following is a standing assumption in this paper.

- ASSUMPTION 2.2. 1. *The EHTC holds. In particular, there exists at least one solution to the LP.*
2. *For any solution  $\xi \in \mathcal{S}_{LP}$  and any  $k \in \mathcal{K}$ ,  $(G\xi)_k = 1$ .*
3. *There is a unique solution,  $(y^*, z^*)$ , to the dual problem (2.8).*

Note that the first part of the above assumption implies that for all  $i \in \mathcal{I}$ ,  $\mathcal{J}_i \neq \emptyset$ . The set  $\mathcal{J}_i$  being empty means no activity involves class  $i$ . This is not compatible with  $(R\xi)_i = \lambda_i > 0$ . This assumption also does not restrict us on the number of solutions to the LP. All our results are valid whether there are multiple or one solution. The meaning of the second part is that all servers are fully loaded. As argued in [15], this could mean that we restrict ourselves to a ‘bottleneck subnetwork’ of critically loaded servers. If an original network had involved more servers, some of which are subcritical and satisfied the EHTC, the additional servers could not have involved any of the classes in the bottleneck subnetwork. Otherwise, one could construct a solution to (2.7) with  $\rho < 1$  from a solution with  $\rho = 1$ . Note that the second part also implies that for any  $k \in \mathcal{K}$ , the set  $\mathcal{J}^k \neq \emptyset$ .

Under the more standard assumption that the HTC holds, activities are classified as *basic* and *non basic* according to whether  $\xi_j > 0$  or not, where  $\xi$  is the unique solution to the LP. In our setting there may be multiple solutions, hence a different terminology is required. Call an

activity  $j \in \mathcal{J}$  *potentially basic* if  $\xi_j > 0$  for some  $\xi \in \mathcal{S}_{\text{LP}}$ . An activity that is not potentially basic is called *always nonbasic*. The corresponding sets of activities are denoted  $\mathcal{A}_{\text{PB}}$  and  $\mathcal{A}_{\text{AN}}$ . Under Assumption 2.2, we are able to prove the following lemma.

- LEMMA 2.3. 1. *The set  $\mathcal{S}_{\text{LP}} = \text{ch}\{\xi^{*,m}, m = 1, \dots, M\}$  is the closed convex hull of a finite number,  $M$ , of extreme points  $\xi^{*,m} \in \mathcal{S}_{\text{LP}}$ , called modes.*
2. *One has  $y^* \cdot \lambda = 1$ .*
  3. *For any server  $k$ , one has  $z_k^* > 0$ .*
  4. *For any activity  $j = (i, k) \in \mathcal{A}_{\text{PB}}$ , one has  $y_i^* \mu_j = z_k^*$ .*
  5. *For any activity  $j = (i, k) \in \mathcal{A}_{\text{AN}}$ , one has  $y_i^* \mu_j < z_k^*$ .*
  6. *For any class  $i \in \mathcal{I}$ , there exists  $j \in \mathcal{J}_i \cap \mathcal{A}_{\text{PB}}$ .*
  7. *For any class  $i \in \mathcal{I}$ , one has  $y_i^* > 0$ .*

We say that the service rates  $\mu$  are *decomposable* if  $\mu_{ik} = \alpha_i \beta_k$  for all  $(i, k) \in \mathcal{J}$  for some constants  $\alpha_i$  and  $\beta_k$ . In this case we will assume without loss of generality  $\sum_k \beta_k = 1$ . In order to motivate the relevance of systems with multiple solutions to the LP, we present the next lemma. On the one hand, decomposability of service rates lead to an explicit solution for the dual. On the other hand, they are linked to non-uniqueness of the LP.

- LEMMA 2.4. 1. *If  $\mu$  are decomposable,*

$$\sum_i \frac{\lambda_i}{\alpha_i} = \sum_k \beta_k = 1.$$

2. *If  $\mu$  are decomposable, the unique solution to (2.8) is given by  $(y^*, z^*)$  with  $(y_i^*)_{i \in \mathcal{I}} = (\alpha_i^{-1})_{i \in \mathcal{I}}$  and  $(z_k^*)_{k \in \mathcal{K}} = (\beta_k)_{k \in \mathcal{K}}$ .*
3. *If  $\mu$  are decomposable and  $\mathcal{J} = \mathcal{I} \times \mathcal{K}$  then the LP has multiple solutions.*
4. *If  $I = K = 2$  and  $\mathcal{J} = \mathcal{I} \times \mathcal{K}$  then the LP has multiple solutions if and only if  $\mu$  are decomposable. In addition, the number of modes  $M$  is at most 2.*

2.3. *Workload control problem, HJB equation and main result.* Based on the notation introduced in the previous section we define the prelimit workload process by

$$\hat{W}_t^n = y^* \cdot \hat{X}_t^n.$$

Here and throughout the paper, we fix  $q \in \mathcal{I}$  such that

$$(2.9) \quad h_q(y_q^*)^{-1} = \min_{i \in \mathcal{I}} h_i(y_i^*)^{-1}.$$

For any  $i \in \mathcal{I}$  and  $j \in \mathcal{J}$ , let  $\sigma_{A,i} := \lambda_i^{1/2} C_{A,i}$ ,  $\sigma_{S,j} := \mu_j^{1/2} C_{S,j}$ , and

$$(2.10) \quad b(\xi) = \sum_i y_i^* (\hat{\lambda}_i - \sum_{j \in \mathcal{J}_i} \hat{\mu}_j \xi_j), \quad \sigma(\xi)^2 = \sum_i (y_i^*)^2 (\sigma_{A,i}^2 + \sum_{j \in \mathcal{J}_i} \sigma_{S,j}^2 \xi_j), \quad \xi = (\xi_j)_{j \in \mathcal{J}} \in \mathcal{S}_{\text{LP}}.$$

Let also

$$(2.11) \quad b_m = b(\xi^{*,m}), \quad \sigma_m = \sigma(\xi^{*,m}), \quad m = 1, \dots, M.$$

Consider a one-dimensional controlled diffusion with controlled drift and diffusion coefficients and reflection at the origin, given by

$$(2.12) \quad Z_t = z + \int_0^t b(\Xi_s) ds + \int_0^t \sigma(\Xi_s) dB_s + L_t,$$

where  $\Xi$  is a control process,  $B$  is a standard BM,  $L$  is a reflection term at zero, and  $z \geq 0$ . A precise definition is as follows.

Given a filtration  $(\mathcal{F}'_t)_{t \geq 0}$ , let  $\mathfrak{X}((\mathcal{F}'_t)_{t \geq 0})$  denote the collection of  $(\mathcal{F}'_t)_{t \geq 0}$ -progressively measurable processes  $\{\Xi\}$  taking values in  $\mathbb{R}^J_+$ , for which one has  $\mathbb{P}(\text{for a.e. } t, \Xi_t \in \mathcal{S}_{\text{LP}}) = 1$ . A tuple  $\mathfrak{G} = (\Omega', \mathcal{F}', (\mathcal{F}'_t)_{t \geq 0}, \mathbb{P}', B, \Xi, Z, L)$  is said to be an *admissible control system for the WCP with initial condition  $z$*  if  $(\Omega', \mathcal{F}', (\mathcal{F}'_t)_{t \geq 0}, \mathbb{P}')$  is a filtered probability space,  $B, \Xi, Z$  and  $L$  are processes defined on it,  $B$  is a standard BM (SBM) and an  $(\mathcal{F}'_t)_{t \geq 0}$ -martingale,  $\Xi \in \mathfrak{X}((\mathcal{F}'_t)_{t \geq 0})$ ,  $Z$  is continuous non-negative and  $(\mathcal{F}'_t)_{t \geq 0}$ -adapted,  $L$  has sample paths in  $C_{\mathbb{R}}^{0,+}[0, \infty)$  and is  $(\mathcal{F}'_t)_{t \geq 0}$ -adapted, and equation (2.12) and the identity  $\int_{[0, \infty)} Z_t dL_t = 0$  are satisfied  $\mathbb{P}'$ -a.s.

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

$$(2.13) \quad J_{\text{WCP}}(z, \mathfrak{G}) = \mathbb{E} \int_0^\infty e^{-\gamma t} Z_t dt, \quad V_{\text{WCP}}(z) = \inf \{ J_{\text{WCP}}(z, \mathfrak{G}) : \mathfrak{G} \in \mathcal{A}_{\text{WCP}}(z) \}.$$

Above,  $\mathbb{E}$  denotes, with a slight abuse of notation, expectation w.r.t. the control system  $\mathfrak{G}$ . A control system  $\mathfrak{G}$  is said to be optimal for the WCP with initial condition  $z$  if  $J_{\text{WCP}}(z, \mathfrak{G}) = V_{\text{WCP}}(z)$ .

It is well known that the dynamic programming equation associated with controlled diffusions as above is of HJB type (see, for example, [9, Section IV.5]). Its role is to characterize the value function of the control problem. In our case the HJB equation takes the following form. For  $(v_1, v_2, \xi) \in \mathbb{R}^2 \times \mathcal{S}_{\text{LP}}$ , let

$$\begin{aligned} \bar{\mathbb{H}}(v_1, v_2, \xi) &= b(\xi)v_1 + \frac{\sigma(\xi)^2}{2}v_2, \\ \mathbb{H}(v_1, v_2) &= \inf_{\xi \in \mathcal{S}_{\text{LP}}} \bar{\mathbb{H}}(v_1, v_2, \xi) = \min_{m=1, \dots, M} \bar{\mathbb{H}}(v_1, v_2, \xi^{*,m}), \end{aligned}$$

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

$$(2.14) \quad \mathbb{H}(u'(z), u''(z)) + z - \gamma u(z) = 0, \quad z \in (0, \infty),$$

and the boundary conditions at 0 and  $\infty$ ,

$$(2.15) \quad u'(0) = 0, \quad |u(z)| < c(1+z), z \in \mathbb{R}_+, \text{ for some constant } c.$$

It is well known that the  $C^2$  smoothness of the value function is tied to the question of existence of a classical solution to the HJB equation such as (2.14). The precise statement in our setting is as follows.

**PROPOSITION 2.5.** *There exists a unique classical solution  $u$  to (2.14)–(2.15). Moreover,  $u = V_{\text{WCP}}$ .*

A version of this diffusion control problem with 2 modes was treated in [27]. There it was shown that the optimal control has a simple form: either one of the modes is always used, or there is a unique threshold, with one mode used below the threshold and the other used above. Those results are used in [1] to develop a policy and prove AO. In Proposition 4.5 we extend the condition for optimality of a single mode policy to the general setting we consider here.

The main result of this paper establishes an asymptotic lower bound.

$$\text{THEOREM 2.6.} \quad \liminf_n \hat{V}^n \geq V_0 := h_q(y_q^*)^{-1} V_{\text{WCP}}(0) = h_q(y_q^*)^{-1} u(0).$$

2.4. *Discussion.* The result raises the question of whether the lower bound is tight, in the sense that there exist policies that asymptotically achieve this bound. When this is the case, one has  $\lim_n \hat{V}^n = V_0$ . In the discussion that follows, the terms *uniqueness* and *multiplicity* refer to the existence of a unique solution, and respectively, multiple solutions to the static allocation LP. A key issue that arises here is related to possible degeneracy of the LP solution: A basic solution of an LP is called *degenerate* if one or more basic variables are zero (cf. Definition 3.1 in [11]).

Consider first the case of uniqueness. As pointed out in [15], in order for CRP to be achieved using basic activities the solution must be nondegenerate. Moreover, when uniqueness holds, our assumption of uniqueness of dual problem solutions boils down to CRP, and therefore Theorem 2.6 recovers the asymptotic lower bound from [3]. In this case, [3] also constructs policies that achieve this bound, and thus the question of tightness is answered affirmatively. However, when the unique LP solution is degenerate, not all servers can communicate in the graph formed by the basic activities. In this case it is expected that the above lower bound does not capture the correct asymptotics.

In the case of multiplicity, the situation is more interesting. Even if all modes (= extreme points of  $\mathcal{S}_{LP}$  = basic optimal solutions) are degenerate, there are elements in  $\mathcal{S}_{LP}$  for which the associated graph (where an edge between class  $i$  and server  $k$  exists when  $\xi_j > 0$  for  $j = (i, k) \in \mathcal{J}$ ) yields communication between all servers. As can be seen in our discussion of the HJB equation and associated optimal policies, the control used in the optimal policy indeed corresponds to extreme points of  $\mathcal{S}_{LP}$ . If none of the extreme points is a degenerate solution to the LP, using a policy that dynamically switches between modes as prescribed by the HJB equation should be enough to yield asymptotic optimality. This statement has been proved in [1] in the case  $I = K = 2$ ,  $J = 4$ , under multiplicity and nondegeneracy, by constructing policies that achieve the lower bound. As a result, the tightness question is settled in this case as well.

The difficulty one faces when some of the modes correspond to degenerate LP solutions is that while using them it may not be possible to achieve the state space collapse needed to have the cost reach the lower bound. Although the presence or absence of degeneracy does not affect our lower bound, a lower bound that is not asymptotically achievable is not as interesting as one that is. However, in contrast to the case of uniqueness, in the case of multiplicity, the interior of  $\mathcal{S}_{LP}$  contains nondegenerate solutions, and it should be true that by carefully choosing controls that approach degenerate modes, a policy achieving the lower bound can be obtained even in this case. We therefore pose the following.

**CONJECTURE 2.7.** *Let the standing assumptions of this paper hold, and assume further that there are multiple solutions to the LP. Then the lower bound of Theorem 2.6 is tight. That is, there exist sequences of policies  $T^n$  such that  $\lim_n \hat{J}^n(T^n) = V_0$ , and consequently  $\lim_n \hat{V}^n = V_0$ .*

We will argue in Corollary 4.6 that the work done under the uniqueness assumption carries over to a case where multiplicity is possible but the WCP is optimally controlled by a single mode. In particular, the tree-based threshold policy introduced by Bell and Williams [3] based on the optimizing mode achieves the lower bound from Theorem 2.6 and thus verifies the above conjecture in this case.

2.5. *Examples.* We consider five examples that yield the extended heavy traffic condition with multiplicity. Three of the examples involve two classes and two servers, and two examples involves three classes and three servers.

EXAMPLES (A) AND (B). Consider a 2-class, 2-server system with four activities. Label the activities  $j = (i, k)$  via the correspondence  $1 = (1, 1)$ ,  $2 = (1, 2)$ ,  $3 = (2, 1)$ ,  $4 = (2, 2)$ . Let

$$\mu_1 = \mu_{11} = 3, \quad \mu_2 = \mu_{12} = 4, \quad \mu_3 = \mu_{21} = 6, \quad \mu_4 = \mu_{22} = 8.$$

Note that these service rates are decomposable, allowing multiple solutions to the LP. (Under our convention that  $\beta_1 + \beta_2 = 1$  we have  $\alpha_1 = 7, \alpha_2 = 14, \beta_1 = 3/7$ , and  $\beta_2 = 4/7$ .) The matrices  $R$  and  $G$  are given by

$$R = \begin{pmatrix} \mu_1 & \mu_2 & 0 & 0 \\ 0 & 0 & \mu_3 & \mu_4 \end{pmatrix} = \begin{pmatrix} 3 & 4 & 0 & 0 \\ 0 & 0 & 6 & 8 \end{pmatrix}, \quad G = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

For the arrival rates, consider two examples, namely

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

Then the linear program takes the form: Minimize  $\rho$  subject to

$$(2.16) \quad \begin{cases} 3\xi_{11} + 4\xi_{12} = \lambda_1, \\ 6\xi_{21} + 8\xi_{22} = \lambda_2, \\ \xi_{11} + \xi_{21} \leq \rho, \\ \xi_{12} + \xi_{22} \leq \rho, \\ \min \xi_{ik} \geq 0. \end{cases}$$

The dual program takes the form: Maximize  $y_1\lambda_1 + y_2\lambda_2$  subject to

$$(2.17) \quad \begin{cases} z_1 + z_2 = 1, \quad \min_k z_k \geq 0, \\ (3y_1, 4y_1, 6y_2, 8y_2) \leq (z_1, z_2, z_1, z_2). \end{cases}$$

Consider the set of constraints (2.16) of the LP, where the third and fourth constraints hold as equalities and  $\rho = 1$ . By a direct calculation in the case of Example (A), the solutions of this set are all given as convex combinations of the two solutions

$$\xi^{*,1} = (1, \frac{1}{2}, 0, \frac{1}{2})^T, \quad \xi^{*,2} = (\frac{1}{3}, 1, \frac{2}{3}, 0)^T.$$

In Example (B), the same is true with

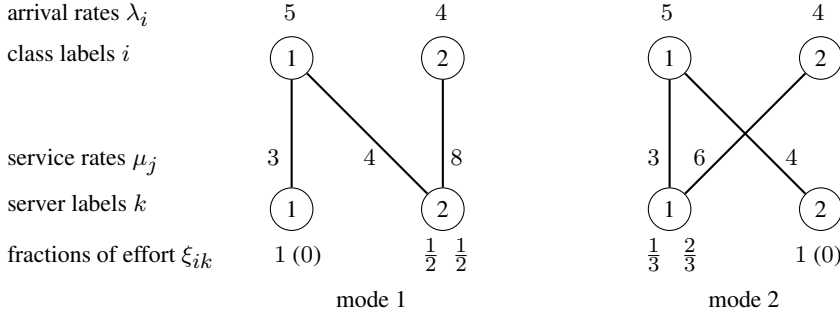
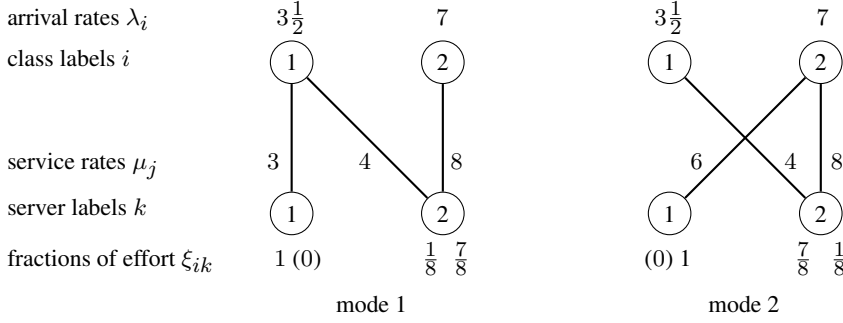
$$\xi^{*,1} = (1, \frac{1}{8}, 0, \frac{7}{8})^T, \quad \xi^{*,2} = (0, \frac{7}{8}, 1, \frac{1}{8})^T.$$

Note that the duals for Examples (A) and (B) differ only in their objective functions - the constraints are the same. It is easy to check that  $y = (\frac{1}{7}, \frac{1}{14})$ ,  $z = (\frac{3}{7}, \frac{4}{7})$  satisfies the constraint (2.17). This yields an objective value of 1 for both Examples (A) and (B). The duality theorem (Theorem A.3) shows that this solution is optimal for the dual in both examples, and, moreover, that  $\rho^* = 1$ . Hence the set of solutions to the LP is given by  $\rho^* = 1$  and  $\mathcal{S}_{LP} = \text{ch}\{\xi^{*,1}, \xi^{*,2}\}$ , and  $\xi^{*,1}, \xi^{*,2}$  are the modes. Figures 1 and 2 present these modes in the context of the graphs formed by classes and servers as nodes, and activities with positive fraction of effort as edges.

These calculations show that, in both examples, the system is critically loaded and there are multiple ways in which the effort can be allocated so as to meet the demand. In other words, the EHTC holds while the HTC does not. Also, in both examples, both modes satisfy  $\sum_k \xi_{ik}^* = 1$ ,  $i = 1, 2$ , expressing that the servers are fully occupied. It is thus seen that Assumption 2.2(1)–(2) hold.

To complete the verification of our standing assumption it remains to show that Assumption 2.2(3) holds, i.e., that the dual program has a unique solution. Our argument works for



Fig 1: *The two modes in Example (A).*Fig 2: *The two modes in Example (B).*

both Examples (A) and (B). Note that in both examples, a strictly positive quantity of each of the four activities is used in at least one of the modes. Thus, by complementary slackness (Lemma A.4(iii)), the constraints in the second line of (2.17) all hold at equality. (This argument is carried out in more detail in the proof of part 4 of Lemma 2.3 in §3.) This yields  $18y_1y_2 = z_1^2$  and  $32y_1y_2 = z_2^2$ , so that  $z_2 = \frac{4}{3}z_1$ . Combined with  $z_1 + z_2 = 1$  this yields a unique solution to the dual.

Going back to the graphs shown in Figures 1 and 2, it is seen that the two examples differ in the way the two modes are paired. In the case of Example (A), the non-basic activity (i.e., the one for which  $\xi_j = 0$ ) switches a server between the two modes, whereas it switches a class in Example (B). Although this distinction has no consequences in this paper, it turns out to be of crucial significance when it comes to implementing AO policies, as in [1]. As is also shown in [1], it is impossible to have a pairing of modes in which the non-basic activity switches both a class and a server. Furthermore, the results of [1], which show that the lower bound of Theorem 2.6 is tight, cover these two examples.

EXAMPLE (C). Keep  $\mu_j$ ,  $j = 1, \dots, 4$  as above, but let  $\lambda_1 = 4$  and  $\lambda_2 = 6$ . In this case the EHTC still holds, with

$$\xi^{*,1} = (0, 1, 1, 0)^T, \quad \xi^{*,2} = (1, \frac{1}{4}, 0, \frac{3}{4})^T.$$

This is an example of a degenerate LP as discussed above. Note that the dual solution is the same as in Examples (A) and (B), and is unique. In particular, this example satisfies our standing assumption and so the lower bound of Theorem 2.6 is valid. Because of the LP degeneracy, the results of [1] do not cover this example, and because, further, there are multiple solutions to the LP, this example falls into the category of problems to which Conjecture 2.7 refers, stating that nevertheless this lower bound is tight.

EXAMPLE (D). We next provide an example that goes beyond uniqueness of solutions to the dual. We consider a 3-class, 3-server system with seven activities. Let

$$\mu_{11} = 3, \quad \mu_{12} = 4, \quad \mu_{21} = 6, \quad \mu_{22} = 8, \quad \mu_{23} = 6, \quad \mu_{32} = 7, \quad \mu_{33} = 6.$$

Although the services rates are not decomposable (because  $\frac{\mu_{22}}{\mu_{23}} \neq \frac{\mu_{32}}{\mu_{33}}$ ), the service rates in the 2 by 2 subsystem consisting of classes 1 and 2, and servers 1 and 2, are decomposable.

Then the matrices  $R$  and  $G$  are given by

$$R = \begin{pmatrix} 3 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 6 & 8 & 6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 7 \end{pmatrix}, \quad G = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

We consider the arrival rates  $\lambda_1 = 5, \lambda_2 = 4, \lambda_3 = 6$ . The LP in this case is: Minimize  $\rho$  subject to

$$(2.18) \quad \begin{cases} 3\xi_{11} + 4\xi_{12} = 5, \\ 6\xi_{21} + 8\xi_{22} + 6\xi_{23} = 4, \\ 7\xi_{32} + 6\xi_{33} = 6, \\ \xi_{11} + \xi_{21} \leq \rho, \\ \xi_{12} + \xi_{22} + \xi_{32} \leq \rho, \\ \xi_{23} + \xi_{33} \leq \rho, \\ \min_{ik} \xi_{ik} \geq 0. \end{cases}$$

The dual program takes the form: Maximize  $5y_1 + 4y_2 + 6y_3$  subject to

$$(2.19) \quad \begin{cases} z_1 + z_2 + z_3 = 1, \quad \min_k z_k \geq 0, \\ (3y_1, 4y_1, 6y_2, 8y_2, 6y_2, 7y_3, 6y_3) \leq (z_1, z_2, z_1, z_2, z_3, z_2, z_3). \end{cases}$$

Let

$$\xi = (\xi_{11}, \xi_{12}, \xi_{21}, \xi_{22}, \xi_{23}, \xi_{32}, \xi_{33}).$$

The optimal objective value is  $\rho^* = 1$ , and the set  $\mathcal{S}_{LP}$  is again a line segment with extreme points

$$\xi^{*,1} = (1, \frac{1}{2}, 0, \frac{1}{2}, 0, 0, 1)^T, \quad \xi^{*,2} = (\frac{1}{3}, 1, \frac{2}{3}, 0, 0, 0, 1)^T.$$

Note that  $\xi_{23}^* = \xi_{32}^* = 0$  in both modes, so both modes correspond to the network separating into two subnetworks.

In this case the dual also has multiple solutions. In particular, if we take  $\eta \in [\frac{3}{10}, \frac{24}{73}]$ , and

$$z_1 = \frac{3}{7}(1 - \eta), \quad z_2 = \frac{4}{7}(1 - \eta), \quad z_3 = \eta,$$

$$y_1 = \frac{1}{7}(1 - \eta), \quad y_2 = \frac{1}{14}(1 - \eta), \quad y_3 = \frac{1}{6}\eta,$$

then  $(y, z)$  solves the dual.

The duality theorem provides an easy verification that the above values of  $\xi, y, z$  are optimal for their respective problems. We simply need to verify that they are feasible and yield the same objective value of 1. This verification follows by simply plugging in the values given above.

We can see that the above 2 modes are the only ones as follows. The fact that  $\xi_{23}^* = \xi_{32}^* = 0$  in all modes follows from complementary slackness: for  $\eta = 0.32 \in (\frac{3}{10}, \frac{24}{73}]$  we have

$6y_2 < z_3$  and  $7y_3 < z_2$ , so complementary slackness implies that  $\xi_{23}^* = \xi_{32}^* = 0$ . It is then immediate from (2.18) that  $\xi_{33}^* = 1$ . With  $\xi_{33}^* = 1, \xi_{23}^* = \xi_{32}^* = 0$ , (2.18) reduces to (2.16) with  $\lambda_1 = 5, \lambda_2 = 4$ . Note that the first 4 entries of  $\xi^{*,1}$  and  $\xi^{*,2}$  match the corresponding values in Example (A).

This system satisfies parts 1 and 2 of Assumption 2.2, but not part 3. Multiplicity of the dual solution, roughly speaking, implies that the workload process is not one-dimensional. PSS of this type are considered in [23], but this particular example is excluded from their treatment because they assume that the solution of the primal LP is unique, and that is not true in this example.

EXAMPLE (E). Finally, we consider the 3-class, 3-server system of Example (D) with slightly modified service rates. Let

$$\mu_{11} = 3, \mu_{12} = 4, \mu_{21} = 6, \mu_{22} = 8, \mu_{23} = 6, \mu_{32} = 8, \mu_{33} = 6.$$

Here, the services rates are decomposable, with  $\alpha_1 = 10, \alpha_2 = 20, \alpha_3 = 20, \beta_1 = 3/10, \beta_2 = 4/10$ , and  $\beta_3 = 3/10$ . The matrices  $R$  and  $G$  are given by

$$R = \begin{pmatrix} 3 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 6 & 8 & 6 & 0 \\ 0 & 0 & 0 & 0 & 8 & 6 \end{pmatrix}, \quad G = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

We consider the same arrival rates as in Example (D):  $\lambda_1 = 5, \lambda_2 = 4, \lambda_3 = 6$ . The LP in this case is: Minimize  $\rho$  subject to

$$(2.20) \quad \begin{cases} 3\xi_{11} + 4\xi_{12} = 5, \\ 6\xi_{21} + 8\xi_{22} + 6\xi_{23} = 4, \\ 8\xi_{32} + 6\xi_{33} = 6, \\ \xi_{11} + \xi_{21} \leq \rho, \\ \xi_{12} + \xi_{22} + \xi_{32} \leq \rho, \\ \xi_{23} + \xi_{33} \leq \rho, \\ \min_{ik} \xi_{ik} \geq 0. \end{cases}$$

The dual program takes the form: Maximize  $5y_1 + 4y_2 + 6y_3$  subject to

$$(2.21) \quad \begin{cases} z_1 + z_2 + z_3 = 1, \quad \min_k z_k \geq 0, \\ (3y_1, 4y_1, 6y_2, 8y_2, 6y_2, 8y_3, 6y_3) \leq (z_1, z_2, z_1, z_2, z_3, z_2, z_3). \end{cases}$$

As before, let

$$\xi = (\xi_{11}, \xi_{12}, \xi_{21}, \xi_{22}, \xi_{23}, \xi_{32}, \xi_{33}).$$

The optimal objective value is  $\rho^* = 1$ , and there are multiple solutions to the LP, with the set  $\mathcal{S}_{\text{LP}}$  having 3 extreme points:

$$\xi^{*,1} = (1, \frac{1}{2}, 0, \frac{1}{2}, 0, 0, 1)^T, \quad \xi^{*,2} = (\frac{1}{3}, 1, \frac{2}{3}, 0, 0, 0, 1)^T, \quad \xi^{*,3} = (1, \frac{1}{2}, 0, 0, \frac{2}{3}, \frac{1}{2}, \frac{1}{3})^T.$$

It follows from the fact that the columns of  $\begin{pmatrix} R \\ G \end{pmatrix}$  corresponding to positive elements of  $\xi$  are linearly independent that these are all extreme points. It can be seen that these are the only extreme points as follows. Turn the constraints in (2.20) into a set of 6 equations by replacing the 3 inequalities with equalities and replacing  $\rho$  by 1. The set of non-negative solutions to this set of equations is precisely  $\mathcal{S}_{\text{LP}}$ . It is easy to see that  $\mathcal{S}_{\text{LP}}$  can be parametrized by  $\xi_{11}$  and  $\xi_{33}$ : The five other  $\xi_{ik}$  can be determined once  $\xi_{11}$  and  $\xi_{33}$  are known. Furthermore, we

must have  $\xi_{11} \in [1/3, 1]$ ,  $\xi_{33} \in [1/3, 1]$ , and  $\xi_{11} + \xi_{33} \geq 4/3$ . These 3 constraints determine a triangle in the positive quadrant, where the extreme points correspond precisely to  $\xi^{*,1}$ ,  $\xi^{*,2}$ , and  $\xi^{*,3}$ .

In this case the dual has a unique solution (which corresponds to  $\eta = \frac{3}{10}$  in Example (D)):

$$\begin{aligned} z_1 &= \frac{3}{10}, & z_2 &= \frac{4}{10}, & z_3 &= \frac{3}{10}, \\ y_1 &= \frac{1}{10}, & y_2 &= \frac{1}{20}, & y_3 &= \frac{1}{20}. \end{aligned}$$

(Uniqueness can be established in a manner similar to Examples (A) and (B).) The duality theorem again provides an easy verification that the above values of  $\xi, y, z$  are optimal for their respective problems.

This system satisfies all three parts of Assumption 2.2, so that Proposition 2.5 applies to it, but it is not covered by [1], which is restricted to 2 class, 2 server systems. It is also worth noting that modes 1 and 2 are both degenerate.

Whereas the examples provided above are concerned only with first order data, the second order data determines the structure of the WCP as well as the HJB equation. In §4.3, Examples (A), (C) and (E) above are further developed, discussing the role of the second order data.

**3. The LP under the EHTC.** The goal of this section is to prove Lemmas 2.3 and 2.4. In the proof of Lemmas 2.3 we use several standard results related to linear programs, such as complementary slackness and the duality theorem. We also use a slightly less standard result, commonly referred to ‘strict complementary slackness’. For all of these we refer to appropriate results in [26]. For ease of reference all of our cited results are reproduced in the appendix.

**PROOF OF LEMMA 2.3.** 1. The set  $\mathcal{S}_{LP}$  is the set of  $\xi$  such that the constraints in (2.7) hold with  $\rho = 1$ . By (9) in Section 7.2 of [26] (Definition A.1), this set is a convex polyhedron. We show that this polyhedron is bounded. Consider activity  $j = (i, k) \in \mathcal{J}$ . Since for any  $\xi \in \mathcal{S}_{LP}$  we have  $(G\xi)_k = 1$ , for any  $k \in \mathcal{K}$ ,  $0 \leq \xi_j \leq 1$ . By Corollary 7.1c of [26] (Lemma A.2), this is enough to show that it is a polytope and thus finitely generated.

2. We first rearrange (2.7) to a more standard form that enables the use of results from [26]. To that effect, let

$$\begin{aligned} \bar{x} &= (\rho, \xi_1, \dots, \xi_J)^T, \\ \bar{c} &= (-1, 0_J), \\ \bar{b}' &= (\lambda, -\lambda, 0_K, 0_J), \end{aligned}$$

and

$$\bar{G} = \begin{pmatrix} 0_I & R \\ 0_I & -R \\ -1_K & G \\ 0_J & -I_J \end{pmatrix}.$$

Then (2.7) becomes:

$$(3.1) \quad \text{maximize } \bar{c} \cdot \bar{x} \text{ subject to } \bar{G}\bar{x} \leq \bar{b}.$$

Then, directly,

$$\bar{c} \cdot \bar{x} = -\rho.$$

In addition,  $\bar{G}\bar{x} \leq \bar{b}$  is equivalent to following the set of conditions:

$$\begin{aligned} R\xi &\leq \lambda, \\ -R\xi &\leq -\lambda, \\ -\rho \cdot \mathbf{1} + G\xi &\leq 0_K \\ -\xi &\leq 0_J. \end{aligned}$$

As a result, the LP (3.1), with the above parameters  $(\bar{b}, \bar{c}, \bar{G})$ , is equivalent to our original LP (2.7). We now use Corollary 7.1g of [26] (Theorem A.3), according to which the dual of (3.1) is given by the LP

$$(3.2) \quad \text{Minimize } y \cdot \bar{b} \text{ subject to } y\bar{G} = \bar{c}, y \geq 0,$$

and the primal and the dual have the same value. (It follows from Assumptions 2.2 (1) and (3) that both sets in (A.1) are non-empty.)

Therefore result 2 will be proved once we show that the LP (3.2) is equivalent to our original dual, (2.8). To this end, write  $y$  as  $y = (y^1, y^2, z, s)$ , where  $y^1, y^2 \in \mathbb{R}_+^I$ ,  $z \in \mathbb{R}_+^K$  and  $s \in \mathbb{R}_+^J$ . Next, note that by plugging in the values for  $y$  and  $\bar{b}$ , we obtain

$$y \cdot \bar{b} = (y^1 - y^2) \cdot \lambda.$$

This equality means that minimizing  $y \cdot \bar{b}$  is the same as maximizing  $(y^2 - y^1) \cdot \lambda$ . In addition,  $y\bar{G} = \bar{c}$  is equivalent to the following set of conditions:

$$\begin{aligned} \sum_k z_k &= 1, \\ (y^2 - y^1)R &= zG - s. \end{aligned}$$

Set  $\bar{y} = y^2 - y^1$ . Note that the sign of  $\bar{y}$  is unknown. Since  $s \geq 0$ , we can replace the last equality by an inequality. This is sufficient to conclude that (3.2) and (2.8) are equivalent. This shows  $y^* \cdot \lambda = \rho^* = 1$ .

3. We appeal to result (36) in chapter 7 of [26] (Theorem A.5), commonly referred to as ‘strict complementary slackness’. By Assumption 2.2, we have  $(G\xi^*)_k = 1$  for all  $k \in \mathcal{K}$ . The dual variable corresponding to the constraint  $(G\xi^*)_k \leq \rho$  is  $z_k^*$ . Since (36) (i) of [26] (Theorem A.5 (a)) does not hold for any solution of the primal, (36) (ii) (Theorem A.5 (b)) must hold. Thus  $z_k^* > 0$  for all  $k$ .

4. In a manner similar to the rearrangement/translation in part 2 above, we can translate our primal and dual respectively into the minimization and maximization problem in (A.1). (In this translation, denoting the variable  $y$  appearing in (A.1) by  $\tilde{y}$  to avoid confusion with our variable  $y$  in (2.8), we have  $\tilde{y} = (\xi, \rho, \tilde{s})$ , where  $\tilde{s} \in \mathbb{R}_+^K$  corresponds to slack variables for the  $K$  constraints  $G\xi \leq \rho \mathbf{1}_K$ .) Invoking (35)(iii) of [26] (Lemma A.4), which is complementary slackness, if there is a solution of (3.1) such that  $\xi_j^* > 0$ , then  $(y^*R)_j = (z^*G)_j$ . In other words,

$$(3.3) \quad y_i^* \mu_j = z_k^* \quad \text{whenever } j = (i, k) \in \mathcal{A}_{\text{PB}}.$$

5. We use the maximization/minimization pair of part 4, and invoke (36) of [26] (Theorem A.5) again, this time to say that if  $\xi_j^* = 0$  for all solutions to (3.1) then the associated constraint  $(y^*R)_j \leq (z^*G)_j$  has positive slack, so that

$$(3.4) \quad y_i^* \mu_j < z_k^* \quad \text{whenever } j = (i, k) \in \mathcal{A}_{\text{AN}}.$$



6. Assume that  $i$  is such that  $\mathcal{J}_i \cap \mathcal{A}_{\text{PB}} = \emptyset$ . Then for any  $\xi \in \mathcal{S}_{\text{LP}}$ , we obtain  $(R\xi)_i = \sum_{j \in \mathcal{J}_i} \mu_j \xi_j = 0$ . Since  $\lambda > 0$  there is a contradiction with  $(R\xi)_i = \lambda_i$  in (2.7), so such an  $i$  cannot exist.

7. By part 6, for every  $i$ , there exists a server  $k$  such that  $j = (i, k) \in \mathcal{A}_{\text{PB}}$ . By part 4,

$$y_i^* = \frac{z_k^*}{\mu_j},$$

and by part 3 the last quantity is positive.  $\square$

PROOF OF LEMMA 2.4. 1. For any  $\xi \in \mathcal{S}_{\text{LP}}$  and  $i \in \mathcal{I}$ ,

$$\begin{aligned} \frac{\lambda_i}{\alpha_i} &= \sum_{j \in \mathcal{J}_i} \frac{\mu_j}{\alpha_i} \xi_j \\ &= \sum_{k, (i,k) \in \mathcal{J}} \beta_k \xi_{ik} \end{aligned}$$

Summing over  $i$ , we get

$$\begin{aligned} \sum_i \frac{\lambda_i}{\alpha_i} &= \sum_i \sum_{k, (i,k) \in \mathcal{J}} \beta_k \xi_{ik} \\ &= \sum_k \beta_k \sum_{j \in \mathcal{J}^k} \xi_j \\ &= \sum_k \beta_k. \end{aligned}$$

The last equality is obtained using Assumption 2.2(2). The claim follows from the convention  $\sum_k \beta_k = 1$  that we have imposed.

2. The choice  $z^* = \beta$  obviously satisfies the condition of summing to 1. For any  $j = (i, k) \in \mathcal{J}$ ,  $(y^* R)_j = \frac{\mu_j}{\alpha_i} = \beta_k = (z^* G)_j$ . In addition, by part 1, the optimal value for the primal is attained by this solution. Since both problems have the same value we have shown that the choice given for  $(y^*, z^*)$  is a solution to the dual. The solution is unique by Assumption 2.2.

3. Let  $\xi^*$  be such that for any  $j \in \mathcal{J}_i$ ,  $\xi_j^* = \frac{\lambda_i}{\alpha_i}$ . One can check that when  $\mathcal{J} = \mathcal{I} \times \mathcal{K}$ ,  $\xi^* \in \mathcal{S}_{\text{LP}}$ . Indeed, for any  $i$ ,

$$(R\xi^*)_i = \sum_{j \in \mathcal{J}_i} \mu_j \xi_j^* = \sum_{k, (i,k) \in \mathcal{J}} \frac{\lambda_i}{\alpha_i} \alpha_i \beta_k = \lambda_i \sum_k \beta_k = \lambda_i,$$

where the last equality comes from the definition of  $\beta$ . Similarly,

$$(G\xi^*)_k = \sum_{j \in \mathcal{J}^k} \xi_j^* = \sum_i \frac{\lambda_i}{\alpha_i} = 1,$$

where the last equality comes from part 1.

Notice that all the components of  $\xi^*$  are positive. We will now perturb this solution to get another element of  $\mathcal{S}_{\text{LP}}$ . Let  $\delta \in \mathbb{R}^J$  such that  $\delta_{ik} = \frac{(-1)^{i+k}}{\beta_k}$  if  $i$  and  $k$  are no larger than 2 and  $\delta_{(i,k)} = 0$  otherwise. For  $\varepsilon > 0$  small enough,  $\xi^* + \varepsilon \delta \in \mathbb{R}_+^J$ . Server 1 diverts a fraction  $\frac{\varepsilon}{\beta_1}$  of its effort from class 2 to class 1, server 2 diverts a fraction  $\frac{\varepsilon}{\beta_2}$  of its effort from class 1 to

class 2. We thus have  $R\delta = G\delta = 0$ , so that this solution satisfies the constraints of the LP. To see this, note that due to decomposability, we have

$$(R\delta)_1 = \frac{\mu_{11}}{\beta_1} - \frac{\mu_{12}}{\beta_2} = \alpha_1 - \alpha_1 = 0,$$

$$(G\delta)_1 = \frac{1}{\beta_1} - \frac{1}{\beta_1} = 0.$$

The second coordinate is obtained in the same way. The other coordinates are also zero because  $\delta_{ik} = 0$  if  $i$  or  $k$  is neither 1 or 2. This proves  $\xi^* + \varepsilon\delta \in \mathcal{S}_{\text{LP}}$  which proves decomposability is sufficient for multiplicity.

4. We have already shown that decomposability is sufficient for multiplicity in general, we now prove the necessary part, that is when  $I = K = 2$ , having multiple solution can only happen if  $\mathcal{J} = \mathcal{I} \times \mathcal{K}$ . We identify  $\mathcal{I}$  and  $\mathcal{J}$  with  $\{1, 2\}$ . Let us have two distinct solutions  $\xi^{(p)}$ ,  $p = 1, 2$ . By Assumption (2.2),  $\sum_{j \in \mathcal{J}^k} \xi_j^{(p)} = 1$  holds for  $k = 1, 2$  and  $p = 1, 2$ . Next, define  $\delta_{ik} = \xi_{ik}^{(1)} - \xi_{ik}^{(2)}$ . Then, since both  $\xi^{(1)}$  and  $\xi^{(2)}$  are solutions to (2.7),

$$(3.5) \quad \begin{aligned} \delta_{11}\mu_{11} + \delta_{12}\mu_{12} &= 0, \\ \delta_{21}\mu_{21} + \delta_{22}\mu_{22} &= 0, \\ \delta_{11} + \delta_{21} &= 0, \\ \delta_{12} + \delta_{22} &= 0. \end{aligned}$$

The last three equations imply

$$(3.6) \quad \delta_{11}\mu_{21} + \delta_{12}\mu_{22} = 0.$$

Since the solutions are distinct, one of the four entries of  $\delta$  is nonzero, and it follows that all four entries of  $\delta$  are nonzero (this uses nonvanishing of  $\mu_{ik}$ ). By (3.5) and (3.6) we have

$$(3.7) \quad \frac{\mu_{11}}{\mu_{12}} = \frac{\mu_{21}}{\mu_{22}}.$$

Hence  $(\mu_{21}, \mu_{22}) = c(\mu_{11}, \mu_{12})$  for some constant  $c > 0$ . This gives the decomposability.

Showing that the number of modes is at most 2 amounts to showing that when the LP has multiple solutions,  $\mathcal{S}_{\text{LP}}$  is a line segment. Using the four equations in (3.5) along with (3.7), it is not hard to show that whenever  $\xi$  and  $\eta$  are in  $\mathcal{S}_{\text{LP}}$ , one has  $\xi - \eta \in \text{span}(\delta^0)$ , where  $\delta^0$  is the vector  $(\delta_{(i,k)}^0)_{(i,k) \in \mathcal{J}} = \left( \frac{\mu_{i1}}{\mu_{ik}} (-1)^{i+k} \right)_{(i,k) \in \mathcal{J}}$ . Since  $\mathcal{S}_{\text{LP}}$  is convex and bounded, this shows that it forms a line segment, and proves the claim.  $\square$

**4. The BCP, WCP and HJB equation.** The main goal of this section is to derive the BCP and WCP describing the asymptotics of the model, and construct an optimal solution to the latter via the HJB equation. By way of doing this, we introduce rescaled processes and develop relations among them, that are used in the remainder of the paper. The derivation of the BCP and WCP appears in §4.1. §4.2 states and proves Lemma 4.3, which constructs a candidate optimal solution to the WCP. §4.3 proves Proposition 2.5 and affirms, in Corollary 4.4, that the aforementioned candidate is optimal. It also addresses, in Proposition 4.5, the special case where there is one mode for which both coefficients  $b$  and  $\sigma$  are smallest, showing that it is optimal to always use this mode.

4.1. *Derivation of the BCP and WCP.* Two control problems formulated in terms of diffusion processes, referred to as the BCP and the WCP, are presented here, aimed at formally describing the asymptotics of the queue length and of the workload, respectively. The WCP has the advantage of lying in dimension 1, and it is this problem that plays a central role in this paper. In earlier work on the PSS model, the BCP was obtained from the prelimit problem and then transformed into a WCP. In this paper, the connection between the prelimit and the WCP is made directly, and therefore the BCP itself has no use in the proofs. It is presented for the sole purpose of comparison to the form it took in earlier work. However, there will be reference to some of the notation introduced and equations obtained along the way to deriving it.

4.1.1. *BCP derivation.* Following standard convention, rescaled processes at the diffusion scale are denoted with hat. In addition to the ones already defined,  $\hat{X}^n$  and  $\hat{W}^n$ , let

$$(4.1) \quad \hat{A}_i^n(t) = n^{-1/2}(A_i^n(t) - \lambda_i^n t),$$

$$\hat{S}_j^n(t) = n^{-1/2}(S_j^n - \mu_j^n t),$$

$$(4.2) \quad \hat{I}_k^n(t) = n^{1/2}I_k^n(t),$$

$$\hat{L}^n(t) = z^* \cdot \hat{I}^n(t).$$

By (2.2) and the balance equation (2.3),

$$(4.3) \quad \hat{X}_i^n(t) = \hat{A}_i^n(t) + n^{-1/2}\lambda_i^n t - \sum_{j \in \mathcal{J}_i} \hat{S}_j^n(T_j^n) - \sum_{j \in \mathcal{J}_i} n^{-1/2}\mu_j^n T_j^n(t).$$

We next introduce a notion of which no rigorous use will be made in this paper, serving only for the derivation of the BCP: A *fluid control process* is a process  $\bar{T}^n$  that takes the form  $\bar{T}_j^n = \int_0^\cdot \bar{\Xi}_j^n(s) ds$ , where  $\bar{\Xi}^n$  is some process taking values in  $\mathcal{S}_{LP}$ . Given any fluid control process  $\bar{T}^n$  we can perform the centering

$$\hat{Y}_j^n(t) = -n^{1/2}(T_j^n(t) - \bar{T}_j^n(t)).$$

Then, by (4.3) and the definition of  $\hat{\lambda}^n$ ,  $\hat{\mu}^n$  one has

$$\begin{aligned} \hat{X}_i^n(t) &= \hat{A}_i^n(t) - \sum_{j \in \mathcal{J}_i} \hat{S}_j^n(T_j^n(t)) + \sum_{j \in \mathcal{J}_i} \mu_j \hat{Y}_j^n(t) \\ &\quad - \sum_{j \in \mathcal{J}_i} \hat{\mu}_j^n T_j^n(t) - \sum_{j \in \mathcal{J}_i} n^{1/2} \mu_j \bar{T}_j^n(t) + n^{-1/2} \lambda_i^n t. \end{aligned}$$

The sum of the last two terms above is equal to

$$\begin{aligned} \int_0^t n^{-1/2} \left( \lambda_i^n - n^{1/2} \sum_{j \in \mathcal{J}_i} \mu_j \bar{\Xi}_j^n(s) \right) ds &= n^{1/2} \int_0^t \left( \lambda_i - \sum_{j \in \mathcal{J}_i} \mu_j \bar{\Xi}_j^n(s) \right) ds + \hat{\lambda}_i^n t \\ &= \hat{\lambda}_i^n t, \end{aligned}$$

where the last equality follows from the fact that  $\bar{\Xi}^n(t)$  takes values in  $\mathcal{S}_{LP}$ . Thus

$$(4.4) \quad \hat{X}_i^n(t) = \hat{A}_i^n(t) - \sum_{j \in \mathcal{J}_i} \hat{S}_j^n(T_j^n(t)) + \hat{\lambda}_i^n t - \sum_{j \in \mathcal{J}_i} \hat{\mu}_j^n T_j^n(t) + \sum_{j \in \mathcal{J}_i} \mu_j \hat{Y}_j^n(t).$$

Next, by (2.4) and the identity  $\sum_{j \in \mathcal{J}^k} \bar{\Xi}_j^n(t) = 1$ ,

$$(4.5) \quad \hat{I}_k^n(t) = n^{1/2} \sum_{j \in \mathcal{J}^k} (\bar{T}_j^n(t) - T_j^n(t)) = \sum_{j \in \mathcal{J}^k} \hat{Y}_j^n(t).$$

By the central limit theorem for renewal processes [4, §17], the processes  $(\hat{A}_i^n, \hat{S}_j^n)$  converge to  $(A_i, S_j)$ , a collection of  $I + J$  mutually independent BMs starting from zero, with zero drift and diffusion coefficients given by the constants  $\lambda_i^{1/2} C_{A_i}$  and  $\mu_j^{1/2} C_{S_j}$ , which we have earlier denoted by  $\sigma_{A_i}$  and  $\sigma_{S_j}$ , respectively. Assuming, without any justification as this point, that  $(\hat{A}^n, \hat{S}^n, \hat{X}^n, T^n, \bar{T}^n, \hat{Y}^n, \hat{I}^n)$  weakly converge to  $(A, S, X, T, \bar{T}, Y, I)$ , we have from (4.4) and (4.5)

$$(4.6) \quad X_i(t) = A_i(t) - \sum_{j \in \mathcal{J}_i} S_j(T_j(t)) + \hat{\lambda}_i t - \sum_{j \in \mathcal{J}_i} \hat{\mu}_j T_j(t) + \sum_{j \in \mathcal{J}_i} \mu_j Y_j(t),$$

$$(4.7) \quad I_k(t) = \sum_{j \in \mathcal{J}^k} Y_j(t).$$

These processes satisfy the constraints

$$(4.8) \quad X_i(t) \geq 0 \text{ for all } t,$$

$$(4.9) \quad I_k \text{ has sample paths in } C_{\mathbb{R}}^{0,+}[0, \infty),$$

$$(4.10) \quad \text{if } T_j \text{ is flat on some interval } (s, t) \text{ then } Y_j \text{ is non-increasing on it.}$$

The BCP consists of minimizing the cost

$$\mathbb{E} \int_0^\infty e^{-\gamma t} h(X(t)) dt,$$

over all tuples  $(T, Y, X, I)$  satisfying (4.6)–(4.10). A precise definition must also account for adaptedness w.r.t. a filtration, as in the definition of a WCP given in §2.3; however, since this problem is presented here only for motivational purposes, we will not formulate it in full detail.

There are similarities between this BCP and those that appear in [15], [2], [3]. An important difference is the appearance of the process  $T$  as a control process, as opposed to it being deterministic in the above references. This has significant consequences already mentioned in the introduction.

**4.1.2. WCP derivation.** We next show how the WCP, which was introduced in §2.3, is obtained directly from the queueing model, and then discuss the role it plays in this paper. Let the process that appears in the definition of the cost (2.6) be denoted by  $\hat{H}^n$ , where

$$(4.11) \quad \hat{H}_t^n = h(\hat{X}_t^n) = h \cdot \hat{X}_t^n.$$

Using (4.3) and the definitions of  $\hat{\lambda}^n, \hat{\mu}^n$ , a straightforward calculation gives, with  $\iota(t) = t$ ,

$$(4.12) \quad \hat{W}^n = \hat{F}^n + n^{1/2} \sum_i y_i^* \lambda_{i\iota} - n^{1/2} \sum_i y_i^* \sum_{j \in \mathcal{J}_i} \mu_j T_j^n,$$

where

$$(4.13) \quad \hat{F}^n = \sum_i y_i^* (\hat{A}_i^n - \sum_{j \in \mathcal{J}_i} \hat{S}_j^n \circ T_j^n + \hat{\lambda}_i^n \iota - \sum_{j \in \mathcal{J}_i} \hat{\mu}_j^n T_j^n).$$

The analysis of the problem via workload considerations is based on two key inequalities. The first, which uses the positivity of  $y^*$  and the definition of  $q$  in (2.9), is the following:

$$(4.14) \quad \hat{H}_t^n = \sum_i h_i y_i^* (y_i^*)^{-1} \hat{X}_i^n(t) \geq h_q (y_q^*)^{-1} \hat{W}_t^n, \text{ with equality when } \sum_{i \neq q} \hat{X}_i^n(t) = 0.$$

The second key inequality is given by the following lemma. To state it we first introduce the *Skorohod map on the half line*, denoted throughout this paper by  $\Gamma : D_{\mathbb{R}}[0, \infty) \rightarrow D_{\mathbb{R}_+}[0, \infty) \times D_{\mathbb{R}_+}^+[0, \infty)$ . It takes a function  $\psi$  to a pair  $(\varphi, \eta)$ , where

$$\varphi(t) = \psi(t) + \eta(t), \quad \eta(t) = \sup_{0 \leq s \leq t} \psi(s)^-, \quad t \geq 0.$$

The corresponding maps  $\psi \mapsto \varphi$  and  $\psi \mapsto \eta$  are denoted by  $\Gamma_1$  and  $\Gamma_2$ , respectively.

LEMMA 4.1. *One has*

$$(4.15) \quad \hat{W}_t^n \geq \Gamma_1[\hat{F}^n](t), \quad t \geq 0.$$

The proof appears at the end of this section.

Inequalities (4.14) and (4.15) allow us to bound the cost  $\hat{J}^n$  defined in (2.6) as follows:

$$(4.16) \quad \hat{J}^n(T^n) = \mathbb{E} \int_0^\infty e^{-\gamma t} \hat{H}_t^n dt \geq h_q(y_q^*)^{-1} \mathbb{E} \int_0^\infty e^{-\gamma t} \Gamma_1[\hat{F}^n](t) dt, \quad T^n \in \mathcal{A}^n.$$

The derivation that follows is concerned with the asymptotics of the RHS of the above inequality.

We take formal limits in (4.12)–(4.13). If we assume that  $(\hat{A}^n, \hat{S}^n, T^n) \Rightarrow (A, S, T)$  then by (4.13),  $\hat{F}^n \Rightarrow F$  defined by

$$(4.17) \quad F(t) = \sum_i y_i^* \left( A_i(t) + \hat{\lambda}_i t - \sum_{j \in \mathcal{J}_i} S_j(T_j(t)) + \hat{\mu}_j T_j(t) \right).$$

Assuming further that  $T$  is of the form  $\int_0^\cdot \Xi(t) dt$  where  $\Xi(t) = (\Xi_j(t))$  takes values in  $\mathcal{S}_{LP}$ , using the definition (2.10) of  $b$  and (4.17), we have

$$(4.18) \quad \begin{aligned} Z_t &:= \Gamma_1[F](t) = \int_0^t \sum_i y_i^* \left( \hat{\lambda}_i - \sum_{j \in \mathcal{J}_i} \hat{\mu}_j \Xi_j(s) \right) ds + M_t + L_t \\ &= \int_0^t b(\Xi_s) ds + M_t + L_t, \end{aligned}$$

where  $Z_t \geq 0$ ,  $L$  has sample paths in  $C_{\mathbb{R}}^{0,+}[0, \infty)$ ,  $\int_{[0, \infty)} Z_t dL_t = 0$ , and

$$(4.19) \quad M_t = \sum_i y_i^* \left( A_i(t) - \sum_{j \in \mathcal{J}_i} S_j(T_j(t)) \right).$$

If one can further show that  $A_i$  and  $S_j \circ T_j$  are all martingales on a common filtration, and moreover,  $\langle S_j \circ T_j \rangle_t = \langle S_j \rangle_{T_j(t)}$  then  $M_t$  is a martingale, and

$$\langle M \rangle_t = \int_0^t \sum_i (y_i^*)^2 \left( \sigma_{A_i}^2 + \sum_{j \in \mathcal{J}_i} \sigma_{S_j}^2 \Xi_j(s) \right) ds = \int_0^t \sigma^2(\Xi(s)) ds,$$

where the definition (2.10) of  $\sigma$  is used. Hence

$$Z_t = \int_0^t b(\Xi_s) ds + \int_0^t \sigma(\Xi_t) dB_t + L_t,$$

for  $B$  a SBM. This is precisely (2.12) with initial condition  $z = 0$ . (Although the case  $z = 0$  reflects our assumption  $\hat{X}^n(0) = 0$ , the formulation of (2.12) with a general initial condition  $z \geq 0$  will be important when we appeal to the HJB equation.) Moreover, because by construction  $Z = \Gamma_1[F]$ , the cost  $J_{WCP}$  is formally the limit of the expected integral on the RHS



of (4.16). This suggests that the value function for the QCP is asymptotically bounded below by  $h_q(y_q^*)^{-1}V_{\text{WCP}}(0)$ , which is the main result of this paper. Indeed, (4.16) is the starting point for its proof in §5.

From the above discussion it also follows that in order for the lower bound to be tight, that is, for there to exist a QCP policy that asymptotically achieves it, the two inequalities (4.14), (4.15) should asymptotically be achieved as equalities, and the RHS of (4.16) should be asymptotic to the optimal WCP solution. For (4.14) to hold as near equality, one must keep  $\hat{X}_i^n$ ,  $i \neq q$  close to zero. As will become clear from the proof below of Lemma 4.1, for (4.15) to hold as equality up to a negligible term, the process  $\hat{L}^n$  (more precisely, the sum  $\hat{L}^n + \hat{L}_{\text{AN}}^n$ , where the latter is defined in the proof of Lemma 4.1), should be kept close to  $\Gamma_2[\hat{F}^n]$ . For  $\hat{L}^n$  this entails that, up to negligible terms, none of the servers should accumulate idleness when there is work in the system, and for  $\hat{L}_{\text{AN}}^n$  this entails that, up to negligible terms, no server should engage in any activity  $j \in \mathcal{A}_{\text{AN}}$ , that is, always non-basic, except possibly when the workload is close to zero. For PSS satisfying the HTC these two points were recognized in [13], who showed that, if sufficient care is not taken in keeping the components  $\hat{X}_i^n$ ,  $i \neq q$  close to zero, the system could be overwhelmingly overloaded, as well as in [2], who have constructed a threshold policy that succeeds in achieving the two goals. When the HTC does not hold there are additional considerations involving the third aspect mentioned above. Namely, for the RHS of (4.16) to be asymptotic to the WCP solution, certain switching control policies need to be implemented. This is the subject of [1].

**PROOF OF LEMMA 4.1.** Any activity involves exactly one class and one server. For that reason, the following three operations are equivalent and can be used interchangeably, depending on what index is needed in the summation:  $\sum_i \sum_{j \in \mathcal{J}_i}$ ,  $\sum_k \sum_{j \in \mathcal{J}^k}$ ,  $\sum_i \sum_k \mathbb{1}_{(i,k) \in \mathcal{J}}$ . By Lemma 2.3 part 4, the last term in (4.12) can be modified using the following identity

$$\begin{aligned}
\sum_i y_i^* \sum_{j \in \mathcal{J}_i} \mu_j T_j^n &= \sum_i \sum_k \mathbb{1}_{(i,k) \in \mathcal{J} \cap \mathcal{A}_{\text{PB}}} y_i^* \mu_{ik} T_{ik}^n + \sum_i \sum_k \mathbb{1}_{(i,k) \in \mathcal{J} \cap \mathcal{A}_{\text{AN}}} y_i^* \mu_{ik} T_{ik}^n \\
&= \sum_k \sum_{j \in \mathcal{J}^k \cap \mathcal{A}_{\text{PB}}} z_k^* T_j^n + \sum_k \sum_i \mathbb{1}_{(i,k) \in \mathcal{J} \cap \mathcal{A}_{\text{AN}}} y_i^* \mu_{ik} T_{ik}^n \\
&= \sum_k z_k^* \sum_{j \in \mathcal{J}^k} T_j^n - \sum_k \sum_i \mathbb{1}_{(i,k) \in \mathcal{J} \cap \mathcal{A}_{\text{AN}}} (z_k^* - y_i^* \mu_{ik}) T_{ik}^n \\
(4.20) \quad &= \sum_k z_k^* \sum_{j \in \mathcal{J}^k} T_j^n - L_{\text{AN}}^n,
\end{aligned}$$

where the last equality defines the process  $L_{\text{AN}}^n$ .

In view of Lemma 2.3 part 5 and the fact that  $T_{ik}^n$  are nonnegative and nondecreasing, so is  $L_{\text{AN}}^n$ . Denote  $\hat{L}_{\text{AN}}^n = n^{1/2} L_{\text{AN}}^n$ . By (4.12), (4.20) and Lemma 2.3 part 2,

$$\hat{W}^n = \hat{F}^n + n^{1/2} \iota - n^{1/2} \sum_k z_k^* \sum_{j \in \mathcal{J}^k} T_j^n + \hat{L}_{\text{AN}}^n.$$

By (2.8),  $\sum_k z_k^* = 1$ , and by (2.4),  $\hat{I}_k^n + n^{1/2} \sum_{j \in \mathcal{J}^k} T_j^n = n^{1/2} \iota$ . Hence

$$\begin{aligned}
\hat{W}^n &= \hat{F}^n + n^{1/2} \iota - \sum_k z_k^* (n^{1/2} \iota - \hat{I}_k^n) + \hat{L}_{\text{AN}}^n \\
&= \hat{F}^n + z^* \cdot \hat{I}^n + \hat{L}_{\text{AN}}^n \\
&= \hat{F}^n + \hat{L}^n + \hat{L}_{\text{AN}}^n.
\end{aligned}$$

It is well known that the Skorohod map  $\Gamma_1$  possesses the following *minimality property* (see [7, Section 2]). Let  $\psi \in D_{\mathbb{R}}[0, \infty)$  and assume that  $\varphi = \psi + \eta$  where  $\eta \in D_{\mathbb{R}}^+[0, \infty)$  and  $\varphi(t) \geq 0$  for all  $t$ . Then  $\hat{\varphi}(t) \geq \Gamma_1[\psi](t)$  for all  $t$ .

Because the process  $\hat{W}^n$  has non-negative sample paths and  $\hat{L}^n$  and  $\hat{L}_{\text{AN}}^n$  have sample paths in  $D_{\mathbb{R}}^+[0, \infty)$ , it follows from the minimality property that  $\hat{W}_t^n \geq \Gamma_1[\hat{F}^n](t)$  for all  $t$ . This completes the proof.  $\square$

REMARK 4.2. We now provide an intuitive explanation for  $\hat{L}_{\text{AN}}^n$ . The process  $\hat{L}_{\text{AN}}^n$  involves a sum over the always non-basic activities. Recall that Lemma 2.3 part 5 states that, for any activity  $j = (i, k) \in \mathcal{A}_{\text{AN}}, y_i^* \mu_{ik} < z_k^*$ , while Lemma 2.3 part 4 states that for any activity  $j = (i, k) \in \mathcal{A}_{\text{PB}}, y_i^* \mu_{ik} = z_k^*$ . Intuitively, with  $y_i^*$  interpreted as the average work of a class  $i$  customer,  $y_i^* \mu_{ik}$  represents the rate at which activity  $j$  removes class  $i$  work from the system. Since each server  $k$  has at least one activity  $j' = (i', k) \in \mathcal{A}_{\text{PB}}$ , for that activity  $y_{i'}^* \mu_{i'k} = z_k^* > y_i^* \mu_{ik}$ , so that activity  $j'$  can be viewed as more ‘efficient’ than activity  $j$ . Thus  $\hat{L}_{\text{AN}}^n(t)$  is the additional amount of work that could have been removed from the system by time  $t$  if the effort expended on activities in  $\mathcal{A}_{\text{AN}}$  had instead been used on activities in  $\mathcal{A}_{\text{PB}}$ .

4.2. *Candidate solution to the WCP.* In the following result, under the assumption of existence of a classical solution to the HJB equation, a candidate solution to the WCP is constructed. This result is later used to prove Proposition 2.5 which states that a classical solution exists, and Corollary 4.4 which verifies the optimality of this candidate.

LEMMA 4.3. 1. *For any measurable map  $\xi^0 : \mathbb{R}_+ \rightarrow \mathcal{S}_{\text{LP}}$  and  $z \geq 0$ , there exists a weak solution to the SDE*

$$(4.21) \quad Z_t = z + \int_0^t b(\xi^0(Z_s)) ds + \int_0^t \sigma(\xi^0(Z_s)) dB_s + L_t,$$

where  $Z_t \geq 0$ ,  $L$  has sample paths in  $C_{\mathbb{R}}^{0,+}[0, \infty)$ , and  $\int_{[0, \infty)} Z_t dL_t = 0$ .

2. *Assume that a classical solution  $u$  to (2.14)–(2.15) exists. Then letting*

$$(4.22) \quad \xi^u(z) = \sum_{m \leq M} \mathbb{1}_{A^m}(z) \xi^{*,m}, \quad \text{where}$$

$$A^m = A_u^m = \{z \in \mathbb{R}_+ : \mathbb{H}(u'(z), u''(z)) = \bar{\mathbb{H}}(u'(z), u''(z), \xi^{*,m})\} \setminus \cup_{m' < m} A^{m'},$$

*there exists a weak solution to (4.21) in the special case  $\xi^0 = \xi^u$ .*

In the literature on 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 bounded measurable map  $\xi^0 : \mathbb{R}_+ \rightarrow \mathcal{S}_{\text{LP}}$ . Thus part 1 of the lemma above shows that a policy  $\xi^0$  induces an admissible control system for the WCP, and part 2 provides a candidate optimal policy.

PROOF. 1. We appeal to Th. 5.5.15 and Rem. 5.5.19 in [17], according to which a weak solution exists to the SDE

$$(4.23) \quad Y_t = y + \int_0^t \tilde{b}(Y_s) ds + \int_0^t \tilde{\sigma}(Y_s) dB_s$$

on the real line, where  $\tilde{b}, \tilde{\sigma} : \mathbb{R}_+ \rightarrow \mathbb{R}$  are measurable,  $\tilde{b}$  bounded, and  $\tilde{\sigma}^2$  bounded away from zero and infinity. In particular, this is true for  $\tilde{b}$  and  $\tilde{\sigma}$  given by  $\tilde{b}(x) = \text{sgn}(x)b(\xi^0(|x|))$ ,

$\tilde{\sigma}(x) = \text{sgn}(x)\sigma(\xi^0(|x|))$ ,  $x \in \mathbb{R}$  and the initial condition  $y = z \geq 0$ , where we define  $\text{sgn}(x) = 1$  if  $x > 0$  and  $\text{sgn}(x) = -1$  if  $x \leq 0$ . These coefficients are indeed bounded because the set  $\mathcal{S}_{\text{LP}}$  is bounded, and moreover  $\tilde{\sigma}^2$  is bounded away from zero because  $\sigma_m > 0$ ,  $1 \leq m \leq M$ . Let then  $(Y, B)$  be a corresponding solution to (4.23). Then by Tanaka's formula ([25] Theorem VI.1.2, p. 222) there exists a process  $L^0$  with sample paths in  $C_{\mathbb{R}}^{0,+}[0, \infty)$  such that

$$|Y_t| = y + \int_0^t \text{sgn}(Y_s)^2 b(\xi^0(|Y_s|)) ds + \int_0^t \text{sgn}(Y_s)^2 \sigma(\xi^0(|Y_s|)) dB_s + L_t^0,$$

showing that  $(Z, L^0, B)$  is a solution to (4.21), where  $Z = |Y|$ . Proposition VI.1.3 on p. 222 of [25] justifies that  $L^0$  satisfies  $\int_{[0, \infty)} Z_t dL_t^0 = 0$ .

2. One only needs to show that  $A^m$  are Borel measurable. To this end, note that the set of  $(v_1, v_2)$  for which  $\mathbb{H}(v_1, v_2) = \mathbb{H}(v_1, v_2, \xi^{*,m})$  is Borel. Since  $u$  is  $C^2$ , so are the sets used in the recursive construction of  $A^m$ , as the inverse images of the above sets under a continuous map.  $\square$

#### 4.3. Classical solution and WCP optimality.

**PROOF OF PROPOSITION 2.5.** The proof is based on results from the PDE literature regarding the existence of classical solutions to (2.14) on a bounded domain with Dirichlet boundary conditions.

Step 1: Existence of classical solutions on a bounded domain. Consider

$$(4.24) \quad \mathbb{H}(u'(z), u''(z)) + z - \gamma u(z) = 0, \quad z \in (0, z_0),$$

with boundary conditions  $u(0) = r_0$ ,  $u(z_0) = r_1$ , where  $z_0 \in (0, \infty)$  and  $r_i \in \mathbb{R}$  are constants. We rely on [10, Theorem 17.18]. To apply this result it suffices to verify conditions [10, (17.62)] (see the discussion on the Bellman equation that follows the statement of [10, Theorem 17.18]). To this end, note that uniform ellipticity holds in our setting because  $\sigma_m > 0$  for all  $m$ . The coefficients denoted  $a, b, c$  and  $f$  in [10] correspond here to  $\sigma(\xi)^2/2$ ,  $b(\xi)$ ,  $-\gamma$  (which do not depend on  $z$ ) and  $-z$ . Clearly these functions and their derivatives w.r.t.  $z$  are bounded on the finite interval. It follows that a classical solution to (4.24) with the Dirichlet boundary conditions  $r_0, r_1$  uniquely exists, denoted in what follows by  $u^{(z_0, r_0, r_1)}$ .

Step 2: A control problem on a bounded domain. Fix  $z_0$  and let  $v = u^{(z_0, V_{\text{WCP}}(0), V_{\text{WCP}}(z_0))}$ . Consider the control problem up to exiting the domain  $(0, z_0)$ , defined analogously to (2.13) by

$$J^{(z_0)}(z, \mathfrak{S}) = \mathbb{E} \left[ \int_0^\tau e^{-\gamma t} Z_t dt + e^{-\gamma \tau} V_{\text{WCP}}(Z_\tau) \right],$$

$$V^{(z_0)}(z) = \inf \{ J^{(z_0)}(z, \mathfrak{S}) : \mathfrak{S} \in \mathcal{A}_{\text{WCP}}(z) \},$$

for  $z \in [0, z_0]$  and  $\tau = \inf \{ t : Z_t \in \{0, z_0\} \}$  (on the event  $\tau = \infty$ , the term  $e^{-\gamma \tau} V_{\text{WCP}}(Z_\tau)$  is defined as 0). Given Step 1 and the existence of a weak solution to (4.21)–(4.22) stated in Lemma 4.3, an argument similar to the one provided in the verification theorem [9, Th. IV.5.1] shows that  $v(z) = V^{(z_0)}(z)$  for  $z \in [0, z_0]$ . We omit the details.

Step 3:  $V_{\text{WCP}}$  is a solution to (2.14)–(2.15). By the dynamic programming principle (see e.g. [18, Theorem 6, page 132] with  $r_t = 0$  and  $\varphi_t = \gamma t$ ), we have  $V_{\text{WCP}}(z) = V^{(z_0)}(z)$  for all  $z \in [0, z_0]$ . Combining with the last step,  $V_{\text{WCP}}(z) = v(z)$  for all  $z \in [0, z_0]$ . Since  $z_0$  is arbitrary, this shows that  $V_{\text{WCP}}$  is a classical solution to (2.14) on all of  $[0, \infty)$ .

Next, by the boundedness of  $b$  and  $\sigma$  in (2.12), it is easy to see that  $\mathbb{E} Z_t \leq c(z+t)$  for some constant  $c$  which does not depend on the control system. Hence by (2.13), we have  $0 \leq V_{\text{WCP}}(z) \leq c(z+1)$ .

As for the boundary condition  $V'_{\text{WCP}}(0) = 0$ , it is a standard fact that this follows from the reflection at the origin. For completeness we provide a proof. Given  $z > 0$  and  $\varepsilon > 0$  let  $\mathfrak{G} = (\Omega', \mathcal{F}', (\mathcal{F}'_t)_{t \geq 0}, \mathbb{P}', B, \Xi, Z, L) \in \mathcal{A}_{\text{WCP}}(z)$  be such that

$$J_{\text{WCP}}(z, \mathfrak{G}) < V_{\text{WCP}}(z) + \varepsilon.$$

By (2.12), one has  $(Z, L) = \Gamma(z + K)$ , where  $K_t = \int_0^t b(\Xi_s) ds + \int_0^t \sigma(\Xi_s) dB_s$ . If we define  $\tilde{\mathfrak{G}}$  using the same probability space and processes  $B$  and  $\Xi$ , but with  $(\tilde{Z}, \tilde{L}) = \Gamma(K)$ , then  $\tilde{\mathfrak{G}} \in \mathcal{A}_{\text{WCP}}(0)$ . Moreover,  $\tilde{Z}_t \leq Z_t$  for all  $t$ . As a result,

$$V_{\text{WCP}}(0) \leq J_{\text{WCP}}(0, \tilde{\mathfrak{G}}) \leq J_{\text{WCP}}(z, \mathfrak{G}) \leq V_{\text{WCP}}(z) + \varepsilon.$$

Since  $\varepsilon$  is arbitrary it can be dropped.

Similarly, let  $\mathfrak{G} = (\Omega', \mathcal{F}', (\mathcal{F}'_t)_{t \geq 0}, \mathbb{P}', B, \Xi, Z, L) \in \mathcal{A}_{\text{WCP}}(0)$  be such that

$$J_{\text{WCP}}(0, \mathfrak{G}) < V_{\text{WCP}}(0) + \varepsilon.$$

Then (with notation as before)  $(Z, L) = \Gamma(K)$ , and defining  $(\tilde{Z}, \tilde{L}) = \Gamma(z + K)$  one has  $\tilde{\mathfrak{G}} \in \mathcal{A}_{\text{WCP}}(z)$ . Denoting  $\Delta_t = \tilde{Z}_t - Z_t$ ,

$$V_{\text{WCP}}(z) - V_{\text{WCP}}(0) \leq J_{\text{WCP}}(z, \tilde{\mathfrak{G}}) - J_{\text{WCP}}(0, \mathfrak{G}) + \varepsilon \leq \mathbb{E} \int_0^\infty e^{-\gamma t} \Delta_t dt + \varepsilon.$$

Let  $\tau = \inf\{t : \tilde{Z}_t = 0\} = \inf\{t : K_t = -z\}$ . Then it is easy to see by the definition of  $\Gamma$  that  $\Delta_t \leq z$  for all  $t$ , while  $\Delta_t = 0$  for all  $t \geq \tau$ . Hence

$$(4.25) \quad 0 \leq V_{\text{WCP}}(z) - V_{\text{WCP}}(0) \leq z \mathbb{E} \int_0^\tau e^{-\gamma t} dt + \varepsilon \leq z \mathbb{E}[\tau \wedge c_0] + \varepsilon,$$

where  $c_0 = \gamma^{-1}$ . For the martingale  $\hat{K}_t = \int_0^t \sigma(\Xi_s) dB_s$  one has  $\langle \hat{K} \rangle_t = \int_0^t \sigma^2(\Xi_s) ds$ , and because  $\sigma$  is bounded away from 0 and  $\infty$ , one has  $0 < c_1 \leq (t_2 - t_1)^{-1} (\langle \hat{K} \rangle_{t_2} - \langle \hat{K} \rangle_{t_1}) \leq c_2 < \infty$ , and thus the map  $t \mapsto \langle \hat{K} \rangle_t$  admits a strictly increasing, Lipschitz inverse  $T$ . By [17, Theorem 3.4.6],  $Y_t := \hat{K}_{T(t)}$  is a SBM. By the boundedness of  $b$ ,  $K_t \leq \hat{K}_t + ct$ . Hence for any  $t > 0$ ,

$$\begin{aligned} \mathbb{P}'(\tau > t) &= \mathbb{P}'(K_s > -z \text{ for all } s \leq t) \\ &\leq \mathbb{P}'(\hat{K}_s + cs > -z \text{ for all } s \leq t) \\ &= \mathbb{P}'(Y_s + cT(s) > -z \text{ for all } s \leq \langle \hat{K} \rangle_t) \\ &\leq \mathbb{P}'(Y_s + cc_1^{-1}s > -z \text{ for all } s \leq c_1 t). \end{aligned}$$

This bound depends only on  $t$  and  $z$ , and converges to zero as  $z \rightarrow 0$ . Hence it follows that  $\mathbb{E}[\tau \wedge c_0] \leq \kappa(z)$  for some function  $\kappa$  which satisfies  $\lim_{z \downarrow 0} \kappa(z) = 0$ . Using this in (4.25) shows that  $z^{-1} |V_{\text{WCP}}(z) - V_{\text{WCP}}(0)| \leq \kappa(z) + z^{-1} \varepsilon$ . Sending  $\varepsilon \downarrow 0$  and then  $z \downarrow 0$  shows that  $V'_{\text{WCP}}(0)$  exists and equals 0. This completes the proof that  $V_{\text{WCP}}$  is a classical solution to (2.14)–(2.15).

Step 4: uniqueness on all of  $[0, \infty)$ . It remains to show that if  $u$  is any classical solution to (2.14)–(2.15) then  $u = V_{\text{WCP}}$ . This is also done in the spirit of [9, Theorem IV.5.1], but here the details are more subtle due to the boundary condition at infinity, and we provide the proof.

Let  $\mathfrak{G} = (\Omega', \mathcal{F}', (\mathcal{F}'_t)_{t \geq 0}, \mathbb{P}', B, \Xi, Z, L)$  be an admissible control system. It will be shown that  $J_{\text{WCP}}(z, \mathfrak{G}) \geq u(z)$ . Since  $u$  is  $C^2$ , a use of Itô's formula gives

$$(4.26) \quad \begin{aligned} e^{-\gamma t} u(Z_t) &= u(z) + \int_0^t e^{-\gamma s} u'(Z_s) [b(\Xi_s) ds + \sigma(\Xi_s) dB_s + dL_s] - \gamma \int_0^t e^{-\gamma s} u(Z_s) ds \\ &\quad + \frac{1}{2} \int_0^t e^{-\gamma s} u''(Z_s) \sigma^2(\Xi(s)) ds. \end{aligned}$$

Because  $\int_0^\infty Z_t dL_t = 0$  and  $u'(0) = 0$ , we have

$$\int_0^t e^{-\gamma s} u'(Z_s) dL_s = \int_0^t e^{-\gamma s} u'(Z_s) \mathbf{1}_{\{Z_s=0\}} dL_s = 0.$$

By equation (2.14), for any member  $\xi$  of  $\mathcal{S}_{LP}$  we have

$$(4.27) \quad \frac{1}{2} \sigma(\xi)^2 u''(y) + b(\xi) u'(y) + y - \gamma u(y) \geq 0, \quad y \in \mathbb{R}_+.$$

Substituting in (4.26), denoting  $\tilde{M}_t = \int_0^t e^{-\gamma s} u'(Z_s) \sigma(\Xi_s) dB_s$ ,

$$(4.28) \quad e^{-\gamma t} u(Z_t) \geq u(z) + \tilde{M}_t - \int_0^t e^{-\gamma s} Z_s ds.$$

We argue that  $\tilde{M}_t$  is in  $L^1$  for any  $t$ , by which the local martingale  $\tilde{M}_t$  is a martingale. First we show that  $\tilde{M}_t^+$  is in  $L^1$  for any  $t$ . In view of (4.28), to bound  $\mathbb{E}[\tilde{M}_t^+]$ , it suffices to estimate  $\mathbb{E}[u(Z_t)]$  and  $\mathbb{E}[\int_0^t e^{-\gamma s} Z_s ds]$ . To bound  $Z_t$ , fix  $t \geq 0$  and denote  $\tau = \sup\{s \in [0, t] : Z_s = 0\}$ . We distinguish two cases. Case (i):  $\tau \in [0, t]$  and  $Z_\tau = 0$ . In this case

$$Z_t = \int_\tau^t b(\Xi_s) ds + \int_\tau^t \sigma(\Xi_s) dB_s,$$

because  $L_t = L_\tau$ . Case (ii):  $Z_s > 0$  for all  $s \in [0, t]$ . Then

$$Z_t = z + \int_0^t b(\Xi_s) ds + \int_0^t \sigma(\Xi_s) dB_s.$$

Since  $Z_t \geq 0$ , in both cases one has

$$\|Z\|_t \leq c + ct + 2 \sup_{\theta \leq t} \left| \int_0^\theta \sigma(\Xi_s) dB_s \right|$$

for a constant  $c$  that does not depend on  $t$ . Because  $\sigma$  is bounded, the second moment of the last term is bounded by  $ct$ , which gives  $\mathbb{E}\|Z\|_t \leq c(1+t)$ . In particular,  $\mathbb{E}[\int_0^t e^{-\gamma s} Z_s ds] < \infty$ . Moreover, by (2.15),

$$(4.29) \quad \mathbb{E}[\|u(Z_\cdot)\|_t] \leq c(1+t).$$

In view of (4.28), this shows not only that  $\mathbb{E}[\tilde{M}_t^+] < \infty$  but also  $\mathbb{E}[\|\tilde{M}^+\|_t] < \infty$ . Next, let  $\tau^N$  be a sequence of stopping times such that  $\tau^N \rightarrow +\infty$  a.s., and  $\tilde{M}_{t \wedge \tau^N}$  is a martingale for every  $N$ . Then  $\mathbb{E}[\tilde{M}_{t \wedge \tau^N}^+] = \mathbb{E}[\tilde{M}_{t \wedge \tau^N}^-]$ . Using this identity and Fatou's Lemma,

$$\mathbb{E}[\tilde{M}_t^-] \leq \liminf_{N \rightarrow +\infty} \mathbb{E}[\tilde{M}_{t \wedge \tau^N}^-] = \liminf_{N \rightarrow +\infty} \mathbb{E}[\tilde{M}_{t \wedge \tau^N}^+] \leq \mathbb{E}[\|\tilde{M}^+\|_t] < \infty.$$

This shows that  $\tilde{M}_t$  is a martingale.

Taking the mean and rearranging in (4.28), for any  $t$ ,

$$(4.30) \quad J_{WCP}(z, \mathfrak{S}) \geq \mathbb{E} \int_0^t e^{-\gamma s} Z_s ds \geq u(z) - e^{-\gamma t} \mathbb{E}[u(Z_t)].$$

In view of (4.29), letting  $t \rightarrow +\infty$  gives  $J_{WCP}(z, \mathfrak{S}) \geq u(z)$ .

To show that  $u = V_{WCP}$  it remains to find  $\mathfrak{S}$  for which  $J_{WCP}(z, \mathfrak{S}) = u(z)$ . To this end, let  $\xi^u$  be as in (4.22). Let an initial condition  $z$  be given. Form a control system  $\mathfrak{S}^u = (\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P}, \Xi, Z, L)$  by letting  $(Z, B, L)$  be a tuple satisfying (4.21) with  $\xi^0 = \xi^u$  and initial condition  $z$ ,  $\Xi_t = \xi^u(Z_t)$  and  $\mathcal{F}_t = \sigma\{(Z_s, L_s, B_s) : s \in [0, t]\}$ ,  $t \geq 0$ . Here we have

assumed, without loss of generality, that the weak solution to (4.21) is defined on our original probability space. Then using  $\Xi_t = \xi^u(Z_t)$  in (2.14) gives

$$b(\Xi_s)u'(Z_s) + \frac{1}{2}\sigma^2(\Xi_s)u''(Z_s) + Z_s - \gamma u(Z_s) = 0.$$

Hence by (4.26), (4.28) holds with equality. Thus  $\mathbb{E} \int_0^t e^{-\gamma s} Z_s ds = u(z) - e^{-\gamma t} \mathbb{E}[u(Z_t)]$ . To take the  $t \rightarrow \infty$  limit, use monotone convergence on the LHS and the bound (4.29) on the RHS. This shows  $J_{\text{WCP}}(z, \mathfrak{S}^u) = u(z)$  and completes the proof.  $\square$

As a corollary of the proposition and its proof we obtain a construction of an optimal solution to the WCP.

**COROLLARY 4.4.** *Denote  $u = V_{\text{WCP}}$ . Let  $\xi^u$  be as in (4.22). Let an initial condition  $z$  be given. Form an admissible control system  $\mathfrak{S}^u = (\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P}, \Xi, Z, L)$  by letting  $(Z, B, L)$  be a tuple satisfying (4.21) with  $\xi^0 = \xi^u$  and initial condition  $z$ ,  $\Xi_t = \xi^u(Z_t)$  and  $\mathcal{F}_t = \sigma\{(Z_s, L_s, B_s) : s \in [0, t]\}$ ,  $t \geq 0$ . Then  $\mathfrak{S}^u$  is optimal for the WCP with initial condition  $z$ .*

The next result is concerned with a special case in which there exists a dominant mode, showing that the WCP's solution is to always use this mode.

**PROPOSITION 4.5.** *Assume there exists a mode, say,  $\xi^{*,1}$ , such that for all  $2 \leq m \leq M$  one has  $b(\xi^{*,1}) \leq b(\xi^{*,m})$  and  $\sigma(\xi^{*,1}) \leq \sigma(\xi^{*,m})$ . Then it is optimal to always use this mode. In other words, the statement of Corollary 4.4 holds with  $\xi^0(z) = \xi^u(z) = \xi^{*,1}$ ,  $z \in \mathbb{R}_+$ .*

**PROOF.** In view of Proposition 2.5 and Corollary 4.4, it suffices to find a classical solution  $u$  to (2.14)–(2.15) which satisfies  $A_u^1 = \mathbb{R}_+$ , where we use the notation of Lemma 4.3. We rely on the explicit expression for the cost associated with the policy that always uses the mode  $\xi^{*,1}$ , (see equation (46) on p. 104 of [27]). This function is given by

$$u(z) = c_1 + c_2 z + \frac{c_2}{c_3} e^{-c_3 z},$$

with explicit constants  $c_i \in (0, \infty)$ . In particular,  $u'(z) \geq 0$  and  $u''(z) \geq 0$  for all  $z$ , and  $\bar{\mathbb{H}}(u'(z), u''(z), \xi^{*,1}) + z - \gamma u(z) = 0$  for all  $z$ . Using the assumption on  $b(\xi^{*,1})$  and  $\sigma(\xi^{*,1})$  and the positivity of  $u'$  and  $u''$  shows that

$$\bar{\mathbb{H}}(u'(z), u''(z), \xi^{*,1}) = \min_{m \leq M} \bar{\mathbb{H}}(u'(z), u''(z), \xi^{*,m}) = \mathbb{H}(u'(z), u''(z)).$$

As a result,  $u$  is a solution to (2.14). Clearly, it also satisfies the boundary conditions (2.15). This shows that  $A_u^1$  of (4.22) is all of  $\mathbb{R}_+$ , with  $u$  a classical solution to (2.14)–(2.15).  $\square$

When it is optimal to use only one mode, one can readily adapt results that were proved when the solution to the LP is unique. The policy introduced in [3] Section 5 is based on a solution to (2.7), as long as the graph of basic activities induced by the solution is a tree. For that reason, we will speak of the ‘‘Bell and Williams policy based on  $\xi$ ’’ to describe a policy constructed in the same way as in [3] Section 5 as if  $\xi$  was the unique solution to (2.7). Note that  $i^*$  from their paper is here denoted  $q$  and we use  $\xi^{*,1}$  as  $x^*$ .

**COROLLARY 4.6.** *Keep the assumption from Proposition 4.5. Assume further that  $\xi^{*,1}$  is non-degenerate, and let Assumption 3.2 from [3] for the arrival and service distributions hold. Then the Bell and Williams policy based on  $\xi^{*,1}$  is AO (the lower bound from Theorem 2.6 is tight).*



We give here proof based on the results and proofs from [3].

PROOF. Under condition 2. of Assumption 2.2, as argued in [15] Section 2, any non degenerate extreme point of  $\mathcal{S}_{LP}$  has  $I + K - 1$  basic activities. Construct a restricted PSS by removing all activities associated with non-basic activities of  $\xi^{*,1}$ . Then  $\xi^{*,1}$  is the unique solution to (2.7) for the restricted PSS. The solution to the original dual is also a solution of the dual for the restricted PSS. With our setup, Assumptions 3.1, 3.2, 3.3, and 3.6 from [3] hold in the restricted PSS. Since Theorem 4.3(iii) of [3] holds, the solution to the dual in the restricted PSS is also unique and Assumption 4.4 from [3] holds. Thus the restricted PSS continues to satisfy our Assumption 2.2, and the graph of basic activities is a tree.

Under the assumption of Corollary 4.4,  $V_{WCP}$  for both the original and restricted PSS is given by

$$\mathbb{E} \int_{[0,\infty)} e^{-\gamma t} Z_t dt,$$

with

$$Z_t = z + tb(\xi^{*,1}) + \sigma(\xi^{*,1})B_t + L_t.$$

This is also the lower bound in Bell and Williams [3]. The results in [3] state that their policy is AO for the restricted PSS, so it must also be optimal for the original problem since it achieves the best possible cost.  $\square$

EXAMPLE (A) - CONTINUED. We go back to Example (A), now considering it with second order data. Assume that the squared coefficients of variation  $C_{A_i}^2$  and  $C_{S_j}^2$  are all 1 except  $C_{S_1}^2 = 4$ . Set both  $\hat{\lambda}_i$  to 0. As for  $\hat{\mu}_j$ , consider two cases. In case (A1), all  $\hat{\mu}_j$  are set to 0. In case (A2), they are all 0 except  $\hat{\mu}_1 = 1$ .

For each of the modes we can compute the drift and squared diffusion coefficients via (2.10)–(2.11). In case (A1) we obtain

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

In case (A2),

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

It is seen that in case (A1),  $b_2 \leq b_1$  and  $\sigma_2 \leq \sigma_1$ . By Proposition 4.5, it is optimal to always use mode 2. Moreover, Corollary 4.6 applies with mode 2.

In case (A2),  $b_1 < b_2$  and  $\sigma_1 > \sigma_2$ , and so none of the modes dominates. As already mentioned, in this case the HJB equation has been solved in [27], showing that both modes must be used. In particular, there exists a switching point  $z^*$  such that it is optimal to use one mode when the state process  $Z$  is below  $z^*$ , and the other when it is above. In the case considered here, mode 2 is used below  $z^*$  and mode 1 above it. In particular, the simplification provided by Corollary 4.6 does not apply. Instead, a certain switching mechanism which mimics the WCP solution is required for the PSS model itself in order to attain AO. We refer to [1] for more information.

EXAMPLES (C) AND (E) - CONTINUED. Recall that in Example (C), because of the LP degeneracy, the results of [1] do not apply. However, there is a case where we can still state an AO result. This is the case where mode 2 (the non-degenerate mode) is dominant in the sense of Proposition 4.5, i.e.,  $b(\xi^{*,2}) \leq b(\xi^{*,1})$  and  $\sigma(\xi^{*,2}) \leq \sigma(\xi^{*,1})$ . In this case it is optimal to only use this mode, and we can apply Corollary 4.6. In other words, under the assumptions of Proposition 4.5, the Bell and Williams policy based on  $\xi^{*,2}$  is in this case AO. A similar remark holds for Example (E) with mode  $\xi^{*,3}$ , which is the only non-degenerate mode in this case.



**5. An asymptotic lower bound.** The goal of this section is to prove Theorem 2.6.

5.1. *Lemmas and proof of Theorem 2.6.* Theorem 2.6 will be proved once it is shown that for every sequence  $\{T^n : n \in \mathbb{N}\}$ , with  $T^n \in \mathcal{A}^n$ , one has  $\liminf_n \hat{J}^n(T^n) \geq V_0$ . A sufficient condition is that for every subsequence along which  $\limsup_n \hat{J}^n(T^n) < \infty$ , one has  $\liminf_n \hat{J}^n(T^n) \geq V_0$ . Let  $\text{Lip}(\mathcal{S}_{\text{LP}})$  denote the collection of paths  $\theta \in C(\mathbb{R}_+ : \mathbb{R}^{\mathcal{J}})$  satisfying  $(t-s)^{-1}(\theta_t - \theta_s) \in \mathcal{S}_{\text{LP}}$  for all  $0 \leq s < t$ , and  $\theta_0 = 0$ . A preliminary lemma, proved in §5.2, is as follows.

LEMMA 5.1. *Let  $\{T^n\}$ ,  $T^n \in \mathcal{A}^n$  be a sequence of admissible controls for the QCP and consider a subsequence (relabelled  $\{T^n\}$ ) along which  $\limsup_{n \rightarrow +\infty} \hat{J}^n(T^n) < +\infty$ . Then the sequence of processes  $T^n$  is C-tight. Consider a further subsequence along which  $(\hat{A}^n, \hat{S}^n, T^n) \Rightarrow (A, S, T)$ . Then  $T$  has sample paths in  $\text{Lip}(\mathcal{S}_{\text{LP}})$  a.s., and if  $\{\mathcal{F}'_t\}_{t \geq 0}$  is a filtration to which  $T$  is adapted then  $T$  is given by  $\int_0^\cdot \Xi_s ds$  for some  $\Xi \in \mathfrak{X}(\{\mathcal{F}'_t\})$ .*

Theorem 2.6 will follow once it is established that for every subsequence as in Lemma 5.1 there exists an admissible system for the WCP,  $\mathfrak{S}$ , such that

$$(5.1) \quad \liminf_n \hat{J}^n(T^n) \geq h_q(y_q^*)^{-1} J_{\text{WCP}}(\mathfrak{S}).$$

We now outline how  $\mathfrak{S}$  is constructed from the tuple  $(A, S, T)$ . The tuple and subsequence along which  $(\hat{A}^n, \hat{S}^n, T^n) \Rightarrow (A, S, T)$  are fixed in the rest of this section.

The admissible system can be constructed, without loss of generality, on the original space  $(\Omega, \mathcal{F}, \mathbb{P})$ . The remaining elements of  $\mathfrak{S}$  to be constructed are a filtration  $\{\mathcal{H}_t\}_{t \geq 0}$  and processes  $(B, Z, L, \Xi)$  satisfying the definition of an admissible system. The pair  $(Z, L)$  is not obtained simply as a limit of  $(\hat{W}^n, \hat{L}^n)$ , because for a general sequence of policies for the QCP it is not true that such a limit satisfies  $\int Z dL = 0$ . Instead, one takes the limit in (4.13) and obtains  $F$  defined in (4.17) as the limit of  $\hat{F}^n$ , and sets  $(Z, L) = \Gamma[F]$ . Later, when it is established that  $\mathfrak{S}$  is an admissible system, the lower bound (5.1) is obtained upon taking limits in the inequality (4.16).

If one can construct a filtration  $\{\mathcal{H}_t\}_{t \geq 0}$  to which  $(T, F)$  (and consequently  $(Z, L)$ ) is adapted and on which the processes  $A_i$  and  $D_j := S_j \circ T_j$  are martingales (compare with the definition of  $D^n$  in (2.2)), then by Lemma 5.1 there exists  $\Xi \in \mathfrak{X}(\{\mathcal{H}_t\}_{t \geq 0})$ , and moreover  $M$  defined in (4.19) is also a martingale. Letting  $B = \int_0^\cdot \sigma(\Xi_s)^{-1} dM_s$ , the calculations in §4.1.2 show that  $B$  is a SBM and a martingale w.r.t. this filtration, and the construction of  $\mathfrak{S}$  is complete. (The precise details appear in Lemma 5.5.)

However, the task of constructing the filtration to which  $(T, F)$  are adapted and on which  $D_j$  are martingales is surprisingly more difficult than when the limiting process  $T$  is deterministic (which is the case in all earlier work on heavy traffic asymptotic control problems). When  $T$  is deterministic there is no question of it being adapted, and moreover  $D_j$  are themselves independent BMs. Thus on the filtration generated by the mutually independent  $A_i, D_j$ , it is immediate that these processes are martingales, and so is  $M$ . Obviously, this direct approach is not fruitful when  $T$  is random.

The approach we develop is based on multi-parameter filtrations, specifically, on a theorem due to Kurtz [20] (presented as Theorem 5.2 below) that translates martingales on a multi-parameter filtration to ones on a single-parameter filtration. (Multi-parameter filtrations were used in the related papers [2] and [3] for other purposes; see e.g. §7 of [2]). In preparation to use this tool for our needs we proceed in two main steps.

The first step is carried out in Lemma 5.3, where it is argued that, for every  $t$  and  $n$ , a certain transformation of the RV  $T_t^n$  (see (5.4)) constitutes a multi-parameter stopping time

on the multi-parameter filtration generated by  $\mathcal{Y}$  and the  $I + J$  processes  $A_i^n, S_j^n$  (specifically, (5.3)). This formal statement expresses the intuitively apparent fact that in order to determine whether the cumulative busyness until time  $t$  in the activities satisfy  $T_j^n(t) \leq u_j$ , it is enough to know the collection  $S_j^n(v_j)$  for  $v_j \leq u_j$  and  $A_i^n(s)$  for  $s \leq t$  (in addition to  $\mathcal{Y}$ ).

It will simplify notation if we use, in this section only, a common notation for the processes  $A^n$  and  $S^n$ , and, similarly, for  $A$  and  $S$ . Recall that activities are indexed by  $\mathcal{J} \subset \mathcal{I} \times \mathcal{K}$ . Introduce  $\mathcal{J}^0 := \{(i, 0), i \in \mathcal{I}\} \cup \mathcal{J}$  and set  $S_{i0}^n = A_i^n$  for  $i \in \mathcal{I}$ . Similarly,  $S_{i0} = A_i, i \in \mathcal{I}$ . This way we have a single index set,  $\mathcal{J}^0$ , where the indices  $j = (i, 0)$  are associated with the arrival processes and  $j = (i, k) \in \mathcal{J}$ , as before, with potential service processes.

The second main step, performed in Lemma 5.4(2), is showing that the  $I + J$ -dimensional limit process  $(A_i, S_j)$  in our original notation, and  $(S_j)$  in our notation for this section, constitutes a multi-parameter martingale on a certain filtration defined in terms of the limit processes  $(S, T)$  (specifically, (5.6)). This is crucially based on the first step. Once this step is established, Theorem 5.2 can be applied, and in several further steps, namely Lemma 5.4(3–5) and Lemma 5.5, the construction of  $\mathfrak{S}$  is completed by the approach outlined above.

Let  $\mathcal{U} = \mathbb{R}_+^{\mathcal{J}^0}$ . In this section, the letter  $u$  is used as a generic member of  $\mathcal{U}$  (rather than a solution to (2.14)). Given  $u \in \mathcal{U}$ , let  $S^n(u)$  denote  $(S_j^n(u_j))$ . Define the partial order  $\preceq$  on  $\mathcal{U}$  by  $u \preceq v$  iff  $u_j \leq v_j$  for all  $j \in \mathcal{J}^0$ . We refer to [20] for more details on multi-parameter filtrations, martingales and stopping times (with the partial order given above). The following is Theorem 6.3(a) in [20].

**THEOREM 5.2 (Kurtz).** *Let  $\mathcal{G}_u$  be a filtration indexed by  $\mathcal{U}$ , and let  $\tilde{S}_j, j \in \mathcal{J}^0$  be a finite collection of independent SBM for which*

$$(5.2) \quad \Phi_\theta^{\mathcal{M}}(u) = \prod_{j \in \mathcal{M}} \exp\{\mathbf{i}\theta_j \tilde{S}_j(u_j) + \frac{1}{2}\theta_j^2 u_j\}$$

*is a  $\{\mathcal{G}_u\}$ -martingale for all  $\mathcal{M} \subset \mathcal{J}^0$ . Let  $\tilde{T}(t), t \in [0, \infty)$  be  $\{\mathcal{G}_u\}$ -stopping times such that  $\tilde{T}$  has sample paths in  $C_{\mathbb{R}}^+[0, \infty)^{I+J}$  a.s. Then, for each  $\mathcal{M} \subset \mathcal{J}^0$ ,  $\prod_{j \in \mathcal{M}} \tilde{S}_j(\tilde{T}_j(t))$  is a continuous local martingale with respect to  $\mathcal{H}_t := \mathcal{G}_{\tilde{T}(t)}$ .*

For  $n \in \mathbb{N}$ , denote by  $\{\mathcal{G}_u^n\}_{u \in \mathcal{U}}$  the multi-parameter filtration

$$(5.3) \quad \mathcal{G}_u^n = \sigma\{\mathcal{Y}, S^n(v) : v \preceq u\}, \quad u \in \mathcal{U}.$$

It will further be convenient to denote, for  $T^n \in \mathcal{A}^n$ , by  $\bar{T}^n$  the  $I + J$  component process

$$(5.4) \quad \bar{T}^n(t) = (t, \dots, t, T_t^n) = (\bar{T}_j^n(t) : j \in \mathcal{J}^0), \quad \text{where } \bar{T}_{i0}^n(t) := t, i \in \mathcal{I}.$$

A similar notation will be used for an  $I + J$  component process  $\bar{T}$ , where  $\bar{T}_t = (t, \dots, t, T_t)$  based on the process  $T$ . Denote  $D_j = S_j \circ \bar{T}_j$ , for all  $j \in \mathcal{J}^0$  and not just those in  $\mathcal{J}$ . The proofs of the next two lemmas appear in §5.2.

**LEMMA 5.3.** *Fix  $n$  and  $T^n \in \mathcal{A}^n$ . Then for every  $t \geq 0$ ,  $\bar{T}_t^n$  is a  $\{\mathcal{G}_u^n\}_{u \in \mathbb{R}_+^{I+J}}$ -stopping time.*

Next we define a multi-parameter filtration based on the processes  $(S, T)$ . Denote by  $\mathbf{e} \in \mathcal{U}$  the vector whose all components are 1. For  $u \in \mathcal{U}$ , let  $\delta^* = \delta^*(u) = u + \mathbf{e}$  and define the function  $\psi^{(u)} : \mathcal{U} \rightarrow \mathcal{U}$  as

$$(5.5) \quad \psi^{(u)}(v) = \begin{cases} v, & v \preceq u, \\ \delta^*, & \text{otherwise.} \end{cases}$$

The choice of the constant  $\delta^*$  is immaterial as long as it lies outside  $\{v : v \preceq u\}$ . Let

$$(5.6) \quad \mathcal{G}_u := \sigma\{S(v), \psi^{(u)}(\bar{T}_t) : v \preceq u, t \geq 0\}, \quad u \in \mathcal{U}.$$

According to the definition of  $\psi^{(u)}$ , the  $\sigma$ -field  $\mathcal{G}_u$  carries the information of whether or not  $\bar{T}_t \preceq u$ , as well as the full information about  $\bar{T}_t$  provided it is  $\preceq u$ . It is immediate that, for every  $t$ ,  $\bar{T}_t$  is a stopping time on this filtration. Hence

$$\mathcal{H}_t = \mathcal{G}_{\bar{T}_t}, \quad t \geq 0,$$

is a well-defined (single parameter) filtration.

The second statement in the result below, that allows us to use Theorem 5.2, is crucially based on Lemma 5.3. Statements 3–5 then follow mostly from Theorem 5.2. Denote by  $S_j^{\text{unit}} = (S_j^{\text{unit}}, j \in \mathcal{J}^0)$  the SBM (of unit diffusivity)  $\sigma_j^{-1}S_j$ , where  $\sigma_j = \sigma_{A,i}$  for  $j = (i, 0)$ , and  $\sigma_j = \sigma_{S,j}$  for  $j \in \mathcal{J}$ .

- LEMMA 5.4. 1.  $\bar{T}_t$  (equivalently,  $T_t$ ) is  $\mathcal{H}_t$ -adapted.  
 2. For every  $u, v \in \mathcal{U}$ ,  $\mathcal{G}_u \perp\!\!\!\perp S(u+v) - S(u)$ . Consequently,  $\{S(u)\}$  is a multi-parameter martingale on  $\{\mathcal{G}_u\}$ , and the tuple  $(\{\mathcal{G}_u\}, \{S_j^{\text{unit}}\}, \bar{T}_t)$  satisfies the hypotheses of Theorem 5.2.  
 3. The processes  $D_j$ ,  $j \in \mathcal{J}^0$  (i.e.,  $A_i$  and  $S_j \circ T_j$ ,  $i \in \mathcal{I}$ ,  $j \in \mathcal{J}$ ) are  $\mathcal{H}_t$ -martingales.  
 4. For distinct indices  $j, j' \in \mathcal{J}^0$  one has  $\langle D_j, D_{j'} \rangle = 0$ .  
 5. Finally, for any  $j \in \mathcal{J}^0$ ,  $\langle D_j \rangle_t = \sigma_j^2 T_j(t)$ .

We now arrive at the main lemma asserting that an admissible system  $\mathfrak{S}$  exists with the desired properties listed earlier. The proof, presented below, is based on Lemmas 5.1–5.4. The proof of the Theorem 2.6 is presented immediately afterwards, and establishes the fact that  $\mathfrak{S}$  indeed satisfies (5.1). Recall that  $\hat{F}^n$  and  $F$  are defined in (4.13) and (4.17).

LEMMA 5.5. Let the assumptions of Lemma 5.1 hold and let  $\{\mathcal{H}_t\}$  be the filtration constructed above. Then along the aforementioned convergent subsequence, where  $(\hat{A}^n, \hat{S}^n, T^n) \Rightarrow (A, S, T)$ , one has  $(\hat{A}^n, \hat{S}^n, T^n, \hat{F}^n) \Rightarrow (A, S, T, F)$ . Moreover, there exist on  $(\Omega, \mathcal{F})$  processes  $(B, Z, L)$  such that the tuple  $\mathfrak{S} := (\Omega, \mathcal{F}, \{\mathcal{H}_t\}_{t \geq 0}, \mathbb{P}, B, \Xi, Z, L)$  forms an admissible control system for the WCP,  $T = \int_0^\cdot \Xi_s ds$ , and

$$(5.7) \quad \begin{aligned} (Z, L) &= \Gamma[F], \\ F_t &= \sum_i y_i^* \left( A_i(t) + \hat{\lambda}_i t - \sum_{j \in \mathcal{J}_i} S_j(T_j(t)) + \hat{\mu}_j T_j(t) \right) \\ &= \int_0^t b(\Xi_s) ds + \int_0^t \sigma(\Xi_s) dB_s. \end{aligned}$$

REMARK 5.6. Note that it is not claimed that  $\hat{W}^n$  converges (or is tight), nor that if it has a limit  $W$  then  $W = Z$ . Both statements are false in general. The most that can be said at this point is that, in view of (4.15), one has  $Z_t \leq W_t$  for all  $t$  whenever  $W$  is a limit of  $\hat{W}^n$ .

PROOF. Since we have  $(\hat{A}^n, \hat{S}^n, T^n) \Rightarrow (A, S, T)$ , it follows by the lemma on random change of time in [4, page 151] that  $\hat{S}_j^n \circ T_j^n \Rightarrow S_j \circ T_j$ , justifying the convergence  $\hat{F}^n \Rightarrow F$ , first assumed above (4.17).

For the construction of  $\Xi$ , note that by Lemma 5.4,  $T_t$  is  $\mathcal{H}_t$  adapted. Hence by Lemma 5.1, there exists  $\Xi \in \mathfrak{X}(\{\mathcal{H}_t\})$  such that  $T = \int_0^\cdot \Xi_s ds$ .

Next,  $B$  is constructed. By (4.17), (4.19) and the definition of  $b$  in (2.10),

$$F_t = \int_0^t b(\Xi_s) ds + M_t, \quad M_t = \sum_i y_i^* \left( A_i(t) - \sum_{j \in \mathcal{J}_i} S_j(T_j) \right).$$

By Lemma 5.4,  $M$  is an  $\mathcal{H}_t$ -martingale, and

$$\langle M \rangle_t = \int_0^t \sum_i (y_i^*)^2 \left( \sigma_{A,i}^2 + \sum_{j \in \mathcal{J}_i} \sigma_{S,j}^2 \Xi_j(s) \right) ds = \int_0^t \sigma^2(\Xi(s)) ds,$$

where the last equality uses the definition of  $\sigma$  in (2.10). Recalling that  $\sigma$  is bounded away from zero, let  $B$  be the  $\mathcal{H}_t$ -local martingale defined as  $B_t = \int_0^t \sigma(\Xi(s))^{-1} dM_s$ . To compute  $\langle B \rangle_t$ , we use Proposition IV.2.7 of [25] (once with  $K = \sigma(\Xi(\cdot))^{-1}$ ,  $M$  as above and  $K \cdot M$  for  $N$ , and a second time with  $M$  for  $N$ ), giving

$$\langle B \rangle_t = \int_0^t \frac{\sigma^2(\Xi(s))}{\sigma^2(\Xi(s))} ds = t.$$

By Levy's characterization (Theorem IV.3.6 of [25]),  $B$  is a SBM. Moreover, we have the identity  $M_t = \int_0^t \sigma(\Xi(s)) dB_s$  (by Proposition IV.2.4 [25]). By construction,  $F$  and  $B$  are  $\mathcal{H}_t$ -adapted. Setting  $(Z, L) = \Gamma[F]$  we thus obtain (5.7) as well as adaptedness of this pair. This shows that  $\mathfrak{S}$  is admissible, completing the proof.  $\square$

PROOF OF THEOREM 2.6. The result will follow once we show that for every subsequence as in Lemma 5.1, one has

$$(5.8) \quad \liminf_n \hat{J}^n(T^n) \geq h_q(y_q^*)^{-1} J_{\text{WCP}}(\mathfrak{S}),$$

where  $\mathfrak{S}$  is the corresponding admissible system from Lemma 5.5. By (4.14) and (4.15), for any finite  $t_0$ ,

$$\hat{J}^n(T^n) \geq \mathbb{E} \int_0^{t_0} e^{-\gamma t} h(\hat{X}^n(t)) dt \geq h_q(y_q^*)^{-1} \mathbb{E} \int_0^{t_0} e^{-\gamma t} \Gamma_1[\hat{F}^n](t) dt.$$

By the weak convergence  $\hat{F}^n \Rightarrow F$  stated in Lemma 5.5, and the continuity of  $\Gamma_1$  (as a map from  $D_{\mathbb{R}}[0, \infty)$  to itself, equipped with the topology of convergence u.o.c.), it follows that  $\Gamma_1[\hat{F}^n] \Rightarrow \Gamma_1[F] = Z$ , where  $Z$  is a member of the tuple  $\mathfrak{S}$  of Lemma 5.5, and we have used (5.7). Because the convergence is u.o.c. and the time interval is finite, Fatou's lemma is applicable, giving

$$(5.9) \quad \liminf_n \hat{J}^n(T^n) \geq h_q(y_q^*)^{-1} \mathbb{E} \int_0^{t_0} e^{-\gamma t} Z_t dt.$$

By (2.13), as  $t_0 \rightarrow \infty$ , the above expectation converges to  $J_{\text{WCP}}(\mathfrak{S})$ . This gives (5.8) and completes the proof.  $\square$

## 5.2. Proof of lemmas.

PROOF OF LEMMA 5.1. The  $C$ -tightness of the processes  $T^n$  is immediate from the fact that their sample paths are 1-Lipschitz. Next, assume  $\limsup \hat{J}^n(T^n) < \infty$ . Define the processes  $\bar{A}^n, \bar{S}^n$  and  $\bar{X}^n$  by

$$\bar{A}^n(t) = \frac{\check{A}(\lambda^n t)}{n}, \quad \bar{S}^n(t) = \frac{\check{S}(\mu^n t)}{n}, \quad \bar{X}^n(t) = \frac{X^n(t)}{n}.$$

By the LLN for renewal processes (see [4, Sec. 14]),  $(\bar{A}^n, \bar{S}^n) \rightarrow (\bar{A}, \bar{S})$  in probability uniformly over compact time sets, where  $\bar{A}(t) = \lambda t$  and  $\bar{S}(t) = \mu t$ . Consider a subsequence along which  $T^n \Rightarrow T$  and note that  $T$  has 1-Lipschitz sample paths a.s. Then using (2.2), (2.3) and denoting  $\bar{X}_i(t) = \bar{A}_i(t) - \sum_{j \in \mathcal{J}_i} \bar{S}_j(T_j(t))$ , we have  $(\bar{A}^n, \bar{S}^n, T^n, \bar{X}^n) \Rightarrow (\bar{A}, \bar{S}, T, \bar{X})$ . First, it is shown that a.s.,  $\sup_{s < t} \Delta(s, t) \in \mathcal{S}_{\text{LP}}$ , where  $\Delta(s, t) = (t - s)^{-1}(T_t - T_s)$ .

To this end, note that if  $\xi \in \mathbb{R}_+^{\mathcal{J}}$  satisfies  $\sum_{j \in \mathcal{J}_i} \xi_j \mu_j = \lambda_i$  for all  $i$  and  $\sum_{j \in \mathcal{J}^k} \xi_j \leq 1$  for all  $k \in \mathcal{K}$ , then it satisfies the conditions (2.7) with  $\rho = 1$ , thus in view of the EHTC  $\rho^* = 1$ ,  $(\xi, 1)$  must be a solution to the LP. In other words,  $\xi \in \mathcal{S}_{\text{LP}}$ .

To apply this to  $\Delta(s, t)$ , note first that by (2.6) and the linearity of  $h(\cdot)$ ,

$$n^{-1/2} \hat{J}^n(T^n) = \mathbb{E} \int_0^\infty e^{-\gamma t} h(\bar{X}^n(t)) dt.$$

By assumption, the LHS converges to zero. Hence by Fatou's lemma,  $\mathbb{E} \int_0^{t_0} h(\bar{X}(t)) dt = 0$  for any  $t_0 > 0$ . By the a.s. continuity of the sample paths of  $\bar{X}$ , the structure  $h(x) = h \cdot x$ ,  $h_i > 0$ , and the non-negativity of  $\bar{X}_i(t)$ , it follows that  $\bar{X}_i(t) = 0$  for all  $t$  a.s. Thus for all  $t$ ,

$$\lambda_i t - \sum_{j \in \mathcal{J}_i} \mu_j T_j(t) = 0.$$

This shows  $\sum_{j \in \mathcal{J}_i} \Delta_j(s, t) \mu_j = \lambda_i$  for all  $i$ . By (2.4),  $\sum_{j \in \mathcal{J}^k} (T_j^n(t) - T_j^n(s)) \leq t - s$  for all  $k$  and  $s < t$ . Consequently the same holds for  $T$ . This shows that  $\Delta(s, t) \in \mathcal{S}_{\text{LP}}$  for all  $s < t$ , a.s., namely  $T$  has sample paths in  $\text{Lip}(\mathcal{S}_{\text{LP}})$  a.s.

Finally, a process  $\Xi$  is constructed. By the above result, the sample paths of  $T$  are a.e. differentiable, with derivatives lying in  $\mathcal{S}_{\text{LP}}$  a.e. To construct the a.e. derivative as an  $\{\mathcal{F}'_t\}$ -progressively measurable process we adopt an argument from a proof in [8]. Recall that the lexicographic order on  $\mathbb{R}^{\mathcal{J}}$  is complete. That is, for every  $A \subset \mathbb{R}^{\mathcal{J}}$  there exists a (unique) greatest lower bound with respect to this order, denoted here  $\widetilde{\inf} A$ . We can then construct

$$\Xi_t = \lim_{\varepsilon \downarrow 0} \widetilde{\inf}_{s \in (t-\varepsilon, t) \cap \mathbb{Q}} (t-s)^{-1}(T_t - T_s),$$

where  $\mathbb{Q}$  is the set of rationals. Similarly to the proof in [8, IV.17],  $\Xi$  is progressively measurable. Moreover, one has a.s., for all  $t$ ,  $\int_0^t \Xi_s ds = T_t$ . This gives  $\Xi \in \mathfrak{X}(\{\mathcal{F}'_t\})$ .  $\square$

**PROOF OF LEMMA 5.3.** We introduce notation special to this proof. We fix  $n$  and remove the superscript  $n$  from the notation. Write  $S_j$  for each of the  $I + J$  processes originally denoted by  $A_i^n$  and  $S_j^n$ . Further, in place of  $\bar{T}(t)$  (originally  $\bar{T}^n(t)$ ) we write  $T(t)$ . There will be no confusion with the processes that are elsewhere denoted by  $S, T$ , etc., because this proof regards the prelimit processes only.

The definition of an admissible control for the QCP requires adaptedness of  $T_t^n$  to  $\mathcal{F}_t^n$  (denoted here as  $T_t$  and  $\mathcal{F}_t$ ). To ease the notation we first present the proof for simplified versions of both  $\mathcal{F}_t$  and  $\mathcal{G}_u$  that do not include  $\mathcal{Y}$  as basis elements for the sigma algebras. That is, for  $\mathcal{F}_t$  and  $\mathcal{G}_u$  we take

$$\mathcal{F}_t = \sigma\{S(T_s) : 0 \leq s \leq t\}, \quad t \in \mathbb{R}_+,$$

$$\mathcal{G}_u = \sigma\{S(v) : 0 \preceq v \preceq u\}, \quad u \in \mathcal{U}.$$

Thus under the assumption that  $T_t \in \mathcal{F}_t$  for all  $t$ , the aim is to show that  $\{T_t \preceq u\} \in \mathcal{G}_u$  for all  $t$  and  $u$ . Only minor modifications are required to account for  $\mathcal{Y}$ , and these are described at the end of the proof.

Some further notation is as follows. For  $j \in \mathcal{J}^0$ , let  $\tau_j(0) = 0$  and let the jump times of  $S_j$  be denoted by  $\tau_j(p) = \inf\{t : S_j(t) \geq p\}$ ,  $p \in \mathbb{N}$ . By assumption,  $\tau_j(0) < \tau_j(1) < \dots$ . The

vector  $(\tau_j(1), \dots, \tau_j(p))$  is denoted by  $\tau_j^{\leq p}$ , where in the case  $p = 0$  this is set to the empty set. For  $N \in \mathbb{Z}_+^{\mathcal{J}^0}$ , the collection  $(\tau_j^{\leq N_j} : j \in \mathcal{J}^0)$  is denoted by  $\tau^{\leq N}$ . Next, let  $\eta(0) = 0$  and for  $p \geq 1$ , let

$$\eta(p) = \inf\{t > \eta(p-1) : \text{for some } j \in \mathcal{J}^0, S_j(T_j(t)) > S_j(T_j(\eta(p-1)))\}.$$

For  $p \in \mathbb{Z}_+$ , let  $N(p)$  denote the  $\mathbb{Z}_+^{\mathcal{J}^0}$ -valued RVs  $N(p) = S(T(\eta(p)))$ . Then the sequence  $0 = \eta(0) < \eta(1) < \dots$  are the jump times of  $\{S_j(T_j(t)), j \in \mathcal{J}^0\}$ , and  $N(p)$  the values of  $S(T(t))$  after each jump. (Note that it is possible that at  $\eta(p)$  multiple trajectories  $S_j(T_j(\cdot))$  jump simultaneously). Also let  $K_t = \max\{p : \eta(p) \leq t\}$  denote the number of jumps by time  $t$ .

It is not hard to see that, given  $t$ , the trajectory  $S(T(s))$ ,  $s \in [0, t]$  can be recovered from  $K_t$  and the collection  $\{(\eta(p \wedge K_t), N(p \wedge K_t)), p \in \mathbb{N}\}$ , and vice versa. As a result, we have the identity

$$\mathcal{F}_t = \sigma\{K_t, \eta(p \wedge K_t), N(p \wedge K_t), p \in \mathbb{N}\}.$$

Similarly, for any  $u \in \mathcal{U}$ , one can recover the trajectory  $\{S(v) : v \preceq u\}$  from the collection  $\{S(u), \tau_j(p \wedge S_j(u_j)), p \in \mathbb{N}, j \in \mathcal{J}^0\}$  and vice versa. Hence we also have the identity

$$(5.10) \quad \mathcal{G}_u = \sigma\{S(u), \tau_j(p \wedge S_j(u_j)), p \in \mathbb{N}, j \in \mathcal{J}^0\}.$$

We use Theorem I.18 in [8], which states the following. If  $X$  is an RV on  $(\Omega, \mathcal{F})$  taking values in a measurable space  $(E, \mathcal{E})$  and  $Y$  an RV on  $(\Omega, \mathcal{F})$  with values in  $(\mathcal{U}, \mathcal{B}_{\mathcal{U}})$  which is  $\sigma(X)$ -measurable then there exists a measurable map  $\varphi$  from  $(E, \mathcal{E})$  to  $(\mathcal{U}, \mathcal{B}_{\mathcal{U}})$  such that  $Y = \varphi \circ X$ . Considering  $E = \mathbb{Z}_+ \times \mathbb{R}_+^{\mathbb{N}} \times (\mathbb{Z}_+^{\mathcal{J}^0})^{\mathbb{N}}$ , endow  $\mathbb{Z}_+$  and  $\mathbb{R}_+$  with the Borel  $\sigma$ -fields and let  $\mathcal{E}$  be the corresponding product  $\sigma$ -field. Then for every  $t$  there exists a measurable map  $\varphi_t : (E, \mathcal{E}) \rightarrow (\mathcal{U}, \mathcal{B}_{\mathcal{U}})$  such that

$$T_t = \varphi_t(K_t, \eta(p \wedge K_t), N(p \wedge K_t), p \in \mathbb{N}).$$

For  $\kappa \in \mathbb{Z}_+$  and  $x = (\zeta, n) \in \mathbb{R}_+^{\kappa} \times (\mathbb{Z}_+^{\mathcal{J}^0})^{\kappa}$ , denote

$$\varphi_t^{(\kappa)}(x) = \varphi_t^{(\kappa)}(\zeta^{\leq \kappa}, n^{\leq \kappa}) = \varphi_t(\kappa, \zeta(p \wedge \kappa), n(p \wedge \kappa), p \in \mathbb{N}).$$

Then we have

$$(5.11) \quad T_t = \varphi_t^{(\kappa)}(\eta^{\leq \kappa}, N^{\leq \kappa}) \quad \text{holds on the event} \quad \{K_t = \kappa\}.$$

Two properties of  $\varphi_t^{(\kappa)}$  can immediately be deduced from properties of  $T_t$ . Namely, the continuity and the monotonicity (w.r.t.  $\preceq$ ) of the sample paths of  $t \mapsto T_t$  imply that, for every  $\kappa$  and  $x$ , the map  $t \mapsto \varphi_t^{(\kappa)}(x)$  must also be continuous and monotone. Another useful fact is as follows. By the definitions of  $\eta(p)$  and  $\tau_j(p)$ ,

$$(5.12) \quad \eta(p+1) = \inf\{s > \eta(p) : \text{for some } j \in \mathcal{J}^0, T_j(s) \geq \tau_j(N_j(p) + 1)\}.$$

Hence using (5.11) and the aforementioned monotonicity,

$$(5.13) \quad \eta(p+1) = \inf\{s > \eta(p) : \text{for some } j \in \mathcal{J}^0, \varphi_{j,s}^{(p)}(\eta^{\leq p}, N^{\leq p}) \geq \tau_j(N_j(p) + 1)\}.$$

Equipped with the maps  $\varphi^{(\kappa)}$  and these properties we proceed with the following steps. Fix  $u$  and  $t$ .

- Construct a process  $\{\tilde{T}_s, s \in \mathbb{R}_+\}$  for which the jump times are defined similarly to (5.13), except that they are based solely on the RVs  $\tau^{\preceq S(u)}$ . From its construction it will be immediate that  $\{\tilde{T}_t \preceq u\}$  is  $\mathcal{G}_u$ -measurable.

- Prove that  $T_t \preceq u$  iff  $\tilde{T}_t \preceq u$ .



- Deduce that  $\{T_t \preceq u\}$  is  $\mathcal{G}_u$ -measurable.

The construction of  $\tilde{T}_s$  is as follows. Recall that  $t, u$  have been fixed and denote  $\Sigma = S(u)$ . Let  $\tilde{\eta}(0) = 0, \tilde{N}(0) = 0 \in \mathbb{Z}_+^{\mathcal{J}^0}$ . For  $p \in \mathbb{Z}_+$  such that  $\tilde{\eta}(p) < \infty$ , define inductively

$$\begin{aligned} G_p &= \{j \in \mathcal{J}^0 : \tilde{N}_j(p) + 1 \leq \Sigma_j\}, \\ J^p(s) &= \{j \in G_p : \varphi_{j,s}^{(p)}(\tilde{\eta}^{\leq p}, \tilde{N}^{\leq p}) \geq \tau_j(\tilde{N}_j(p) + 1)\}, \quad s > \tilde{\eta}(p), \\ \tilde{\eta}(p+1) &= \inf\{s > \tilde{\eta}(p) : J^p(s) \neq \emptyset\} \\ \tilde{N}(p+1) &= \tilde{N}(p) + \sum_{j \in J^p(\tilde{\eta}(p+1))} e_j \quad \text{provided } \tilde{\eta}(p+1) < \infty. \end{aligned}$$

Note that after at most  $|\Sigma|$  steps,  $G_p$  becomes empty, and as a result  $\tilde{\eta}(p_0 + 1) = \infty$  for some finite  $p_0$ . Based on the above construction, define, for  $0 \leq p \leq p_0$ ,

$$\tilde{T}_s = \varphi_s^{(p)}(\tilde{\eta}^{\leq p}, \tilde{N}^{\leq p}), \quad \text{for } s \in [\tilde{\eta}(p), \tilde{\eta}(p+1)).$$

This defines  $\tilde{T}_s$  for all  $s \in \mathbb{R}_+$ .

If instead of  $G_p$  we had used  $\mathcal{J}^0$  in the definition of  $\tilde{\eta}(p+1)$ , we would have recovered  $T_s$  itself as can be seen by (5.13). The use of  $G_p$  achieves the goal that only the RVs  $\Sigma = S(u)$  and  $\tau^{\preceq \Sigma}$  are involved in the construction. As a result, the whole process  $\{\tilde{T}_s, s \in \mathbb{R}_+\}$  is measurable on  $\sigma\{S(u), \tau^{\preceq S(u)}\}$ , hence by (5.10), it is measurable on  $\mathcal{G}_u$ . Consequently, the event  $\{\tilde{T}_t \preceq u\}$  is  $\mathcal{G}_u$ -measurable.

The proof that  $T_t \preceq u$  iff  $\tilde{T}_t \preceq u$  proceeds in two steps. First, assume

$$(5.14) \quad T_t \preceq u.$$

It is argued by induction that, for  $p \in \mathbb{Z}_+$ ,

$$(5.15) \quad \begin{aligned} &\text{if } \tilde{\eta}(p) \wedge \eta(p) \leq t \text{ then} \\ &\text{(a) } (\tilde{\eta}^{\leq p}, \tilde{N}^{\leq p}) = (\eta^{\leq p}, N^{\leq p}), \\ &\text{(b) for } s \in [\eta(p), \tilde{\eta}(p+1) \wedge \eta(p+1)) \text{ one has } \tilde{T}_s = T_s. \end{aligned}$$

In particular, (5.15) gives  $\tilde{T}_s = T_s, s \in [0, t]$ , hence  $\tilde{T}_t \preceq u$ .

Before proving (5.15), notice that, for every  $k$ , (5.15)(a) implies (5.15)(b), because by (5.11) and the definition of  $\tilde{T}_s$ , the processes  $T_s$  on the interval  $[\eta(p), \eta(p+1))$  and  $\tilde{T}_s$  on the interval  $[\tilde{\eta}(p), \tilde{\eta}(p+1))$  are given by the same formula, and since  $\eta(p) = \tilde{\eta}(p)$ , they must be equal for all  $s \in [\eta(p), \tilde{\eta}(p+1) \wedge \eta(p+1))$ . Therefore it suffices to prove (5.15)(a). Also note that, by the definition of  $\Sigma$ ,

$$(5.16) \quad \tau_j(\Sigma_j) \leq u_j < \tau_j(\Sigma_j + 1), \quad j \in \mathcal{J}^0.$$

We now prove (5.15)(a). First, for  $p = 0$ ,  $(\tilde{\eta}(0), \tilde{N}(0)) = (\eta(0), N(0))$  by definition. Next, assume (5.15)(a) holds for  $p$  and consider  $p+1$ . Thus, under the condition

$$\tilde{\eta}(p+1) \wedge \eta(p+1) \leq t$$

we must show

$$(5.17) \quad (\tilde{\eta}(p+1), \tilde{N}(p+1)) = (\eta(p+1), N(p+1)).$$

By the induction assumption,  $T_s = \tilde{T}_s = \varphi_s^{(p)}(\tilde{\eta}^{\leq p}, \tilde{N}^{\leq p})$  for  $s \in [\eta(p), \tilde{\eta}(p+1) \wedge \eta(p+1))$ . Comparing (5.13) to the definition of  $\tilde{\eta}(p+1)$  it is seen that  $\eta(p+1) \leq \tilde{\eta}(p+1)$  because  $G_p \subset \mathcal{J}^0$ . Thus

$$(5.18) \quad \eta(p+1) \leq t.$$



Arguing by contradiction, assume that (5.17) is false. Then either (a)  $\eta(p+1) < \tilde{\eta}(p+1)$  or (b)  $\eta(p+1) = \tilde{\eta}(p+1)$  but  $N(p+1) \neq \tilde{N}(p+1)$ . In both cases, by (5.12), there exists  $j \notin G_p$  such that one has  $T_j(\eta(p+1)) = \tau_j(N_j(p) + 1)$ . Since the processes  $\tilde{T}$  and  $T$  are continuous and equal on  $[\eta(p), \eta(p+1))$ , we have

$$(5.19) \quad \tilde{T}_j(\eta(p+1)) = T_j(\eta(p+1)) = \tau_j(N_j(p) + 1).$$

By the induction assumption,  $\tilde{N}_j(p) = N_j(p)$ , hence  $T_j(\eta(p+1)) = \tau_j(\tilde{N}_j(p) + 1)$ . Also, since  $j \notin G_p$ ,  $\tilde{N}_j(p) + 1 > \Sigma_j$ . That is,  $\tilde{N}_j(p) + 1 \geq \Sigma_j + 1$ . We therefore obtain

$$\begin{aligned} \tau_j(\Sigma_j + 1) &\leq \tau_j(\tilde{N}_j(p) + 1) \\ &= T_j(\eta(p+1)) && \text{by (5.19)} \\ &\leq T_j(t) && \text{by (5.18)} \\ &\leq u_j, && \text{by (5.14)} \\ &< \tau_j(\Sigma_j + 1), && \text{by (5.16)} \end{aligned}$$

a contradiction. This shows that (5.15) holds under the assumption  $T_t \preceq u$ . Consequently,  $T_t \preceq u$  implies  $\tilde{T}_t \preceq u$ .

Next assume

$$(5.20) \quad \tilde{T}_t \preceq u.$$

Under this condition we also prove that (5.15) holds. The proof follows the same lines as above, except the last chain of inequalities, that now reads

$$\begin{aligned} \tau_j(\Sigma_j + 1) &\leq \tau_j(\tilde{N}_j(p) + 1) \\ &= \tilde{T}_j(\eta(p+1)) && \text{by (5.19)} \\ &\leq \tilde{T}_j(t) && \text{by (5.18)} \\ &\leq u_j, && \text{by (5.20)} \\ &< \tau_j(\Sigma_j + 1), && \text{by (5.16)} \end{aligned}$$

that again gives a contradiction, establishing (5.15). As a result,  $\tilde{T}$  and  $T$  are equal on  $[0, t]$ , and thus  $T_t \preceq u$ . This completes the proof that  $\tilde{T}_t \preceq u$  iff  $T_t \preceq u$ . We deduce that  $\{T_t \preceq u\}$  is  $\mathcal{G}_u$ -measurable.

Finally, accounting for the RV  $\mathcal{Y}$  by adding it to the definitions of both  $\mathcal{F}_t$  and  $\mathcal{G}_u$ , as in  $\mathcal{F}_t = \sigma\{\mathcal{Y}, S(T_s) : 0 \leq s \leq t\}$ , presents no difficulty. Apart from applying this change, augmenting the measurable space  $(E, \mathcal{E})$  and updating the functions  $\varphi^{(\kappa)}$  in an obvious way, the proof requires no adaptation.  $\square$

**PROOF OF LEMMA 5.4.** We keep the convention of writing  $A_i$  as  $S_{(i,0)}$ . However, unlike in the proof of the previous lemma,  $S^n$  (and similarly,  $\bar{T}^n$ ,  $\mathcal{G}_u^n$ , etc.) and  $S$  are distinct objects.

1. This part is a direct consequence of the fact that  $\bar{T}_t$  is a  $\mathcal{G}_u$ -stopping time for every  $t$ , and  $\mathcal{H}_t = \mathcal{G}_{\bar{T}_t}$ . Indeed, to show that  $\bar{T}_t$  is  $\mathcal{H}_t$ -measurable, it suffices to show that for every  $v \in \mathcal{U}$ ,  $\{\bar{T}_t \preceq v\} \in \mathcal{H}_t$ . But  $\mathcal{H}_t = \{G \in \mathcal{G}_0 : G \cap \{\bar{T}_t \preceq u\} \in \mathcal{G}_u\}$  and  $\{\bar{T}_t \preceq v\} \cap \{\bar{T}_t \preceq u\} = \{\bar{T}_t \preceq v \wedge u\}$ , where  $v \wedge u$  denotes componentwise minimum. Since the latter event is in  $\mathcal{G}_u$ , statement 1 is proved.

2. We first prove the independence

$$(5.21) \quad \mathcal{G}_u \perp\!\!\!\perp S(y) - S(x), \quad \text{for all } 0 \preceq u \preceq x \preceq y \in \mathcal{U}.$$

By the continuity of the sample paths of  $S$ , it suffices to show this independence under the additional restriction that for all  $j \in \mathcal{J}^0$ ,  $u_j < x_j$ . Fix such  $u, x, y$ . Denote  $u_\varepsilon = u + \varepsilon \mathbf{e}$ ,  $\varepsilon > 0$ . Fix  $\varepsilon \in (0, 1)$  so small that  $u_\varepsilon \preccurlyeq x$ . Denote  $\Delta = S(y) - S(x)$  and  $\Delta^n = \hat{S}^n(y) - \hat{S}^n(x)$ .

The proof of (5.21) is based on two facts. First, one can find a sequence of RVs  $\tilde{\Delta}^n$  such that  $\Delta^n - \tilde{\Delta}^n \rightarrow 0$  in probability and at the same time, for every  $n$ ,

$$(5.22) \quad (\Upsilon, \{S^n(v) : v \preccurlyeq u_\varepsilon\}) \perp\!\!\!\perp \tilde{\Delta}^n.$$

Second, there exists a continuous approximation of the function  $\psi^{(u)}$  with certain properties. More precisely, there exist three functions mapping  $\mathcal{U}$  to itself, denoted  $\psi_{\text{cts}}^{(u)}$ ,  $g_1$ ,  $g_2$ , where  $\psi_{\text{cts}}^{(u)}$  is continuous and  $g_1, g_2$  are Borel, such that

$$(5.23) \quad \psi_{\text{cts}}^{(u)} = g_1 \circ \psi^{(u_\varepsilon)}, \quad \psi^{(u)} = g_2 \circ \psi_{\text{cts}}^{(u)}.$$

We first show (5.21) based on these facts and then prove these facts. To this end, let us show that, for every  $t$ , the RV  $\psi_{\text{cts}}^{(u)}(\bar{T}_t^n)$  is measurable on  $\mathcal{G}_{u_\varepsilon}^n$ . By the first part of (5.23), it suffices to show that  $\psi^{(u_\varepsilon)}(\bar{T}_t^n)$  is  $\mathcal{G}_{u_\varepsilon}^n$ -measurable. To show this we use the following notation. For  $v \in \mathcal{U}$  denote  $R_v = \{w \in \mathcal{U} : w \preccurlyeq v\}$  and  $R_v^c$  its complement in  $\mathcal{U}$ . Because the Borel  $\sigma$ -field on  $\mathcal{U}$  is generated by the rectangles  $R_v$ ,  $v \in \mathcal{U}$ , it suffices to show that

$$\eta(n, t, u_\varepsilon, v) := \{\psi^{(u_\varepsilon)}(\bar{T}_t^n) \in R_v\} \in \mathcal{G}_{u_\varepsilon}^n$$

for all  $v \in \mathcal{U}$ . By Lemma 5.3, for  $v \in R_\varepsilon$  one has  $\{\bar{T}_t^n \preccurlyeq v\} \in \mathcal{G}_v^n \subset \mathcal{G}_{u_\varepsilon}^n$ , where the fact that  $\mathcal{G}^n$  forms a filtration is used. Now, by the definition of  $\psi^{(u_\varepsilon)}$ , we can write

$$\begin{aligned} \eta(n, t, u_\varepsilon, v) &= \begin{cases} \{\bar{T}_t^n \in R_v\} & \text{on } \{\bar{T}_t^n \in R_{u_\varepsilon}\} \\ \{u_\varepsilon + \mathbf{e} \in R_v\} & \text{on } \{\bar{T}_t^n \in R_{u_\varepsilon}\}^c \end{cases} \\ &= \{\bar{T}_t^n \preccurlyeq v \wedge u_\varepsilon\} \cup [\{u_\varepsilon + \mathbf{e} \preccurlyeq v\} \cap \{\bar{T}_t^n \preccurlyeq u_\varepsilon\}^c]. \end{aligned}$$

Since  $v \wedge u_\varepsilon \preccurlyeq u_\varepsilon$ , the above expression shows that  $\eta(n, t, u_\varepsilon, v) \in \mathcal{G}_{u_\varepsilon}^n$ . Hence follows the claim that  $\psi_{\text{cts}}^{(u)}(\bar{T}_t^n)$  is  $\mathcal{G}_{u_\varepsilon}^n$ -measurable.

Using this in (5.22) and recalling that  $\hat{S}^n$  is a normalization of  $S^n$  gives

$$(5.24) \quad (\Upsilon, \{\hat{S}^n(v) : v \preccurlyeq u_\varepsilon\}, \{\psi_{\text{cts}}^{(u)}(\bar{T}_t^n), t \geq 0\}) \perp\!\!\!\perp \tilde{\Delta}^n.$$

We can now take limits. By Lemma 5.1,

$$(\{\hat{S}^n(v) : v \preccurlyeq u_\varepsilon\}, \{\bar{T}_t^n : t \geq 0\}, \Delta^n) \Rightarrow (\{S(v) : v \preccurlyeq u_\varepsilon\}, \{\bar{T}_t : t \geq 0\}, \Delta).$$

Thus by the continuity of  $\psi_{\text{cts}}^{(u)}$  and the fact  $\Delta^n - \tilde{\Delta}^n \rightarrow 0$ ,

$$(\{\hat{S}^n(v) : v \preccurlyeq u_\varepsilon\}, \{\psi_{\text{cts}}^{(u)}(\bar{T}_t^n) : t \geq 0\}, \tilde{\Delta}^n) \Rightarrow (\{S(v) : v \preccurlyeq u_\varepsilon\}, \{\psi_{\text{cts}}^{(u)}(\bar{T}_t) : t \geq 0\}, \Delta).$$

By the independence stated in (5.24), this shows

$$(\{S(v) : v \preccurlyeq u_\varepsilon\}, \{\psi_{\text{cts}}^{(u)}(\bar{T}_t) : t \geq 0\}) \perp\!\!\!\perp \Delta.$$

We can now use the second part of (5.23) to obtain

$$(\{S(v) : v \preccurlyeq u\}, \{\psi^{(u)}(\bar{T}_t) : t \geq 0\}) \perp\!\!\!\perp \Delta.$$

By the definition of  $\mathcal{G}_u$  in (5.6), this shows that  $\mathcal{G}_u \perp\!\!\!\perp \Delta$ , as stated in (5.21).

To complete the proof of (5.21), the two aforementioned facts need to be established. To construct  $\tilde{\Delta}^n$ , let

$$\tau_j^n := \inf \{z > v_j, S_j^n(z) > S_j^n(x_j)\}, \quad j \in \mathcal{J}^0.$$

Let  $(\tau^n) = (\tau_j^n)_{j \in \mathcal{J}^0}$  and  $\tilde{\Delta}^n = \hat{S}^n(\tau^n + y - x) - \hat{S}^n(\tau^n)$ , which, according to our convention, is a shorthand notation for  $\tilde{\Delta}^n = (\tilde{\Delta}_j^n)_{j \in \mathcal{J}^0}$ ,  $\tilde{\Delta}_j^n = \hat{S}_j^n(\tau_j^n + y_j - x_j) - \hat{S}_j^n(\tau_j^n)$ . Since  $\lambda_i^n$  and  $\mu_j^n$  grow to  $\infty$ , it is easy to see that  $\tau_j^n \rightarrow x_i$  in probability. Because  $\hat{S}^n$  are  $C$ -tight, this implies the first desired property of  $\tilde{\Delta}^n$ , namely  $\Delta^n - \tilde{\Delta}^n \rightarrow 0$  in probability. To prove the second property of  $\tilde{\Delta}^n$ , (5.22), due to the independence of  $\Upsilon$  and  $S^n$ , it suffices to show that

$$(5.25) \quad S^n(v) \perp\!\!\!\perp \bar{W}^n := S^n(\tau^n + y - x) - S^n(\tau^n),$$

for every  $v \preceq u_\varepsilon$ . To this end we introduce some notation. The accelerated interarrival and service times denoted in (2.1) by  $a_i^n(p)$  and  $u_j^n(p)$  will now be denoted  $u_j^n(p)$ ,  $j \in \mathcal{J}^0$ ,  $p = 1, 2, \dots$ , in accordance with the relabeling convention used in this section. For  $l \in \mathbb{Z}_+^{\mathcal{J}^0}$ , introduce the multi-parameter filtration

$$\bar{\mathcal{F}}_l^n = \sigma\{u_j^n(p_j), p_j \leq l_j + 1, j \in \mathcal{J}^0\}.$$

Then it is easy to see that  $\mathcal{T}^n := S^n(x)$  is a (multi-parameter) stopping time w.r.t.  $\bar{\mathcal{F}}^n$ . We now apply Lemma 7.6 of [2], which states that the conditional distribution of the sequence  $(u_j^n(\mathcal{T}_j^n + p), j \in \mathcal{J}^0)_{p=2,3,\dots}$  given the stopped  $\sigma$ -field  $\bar{\Sigma}^n := \bar{\mathcal{F}}_{\mathcal{T}^n}^n$  is the same as the (unconditioned) distribution of the original family of i.i.d. random variables  $(u_j^n(p), j \in \mathcal{J}^0)_{p=1,2,\dots}$ . (The lemma in [2] and its proof are concerned with the special case  $|\mathcal{J}^0| = 3$  but are valid in general). Now, for each  $j \in \mathcal{J}^0$ , the RV  $\bar{W}_j^n = S_j^n(\tau_j^n + y_j - x_j) - S_j^n(\tau_j^n)$  is determined by the sequence  $(u_j^n(\mathcal{T}_j^n + p), p = 2, 3, \dots)$ . Therefore the conditional distribution of  $\bar{W}^n$  given  $\bar{\Sigma}^n$  is the same as the distribution of  $S^n(y - x)$ . On the other hand, for each  $v \preceq x$ , (in particular for each  $v \preceq u_\varepsilon$ ), the RV  $S^n(v)$  is measurable w.r.t.  $\bar{\Sigma}^n$ . As a consequence, (5.25) holds.

Next we describe the construction of  $\psi_{\text{cts}}^{(u)}$ . For  $v \in \mathcal{U}$ , set

$$\psi_{\text{cts}}^{(u)}(v) = (\psi_{\text{cts},j}^{(u)}(v)), \quad \psi_{\text{cts},j}^{(u)}(v) = (v_j + \varepsilon^{-1} \max(v_g - u_g)^+ \wedge (u_j + 1)).$$

Note that

$$\psi_{\text{cts}}^{(u)} = \psi_{\text{cts}}^{(u)} \circ \psi^{(u_\varepsilon)},$$

because in  $R_{u_\varepsilon}$ , the function  $\psi^{(u_\varepsilon)}$  agrees with the identity map, and if  $v \in R_{u_\varepsilon}^c$  then  $\psi_{\text{cts}}^{(u)}(v) = u + \mathbf{e}$  whereas

$$\psi_{\text{cts}}^{(u)} \circ \psi^{(u_\varepsilon)}(v) = \psi_{\text{cts}}^{(u)}(u_\varepsilon + \mathbf{e}) = u + \mathbf{e}.$$

Also,

$$\psi^{(u)} = \psi^{(u)} \circ \psi_{\text{cts}}^{(u)}$$

holds, because in  $R_u$ , both sides of the equality agree with the identity map, and any member of  $R_u^c$  is sent by  $\psi_{\text{cts}}^{(u)}$  to a member of  $R_u^c$ , resulting in  $u + \mathbf{e}$  on both sides of the equation. These two identities validate (5.23). This completes the proof of the first part of statement 2.

As for the second part of this statement, by the definition of  $\mathcal{G}_u$  in (5.6), for any  $t$ ,  $\bar{T}_t$  is a  $\mathcal{G}_u$  stopping time, and  $\bar{T}_t$  has non-decreasing continuous sample paths a.s. Moreover,  $S^{\text{unit}}$  is a SBM. Thus to show that the hypotheses of Theorem 5.2 hold, it remains to show that for every subset  $\mathcal{M} \subset \mathcal{J}^0$ ,  $\Phi_\theta^{\mathcal{M}}(u)$  of (5.2), with  $S^{\text{unit}}$  in place of  $\tilde{S}$ , is a  $\mathcal{G}_u$ -martingale. To this end, note by (5.21) that for  $u, v \in \mathcal{U}$ ,  $S^{\text{unit}}(u + v) - S^{\text{unit}}(u)$  is independent of  $\mathcal{G}_u$ . Hence, recalling  $\mathcal{M} \subset \mathcal{J}^0$  and

$$\Phi_\theta^{\mathcal{M}}(u) = \prod_{j \in \mathcal{M}} \exp\left(\mathbf{i}\theta_j S_j^{\text{unit}}(u_j) + \frac{1}{2}\theta_j^2 u_j\right),$$

we have

$$\begin{aligned} \mathbb{E} [\Phi_\theta^{\mathcal{M}}(u+v) | \mathcal{G}_u] &= \Phi_\theta^{\mathcal{M}}(u) \mathbb{E} \prod_{j \in \mathcal{M}} \exp \left\{ \mathbf{i} \theta_j (S_j^{\text{unit}}(u_j + v_j) - S_j^{\text{unit}}(u_j)) + \frac{1}{2} \theta_j^2 v_j \right\} \\ &= \Phi_\theta^{\mathcal{M}}(u), \end{aligned}$$

where the last identity follows from the fact that  $S_j^{\text{unit}}(u_j + v_j) - S_j^{\text{unit}}(u_j) \sim \mathcal{N}(0, \text{diag}(v_j))$ . This shows that the tuple satisfies the hypotheses of Theorem 5.2 and completes the proof of statement 2.

3 and 4. We deduce from Theorem 5.2 that  $D_j = S_j \circ \bar{T}_j$ ,  $j \in \mathcal{J}^0$  are continuous local martingales w.r.t.  $\mathcal{H}_t$ . Moreover, for every distinct  $j, j' \in \mathcal{J}^0$ ,  $D_j D_{j'}$  is also a continuous local martingale. Because the process  $\langle D_j, D_{j'} \rangle$ , by its definition, is the unique predictable process starting at zero which makes  $D_j D_{j'} - \langle D_j, D_{j'} \rangle$  a local martingale, it follows that  $\langle D_j, D_{j'} \rangle = 0$ .

5. The proof is based on Proposition V.1.5(i) of [25], according to which if there exists a single-parameter filtration with respect to which, for every  $t$  and  $j$ ,  $\bar{T}_j(t)$  is a stopping time and  $S_j$  a martingale, then  $\langle S_j \circ \bar{T}_j \rangle_t = \langle S_j \rangle_{\bar{T}_j(t)}$ .

Toward finding such a filtration, we argue as follows. It suffices to obtain the result on a finite time interval  $t \in [0, c_1]$ , where  $c_1$  is arbitrary. Fix  $c_1$  and let  $c_2 > c_1$ . Fix  $j^0$ . For  $j \in \mathcal{J}^0$ , let  $e_j \in \mathcal{U}$  denote the unit vector whose  $p$ -th component is  $\mathbb{1}_{p=j}$ . Consider the single-parameter filtration  $\mathcal{G}_b^{(j^0)}$ ,  $b \in [0, c_1]$  defined in terms of  $\mathcal{G}_u$  as

$$\mathcal{G}_b^{(j^0)} = \mathcal{G}_{u(j^0, b)} \quad u(j^0, b) = b e_{j^0} + c_2 \sum_{j \neq j^0} e_j, \quad b \in [0, c_1].$$

We use the fact that for all  $t \in [0, c_1]$  and  $j \in \mathcal{J}^0$ ,  $\bar{T}_j(t) < c_2$ . Recall that  $\bar{T}_t$  is a (multi-parameter) stopping time on  $\mathcal{G}_u$ . Thus for  $t \in [0, c_1]$  and  $b \in [0, c_1]$ ,

$$\{\bar{T}_{j^0}(t) \leq b\} = \{\bar{T}_{j^0}(t) \leq b, \bar{T}_j(t) \leq c_2 \text{ for all } j \neq j^0\} = \{\bar{T}_t \preceq u(j^0, b)\} \in \mathcal{G}_{u(j^0, b)} = \mathcal{G}_b^{(j^0)}.$$

Hence  $\bar{T}_{j^0}(t)$  is a  $\{\mathcal{G}_b^{(j^0)}\}$ -stopping time for every  $t \in [0, c_1]$ . Moreover, using (5.21) with  $x = u(j^0, a)$  and  $y = u(j^0, b)$ , where  $0 \leq a \leq b \leq c_1$ , shows that  $\mathcal{G}_a^{(j^0)} \perp\!\!\!\perp S_{j^0}(b) - S_{j^0}(a)$ , so that the (zero drift) BM  $S_{j^0}$  is a martingale on this filtration. Clearly,  $t \mapsto \bar{T}_{j^0}(t)$  is a.s. continuous non-decreasing. Hence the assumptions of Proposition V.1.5(i) of [25] are verified, and so  $\langle S_{j^0} \circ \bar{T}_{j^0} \rangle_t = \langle S_{j^0} \rangle_{\bar{T}_{j^0}(t)} = \sigma_{j^0}^2 \bar{T}_{j^0}(t)$  for all  $j^0 \in \mathcal{J}^0$ .  $\square$

## APPENDIX A: SOME USEFUL LINEAR PROGRAMMING RESULTS

For ease of reference we reproduce here several results related to linear programming from Chapter 7 of [26]. We begin with the definitions of a convex polyhedron and a convex polytope in Section 7.2 of [26].

**DEFINITION A.1.** (i) A set  $P$  of vectors in  $\mathbb{R}^m$  is called a convex polyhedron if  $P = \{x \mid Ax \leq b\}$  for some matrix  $A \in \mathbb{R}^{n \times m}$  and column vector  $b \in \mathbb{R}^n$ .

(ii) A set of vectors is a convex polytope if it is the convex hull of finitely many vectors.

The following lemma is Corollary 7.1c in [26].

**LEMMA A.2.** A set  $P$  is a convex polytope if and only if  $P$  is a bounded convex polyhedron.

For the remainder of this appendix we let  $A \in \mathbb{R}^{n \times m}$  be a matrix,  $b \in \mathbb{R}^n$  be a column vector and  $c \in \mathbb{R}^m$  be a row vector. We next state what is commonly referred to as the duality theorem of linear programming, which is Corollary 7.1g of [26].

THEOREM A.3. *We have*

$$(A.1) \quad \max_{x \in \mathbb{R}^m} \{cx \mid Ax \leq b\} = \min_{y \in \mathbb{R}^n} \{yb \mid yA = c, y \geq 0\}.$$

*provided that both sets in (A.1) are non-empty.*

Part (iii) of the next lemma is commonly referred to as complementary slackness, and is (35)(iii) in [26].

LEMMA A.4. *Assume that both optima in (A.1) are finite, and let  $x_0, y_0$  be feasible solutions, respectively, of the maximization and minimization problems (so that  $Ax_0 \leq b$ ,  $y_0A = c$ , and  $y_0 \geq 0$ ). Then the following are equivalent:*

- (i)  $x_0$  and  $y_0$  are optimal solutions in (A.1);
- (ii)  $cx_0 = y_0b$ ;
- (iii)  $y_0(b - Ax_0) = 0$ .

Finally, we close with strict complementary slackness, which is (36) in [26].

THEOREM A.5. *Assume that both both sets in (A.1) are non-empty. For each  $i = 1, \dots, n$ , exactly one of the following holds:*

- (a) *The maximum in (A.1) has a solution  $x^0$  such that  $(Ax^0)_i < b_i$*
- (b) *The minimum in (A.1) has a solution  $y^0$  such that  $y_i^0 > 0$ .*

**Funding.** RA is supported by ISF grant 1035/20.

## REFERENCES

- [1] ATAR, R., CASTIEL, E. and REIMAN, M. (2022). Asymptotic optimality of switched control policies in a simple parallel server system under an extended heavy traffic condition. *Submitted*.
- [2] BELL, S. L. and WILLIAMS, R. J. (2001). 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** 608–649.
- [3] BELL, S. L. and WILLIAMS, R. J. (2005). Dynamic scheduling of a parallel server system in heavy traffic with complete resource pooling: asymptotic optimality of a threshold policy. *Electron. J. Probab.* **10** no. 33, 1044–1115.
- [4] BILLINGSLEY, P. (1999). *Convergence of Probability Measures*, second ed. *Wiley Series in Probability and Statistics: Probability and Statistics*. John Wiley & Sons Inc., New York. A Wiley-Interscience Publication.
- [5] BUDHIRAJA, A. and GHOSH, A. P. (2006). Diffusion approximations for controlled stochastic networks: an asymptotic bound for the value function. *Ann. Appl. Probab.* **16** 1962–2006.
- [6] BUDHIRAJA, A. and GHOSH, A. P. (2012). Controlled stochastic networks in heavy traffic: convergence of value functions. *Ann. Appl. Probab.* **22** 734–791.
- [7] CHEN, H. and MANDELBAUM, A. (1991). Leontief systems, RBVs and RBMs. In *Applied stochastic analysis (London, 1989). Stochastics Monogr.* **5** 1–43. Gordon and Breach, New York.
- [8] DELLACHERIE, C. and MEYER, P.-A. (1975). *Probabilités et potentiel*. Hermann, Paris. Chapitres I à IV, Édition entièrement refondue, Publications de l’Institut de Mathématique de l’Université de Strasbourg, No. XV, Actualités Scientifiques et Industrielles, No. 1372.
- [9] FLEMING, W. H. and SONER, H. M. (2006). *Controlled Markov Processes and Viscosity Solutions* **25**. Springer Science & Business Media.
- [10] GILBARG, D. and TRUDINGER, N. S. (2015). *Elliptic Partial Differential Equations of Second Order* **224**. Springer.

- [11] GOLDFARB, D. and TODD, M. J. (1989). Linear Programming. *Handbooks in Operations Research and Management Science* **1** 73–170.
- [12] HARRISON, J. M. (1988). Brownian models of queueing networks with heterogeneous customer populations. In *Stochastic differential systems, stochastic control theory and applications* (Minneapolis, Minn., 1986). *IMA Vol. Math. Appl.* **10** 147–186. Springer, New York.
- [13] HARRISON, J. M. (1998). Heavy traffic analysis of a system with parallel servers: asymptotic optimality of discrete-review policies. *Annals of Applied Probability* 822–848.
- [14] HARRISON, J. M. (2000). Brownian models of open processing networks: canonical representation of workload. *Ann. Appl. Probab.* **10** 75–103.
- [15] HARRISON, J. M. and LÓPEZ, M. J. (1999). Heavy traffic resource pooling in parallel-server systems. *Queueing Systems Theory Appl.* **33** 339–368.
- [16] HARRISON, J. M. and VAN MIEGHEM, J. A. (1997). Dynamic control of Brownian networks: state space collapse and equivalent workload formulations. *Ann. Appl. Probab.* **7** 747–771.
- [17] KARATZAS, I. and SHREVE, S. E. (1991). *Brownian Motion and Stochastic Calculus*, second ed. *Graduate Texts in Mathematics* **113**. Springer-Verlag, New York.
- [18] KRYLOV, N. V. (2008). *Controlled Diffusion Processes* **14**. Springer Science & Business Media.
- [19] KRYLOV, N. V. and LIPTSER, R. (2002). On diffusion approximation with discontinuous coefficients. *Stochastic Processes and Their Applications* **102** 235–264.
- [20] KURTZ, T. G. (1980). Representations of Markov Processes as Multiparameter Time Changes. *Ann. Probab.* **8** 682–715.
- [21] LAWS, C. (1992). Resource pooling in queueing networks with dynamic routing. *Advances in Applied Probability* **24** 699–726.
- [22] MANDELBAUM, A. and STOLYAR, A. L. (2004). Scheduling flexible servers with convex delay costs: heavy-traffic optimality of the generalized  $c\mu$ -rule. *Oper. Res.* **52** 836–855.
- [23] PESIC, V. and WILLIAMS, R. (2016). Dynamic scheduling for parallel server systems in heavy traffic: Graphical structure, decoupled workload matrix and some sufficient conditions for solvability of the Brownian control problem. *Stochastic Systems* **6** 26–89.
- [24] RATH, J. H. (1975). Controlled queues in heavy traffic. *Advances in Applied Probability* **7** 656–671.
- [25] REVUZ, D. and YOR, M. (1999). *Continuous Martingales and Brownian Motion*, third ed. *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]* **293**. Springer-Verlag, Berlin.
- [26] SCHRIJVER, A. (1986). *Theory of Linear and Integer Programming*. *Wiley-Interscience Series in Discrete Mathematics*. John Wiley & Sons, Ltd., Chichester.
- [27] SHENG, D. (1978). *Some Problems in the Optimal Control of Diffusions*. PhD Thesis, Stanford University, California.
- [28] STOLYAR, A. L. (2005). Optimal routing in output-queued flexible server systems. *Probability in the Engineering and Informational Sciences* **19** 141–189.