Sub-diffusive load-balancing in time-varying queueing systems

Rami Atar*† Isaac Keslassy* Gal Mendelson*†

Abstract

Load-balancing algorithms for systems that operate in heavy traffic are known to lead, under suitable conditions, to state space collapse (SSC). This term refers to the phenomenon where imbalance is negligible compared to queue lengths. Specifically, whereas queue lengths behave diffusively, the size of imbalance is at a sub-diffusive scale: denoting by n the usual scaling parameter, the former and the latter are of order $O(n^{1/2})$ and $o(n^{1/2})$, respectively. In this paper we consider load balancing for time-varying systems. SSC results and the standard techniques on which they are based do not apply to these systems, which (a) are not in heavy traffic, thus queue lengths may reach levels as high as $O(n)$, and (b) have time-varying traffic intensities that cause transitions between underloaded, critically loaded and overloaded regimes. Our results extend SSC far beyond the heavy traffic setting, by establishing sub-diffusive (i.e., $o(n^{1/2})$) balance for time-varying systems.

To exhibit the breadth of the described phenomenon, the results address three load-balancing models. The first is *the-power-of-d-choices* (SQ(d)), where arrivals from a single stream are routed to the shortest among d randomly-chosen queues, where $1 < d \leq N$, and N denotes the fixed number of queues in the system. The second is *redundancy-d* (R(d)), where jobs are replicated d times, routed simultaneously to d randomly-chosen queues, and all but the first replica to be admitted into service are canceled. The third model is *longest queue first* (LQF), where a single resource is shared by N job classes, and the job that receives service is always selected from the queue that is longest.

As an application of these results, asymptotic optimality of SQ(d) and R(d) is shown, with an optimality guarantee of order $o(n^{1/2})$ in the aforementioned framework, where in particular queue sizes may reach $O(n)$. Moreover, in the special case of the standard heavy traffic setting, the results are shown to yield new, explicit sufficient conditions for SSC.

AMS subject classification: 60F170, 60J60, 60K25, 93E20.

Keywords: Randomized load-balancing, time-varying queues, join the shortest queue, longest queue first, power of choice, task redundancy, redundancy routing, job cancellations, diffusion limits, heavy traffic, state space collapse

1 Introduction

In this paper we study the performance of load-balancing algorithms in time-varying queueing systems. Since the goal of load balancing algorithms is to distribute load evenly among a number of different channels, the degree to which delays or queue lengths equalize gives a good indication on their performance. Known results in this context are obtained via a heavy traffic analysis,

*The Viterbi Faculty of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel

†Research supported in part by the ISF (grant 1184/16)

which is concerned with the asymptotic behavior of the delay or queue length at the *diffusion scale* under a *critical load* condition. Consider, for example, the well-known *join-the-shortest-queue* policy, denoted in the paper by SQ(N), where arrivals from a single stream are routed to the shortest among N queues, each associated with a single server. Under this policy, it is known that equalization is achieved in heavy traffic at a scale smaller than that of the stochastic fluctuations. That is, denoting by n the heavy traffic scaling parameter, arrival and service rates scale like n, and queue lengths are diffusive, i.e., scale like $n^{1/2}$, whereas the *queue length deviation process*, defined as the difference between the greatest and smallest queue length as it varies over time, is sub-diffusive, i.e., of order $o(n^{1/2})$. This result is known since Reiman's paper [57] and has been referred to as *state space collapse* (SSC). It is valid also for *the-power-of-d-choices* policy, denoted SQ(d), in which arrivals are routed to the shortest among d randomly chosen queues out of the N queues, for $1 < d \leq N$. This is true even if servers are heterogeneous, under some further conditions, as shown by Chen and Ye [18].

The goal of this paper is to argue that sub-diffusivity of the deviation process (SDDP) is not limited to heavy traffic. Our main results establish SDDP in settings that accommodate server heterogeneity and time-varying arrival and service intensities. These systems do not adhere to the critical load condition and need not exhibit diffusive fluctuations away from zero for the queue length process. In particular, the queue length may reach levels as high as $O(n)$, and thus one might expect the imbalance to be negligible with respect to n, but not to $n^{1/2}$. Compared to this, the results we obtain are much sharper, as the scale of imbalance is shown to be $o(n^{1/2})$ for broad classes of time-varying traffic intensities, under which the system may undergo transitions between underloaded, critical and overloaded regimes. To our knowledge, this is the first time SDDP has been established for any time-varying system. This includes the most basic load balancing model, namely SQ(N).

To exhibit the breadth of this phenomenon, our results address three basic, well-known load-balancing models. The basic structures of their queueing models are depicted in Figure 1. The first two models address balancing among a fixed number, N, of resources, by routing a stream of arrivals to one of the resources according to the state of the system. One is the aforementioned SQ(d), and another is *join the least workload*, LW(d) (here and throughout, $1 < d \leq N$, and N denotes the total number of queues). Under LW(d), arrivals are routed to a queue containing the least work (measured in time units) among d queues that, again, are chosen at random. Our results on the latter policy directly apply to another load balancing scheme, namely *Redundancy-d* (R(d)), where arriving jobs are replicated d times, and the replicas, referred to as *tasks*, are sent to randomly chosen queues. When the first among the d tasks is accepted to service, the other $d-1$ replicas are canceled. In a sense that is made precise in §2.2, R(d) and LW(d) are mathematically equivalent. (This relation has been noticed before in [49] for the case $d = N$, but was not proven.) The third model studied in this paper consists of a single resource shared by N job classes, where each class has its own arrival stream and the job that receives service is always selected from the queue that is longest. Within each class, service is given by order of arrival. This discipline is referred to as *longest queue first* (LQF).

In all three models the deviation process is a natural performance measure, and it is of prime importance to characterize its scale. As mentioned above, our first main contribution, SDDP, is to show that its scale is sub-diffusive. (We reserve the term SSC to the context in which it is used in the literature, that is, when the critical load condition holds.) By no means does SDDP hold in complete generality; we provide a simple counterexample to make this point. The conditions under

which we establish SDDP are formulated in terms of the range of values that the arrival and service intensities take. For example, under SQ(d), assume that the rate of arrivals varies within an interval of the form $[\bar{\lambda}_{min}n, \bar{\lambda}_{max}n]$, and the rates of service all vary within an interval $[\bar{\mu}_{min}n, \bar{\mu}_{max}n]$. We find a condition of the form $\bar{\mu}_{max} - \bar{\mu}_{min} < \bar{\lambda}_{min}\varphi^*$ that assures SDDP, where φ^* is a constant that depends on N and d. The SDDP result obtained in the case of LW(d) (and relevant also to R(d)) addresses the *workload deviation process*, defined in terms of workload, analogously to the queue length deviation process. The one for LQF again refers to queue length. Let us reiterate that the reason for putting together results on three different models in this paper is the aim to show that the described phenomenon is broad; indeed, it occurs under three different policies. Although the intuition is similar in the three models, the details of the proof turn out to be quite different in each case.

Whereas our main goal is to study SDDP, our results have further significant implications. The first is a set of asymptotic optimality (AO) results. In a heavy traffic setting, AO usually refers to showing that a certain policy performs better than any other policy in the sense that the performance measure (such as the sum of all queue lengths or the total workload) under this policy is stochastically dominated by that under any other policy up to an error of order $o(n^{1/2})$. The most relevant result of this kind is [18], which proves AO of a balanced routing policy proposed in that paper, in a setting that allows for server heterogeneity. Their technique also covers SQ(d). Our AO results are similar in that they cover SQ(d) (as well as R(d)/LW(d)), and that the optimality guarantee is $o(n^{1/2})$. The fundamental difference is that we achieve these guarantees for time-varying systems not restricted to heavy traffic. Again, the setting accommodates changes in behavior as far as the load criticality is concerned, and the optimality guarantees are not sensitive to that. To the best of our knowledge, these are the first to address $o(n^{1/2})$ optimality guarantees for time-varying systems. The only paper the authors are aware of, that establishes AO in time-varying queues, is [19]. In this paper, a large family of queueing systems with a fixed number of stations and routing dynamics is studied. AO is considered and proved there in the *fluid scale*, and thus the optimality guarantees provided are $o(n)$.

Another set of results implied by the SDDP is obtained by simply specializing to the heavy traffic setting. In heavy traffic, SSC is obtained, as well as convergence of the diffusion-scale queue length (workload in the case of R(d)/LW(d)). For heterogeneous servers, the only paper the authors are aware of, proving SSC, is the aforementioned paper [18]. However, the conditions under which SSC holds under [18] and under our results are distinct, and it is not clear whether one set of conditions contains the other.

The seminal papers by Bramson [11] and Williams [69] have played a pivotal role in the area by setting the ground for a large number of results on SSC of various queueing models. However, their methodology is restricted to critically loaded systems, and is not valid for time-varying queues of the kind treated in this paper. Thus our results do not rely on this set of techniques, and use in fact quite different methods.

1.1 Prior work

1.1.1 Time-varying queues

Many real-life queueing systems have time-varying characteristics. This has been recognized in call center queueing systems [15] and in data center traffic studies [37]. In wireless networks, multiple users transmit data on a shared channel, the capacity of which varies with time randomly and

asynchronously for different users [3], [62].

Time-varying queueing systems have been thoroughly analyzed. These systems may be non-stationary, operate for a fixed period of time, and may encounter light, heavy or overloading traffic during their operation (see [19] and references therein). Work reported in [50], [51] and [52] provided a heuristic analysis of transportation systems. In [38], a method to approximate the transient distribution of the queue length process was developed, using the stationary measure of an appropriate Markov chain. It became known as the point-wise stationary approximation. This approach was later justified in [46], [47] and [48], using a technique called uniform acceleration as an asymptotic expansion of the transient distribution. The authors of [45] considered the time-varying $M_t/M_t/1$ queue and developed fluid and diffusion limit theorems using strong approximation methods. The limiting queue length approximating process is characterized via a directional derivative of the Skorohod map, as the system moves through under-, over- and critically-loaded regimes. For further results in this direction, and analysis of more complicated systems, see [3], [31], [42], [44], and [68]. Recent work in [33] and [34] considered transitory queueing systems that operate for a finite period of time or for a finite number of arriving users, and provided fluid and diffusion limit results using population acceleration techniques.

1.1.2 Load balancing policies

SQ(d) has attracted a considerable amount of attention in recent years. We mention several related results and refer the reader to [9], [18], and references therein. In an asymptotic regime corresponding to $N \to \infty$ and sub-critical load, some randomized schemes (e.g., SQ(d)) are well known to result in dramatic improvement of resource sharing, as expressed by the fact that the equilibrium tail decay rate is doubly exponential under load-balancing whereas this rate is only exponential otherwise (see [64] and the recent developments in [12], and references therein). SQ(d) has recently been considered under a heavy traffic regime in [24], in a setting where $N \to \infty$. In that, it is similar to the setting of [64]. Their main result is the scaling limit description of queue lengths in terms of a deterministic infinite system of equations. For a fixed number of homogeneous servers, the authors of [32] analyze a policy related to but distinct from SQ(d) in heavy traffic. Specifically, with a fixed probability, jobs are routed to a queue chosen uniformly at random, and otherwise, to the shorter among two neighboring queues, where the pair is again chosen uniformly at random. It is shown that the heavy traffic limit is identical to that attained under SQ(d) with $d = N$.

LQF has been studied mainly in the context of switch scheduling and wireless communications. Existing work is concerned with stability (e.g., [6], [43], [55], and [65]) and performance analysis (e.g., [5], [7], [40], [41], [61] and [71]). To the best of our knowledge, LQF was not analyzed at the diffusion scale.

Our original motivation for studying load balancing stemmed from our interest in redundancy routing and its recent technological uses. We provide a more detailed review on R(d) since the literature concerned with its analytical analysis is relatively scarce. In recent years there has been an extensive use of large-scale systems such as data centers to provide end-users low latency of response to requests. Low latency corresponds to the high quality of users' experience, which in turn affects company revenue ([13], [60]). To achieve low response time, computations, or *requests*, are broken down into sub-computations, or *jobs*, which run in parallel on different servers. The outcomes are then added up to form a response to the request. An example of an implementation of this paradigm is MapReduce [21], where the *map* phase corresponds to the breaking down to

sub-computations and the *reduce* phase to the collection of the results. In the queueing literature, models in which a response is formed when all jobs complete are referred to as *fork-join networks* [54].

The slowest job, referred to in the literature as an *outlier* or *straggler*, determines the overall response time. This type of behavior may cause substantial problems in large-scale systems ([2], [66], [70], [72]), because the more jobs are computed in parallel, the greater the probability that one of them is slow. R(d) is often used to address this issue as follows. In addition to having requests broken down into jobs, every job is replicated into several identical tasks that are routed to distinct servers. The job is completed when the *first* of its tasks is finished being processed. In different versions of this model, replicas can be canceled at any time, or not at all. This method is discussed in [20], implemented in sophisticated schedulers in parallel computing systems, and is reported to greatly reduce the effect of stragglers in [1], [2] and [72]. The intuition behind the efficiency of R(d) is that jobs that wait for their turn in several queues simultaneously have a higher probability of completing faster. The tradeoff lies in wasted capacity (in models where replicas are not canceled, or where cancellation is delayed) and in the possible overhead incurred by performing cancellations and keeping track of jobs and their replicas. The idea of redundancy can be applied to a large variety of networks, not only ones that model data centers, and seems to be significant to explore in a broader context.

There is a substantial body of research on redundancy routing in the context of application-specific, sophisticated schedulers, and simulation-based performance analysis ([2], [53]). As for theoretical analysis is concerned, few results exist (note that redundancy is different from fork-join networks and from networks with flexible servers; see [29] for a thorough comparison). The main concern in the existing literature is the effect of redundancy on the mean delay of jobs in various queueing models ([10], [35], [39], [58], [59], [63]). Other papers study the tradeoff between latency and resource usage when scheduling a fixed number of tasks ([36]). In [29], for a queueing model where classes of costumers may choose whether to use redundancy, the limiting distribution of the state of the queueing system is found, as well as that of the delay. The paper analyzes the effect of redundant classes on non redundant classes, and empirically compares redundancy to other choices that may reduce delay, such as SQ(d). The authors of [28] study several implications of relaxing the common assumption that the processing times of replicas of the same job in different servers are independent. In [30], exact expressions are derived for the mean delay of jobs when redundancy is used for a fixed number of servers, as well as expressions for the delay distribution as the number of servers goes to infinity. It is shown that the largest marginal benefit from using redundancy is obtained when replicating each job twice as opposed to not replicating at all.

The cancellation mechanism in the R(d) model is such that replicas of a job are canceled when the first *begins* processing, as opposed to cancelling *after* the first task is completed, which is more often found in the literature ([29], [35], [36], [39]). The motivation for working with the former is that cancelling a task while it is being processed may incur intolerable delay or overhead, and is often impossible in practice. Another consideration is that under the former option no capacity is wasted on tasks that are to be canceled.

1.2 Organization and notation

The organization of this paper is as follows. This section concludes with the introduction of some notation that is used in the sequel. In §2 the models are described and the main results are stated,

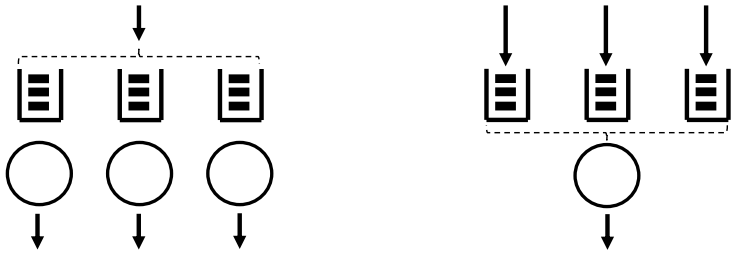

Figure 1: Basic structure of the queueing model for SQ(d), LW(d) and R(d) (left) and for LQF (right).

where §2.1, §2.2 and §2.3, respectively, are devoted to SQ(d), R(d)/LW(d) and LQF. The proofs, which appear in §3, are provided, respectively, in §3.1, §3.2 and §3.3.

Notation. For $a, b \in \mathbb{R}$, the maximum (resp., minimum) is denoted by $a \vee b$ (resp., $a \wedge b$), and $a^+ = a \vee 0$, $a^- = (-a) \vee 0$. For $x \in \mathbb{R}^k$ (k a positive integer), $\|x\|$ denotes the ℓ_1 norm. Denote $\mathbb{R}_+ = [0, \infty)$, and for $f : \mathbb{R}_+ \to \mathbb{R}^k$, $\|f\|_T = \sup_{t\in[0,T]} \|f(t)\|$, and, for $\theta > 0$,

$$w_T(f, \theta) = \sup_{0\le s<u\le s+\theta\le T} \|f_u - f_s\|. \tag{1}$$

For a Polish space $\mathcal{S}$, let $\mathcal{C}_\mathcal{S}[0, T]$ and $\mathcal{D}_\mathcal{S}[0, T]$ denote the set of continuous and, respectively, cadlag functions $[0, T] \to \mathcal{S}$. Write $\mathcal{C}_\mathcal{S}$ and $\mathcal{D}_\mathcal{S}$ for the case where $[0, T]$ is replaced by $\mathbb{R}_+$. Endow $\mathcal{D}_\mathcal{S}$ with the Skorohod J_1 topology. Write $X_n \Rightarrow X$ for convergence in distribution.

A sequence of processes X_n with sample paths in $\mathcal{D}_\mathcal{S}$ is said to be *$\mathcal{C}$-tight* if it is tight and every subsequential limit has, with probability 1, sample paths in $\mathcal{C}_\mathcal{S}$. For $m \in \mathbb{R}$ and $\sigma \in \mathbb{R}$, an (m, σ^2)-*Brownian motion* (BM) is a 1-dimensional BM starting from zero, having drift m and infinitesimal covariance σ^2. The Skorohod map Γ from $\mathcal{D}(\mathbb{R}_+ : \mathbb{R})$ to itself is defined by

$$\Gamma[\phi](t) = \phi_t - \inf_{s\le t}(\phi_s \wedge 0), \qquad t \ge 0.$$

If $\{\beta_t\}$ is an (m, σ^2)-BM for some $m \in \mathbb{R}$ and $\sigma \in (0, \infty)$, then the process $\{\beta^0_t\}$ defined by the pathwise transformation $\beta^0 = \Gamma[\beta]$ is referred to as *an (m, σ^2)-reflecting Brownian motion (RBM).*

Given a time interval $J = [t_1, t_2] \subset \mathbb{R}_+$ we write $f[t_1, t_2] = f[J] = f(t_2) - f(t_1)$, for any function f defined on $\mathbb{R}_+$. We use shorthand notation for integration as follows: $\mathfrak{I}f(t) = \int_0^t f(u)du$.

2 Models and results

2.1 Results on SQ(d)

2.1.1 Model and scaling

We begin by describing the SQ(d) model and the scaling under consideration. A sequence of models indexed by $n \in \mathbb{N}$ is defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ as follows. A fixed number $N \in \mathbb{N}$ ($N > 1$) of servers, labeled by $\{1, \ldots, N\}$, have infinite sized buffers, one dedicated to each server.

Each server is non-idling and offers service on a first-come-first-served basis. There is a single stream of arriving jobs. The kth job to arrive (after time zero) is referred to as *job* k. For simplicity, we assume that the system starts empty. The stream is modeled by an inhomogeneous Poisson process of rate $\lambda^n(\cdot)$, where λ^n is a deterministic, Borel measurable, locally integrable function $\mathbb{R}_+ \to \mathbb{R}_+$. To this end, a rate-1 Poisson process A is given, and the arrival counting process A^n is defined via the relation

$$A^n(t) = A\Big(\int_0^t \lambda^n(s)ds\Big), \quad t \geq 0. \tag{2}$$

For $i \in \{1, \ldots, N\}$, let $\{T_i(l) : l \in \mathbb{N}\}$ be a sequence of strictly positive i.i.d. RVs with mean 1 and variance $0 < V_{T_i(1)} < \infty$. It is assumed that the size of the kth job served by server i is given by $T_i^n(k)$; that is, if that server works at *constant* rate μ_i^n then it takes $T_i^n(k) := T_i(k)/\mu_i^n$ units of time to process. However, our model will allow for the rates of service to vary over time. To this end, let $\{S_i\}$ be independent rate-1 renewal processes with inter-event times given by $\{T_i(k)\}$, namely

$$S_i(t) = \sup\Big\{l \geq 0 : \sum_{k=1}^{l} T_i(k) \leq t\Big\}. \tag{3}$$

Let deterministic, Borel measurable, locally integrable functions μ_i^n, $i = 1, \ldots, N$ be given. Server i works at rate $\mu_i^n(t)$ at the time when it has been busy for t units of time. Thus the potential service process is given by

$$S_i^n(t) = S_i\Big(\int_0^t \mu_i^n(s)ds\Big), \quad t \geq 0, \tag{4}$$

namely, $S_i^n(t)$ is the number of job departures from queue i by the time the corresponding server has been busy for t units of time. The $N + 1$ processes $\{T_i\}$ and A are assumed to be mutually independent.

Denote by $I_i^n(t)$ the cumulative idle time of server i at time t. Next, A_i^n and D_i^n are counting processes for arrivals into buffer i, and departures from buffer i, respectively. Let $Q_i^n(t)$ denote the queue length of the ith queue in the nth system at time t (this includes the job being processed at that time, if there is one), and denote $Q^n = (Q_1^n, \ldots, Q_N^n)$. The relations between the processes A_i^n, S_i^n, D_i^n, I_i^n and Q_i^n are expressed by the following equations:

$$D_i^n(t) = S_i^n(t - I_i^n(t)) \tag{5}$$

$$Q_i^n(t) = A_i^n(t) - D_i^n(t), \tag{6}$$

and the non-idling property

$$\int_{[0,\infty)} Q_i^n(t) dI_i^n(t) = 0. \tag{7}$$

For integer $1 < d \leq N$, let $\mathcal{B}_d := \{b \subset \{1, ..., N\}, |b| = d\}$ be the set of all d-size subsets of $\{1, ..., N\}$. Let $\{B_k\}$ be an i.i.d. sequence of set valued random variables (RVs) distributed uniformly over $\mathcal{B}_d$. The subset B_k of the set of all servers $\{1, \ldots, N\}$ is associated with job k: it is used to determine which buffer this job is to be sent to. Specifically, job k is routed to the buffer $j_k \in B_k$ that contains the least number of jobs at the moment of its arrival. Ties are broken by prioritizing buffer i over buffer j whenever $1 \leq i < j \leq N$. Thus, if τ_k^n denotes the time of arrival of the kth job in the nth system, the job is routed to the buffer

$$j_k = \min\{i \in B_k : Q_i^n(\tau_k^n-) \leq Q_j^n(\tau_k^n-) \text{ for all } j\}, \tag{8}$$

and the arrival processes A_i^n satisfy

$$A_i^n(t) = \sum_{k=1}^{A^n(t)} \mathbb{1}_{\{j_k=i\}}. \tag{9}$$

Given the primitives A, $\{T_i\}$ and $\{B_k\}$, equations (2)–(9) uniquely determine the processes A_i^n, D_i^n, I_i^n and Q_i^n for $i \in \{1, \ldots, N\}$. We thus use this set of equations as the rigorous definition of these processes. We refer to this model as SQ(d).

2.1.2 SDDP result

In what follows, we provide a sufficient condition for SDDP. Essentially, SDDP holds if the input rate difference between servers with short and long queues is enough to overcome their service rate difference, thus pushing their queue lengths towards equalization. To state our condition, first, we look at the maximal and minimal arrival and service rates during a time interval $[0, T]$. To this end, let scaled versions of λ^n and μ_i^n be given by

$$\bar{\lambda}^n(t) = n^{-1}\lambda^n(t), \qquad \bar{\mu}_i^n(t) = n^{-1}\mu_i^n(t), \qquad t \geq 0, \tag{10}$$

and denote

$$\bar{\lambda}_{min}(T) = \inf_n \inf_{t\in[0,T]} \bar{\lambda}^n(t), \qquad \bar{\lambda}_{max}(T) = \sup_n \sup_{t\in[0,T]} \bar{\lambda}^n(t), \tag{11}$$

$$\bar{\mu}_{min}(T) = \min_i \inf_n \inf_{t\in[0,T]} \bar{\mu}_i^n(t), \quad \bar{\mu}_{max}(T) = \max_i \sup_n \sup_{t\in[0,T]} \bar{\mu}_i^n(t). \tag{12}$$

Second, the condition involves the fraction of the arrival stream a server receives depending on its relative queue size. For the ith shortest queue, this fraction is given by

$$\varphi_i = \frac{\binom{N-i}{d-1}}{\binom{N}{d}} = \begin{cases} \frac{d}{N}\frac{N-i}{N-1}\cdots\frac{N-i-(d-2)}{N-(d-1)}, & \text{if } 1 \leq i \leq N-d+1, \\ 0, & \text{if } i > N-d+1. \end{cases} \tag{13}$$

Define

$$\varphi^* = \min_{1\leq i\leq N-1} \left\{ \left(\frac{1}{i} \sum_{j:1\leq j\leq i} \varphi_j \right) - \varphi_{i+1} \right\}. \tag{14}$$

Our main result on the SQ(d) model establishes SDDP.

Theorem 2.1 (SQ(d) SDDP)**.** *Fix T and assume that $\bar{\lambda}_{max}(T) < \infty$ and $\bar{\mu}_{max}(T) < \infty$. Assume, moreover, that*

$$\bar{\mu}_{max}(T) - \bar{\mu}_{min}(T) < \bar{\lambda}_{min}(T)\,\varphi^*. \tag{15}$$

Then, as $n \to \infty$,

$$n^{-1/2} \max_{1\leq i,j\leq N} \|Q_i^n - Q_j^n\|_T \to 0 \text{ in probability.}$$

In the special case where the service rates are constant over time and equal to each other, the l.h.s. of (15) vanishes, and therefore this condition is clearly valid. In general, the l.h.s. of (15) may be regarded as a measure of heterogeneity, namely the degree to which service rates vary with time and across servers. The condition (15) specifies the maximal range of this quantity under which our result is valid.

Note that we do not claim that this condition is necessary. Yet, it is interesting to examine how it varies as one changes d and N. First, one may verify that φ^* increases with d. Therefore our sufficient condition allows for a greater degree of service-rate heterogeneity as d increases. Also, φ^* decreases with N. Hence the aforementioned condition allows for a smaller degree of heterogeneity with increasing N. The first observation supports the intuition that the policy achieves better balance when the fraction of sampled queues increases. The second observation supports the same intuitive claim, since when N is increased, the fraction of sampled queues, d/N, decreases.

Remark 2.2. *Some simplified expressions for φ^* are available for certain values of d:*

$$\varphi^* = \begin{cases} \frac{2}{N(N-1)}, & \text{if } d = 2, \\ \frac{1}{N-1}, & \text{if } d = N-1, \\ 1, & \text{if } d = N. \end{cases}$$

The following simple example addresses the well-known fact that SDDP need not always hold, even under critical load.

Example 2.3. *Let $N = 3$ and $d = 2$. Assume that the service and arrival rates do not vary in time, and are given by $\lambda^n(t) = \lambda n$ and $\mu_i^n(t) = \mu_i n$. Let $\mu_1 = \mu_2 = 1$, $\mu_3 = 10$ and $\lambda = 12$ (note that the critical load condition holds). Then the shortest, mid-size and longest queues receive 2/3, 1/3 and 0 fraction of the incoming traffic, respectively. The fastest server processes at rate $10n$, i.e., more than the maximal possible input rate $\frac{2}{3}12n = 8n$. By standard tools one can show that its diffusion-scale perturbations away from zero converge to zero. On the other hand, the remaining two servers, whose total capacity is $2n$, have to process jobs that arrive at rate $4n$ at least. Again, it is easy to see that their diffusion-scale sum diverges. Thus the difference between Q_i^n is of order n, and therefore SDDP does not hold. We note that this example is related to questions of instability that have been addressed in [27].*

2.1.3 Applications to AO and SSC

Next we present an AO result that follows from Theorem 2.1. We specialize to the case of constant service rates. The arrival rate however, may still vary with time. Recall that the service times $T_i^n(k)$ are given by $T_i(k)/\mu_i^n$. We further assume that for different i the primitives $T_i(k)$ have the same distribution, i.e., the service times at the different stations can be obtained as accelerated versions of a single i.i.d. sequence, where acceleration depends on the server.

To state the result we need to introduce the notion of *nominal workload*. Whereas the workload at buffer i, denoted by $W_i^n(t)$ is, by definition, the time that server i takes to process all jobs present in the buffer at that time, the nominal workload in this buffer is given by $\mu_i^n W_i^n(t)$. It represents the time it takes a *nominal* server, i.e., a server that processes at rate 1, to complete this work. The *total nominal workload* is defined as

$$Z^n(t) = \sum_i \mu_i^n W_i^n(t). \tag{16}$$

We also need the notion of an arbitrary sequence of policies. Fix n. Consider the definition of the model in §2.1 with the process j_k (defined in (8)) replaced by an arbitrary process taking values in $\{1,\ldots,N\}$ (still denoted by j_k). Leave all other ingredients of the model as they are in §2.1. We identify the term 'policy' with the routing process j_k, and refer to the resulting processes $A_i^n, D_i^n, I_i^n, Q_i^n$ and W_i^n as the processes corresponding to this policy. Note that this is a very broad definition of a policy, since it allows, for example, for the routing process j_k to depend on future events.

Proposition 2.4 (SQ(d) Asymptotic Optimality)**.** *Fix T, and let the hypotheses of Theorem 2.1 hold with constant service rates. Assume, moreover, that for different i the primitives $T_i(k)$ have the same distribution. Let an arbitrary sequence of policies be given, and denote by $\tilde{Z}^n(t)$ the corresponding total nominal workload process. Keep the notation $Z^n(t)$ for SQ(d). Then there exists a sequence of RVs δ_n (that does not depend on the given sequence of policies) converging to zero in probability, such that, for all n and $t \in [0,T]$,*

$$n^{-1/2} Z^n(t) \le n^{-1/2} \tilde{Z}^n(t) + \delta_n.$$

Remark 2.5. *The total nominal workload* (16) *is an interesting objective function to be optimized since it is closely related to delay. The workload at a given buffer, k, at a given time, t, is the delay that a new customer will experience if it joins queue k at time t. It is therefore natural to consider as a performance measure a weighted sum of workload over all buffers. The result we present does not address a weighted sum with arbitrary weights but the special case, that is still an interesting one, where all weights equal 1, as is considered, for example, in [18].*

Finally, we further specialize to the conventional heavy traffic setting. Thus the arrival and service intensities do not vary with time, and we assume that there exist constants $\lambda \in (0,\infty)$, $\hat{\lambda} \in \mathbb{R}$, $\mu_i \in (0,\infty)$ and $\hat{\mu}_i \in \mathbb{R}$, $i \in \{1,\ldots,N\}$, such that

$$\lim_{n\to\infty} n^{-1/2}(\lambda^n - n\lambda) = \hat{\lambda}, \tag{17}$$

$$\lim_{n\to\infty} n^{-1/2}(\mu_i^n - n\mu_i) = \hat{\mu}_i, \qquad i = 1,\ldots,N. \tag{18}$$

Moreover, a critical load condition is assumed. The first order terms in the n-scale arrival rate and total processing rate are given by λ and $\sum_{j=1}^N \mu_j$, respectively. Therefore, criticality corresponds to

$$\lambda = \sum_{i=1}^N \mu_i. \tag{19}$$

We denote $\mu_{max} = \max_i \mu_i$ and $\mu_{min} = \min_i \mu_i$. Denote the diffusion-scaled queue lengths by

$$\hat{Q}_i^n = n^{-1/2} Q_i^n, \qquad i = 1,\ldots,N, \tag{20}$$

and denote $\hat{Q}^n = (\hat{Q}_1^n,\ldots,\hat{Q}_N^n)$. We will say that *the queue lengths exhibit diffusive behavior* if, for each T, the sequence of RVs $\|\hat{Q}^n\|_T$ is tight. It is clear that in the case $d = 1$ (which is excluded from our analysis and corresponds to uniformly random routing of tasks), queue lengths are non-diffusive under our assumptions. This is simply due to the fact that for the slowest server, $\lambda/N > \mu_{min}$ (except when the μ_i's are all equal). As a result, $\|Q^n\|_T$ are of order n rather than $\sqrt{n}$. For $d \ge 2$

there may be a dramatic improvement in performance, in the sense that the queue lengths exhibit diffusive behavior.

Denote $\hat{\mu}_0 = N^{-1}(\hat{\lambda} - \sum_i \hat{\mu}_i)$ and $\sigma_0^2 = N^{-2}(\lambda + \sum_{i=1}^N \mu_i V_{T_i(1)})$. The next result shows SSC and convergence.

Proposition 2.6 (SQ(d) in heavy traffic)**.** *Assume* (17), (18), (19) *hold and*

$$\mu_{max} - \mu_{min} < \lambda\varphi^*. \tag{21}$$

Let β^0 *be a* $(\hat{\mu}_0, \sigma_0^2)$*-RBM. Then, as* $n \to \infty$*,*

$$(\hat{Q}_1^n, \hat{Q}_2^n, \ldots, \hat{Q}_N^n) \Rightarrow (\beta^0, \beta^0, \ldots, \beta^0).$$

A similar result in presence of server heterogeneity has been shown in [18] (in fact, for general renewal arrivals). However, the condition (21) and the condition on the rates assumed in [18] are distinct, and it is not clear if one of them implies the other.

2.2 Results on R(d) and LW(d)

2.2.1 Model and scaling

In the LW(d) model, for each of the arriving jobs, d servers are chosen uniformly at random and every arriving job is sent to the server with the minimal amount of workload, defined as the time it will take the server to finish all its existing work (to this end, it is assumed that workload in the buffers is known to the decision maker). The precise setting is as follows. First, we keep some elements of the model from §2.1, namely we consider a sequence of models indexed by $n \in \mathbb{N}$, with $N > 1$ non-idling servers with infinite size buffers. Service is given on a first-come-first-served basis and we assume the system starts empty. The job arrival process is an inhomogeneous Poisson process A^n with rate $\lambda^n(\cdot)$, as in (2), and the RVs B_k are as before. Also, j_k denotes the server chosen for job k, but when using the LW(d) policy.

The service requirements follow a slightly different structure. Unlike the treatment of the SQ(d) model, here we assume that the service rates are constant. The reason for this assumption is that when service rates vary with time, a routing scheme based on workload (that we have defined as the time it will take the server to complete current assigned work) has to rely on calculations involving future service rates of all servers, making the model equations and their analysis rather complex. Thus we are given an i.i.d. sequence $\{T(k)\}$ of real-valued, strictly positive RVs, with mean 1 and variance $0 < V_{T(1)} < \infty$. The service duration associated with job k, on the event that it is served by server i, is then given by

$$T_i^n(k) = T(k)/\mu_i^n, \tag{22}$$

where μ_i^n is the *constant* mean service rate of server i. We assume the processes A, $\{T(k)\}$ and $\{B_k\}$ are mutually independent.

Denote by $W_i^n(t)$ the workload of server i at time t. To reiterate, it is defined as the time, that it will take server i to finish its existing work at time t. Let $W^n = (W_1^n, \ldots, W_N^n)$. Denote by $W_i^{A,n}(t)$ the cumulative work that arrived to server i until time t, and let $I_i^n(t)$ denote the cumulative idle time. Then the relations between the processes A^n, T^n, W^n, $W^{A,n}$ and I^n are given

via the following equations:

$$I_i^n(t) = \int_0^t 1_{\{W_i^n(s)=0\}} ds, \tag{23}$$

$$W_i^n(t) = W_i^{A,n}(t) - (t - I_i^n(t)) = W_i^{A,n}(t) - \int_0^t 1_{\{W_i^n(s)>0\}} ds, \tag{24}$$

$$W_i^{A,n}(t) = \sum_{k=1}^{A^n(t)} T_i^n(k) \mathbb{1}_{\{j_k=i\}}. \tag{25}$$

Job k is routed to the server $j_k \in B_k$ with the least amount of workload. Ties are broken by prioritizing buffer i over buffer j whenever $1 \le i < j \le N$. Thus, if τ_k^n denotes the time of arrival of the kth job in the nth system, the job is routed to the buffer

$$j_k = \min\{i \in B_k : W_i^n(\tau_k^n-) \le W_j^n(\tau_k^n-) \text{ for all } j\}, \tag{26}$$

and the arrival processes A_i^n satisfy the relation given in (9), now with j_k chosen according to (26). Given the primitives A^n, $\{T(k)\}$ and $\{B_k\}$, equations (22)–(26), together with (2) and (9), uniquely define the unknowns W_i^n, $W_i^{A,n}$ and I_i^n for $i \in \{1, \dots, N\}$. As in §2.1, we use the aforementioned set of equations as a means of rigorously defining these processes.

2.2.2 R(d) and its equivalence to LW(d)

We now argue that the LW(d) policy, in which the decision maker is aware of the workload of the servers, is mathematically equivalent to a policy based on redundancy and cancellations, where the workload information is not available to the decision maker.

Recall that in the LW(d) model, when a job arrives, d servers are chosen uniformly at random and the job joins the buffer corresponding to the server with the least amount of workload. In R(d), instead of selecting one out of the d buffers, the job is replicated d times, and these replicas are sent to the d (again, randomly chosen) buffers. When the first of the d replicas of a given job reaches a server, it is accepted to service and all remaining replicas are canceled (removed from the system). Ties are broken, as before, according to the ordering of the buffers. Note that this policy does not use any information on the workload. The times a job begins and completes service are defined as the corresponding times for the replica that makes it to the server.

We argue that R(d) is equivalent to LW(d). More precisely,

Proposition 2.7. *If the same stochastic primitives are used under both policies then, for each job, the buffer selected by LW(d) is the same as the buffer containing the replica that eventually makes it to a server under R(d).*

It is an immediate consequence of this claim that each job starts and completes service at the same time under both policies. The proof appears in §3.2.

An intuitive explanation is as follows. In terms of workload, only the replica that will not be canceled adds work to the server it joined. The workloads of the other $d - 1$ servers remain unchanged. Interestingly, the replica that eventually receives service (i.e., reaches a server first) is the one that joined the buffer corresponding to the server with the least amount of workload. Essentially, R(d) is a method of implementing LW(d) without observing the workload.

We proceed with presenting the LW(d) model. By the equivalence discussed above, all subsequent results on LW(d) apply to R(d) as well.

2.2.3 SDDP result and its applications

Let $\bar{\lambda}_{min}(T)$ and $\bar{\lambda}_{max}(T)$ be defined as in (11). Let μ_i^n satisfy (18) for some constants $\mu_i \in (0,\infty)$ and $\hat{\mu}_i \in \mathbb{R}$. Let $\mu_{max} = \max_i \mu_i$ and $\mu_{min} = \min_i \mu_i$. Our main result on LW(d) (equivalently, R(d)) is the following.

Theorem 2.8 (R(d)/LW(d) SDDP)**.** *Fix T. Assume that $\bar{\lambda}_{min}(T) > 0$ and $\bar{\lambda}_{max}(T) < \infty$. Assume moreover that*

$$\frac{\mu_{max}}{\mu_{min}} < \frac{N-1}{N-d}. \tag{27}$$

Then, as $n \to \infty$,

$$n^{1/2} \max_{1 \le i,j \le N} \|W_i^n - W_j^n\|_T \to 0 \text{ in probability.}$$

As in the case of SQ(d), the above result implies AO at sub-diffusive scale. Recall the definition (16) of the total nominal workload, $Z^n(t)$. The definition of an arbitrary policy is analogous to the one given before Proposition 2.4.

Proposition 2.9 (R(d)/LW(d) Asymptotic Optimality)**.** *Fix $T > 0$, and let the hypotheses of Theorem 2.8 hold. Let an arbitrary sequence of policies be given, and denote by $\tilde{Z}^n(t)$ the corresponding total nominal workload process. Keep the notation $Z^n(t)$ for LW(d). Then there exists a sequence of RVs δ_n (that does not depend on the given sequence of policies) converging to zero in probability, such that, for all n, $t \in [0,T]$ and $\omega \in \Omega$,*

$$n^{-1/2} Z^n(t) \le n^{-1/2} \tilde{Z}^n(t) + \delta_n.$$

Once again, we examine our main result under the conventional heavy traffic assumptions. Denote $\hat{W}^n = n^{1/2} W^n$. Assume that the constant rates λ^n and μ_i^n satisfy (17), (18) and (19).

Proposition 2.10 (R(d)/LW(d) in heavy traffic)**.** *Assume that* (27) *holds. Let β^0 be a (μ_w, σ_w^2)-RBM, where $\mu_w = N\hat{\mu}_0 \lambda^{-1}$ and $\sigma_w^2 = (V_{T(1)} + 1)\lambda^{-1}$. Then, as $n \to \infty$,*

$$(\hat{W}_1^n, \hat{W}_2^n, \ldots, \hat{W}_N^n) \Rightarrow (\beta^0, \beta^0, \ldots, \beta^0).$$

The heuristic by which SSC implies diffusive behavior is as in the previous subsection, and in fact the proof of Proposition 2.10 based on Theorem 2.8 is similar to that of Proposition 2.6 based on Theorem 2.1. On the other hand, as far as the SDDP (and specifically SSC) results are concerned, we use a different argument for SQ(d) and LW(d), and this results in different sufficient conditions. Neither of the conditions (15) and (27) is stronger than the other. For example, one can check that for $(N, d, \mu) = (4, 3, (1, 1, 2, 3))$, the condition (15) for SSC under SQ(d) holds while the condition (27) for LW(d) does not. On the other hand, for $(N, d, \mu) = (5, 4, (5, 5, 6, 7, 19))$, (15) does not hold while (27) does. An interesting question left open is to find necessary and sufficient conditions for SSC under SQ(d) or LW(d).

2.3 Results on LQF

2.3.1 Model and scaling

Consider a single server that processes jobs belonging to N classes, where each class has a dedicated infinite capacity buffer in which a queue can form. The stream of arriving jobs into each buffer is

modeled by a modulated renewal process. To this end, for $i \in \{1, \ldots, N\}$, let $\{E_i(l) : l \in \mathbb{N}\}$ be a sequence of strictly positive i.i.d. RVs with mean 1 and variance $0 < V_{E_i(1)} < \infty$. Let $\{A_i\}$ be renewal processes with inter-event times given by $E_i(k)$, namely

$$A_i(t) = \sup\Big\{l \geq 0 : \sum_{k=1}^{l} E_i(k) \leq t\Big\}. \tag{28}$$

The arrival counting processes $\{A_i^n\}$ are assumed to be given by

$$A_i^n(t) = A_i\Big(\int_0^t \lambda_i^n(s)ds\Big),\ t \geq 0, \tag{29}$$

where $\{\lambda_i^n\}$ are deterministic functions.

For $i \in \{1, \ldots, N\}$, let $\{T_i(l) : l \in \mathbb{N}\}$ be a sequence of strictly positive i.i.d. RVs with mean 1 and variance $0 < V_{T_i(1)} < \infty$. Let $\{S_i\}$ be renewal processes with inter-event times given by $T_i(k)$, namely

$$S_i(t) = \sup\Big\{l \geq 0 : \sum_{k=1}^{l} T_i(k) \leq t\Big\}. \tag{30}$$

It is assumed that the size of the kth job of class i in the nth system is given by $T_i^n(k)$; that is, if the server works at *constant* rate μ_i^n then it takes $T_i^n(k) := T_i(k)/\mu_i^n$ units of time to process. As in the SQ(d) model, we assume that the rates of service may vary over time. To this end, it is assumed that deterministic functions $\mu_i^n(\cdot)$, $i = 1, \ldots, N$, are given. The server works at rate $\mu_i^n(t)$ at the time that it has been busy with class i for t units of time. Thus the potential service processes are given by

$$S_i^n(t) = S_i\Big(\int_0^t \mu_i^n(s)ds\Big), \quad t \geq 0. \tag{31}$$

The $2N$ processes $\{E_i\}$ and $\{T_i\}$ are assumed to be mutually independent.

Let $Q_i^n(t)$ denote the queue length in the ith buffer in the nth system at time t (including the class-i job being processed at that time, if there is one). Let $Q^n = (Q_1^n, \ldots, Q_N^n)$ denote the queue length process. Denote by $A_i^n(t)$ and $D_i^n(t)$ the counting processes corresponding to arrivals and departures respectively, from buffer i until time t. Denote by $B_i^n(t)$ the cumulative amount of time the server has served class-i jobs until time t. Assuming as before that the system is initially empty, the relations between these processes are given by

$$Q_i^n(t) = A_i^n(t) - D_i^n(t), \tag{32}$$

$$D_i^n(t) = S_i^n(B_i^n(t)). \tag{33}$$

The policy acts as follows. When there are jobs in the system, the job at the head of the line of the class with the longest queue receives service. Thus, within each class, service is given by the order of arrival, and the policy is non-idling and preemptive. The server resumes working on a previously preempted job from where it has left off. Let $LQ^n(t)$ denote the label of the class with the longest queue at time t, where ties are broken by prioritizing lower indexes. Then

$$LQ^n(t) = \min\{i \in \{1, \ldots, N\} : Q_i^n(t-) \geq Q_j^n(t-) \text{ for all } j\}, \tag{34}$$

$$B_i^n(t) = \int_0^t 1_{\{LQ^n(s)=i, Q_i^n(s)>0\}} ds. \tag{35}$$

Given the primitive processes $\{E_i\}$ and $\{T_i\}$, equations (28)–(35) uniquely determine the processes A_i, D_i, Q_i and B_i for $i \in \{1, \ldots, N\}$. We refer to this model as LQF.

2.3.2 SDDP result and its applications

Scaled versions of λ_i^n and μ_i^n are given by

$$\bar{\lambda}_i^n(t) = n^{-1}\lambda_i^n(t), \qquad \bar{\mu}_i^n(t) = n^{-1}\mu_i^n(t), \qquad t \geq 0. \tag{36}$$

Denote

$$\bar{\lambda}_{min}(T) = \min_i \inf_n \inf_{t\in[0,T]} \bar{\lambda}_i^n(t), \qquad \bar{\lambda}_{max}(T) = \max_i \sup_n \sup_{t\in[0,T]} \bar{\lambda}_i^n(t), \tag{37}$$

$$\bar{\mu}_{min}(T) = \min_i \inf_n \inf_{t\in[0,T]} \bar{\mu}_i^n(t), \qquad \bar{\mu}_{max}(T) = \max_i \sup_n \sup_{t\in[0,T]} \bar{\mu}_i^n(t). \tag{38}$$

Theorem 2.11 (LQF SDDP)**.** *Fix T and assume that $\bar{\lambda}_{max}(T) < \infty$ and $\bar{\mu}_{max}(T) < \infty$. Assume moreover that*

$$\bar{\lambda}_{max}(T) - \bar{\lambda}_{min}(T) < N^{-1}\bar{\mu}_{min}(T). \tag{39}$$

Then, as $n \to \infty$,

$$n^{-1/2} \max_{1\leq i,j\leq N} \|Q_i^n - Q_j^n\|_T \to 0 \text{ in probability.}$$

To state a consequence regarding the conventional heavy traffic setting, assume $\lambda_i^n(t) = \lambda_i^n$ and $\mu_i^n(t) = \mu_i^n$. Assume moreover there exist parameters λ_i, $\hat{\lambda}_i$ that satisfy

$$\lim_{n\to\infty} n^{-1/2}(\lambda_i^n - n\lambda_i) = \hat{\lambda}_i, \tag{40}$$

and μ_i, $\hat{\mu}_i$ that satisfy (18). Assume the critical loading condition

$$\sum_i \lambda_i {\mu_i}^{-1} = 1. \tag{41}$$

Denote $c_l = (\sum_i \mu_i{}^{-1})^{-1}$, $\mu_l = c_l \sum_i (\hat{\lambda}_i \mu_i{}^{-1} - \lambda_i \hat{\mu}_i \mu_i{}^{-2})$ and $\sigma_l^2 = {c_l}^2 \sum_i \lambda_i \mu_i^{-2}(V_{E_i(1)} + V_{T_i(1)})$. The next result shows SSC and convergence.

Proposition 2.12 (LQF in heavy traffic)**.** *Assume* (40)*,* (18) *and* (41) *hold, and*

$$\lambda_{max} - \lambda_{min} < N^{-1}\mu_{min}. \tag{42}$$

Let β^0 be a (μ_l, σ_l^2)-RBM. Then, as $n \to \infty$,

$$(\hat{Q}_1^n, \hat{Q}_2^n, \ldots, \hat{Q}_N^n) \Rightarrow (\beta^0, \beta^0, \ldots, \beta^0).$$

3 Proofs

3.1 Proofs for SQ(d)

In this subsection we analyze the SQ(d) model and prove Theorem 2.1 and Proposition 2.6. The proof of Proposition 2.4 is based on that of Proposition 2.9, and is therefore deferred to the next section.

We begin by introducing some notation used in the proof. Centered and scaled versions of the primitive processes A and S_i are given by

$$\hat{A}^n(t) = n^{-1/2}(A(nt) - nt), \quad \hat{S}_i^n(t) = n^{-1/2}(S_i(nt) - nt). \tag{43}$$

These processes jointly converge to mutually independent BMs with zero drift and infinitesimal variance 1 and $\mu_i V_{T_i(1)}$, respectively (see §17 of [8]). Recalling from section 1.2 the notation $\mathfrak{I}$ for integration, by (2) and (4), the processes A^n and S_i^n are given as $A^n = A \circ \mathfrak{I}\lambda^n$ and $S_i^n = S_i \circ \mathfrak{I}\mu_i^n$, respectively. Therefore

$$n^{-1/2}(A^n(t) - \mathfrak{I}\lambda^n(t)) = n^{-1/2}(A(n\mathfrak{I}\bar{\lambda}^n(t)) - n\mathfrak{I}\bar{\lambda}^n(t)) = (\hat{A}^n \circ \mathfrak{I}\bar{\lambda}^n)(t), \tag{44}$$

and

$$n^{-1/2}(S_i^n(t) - \mathfrak{I}\mu_i^n(t)) = (\hat{S}_i^n \circ \mathfrak{I}\bar{\mu}_i^n)(t), \tag{45}$$

where $\bar{\lambda}^n(t)$ and $\bar{\mu}_i^n(t)$ are defined in (10).

Given n and t, let $l(t) = l^n(t)$ denote the unique permutation of $\{1, \ldots, N\}$ for which

$$Q^n_{l_1(t)}(t) \le Q^n_{l_2(t)}(t) \le \cdots \le Q^n_{l_N(t)}(t),$$

and whenever $Q_i^n(t) = Q_j^n(t)$ and $i < j$, one has $l_i(t) < l_j(t)$. The inverse permutation is denoted by $L(t) = (L_1(t), \ldots, L_N(t))$. For example, if $N = 4$, and $Q^n(t) = (9, 7, 8, 7)$, then $l(t) = (2, 4, 3, 1)$ and $L(t) = (4, 1, 3, 2)$. In addition, for any vector $v \in \mathbb{R}^N$ we use the notation $v_{(i)}$ for the ith smallest coordinate of v, counting multiplicity. That is, $\{v_{(i)}\}$ satisfy

$$v_{(1)} \le \cdots \le v_{(N)},$$

such that for each $i \in \{1, \ldots, N\}$, $v_{(i)} = v_j$ for exactly one j. In particular, one has $Q^n_{l_i(t)}(t) = Q^n_{(i)}(t)$ and $Q^n_{(L_i(t))}(t) = Q_i^n(t)$.

By thinning of (modulated) Poisson processes [23], the input stream of jobs to the buffers, *ordered according to the queue sizes,* form N independent (modulated) Poisson processes, with time-dependent rates given by

$$\Lambda_i^n(t) = \lambda^n(t) \frac{\binom{N-i}{d-1}}{\binom{N}{d}} = \lambda^n(t)\varphi_i. \tag{46}$$

Specifically, the shortest queue arrival intensity is $\Lambda_1^n(t)$, whereas the $d-1$ longest queues do not receive any arrivals (by (13), $\varphi_i = 0$ for $i \in \{N-d, \ldots, N\}$). The intensity of arrivals into queue i is thus given by the stochastic process

$$\lambda_i^n(t) := \Lambda^n_{L_i(t)}(t) = \lambda^n(t)\varphi_{L_i(t)}. \tag{47}$$

We express this by stating that given any fixed subset $S \subset \{1, \ldots, N\}$, there exists a standard Poisson process π such that

$$\sum_{i \in S} A_i^n(t) = \pi\Big(\int_0^t \sum_{i \in S} \lambda_i^n(s) ds \Big). \tag{48}$$

Proof of Theorem 2.1. The crux of the argument is to identify and analyze, for given T and ε, an event containing the event $\{\max_{i,j} \|\hat{Q}_i^n - \hat{Q}_j^n\|_T > \varepsilon\}$, whose probability can be estimated effectively. In doing so, the $\mathcal{C}$-tightness of the rescaled processes $\hat{A}^n$ and $\hat{S}_i^n$ (which follows from their convergence to BMs) plays a key role.

For $T > 0$ and $\varepsilon > 0$, denote

$$\tau := \tau^n = \inf \Big\{ t : \text{ there exists } i < N \text{ such that } \sup_{t \le T} \Big[\hat{Q}_{(i+1)}^n(t) - \hat{Q}_{(i)}^n(t) \Big] > \varepsilon 4^i \Big\}. \tag{49}$$

If for some $T > 0$ and $\varepsilon' > 0$,

$$\{ \max_{i,j} \|\hat{Q}_i^n - \hat{Q}_j^n\|_T > \varepsilon' \} \tag{50}$$

holds, then there exists $t \le T$ for which $\hat{Q}_{(N)}^n(t) - \hat{Q}_{(1)}^n(t) > \varepsilon'$, and consequently, there exists $i < N$ for which

$$\hat{Q}_{(i+1)}^n(t) - \hat{Q}_{(i)}^n(t) > (N-1)^{-1} \varepsilon'.$$

Thus if we let $\varepsilon = 4^{-N}(N-1)^{-1}\varepsilon'$ then $\tau \le T$ also holds on the event (50), and consequently, for any $T' > T$, $\tau < T'$ holds on this event. As a result, in order to show that, as $n \to \infty$, $\mathbb{P}(\{\max_{i,j} \|\hat{Q}_i^n - \hat{Q}_j^n\|_T > \varepsilon) \to 0$ for arbitrary T and ε, it suffices to prove that $\mathbb{P}(\tau < T) \to 0$ for arbitrary T and ε.

Fixing T and ε, we analyze the event $\{\tau < T\}$. The processes were constructed in such a way that $\hat{Q}^n$ have right-continuous sample paths. Therefore, on the event $\{\tau < T\}$ there exists $i < N$ such that $\hat{Q}_{(i+1)}^n(\tau) - \hat{Q}_{(i)}^n(\tau) > \varepsilon 4^i$. Fix such i. Because $\hat{Q}^n(0) = 0$, there exist times t earlier than τ when $\hat{Q}_{(i+1)}^n(t) - \hat{Q}_{(i)}^n(t) \le \varepsilon 4^{i-1}$. Let then

$$\sigma := \sigma^n = \sup\{ t < \tau : \hat{Q}_{(i+1)}^n(t) - \hat{Q}_{(i)}^n(t) \le \varepsilon 4^{i-1} \},$$

and set $J = [\sigma, \tau]$. Note that the processes $Q^n(t)$ can only jump by 1, therefore by (20), the jumps of the processes $\hat{Q}_i^n$ are all of size $n^{-1/2}$. Assume that n is sufficiently large so that these jumps are of size smaller than $\varepsilon/2$. Since τ and σ are defined in such a way that $\hat{Q}_{(i)}^n$ and $\hat{Q}_{(i+1)}^n$ are kept apart throughout the interval J at a distance greater than twice the size of the jumps, it follows that the collection of indices of the i shortest queues does not vary over that time interval (however, the ordering within this collection may change). That is, the set $M(t) := \{l_1(t), \ldots, l_i(t)\}$ remains fixed for all $t \in [\sigma, \tau]$. We denote this set by M. Let also $m = l_{i+1}(\sigma)$. We shall focus on the evolution of the queues $\hat{Q}_j^n$, $j \in M$ and on the queue $\hat{Q}_m^n$, over the time interval J. Figure 2 illustrates our construction.

First, note that m is not a member of M, and because $M(t)$ does not vary with t during J, it is not a member of $M(\tau)$. Hence

$$\hat{Q}_m^n(\tau) \ge \hat{Q}_{(i+1)}^n(\tau) \ge \hat{Q}_{(i)}^n(\tau) + \varepsilon 4^i.$$

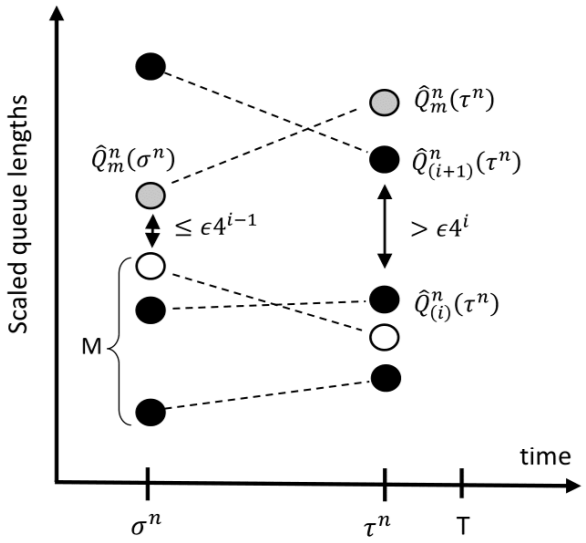

Figure 2: Illustration of the construction for SQ(d). The circles represent the scaled queue lengths in the buffers. The white (gray) circle corresponds to the ith $(i+1)$ smallest scaled queue length at time σ^n. The set M of i smallest queue lengths does not change during the interval.

Therefore, if $\hat{Q}^n_M(t) = i^{-1}\sum_{j\in M}\hat{Q}^n_j(t)$ denotes the arithmetic mean over the set M,

$$\hat{Q}^n_m(\tau) \geq \hat{Q}^n_M(\tau) + \varepsilon 4^i. \tag{51}$$

By the definition of σ and the bound on the jump sizes,

$$\hat{Q}^n_m(\sigma) - \hat{Q}^n_{(i)}(\sigma) \leq \varepsilon 4^{i-1} + \varepsilon.$$

Since $\sigma < \tau$, it follows from the definition of τ that $\hat{Q}^n_{(i)}(\sigma) - \hat{Q}^n_{(1)}(\sigma) \leq \varepsilon(4 + 4^2 + \cdots + 4^{i-1})$ $= \varepsilon(4^i - 4)/3$. Thus

$$\hat{Q}^n_m(\sigma) - \hat{Q}^n_M(\sigma) \leq \hat{Q}^n_m(\sigma) - \hat{Q}^n_{(1)}(\sigma) \leq \varepsilon(4^{i-1} + 1 + 4^i/3 - 4/3) < \varepsilon\frac{7}{12}4^i. \tag{52}$$

Combining (51) and (52), multiplying by $n^{1/2}$ and recalling from section 1.2 our notation $f[J]$ for $f(t_2) - f(t_1)$ where $J = [t_1, t_2]$, we obtain

$$n^{1/2}\varepsilon \leq Q^n_m[J] - Q^n_M[J]. \tag{53}$$

The quantity i, the index m and the set M are all random. Appealing to the union bound, we have

$$\mathbb{P}(\tau < T) \leq \sum_{i_0=1}^{N-1}\sum_{m_0=1}^{N}\sum_{M_0:|M_0|=i_0}\mathbb{P}(\Omega^n(i_0, m_0, M_0)), \quad \text{where}$$

$$\Omega^n(i_0, m_0, M_0) = \Big\{\text{there exist } s,t \in [0,T], s<t, \text{ such that } n^{1/2}\varepsilon \leq Q^n_{m_0}[s,t] - Q^n_{M_0}[s,t],$$
$$\text{and } Q^n_{m_0}(u) > Q^n_{(i_0)}(u) \geq Q^n_j(u) \text{ for all } j \in M_0, u \in [s,t]\Big\}. \tag{54}$$

To prove the result, it thus suffices to show that each summand $\mathbb{P}(\Omega^n(i_0, m_0, M_0))$ converges to zero with n. We focus on fixed, deterministic i_0, m_0 and M_0, and, with a slight abuse of notation, refer to them again as i, m and M. We also write J for the corresponding time interval $[s,t]$, and Ω^n for $\Omega^n(i,m,M)$.

On the event Ω^n, using (6), we have

$$Q^n_m[J] - Q^n_M[J] = A^n_m[J] - D^n_m[J] - A^n_M[J] + D^n_M[J], \tag{55}$$

where $X_M = i^{-1}\sum_{j\in M} X_j$ for $X = Q^n, A^n$ and D^n. On this event, server m is non-idling during $[s,t]$. Therefore, if we denote for $j \in \{1,\dots,N\}$, $J_j = [s - I^n_j(s), t - I^n_j(s)]$, we have $D^n_m[J] = S^n_m[J_m]$. Servers $j \in M$ need not be non-idling (may idle) during this time window, but since S^n_j has non-decreasing sample paths, we still have a valid inequality. Namely, we have by (5) that $D^n_j[J] \le S^n_j[J_j]$ for $j \in M$. Hence by (54) and (55),

$$n^{1/2}\varepsilon \le A^n_m[J] - S^n_m[J_m] - A^n_M[J] + i^{-1}\sum_{j\in M} S^n_j[J_j]. \tag{56}$$

We appeal to (48) twice: with $S = \{m\}$ and with $S = M$. We emphasize that m and M are now deterministic. Thus there exist standard Poisson processes π_1 and π_2 such that

$$A^n_m(u) = \pi_1(nf^n_1(u)), \qquad f^n_1(u) = \int_0^u \frac{\Lambda^n_{L_m(v)}(v)}{n}dv = \int_0^u \bar\lambda^n(v)\varphi_{L_m(v)}dv,$$

where identities (10) and (47) are used, and, denoting $\Lambda^n_{\{1,i\}}(u) = \sum_{j\le i}\Lambda^n_j(u)$ and $\varphi_{\{1,i\}} = \sum_{j\le i}\varphi_j$,

$$A^n_M(u) = i^{-1}\pi_2(nf^n_2(u)), \qquad f^n_2(u) = \int_0^u \frac{\Lambda^n_{\{1,i\}}(v)}{n}dv = \int_0^u \bar\lambda^n(v)dv\,\varphi_{\{1,i\}},$$

where we used the fact that M is equal to the set $\{1,\dots,i\}$. Thus if we let

$$\hat\pi^n_k(u) = n^{-1/2}(\pi_k(nu) - nu), \qquad k = 1,2, \tag{57}$$

then dividing by $n^{1/2}$ in (56), appealing to (45), gives

$$\varepsilon \le \hat\pi^n_1 \circ f^n_1[J] - \hat S^n_m \circ \mathfrak{I}\bar\mu^n_m[J_m] - i^{-1}\hat\pi^n_2 \circ f^n_2[J] + i^{-1}\sum_{j\in M}\hat S^n_j \circ \mathfrak{I}\bar\mu^n_j[J_j] + Y^n, \tag{58}$$

where

$$Y^n = n^{1/2}\Big\{f^n_1[J] - i^{-1}f^n_2[J] - \int_{J_m}\bar\mu^n_m(u)du + i^{-1}\sum_{j\in M}\int_{J_j}\bar\mu^n_j(u)du\Big\}. \tag{59}$$

For a bound on Y^n, note that for $v \in J$, $\varphi_{L_m(v)} \le \varphi_{i+1}$, hence

$$f^n_1[J] - i^{-1}f^n_2[J] \le (\varphi_{i+1} - i^{-1}\varphi_{\{1,i\}})\int_J \bar\lambda^n(v)dv = -\varphi^*\int_J \bar\lambda^n(v)dv \le -\varphi^* \inf_{u\in[0,T]}\bar\lambda^n(u)(t-s).$$

Moreover, using $|J_m| = t - s$ and $|J_j| \le t - s$, $j \in M$,

$$-\int_{J_m}\bar\mu^n_m(u)du + i^{-1}\sum_{j\in M}\int_{J_j}\bar\mu^n_j(u)du \le \Big[\max_{k\le N}\sup_{u\in[0,T]}\bar\mu^n_k(u) - \min_{k\le N}\inf_{u\in[0,T]}\bar\mu^n_k(u)\Big](t-s).$$

Hence by (15), for some $\varepsilon_0 > 0$, $Y^n \le -\varepsilon_0 n^{1/2}(t-s)$ on the event Ω^n. As a result, using the notation in (1), on the same event we have

$$\varepsilon \le \sum_{k=1}^{2} w_T(\hat{\pi}_k^n \circ f_k^n, t-s) + \sum_{j=1}^{N} w_T(\hat{S}_j^n \circ \mathfrak{I}\bar{\mu}_j^n, t-s) - \varepsilon_0 n^{1/2}(t-s). \tag{60}$$

Fix a sequence $\{r_n\}$ such that $r_n \to 0$ and $n^{1/2} r^n \to \infty$. Considering the two cases $t-s \le r_n$ and $t-s > r_n$, it follows from (60) that $\mathbb{P}(\Omega^n) \le p_1^n + p_2^n$, where

$$p_1^n = \mathbb{P}\Big(\sum_{k=1}^{2} w_T(\hat{\pi}_k^n \circ f_k^n, r_n) + \sum_{j=1}^{N} w_T(\hat{S}_j^n \circ \mathfrak{I}\bar{\mu}_j^n, r_n) \ge \varepsilon \Big),$$

$$p_2^n = \mathbb{P}\Big(2\sum_{k=1}^{2} \|\hat{\pi}_k^n \circ f_k^n\|_T + 2\sum_{j=1}^{N} \|\hat{S}_j^n \circ \mathfrak{I}\bar{\mu}_j^n\|_T \ge \varepsilon_0 n^{1/2} r_n \Big).$$

The processes $\hat{\pi}_k^n$ and $\hat{S}_j^n$ are $\mathcal{C}$-tight, as processes that converge to BM. Moreover, by the definition of f_k^n and the assumed uniform bound on $\bar{\lambda}^n(t)$, there exists a deterministic constant c_1 (independent of n) such that f_k^n are bounded by c_1 on the time interval $[0,T]$, as well as c_1-Lipschitz on it. By the assumed uniform bound on $\bar{\mu}_i^n(t)$, a similar assertion holds for the processes $\mathfrak{I}\bar{\mu}_i^n$. It follows that $\hat{\pi}_k^n \circ f_k^n$ and $\hat{S}_j^n \circ \mathfrak{I}\bar{\mu}_j^n$ are also $\mathcal{C}$-tight. As a result, both p_1^n and p_2^n converge to zero. This shows that $\mathbb{P}(\Omega^n(i,m,M)) \to 0$ for every i, m, M, and therefore by (54) completes the proof. □

Proof of Proposition 2.6. Let $Q_S^n = \sum_{i=1}^N Q_i^n$ and $\hat{Q}_S^n = n^{-1/2} Q_S^n$ denote the sum and normalized sum of queue lengths. It follows from Theorem 2.1 that SSC holds, i.e.,

$$\hat{Q}_i^n - N^{-1}\hat{Q}_S^n \Rightarrow 0, \tag{61}$$

as $n \to \infty$, for every i. Hence it suffices to show that $\hat{Q}_S^n \Rightarrow N\beta^0$.

To this end, we use relations (20), (5), (6) and (43), with which we can write the balance equation $\hat{Q}_S^n = \hat{X}^n + \hat{Y}^n$, where, denoting $\bar{\lambda}^n = n^{-1}\lambda^n$ and $\bar{\mu}_i^n = n^{-1}\mu_i^n$,

$$\hat{X}^n(t) = \hat{A}^n(\bar{\lambda}^n t) - \sum_{i=1}^{N} \hat{S}_i^n(\bar{\mu}_i^n(t - I_i^n(t))) + v^n t, \qquad \hat{Y}^n = n^{-1/2}\sum_{i=1}^{N} \mu_i^n I_i^n(t), \tag{62}$$

$$v^n = n^{-1/2}\Big(\lambda^n - \sum_{i=1}^{N} \mu_i^n\Big). \tag{63}$$

By (17), (18) and (19), $v^n \to N\hat{\mu}_0 = \hat{\lambda} - \sum_i \hat{\mu}_i$.

Next, we identify a sequence of events whose probability converges to 1 by virtue of the SSC, on which $\hat{Q}_S^n$ is given, up to a small error term, as the image of $\hat{X}^n$ under the Skorohod map. Fix $T > 0$. It follows from (61) that for every $\varepsilon > 0$, $\mathbb{P}(\Omega^{n,\varepsilon}) \to 1$ as $n \to \infty$, where

$$\Omega^{n,\varepsilon} = \{\text{for all } t \in [0,T], \hat{Q}_S^n(t) > \varepsilon \text{ implies } \min_i \hat{Q}_i^n(t) > 0\}. \tag{64}$$

As a result, there exists a sequence $\varepsilon_n \to 0$, $\varepsilon_n > 0$, such that $\mathbb{P}(\Omega^{n,\varepsilon_n}) \to 1$ as $n \to \infty$. Fix such a sequence. Denote $\Omega^n = \Omega^{n,\varepsilon_n}$. Define

$$\tilde{Q}_S^n = (\hat{Q}_S^n - \varepsilon_n)^+, \qquad e_n = \tilde{Q}_S^n - \hat{Q}_S^n. \tag{65}$$

Note that $\|e_n\|_T \leq \varepsilon_n$. Now, by (7), $\int Q_i^n(t) dI_i^n(t) = 0$ for all i. Moreover, $\tilde{Q}_S^n(t) > 0$ if and only if $\hat{Q}_S^n(t) > \varepsilon_n$. As a result, on the event Ω^n, we also have $\int \tilde{Q}_S^n(t) dI_i^n(t) = 0$ for all i, and therefore $\int \tilde{Q}^n(t) d\hat{Y}^n(t) = 0$.

If we let $\tilde{X}^n = \hat{X}^n + e_n$ then the following relations hold on the event Ω^n:

$$\begin{aligned}
&\hat{Y}^n \text{ has continuous, nonnegative and nondecreasing sample paths,}\\
&\tilde{Q}_S^n \text{ is non negative,}\\
&\tilde{Q}_S^n = \tilde{X}^n + \hat{Y}^n,\\
&\int \tilde{Q}^n(t) d\hat{Y}^n(t) = 0.
\end{aligned} \tag{66}$$

These relations imply that, on the event Ω^n, $\tilde{X}^n$ determines $\tilde{Q}^n$ in terms of the Skorohod map, namely $\tilde{Q}_S^n = \Gamma[\tilde{X}_S^n]$ (see [17], Ch. 6, p. 128).

Let us argue that the RVs $\|\hat{Y}^n\|_T$ are tight. We have on Ω^n, $\hat{Y}^n(t) = \tilde{Q}_S^n(t) - \tilde{X}^n(t) = -\inf_{s \leq t} \tilde{X}^n(s) \wedge 0$. Moreover, using (62) and letting $c_1 := \sup_n[\bar{\lambda}^n \vee \max_i \bar{\mu}_i^n] < \infty$,

$$\|\tilde{X}^n\|_T \leq \|\hat{X}^n\| + \varepsilon_n \leq \|\hat{A}^n\|_{c_1 T} + \sum_i \|\hat{S}_i^n\|_{c_1 T} + |v^n| T.$$

Hence, by the weak convergence of $\hat{A}^n$ and $\hat{S}_i^n$, the convergence of v^n and the fact $\mathbb{P}(\Omega^n) \to 1$, the tightness of $\|\hat{Y}^n\|_T$ follows. As a result, noting the relation (62) of I_i^n to $\hat{Y}^n$ and recalling that μ_i^n are asymptotic to $\mu_i n$, it follows that $I_i^n \Rightarrow 0$ for every i. Using this in the equation for $\hat{X}^n$ in (62) and, again, the convergence of $\hat{A}^n(\bar{\lambda}^n \cdot) - \sum \hat{S}_i^n(\bar{\mu}_i^n \cdot)$ to a $(0, \sigma^2)$-BM with $\sigma^2 = \lambda + \sum_{i=1}^N \mu_i V_{T_i(1)}$, it follows that $\hat{X}^n \Rightarrow \beta$, where β is an $(N\hat{\mu}_0, \sigma^2)$-BM. The continuous mapping theorem then applies for $\tilde{Q}_S^n$, and in turn for $\hat{Q}_S^n$, giving $\hat{Q}_S^n \Rightarrow \Gamma[\beta]$. Thus $N^{-1}\hat{Q}_S^n$ converges to a $(\hat{\mu}_0, \sigma_0^2)$-RBM. This completes the proof. □

3.2 Proofs for R(d)/LW(d)

Here we prove Theorem 2.8 and then Propositions 2.9, 2.10 and 2.7. The proof of Proposition 2.4 is based on that of Proposition 2.9, and is also presented in this section. As far as Theorem 2.8 is concerned, there are similarities to the approach taken in §3.1, but the details are different, and in particular, the construction of an interval $[\sigma, \tau]$ is different. Recall that $\hat{W}_i^n = n^{1/2} W_i^n$. For the order statistics under LW(d), we use notation similar to that of §3.1, working with $\hat{W}^n$ in place of $\hat{Q}^n$. Thus l and L are processes defined analogously to those in §3.1, such that for every t and i, $\hat{W}_{l_i(t)}^n(t) = \hat{W}_{(i)}^n(t)$ and $\hat{W}_{(L_i(t))}^n(t) = \hat{W}_i^n(t)$. We also keep the notation A_i^n from (46) and $\lambda_i^n(t)$ from (47), where now the latter stands for the stochastic intensity of the arrival process into queue i under LW(d). Specifically, if A_i^n is the counting process for jobs routed to i, then relation (48) holds here as well.

Proof of Theorem 2.8. (R(d)/LW(d) SDDP). Recall that the vector $T^n(k)$ gives the service duration of job k. That is, provided that job k is routed to server i, the duration is given by

$T_i^n(k) = T(k)/\mu_i^n$. Let $k_i^*(j)$ denote the index of the jth job that is processed by server i. That is, $k_i^*(j) = k$ if job k is the jth job routed to server i. Then, instead of using (25) for $W_i^{A,n}$, we can write

$$W_i^{A,n}(t) = \sum_{j=1}^{A_i^n(t)} T_i^n(k_i^*(j)). \tag{67}$$

A use of (48) with $S = \{i\}$ gives the following. For each i there exists a standard Poisson process π_i such that the arrival process into buffer i is given by

$$A_i^n(u) = \pi_i(nf_i^n(u)), \qquad f_i^n(u) = \int_0^u \frac{\lambda_i^n(v)}{n} dv = \int_0^u \frac{\Lambda_{L_i(v)}^n}{n} dv. \tag{68}$$

The workload arrival process $W_i^{A,n}$ of (67) can be written as a time change of a renewal reward process. Namely, if we let

$$R_i^n(u) = \sum_{j=1}^{\pi_i(nu)} T_i^n(k_i^*(j)), \tag{69}$$

then $W_i^{A,n} = R_i^n \circ f_i^n$. Indeed, by construction, the sequence $T_i^n(k_i^*(j))$, $j \in \mathbb{N}$ is i.i.d. for each i, with common distribution identical to that of $T_i^n(1)$, and π_i, which depends only on the arrival mechanism, is independent of this sequence. Therefore R_i^n are renewal reward processes. Recall the notation $\bar{\mu}_i^n = \mu_i^n/n$ and $\bar{\theta}_i^n = n/\mu_i^n$. Let

$$\hat{R}_i^n(u) = n^{1/2}(R_i^n(u) - \bar{\theta}_i^n u) = \bar{\theta}_i^n n^{-1/2} \left(\sum_{j=1}^{\pi_i(nu)} T(k_i^*(j)) - nu \right).$$

Using the FCLT for renewal reward processes, Theorem 7.4.1 of [67], and the fact that $\bar{\theta}_i^n$ is asymptotic to $1/\mu_i$, the sequence $\hat{R}_i^n$ converges weakly to a $(0, \sigma_i^2)$-BM, with $\sigma_i^2 = \mu_i^{-2}(V_{T(1)} + 1)$. In this proof, the specific parameters of the limiting BM are not used; however, the consequential $\mathcal{C}$-tightness of $\hat{R}_i^n$ shall be used.

Using (24) we can write an equation for $\hat{W}_i^n = n^{1/2} W_i^n$, namely

$$\begin{aligned} \hat{W}_i^n(u) &= n^{1/2} R_i^n \circ f_i^n(u) - n^{1/2} \int_0^u 1_{\{\hat{W}_i^n(v) > 0\}} dv \\ &= \hat{R}_i^n \circ f_i^n(u) + n^{1/2} \bar{\theta}_i^n f_i^n(u) - n^{1/2} \int_0^u 1_{\{\hat{W}_i^n(v) > 0\}} dv. \end{aligned} \tag{70}$$

Fix $\varepsilon > 0$ and $T > 0$. Define

$$\tau = \tau^n = \inf\{t : \hat{W}_{(N)}^n(t) - \hat{W}_{(1)}^n(t) > 2\varepsilon\}, \tag{71}$$

and let $\Omega^n = \{\tau^n < T\}$. Our goal is to show that $\mathbb{P}(\Omega^n) \to 0$ as $n \to \infty$.

Let $i_1 = l_i(\tau)$ and $i_2 = l_2(\tau)$. Thus i_1 [resp., i_2] is the index of the server with the minimal [maximal] workload at τ. Define

$$\sigma = \sigma^n = \sup\{t < \tau : \hat{W}_{i_2}^n(t) - \hat{W}_{i_1}^n(t) \le \varepsilon\}.$$

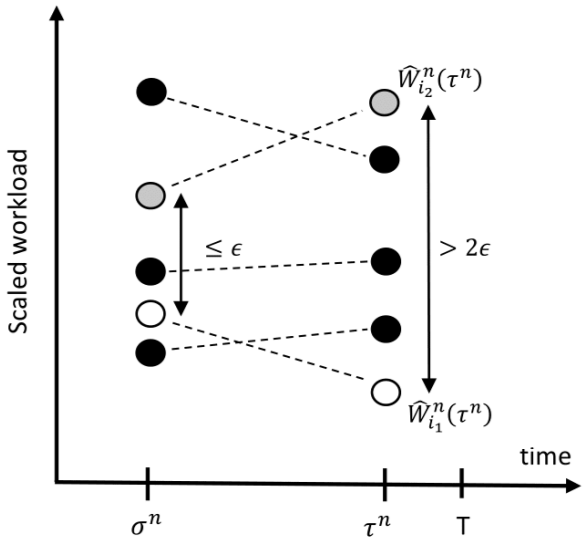

Figure 3: Illustration of the construction for LW(d). The circles represent the scaled workload in the buffers. The white (gray) circle corresponds to buffer i_1 (i_2), with the smallest (largest) workload at time τ^n.

Then on the event Ω^n we have $\hat{W}^n_{i_2}(\tau)-\hat{W}^n_{i_1}(\tau) > 2\varepsilon$, $\hat{W}^n_{i_2}(\sigma-)-\hat{W}^n_{i_1}(\sigma-) \leq \varepsilon$. We denote $J=[\sigma,\tau]$. In the case $\sigma=\tau$, J is identical to $\{\tau\}$. Figure 3 illustrates our construction.

In this proof we slightly modify the notation $f[\tilde{J}]$ for a function f possessing left limits defined on $\mathbb{R}_+$ and an interval $\tilde{J}=[s,t]$, namely we shall write $f[\tilde{J}]$ for $f(t)-f(s-)$. With this notation, we have

$$\hat{W}^n_{i_2}[J]-\hat{W}^n_{i_1}[J] \geq \varepsilon, \text{ and } \hat{W}^n_{i_2}(u) > \hat{W}^n_{i_1}(u) \text{ for all } u \in J=[\sigma,\tau].$$

Again, it is more convenient to work with deterministic indices in place of i_1 and i_2 that are random, and this may be achieved using the union bound. Indeed, we have

$$\begin{aligned}
&\mathbb{P}(\tau < T) \leq \sum_{j_1=1}^{N} \sum_{j_2 \neq j_1} \mathbb{P}(\Omega^n(j_1,j_2)), \quad \text{where} \\
&\Omega^n(j_1,j_2) = \Big\{ \text{there exist } s,t \in [0,T],\ s \leq t, \text{ such that, with } J=[s,t],\ \hat{W}^n_{j_2}[J]-\hat{W}^n_{j_1}[J] \geq \varepsilon, \\
&\qquad \text{and } \hat{W}^n_{j_2}(u) > \hat{W}^n_{j_1}(u) \text{ for all } u \in J \Big\}. \qquad (72)
\end{aligned}$$

The result will follow once we show that each $\mathbb{P}(\Omega^n(j_1,j_2)) \to 0$. We thus fix a deterministic pair (j_1,j_2), and, without loss of generality assume that $(j_1,j_2)=(1,2)$. The corresponding time interval is denoted by $J=[s,t]$, while $\Omega^n(1,2)$ is abbreviated as Ω^n.

To this end, note that by (70) and (72),

$$\begin{aligned}
\varepsilon &\leq \hat{R}^n_2 \circ f^n_2[J] - \hat{R}^n_1 \circ f^n_1[J] + \Delta^n + n^{1/2} \int_s^t (1_{\{\hat{W}^n_1(u)>0\}} - 1_{\{\hat{W}^n_2(u)>0\}}) du \\
&\leq \hat{R}^n_2 \circ f^n_2[J] - \hat{R}^n_1 \circ f^n_1[J] + \Delta^n, \qquad (73)
\end{aligned}$$

where we denote $\Delta^n = n^{1/2}(\bar{\theta}_2^n f_2^n[J] - \bar{\theta}_1^n f_1^n[J])$, and in the second inequality we used the fact that server 2 is non-idling on the interval $[s,t]$, under the event Ω^n. Denoting $\bar{A}_i^n = n^{-1}A_i^n$, using (68) and then (47), the term Δ^n is given by

$$\begin{aligned}\Delta^n &= n^{1/2}\int_s^t (\bar{\theta}_2^n \bar{A}_{L_2(u)}^n - \bar{\theta}_1^n \bar{A}_{L_1(u)}^n)du \\ &= n^{1/2}\int_s^t \bar{\lambda}^n(u)(\bar{\theta}_2^n \varphi_{L_2(u)} - \bar{\theta}_1^n \varphi_{L_1(u)})du. \end{aligned} \tag{74}$$

It follows from (72) that during $[s,t]$, $L_2(u) > L_1(u)$ and therefore either $\varphi_{L_2(u)} < \varphi_{L_1(u)}$, or $\varphi_{L_2(u)} = \varphi_{L_1(u)} = 0$ for all $u \in [s,t]$. Hence whenever $\varphi_{L_1(u)} = 0$, the integrand in (74) is zero. Thus

$$\Delta^n = n^{1/2}\int_s^t \bar{\lambda}^n(u)(\bar{\theta}_2^n \varphi_{L_2(u)} - \bar{\theta}_1^n \varphi_{L_1(u)})1_{\{\varphi_{L_1(u)}>0\}}du. \tag{75}$$

Next, we derive an upper bound on Δ^n. First, consider the case where $\varphi_{L_1(u)} > \varphi_{L_2(u)} > 0$. By (13),

$$\frac{\varphi_i}{\varphi_{i+1}} = \frac{N-i}{N-i-(d-1)} = 1 + \frac{d-1}{N-i-(d-1)}, \qquad 1 \le i \le N-d.$$

Therefore φ_i is decreasing in i for $i \le N-d+1$ and for any i and j such that $\varphi_i > \varphi_j > 0$,

$$\frac{\varphi_i}{\varphi_j} \ge \frac{\varphi_1}{\varphi_2} = \frac{N-1}{N-d}.$$

Thus, whenever $\varphi_i > \varphi_j > 0$, using the fact that $\bar{\theta}_i^n$ is asymptotic to μ_i^{-1},

$$\begin{aligned}\bar{\theta}_2\varphi_j - \bar{\theta}_1\varphi_i &\le \mu_{min}^{-1}\varphi_j - \mu_{max}^{-1}\varphi_i \\ &= \frac{\varphi_j}{\mu_{max}}\Big(\frac{\mu_{max}}{\mu_{min}} - \frac{\varphi_i}{\varphi_j}\Big) \\ &\le \frac{\varphi_j}{\mu_{max}}\Big(\frac{\mu_{max}}{\mu_{min}} - \frac{N-1}{N-d}\Big) \\ &\le -2\tilde{C}_1 < 0,\end{aligned}$$

where on the last line we have used (27), and $\tilde{C}_1 > 0$ is a suitable constant. For the case where $\varphi_{L_1(u)} > \varphi_{L_2(u)} = 0$, we obtain

$$\mu_2^{-1}\varphi_{L_2(u)} - \mu_1^{-1}\varphi_{L_1(u)} = -\mu_1^{-1}\varphi_{L_1(u)} \le -\mu_{max}^{-1}\varphi_{N-d+1} := -2\tilde{C}_2 < 0. \tag{76}$$

As a result, for all large n, using the lower bound $\bar{\lambda}_{min}(T)$ on $\bar{\lambda}^n$ given in Theorem 2.8,

$$\Delta^n \le -Cn^{1/2}\int_s^t 1_{\{\varphi_{L_1(u)}>0\}}du, \tag{77}$$

where $C = \bar{\lambda}_{min}(T)(\tilde{C}_1 \wedge \tilde{C}_2)$. Combining (73) and (77) we have on Ω^n,

$$Cn^{1/2}\int_s^t 1_{\{\varphi_{L_1(u)}>0\}}du + \varepsilon \le \hat{R}_2^n \circ f_2^n[J] - \hat{R}_1^n \circ f_1^n[J]. \tag{78}$$

Fix two sequences $r_n, \hat{r}_n$ that converge to zero in such a way that $n^{1/2}\hat{r}_n \to \infty$ and $\hat{r}_n/r_n \to 0$. Then $\mathbb{P}(\Omega^n) = \sum_{k=1}^3 \mathbb{P}(\Omega_k^n)$, where

$$\Omega_1^n = \Omega^n \text{ holds with } t-s \le r_n,$$

$$\Omega_2^n = \Omega^n \text{ holds with } t-s > r_n \text{ and } \int_s^t 1_{\{\varphi_{L_1(u)}>0\}} du > \hat{r}_n,$$

$$\Omega_3^n = \Omega^n \text{ holds with } t-s > r_n \text{ and } \int_s^t 1_{\{\varphi_{L_1(u)}>0\}} du \le \hat{r}_n.$$

For $\mathbb{P}(\Omega_1^n)$, using (78), we have

$$\mathbb{P}(\Omega_1^n) \le \mathbb{P}\Big(\sum_{i=1}^2 w_T(\hat{R}_i^n \circ f_i^n, r_n) \ge \varepsilon\Big).$$

The processes $\hat{R}_i^n$ are $\mathcal{C}$-tight as processes converging to BMs, and by (68) and the bounds on $\bar{\lambda}^n$ given in Theorem 2.8, f_i^n are uniformly Lipschitz and uniformly bounded. It follows that $\hat{R}_i^n \circ f_i^n$ are also $\mathcal{C}$-tight, hence $\mathbb{P}(\Omega_1^n) \to 0$.

On the event Ω_2^n, using (78) again,

$$2\sum_{i=1}^2 \|\hat{R}_i^n \circ f_i^n\|_T \ge Cn^{1/2} \int_s^t 1_{\{\varphi_{L_1(u)}>0\}} du \ge Cn^{1/2}\hat{r}_n.$$

Hence, by the tightness of the RVs on the LHS, $\mathbb{P}(\Omega_2^n) \to 0$.

On the event Ω_3^n, we have

$$\begin{aligned} f_1^n[J] = f_1^n[s,t] = \int_s^t \bar{\Lambda}_{L_1(v)}^n dv &= \int_s^t \bar{\lambda}^n(v)\varphi_{L_1(v)} dv \\ &= \int_s^t \bar{\lambda}^n(v)\varphi_{L_1(v)} 1_{\{\varphi_{L_1(u)}>0\}} dv \le \bar{\lambda}_{max}(T)\hat{r}_n, \qquad t-s > r_n, \end{aligned}$$

where we used the bound on $\bar{\lambda}^n$ stated in Theorem 2.8, and the fact that $\varphi_i \le 1$, for all i. Since $\varphi_{L_1(u)} \ge \varphi_{L_1(u)}$ during $[s,t]$, we also have $f_2^n[J] \le f_1^n[J] \le \bar{\lambda}_{max}(T)\hat{r}_n$. Using (78) for the final time, we obtain

$$\mathbb{P}(\Omega_3^n) \le \mathbb{P}\Big(w_T(\hat{R}_2^n, \bar{\lambda}_{max}(T)\hat{r}_n) \ge \varepsilon\Big).$$

Hence the $\mathcal{C}$-tightness of $\hat{R}_2^n$ gives $\mathbb{P}(\Omega_3^n) \to 0$. This completes the proof. □

Proof of Proposition 2.9. Let $\langle \cdot, \cdot \rangle$ denote the usual scalar product in $\mathbb{R}^N$. Denote $\tilde{\mu}_i^n = \bar{\mu}_i^n / \langle \bar{\mu}^n, 1 \rangle$. Then $\langle \tilde{\mu}^n, W^n(t) \rangle$ is the average workload in the different buffers, with weights proportional to the respective service rates. To relate this quantity to Z^n, note by (16) that

$$Z^n = \langle \mu^n, W^n \rangle = n \langle \bar{\mu}^n, W^n \rangle = n \langle \bar{\mu}^n, 1 \rangle \, \langle \tilde{\mu}^n, W^n(t) \rangle.$$

Since $\langle \bar{\mu}^n, 1 \rangle$ converges to a positive constant, it suffices to prove that, for all n and all $t \in [0,T]$,

$$n^{1/2} \langle \tilde{\mu}^n, W_i^n(t) \rangle \le n^{1/2} \langle \tilde{\mu}^n, \tilde{W}_i^n(t) \rangle + \delta_n, \tag{79}$$

where W_i^n and $\tilde{W}_i^n$ correspond to LW(d) and an arbitrary sequence of policies, respectively, and δ_n is as in the statement of the result.

To this end, fix an arbitrary sequence of policies. The balance equation for the workload, (24), is valid, and therefore we have

$$\tilde{W}_i^n(t) = \tilde{W}_i^{A,n}(t) - \int_0^t 1_{\{\tilde{W}_i^n(s)>0\}} ds.$$

Recall that $\tilde{\mu}_i^n = \bar{\mu}_i^n / \langle \bar{\mu}^n, 1 \rangle$. It follows that

$$\begin{aligned}
\tilde{U}^n(t) &:= n^{1/2} \langle \tilde{\mu}^n, \tilde{W}^n(t) \rangle \\
&= n^{1/2} \langle \tilde{\mu}^n, \tilde{W}^{A,n}(t) \rangle - n^{1/2} \sum_i \tilde{\mu}_i^n \int_0^t 1_{\{\tilde{W}_i^n(s)>0\}} ds \\
&= X^n(t) + \tilde{Y}^n(t),
\end{aligned}$$

where we set

$$X^n(t) = n^{1/2} \langle \tilde{\mu}^n, \tilde{W}^{A,n}(t) \rangle - n^{1/2} t, \qquad \tilde{Y}^n(t) = n^{1/2} \sum_i \tilde{\mu}_i^n \int_0^t 1_{\{\tilde{W}_i^n(s)=0\}} ds, \tag{80}$$

and used the fact that $\langle \tilde{\mu}^n, 1 \rangle = 1$. Now, $\langle \tilde{\mu}^n, \tilde{W}^{A,n}(t) \rangle = \langle \mu^n, 1 \rangle^{-1} \sum_{k=1}^{A^n(t)} T(k)$. Hence X^n does not depend on the sequence of policies. We can now use the minimality property of the Skorohod map (see Section 2 of [16]). It states that if $(\varphi, \eta) \in \mathcal{D}_{\mathbb{R}}^2$, η is nondecreasing and nonnegative, and $\varphi + \eta$ is nonnegative then

$$\varphi(t) + \eta(t) \geq \Gamma[\varphi](t), \qquad t \geq 0.$$

As a result we obtain $\tilde{U}^n(t) \geq \Gamma[X^n](t)$ for all n and $t \geq 0$.

Next, consider the LW(d) policy, for which the workload process is denoted by W^n. Denote $U^n = n^{1/2} \langle \tilde{\mu}^n, W^n(t) \rangle$. Also, let Y^n be defined as in (80) with W^n in place of $\tilde{W}^n$. Then as a special case of the above, we have $U^n = X^n + Y^n$. Set

$$\Delta^n = n^{1/2} \max_{1 \leq i,j \leq N} \|W_i^n - W_j^n\|_T,$$

and recall that, by Theorem 2.8, $\Delta^n \to 0$ in probability. Let $\zeta^n(t) = (U^n(t) - 2\Delta^n)^+$, for $t \in [0,T]$. Then ζ^n is a nonnegative process, and one has

$$\zeta^n(t) = X^n(t) + \varepsilon^n(t) + Y^n(t), \qquad t \in [0,T],$$

for $\|\varepsilon^n\|_T \leq 2\Delta^n$. We now argue that $\int_0^T \zeta^n(t) dY^n(t) = 0$. Indeed, if for some $t \in [0,T]$ one has $\zeta^n(t) > 0$ then $U^n(t) > 2\Delta^n$. Since $\tilde{\mu}_i^n$ sum to 1, U^n is a weighted average of W^n, hence $n^{1/2} \max_i W_i^n(t) > 2\Delta^n$. By the definition of Δ^n, it follows that $n^{1/2} \min_i W_i^n(t) > \Delta^n$. By the definition of Y^n, it follows that the right-derivative of Y^n at t is zero. Consequently, $\int_0^T \zeta^n(t) dY^n(t) = 0$.

The above argument shows that ζ^n is given by $\Gamma[X^n + \varepsilon^n]$ on the time interval $[0,T]$. By the Lipschitz property of Γ (with Lipschitz constant 2), we have

$$\zeta^n(t) \leq \Gamma[X^n](t) + 2\|\varepsilon^n\|_T, \qquad t \in [0,T].$$

This shows

$$U^n(t) \le \zeta^n(t) + 2\Delta^n \le \Gamma[X^n](t) + 2\|\varepsilon^n\|_T + 2\Delta^n \le \Gamma[X^n](t) + 6\Delta^n, \qquad t \in [0,T].$$

Combined with the lower bound on $\tilde{U}^n$, this gives $U^n(t) \le \tilde{U}^n(t) + \delta^n$, upon setting $\delta^n = 6\Delta^n$, showing (79). The result follows. □

Proof of Proposition 2.4. This proof is based on the proof of Proposition 2.9. We keep the notation X^n, Y^n, U^n as in the proof of Proposition 2.9, but define Δ^n, ζ^n and ε^n in a different way. The lower bound $\tilde{U}_n(t) \ge \Gamma[X^n](t)$ for an arbitrary sequence of policies, provided in the above proof, is valid.

For an upper bound in the case of SQ(d), let

$$\Delta^n = n^{-1/2} \max_{1 \le i,j \le N} \|Q_i^n - Q_j^n\|_T,$$

and note that by Theorem 2.1 $\Delta^n \to 0$ in probability. Let $\zeta^n(t) = (U^n(t) - c_n)^+$, by which

$$\zeta^n(t) = X^n(t) + \varepsilon(t) + Y^n(t), \qquad t \in [0,T],$$

with $\|\varepsilon^n\|_T \le c_n$. Here, $c_n > 0$ are constants, whose values are to be determined later in the proof. Our goal is to argue that c_n can be chosen so that $c_n \to 0$, whereas as $n \to \infty$,

$$\mathbb{P}(\Omega^n) \to 0, \quad \text{where} \quad \Omega^n = \Big\{ \int_0^T \zeta^n(t) dY^n(t) > 0 \Big\}. \tag{81}$$

Indeed, once this goal is achieved, the proof can be completed precisely as that of Proposition 2.9.

To this end, note that on the event Ω^n there exists (random) $t \in [0,T]$ such that $\zeta^n(t) > 0$ and $\min_i W_i^n(t) = 0$ (where we used (80)). Note that $\zeta^n(t) > 0$ implies $U^n(t) > c_n$, and since $U^n = n^{1/2}\langle \tilde{\mu}^n, W^n \rangle$, namely U^n is given as a mean of the terms $n^{1/2}W_i^n$ (with weights $\tilde{\mu}_i^n$), it also implies $n^{1/2} \max_i W_i^n(t) > c_n$. On the other hand, $\min_i W_i^n(t) = 0$ implies $\min_i Q_i^n(t) = 0$, and so by the definition of Δ^n, $\max_i Q_i^n(t) \le n^{1/2}\Delta^n$. Hence

$$\Omega^n \subset \cup_i \Omega_i^n, \quad \Omega_i^n = \{\text{there exists } t \in [0,T] : W_i^n(t) > c_n n^{-1/2},\ Q_i^n(t) \le n^{1/2}\Delta^n\}.$$

Recall that the service times of class-i jobs are given by $T_i(k)/\mu_i^n$ and that the nominal workload at buffer i is given by $\mu_i^n W_i^n(t)$. Hence, denoting the partial sums for the unnormalized service times by $\Sigma_i(k) = \sum_{l \le k} T_i(l)$, a simple balance equation for the nominal workload at buffer i gives

$$\mu_i^n W_i^n(t) \le \Sigma_i(A_i^n(t)) - \Sigma_i(D_i^n(t)),$$

where we recall that A_i^n and D_i^n correspond to arrival to and departure from queue i. (The above is an inequality rather than equality, since the RHS does not take into account the job being processed at time t). Since $A_i^n - D_i^n = Q_i^n$, we have

$$\Omega_i^n \subset \{\text{there exist } 0 \le u \le v \le A_i^n(T) : \Sigma_i(v) - \Sigma_i(u) > c_n \mu_i^n n^{-1/2},\ v - u < n^{1/2}\Delta^n\}.$$

Recall that $E[T_i(k)] = 1$ and that for each i, $T_i(k)$ are i.i.d. Then $\hat{\Sigma}_i^n(t) = n^{-1/2}(\Sigma_i([nt]) - nt)$ converges to a standard BM. Letting $C_0 = \inf_{n,i} n^{-1}\mu_i^n > 0$, we have on the event Ω_i^n, using $A_i^n \le A^n$ and $\bar{A}^n = n^{-1}A^n$,

$$\{\text{there exist } 0 \le u \le v \le \bar{A}^n(T) : \Sigma_i(nv) - \Sigma_i(nu) > C_0 c_n n^{1/2},\ v - u < n^{-1/2}\Delta^n\}.$$

Now, under both the inequalities above we also have

$$\Sigma_i(nv) - \Sigma_i(nu) - (nv - nu) > C_0 c_n n^{1/2} - n^{1/2}\Delta^n = n^{1/2}(C_0 c_n - \Delta^n).$$

Hence $\hat{\Sigma}_i^n(v) - \hat{\Sigma}_i^n(u) > (C_0 c_n - \Delta^n)$. Thus, on the event Ω^n one must have

$$\max_i w_{\bar{A}^n(T)}(\hat{\Sigma}_i^n, n^{-1/2}\Delta^n) > (C_0 c_n - \Delta^n). \tag{82}$$

The fact that $\Delta^n \to 0$ in probability and $\hat{\Sigma}_i^n$ are $\mathcal{C}$-tight implies that for every finite K and $\varepsilon > 0$,

$$\mathbb{P}(\max_i w_K(\hat{\Sigma}_i^n, n^{-1/2}\Delta^n) > \varepsilon) \to 0.$$

Hence by the tightness of the RVs $\bar{A}^n(T)$,

$$\mathbb{P}(\max_i w_{\bar{A}^n(T)}(\hat{\Sigma}_i^n, n^{-1/2}\Delta^n) > \varepsilon) \to 0.$$

Therefore there exists a sequence $\varepsilon_n > 0$, $\varepsilon_n \to 0$ such that

$$c_n^* := \mathbb{P}(\max_i w_{\bar{A}^n(T)}(\hat{\Sigma}_i^n, n^{-1/2}\Delta^n) > \varepsilon_n) \to 0. \tag{83}$$

Since $\Delta^n \to 0$ in probability, there exists a sequence $\hat{c}_n > 0$, $\hat{c}_n \to 0$ such that $\mathbb{P}(\Delta^n > \hat{c}_n) < 1/n$ for all n. Fix such a sequence. Set now $c_n = C_0^{-1}(\varepsilon_n + \hat{c}_n)$. Then $c_n \to 0$, and we obtain from (82) and (83),

$$\begin{aligned}
\mathbb{P}(\Omega^n) &\le \mathbb{P}(\max_i w_{\bar{A}^n(T)}(\hat{\Sigma}_i^n, n^{-1/2}\Delta^n) > (C_0 c_n - \Delta^n)) \\
&\le \frac{1}{n} + \mathbb{P}(\max_i w_{\bar{A}^n(T)}(\hat{\Sigma}_i^n, n^{-1/2}\Delta^n) > (C_0 c_n - \hat{c}_n)) \\
&= \frac{1}{n} + c_n^* \to 0.
\end{aligned}$$

This establishes (81) and completes the proof. □

Proof of Proposition 2.10. Based on Theorem 2.8, the proof is similar to that of Proposition 2.6, and thus most details are omitted. However, the first step is different. Rather than working with the mean, we follow the first step of the proof of Proposition 2.6 and define the mean with respect to the vector $\bar{\mu}^n$. That is, as before, let $\tilde{\mu}_i^n = \bar{\mu}_i^n / \langle \bar{\mu}^n, 1 \rangle$. Then Theorem 2.8 implies SSC, in the sense that for each i, $\hat{W}_i^n - \langle \tilde{\mu}^n, \hat{W}^n \rangle \Rightarrow 0$. Since $\tilde{\mu}^n$ converges to a strictly positive vector, it suffices to prove that $\langle \tilde{\mu}^n, \hat{W}^n \rangle \Rightarrow \beta^0$.

To this end, note that it follows from (24) that

$$\begin{aligned}
\langle \tilde{\mu}^n, \hat{W}^n(t) \rangle &= n^{1/2} \langle \tilde{\mu}^n, W^n(t) \rangle \\
&= n^{1/2} \langle \tilde{\mu}^n, W^{A,n} \rangle - n^{1/2} \sum_i \tilde{\mu}_i^n \int_0^t 1_{\{W_i^n(s) > 0\}} ds \\
&= n^{1/2} \langle \tilde{\mu}^n, W^{A,n} \rangle - n^{1/2} t + Y^n(t),
\end{aligned}$$

where we used the fact that $\langle \tilde{\mu}^n, 1\rangle = 1$ and set

$$Y^n(t) = n^{1/2} \sum_i \tilde{\mu}_i^n \int_0^t 1_{\{W_i^n(s)=0\}} ds.$$

But $\langle \tilde{\mu}^n, W^{A,n}\rangle = \langle \mu^n, 1\rangle^{-1} \sum_{k=1}^{A^n(t)} T(k)$ is a renewal-reward process. Appealing again to the FCLT (Theorem 7.4.1 of [67]) shows that $K^n(t) := n^{-1/2}\big[\sum_{k=1}^{A(nt)} T(k) - nt\big]$ (where we recall that $A(\cdot)$ is a standard Poisson process) converges to a $(0, \sigma^2)$-BM, where $\sigma^2 = V_{T(1)} + 1$. Hence, recalling that $A^n(t) = A(\lambda^n t)$, and the notation v^n of (63),

$$\begin{aligned} n^{1/2}\langle \tilde{\mu}^n, W^{A,n}\rangle - n^{1/2}t &= \frac{n^{1/2}}{\langle \mu^n, 1\rangle}\Big[\sum_{k=1}^{A(\lambda^n t)} T(k) - \langle \mu^n, 1\rangle t\Big] \\ &= \frac{n}{\langle \mu^n, 1\rangle}\frac{1}{\sqrt{n}}\Big[\sum_{k=1}^{A(\lambda^n t)} T(k) - \lambda^n t\Big] + \frac{n}{\langle \mu^n, 1\rangle} v^n t \\ &= \frac{1}{\langle \bar{\mu}^n, 1\rangle} K^n(\bar{\lambda}^n t) + \frac{1}{\langle \bar{\mu}^n, 1\rangle} v^n t \end{aligned}$$

converges to a BM with drift $N\hat{\mu}_0\lambda^{-1}$ and diffusion coefficient $(V_{T(1)} + 1)^{1/2}\lambda^{-1/2}$. The remainder of the proof now follows along the lines of that of Proposition 2.6. □

Proof of Proposition 2.7. We introduce some notation special to this proof. Consider the R(d) setting. Recall that job k is replicated d times, and the replicas are sent to the buffers specified by the set B_k. Moreover, all but one of these replicas are canceled. We introduce a scheme that marks replicas. That is, upon the arrival of a job, one of the corresponding replicas is marked, and all others remain unmarked. The marking scheme uses information on the service durations (given in (22)) of the various replicas (including residual service durations of replicas that are in service), information that is not available to the decision maker under R(d). However, the marks do not interfere with the operation of the policy, therefore the underlying stochastic processes under R(d) do not vary as a result of defining the marks.

The marks are constructed in a recursive way. Initially, the system is empty, therefore there are no marks. When replicas of job k are routed to the set B_k, the replica to be marked is determined based on past markings. To this end, a computation is carried out for each of these replicas, as follows. For each replica, one sums the service durations of all the marked replicas in the corresponding buffer. If there is a replica in service and it is marked, its residual service duration is added to the sum (thus unmarked replicas are ignored in this calculation). We refer to this sum as the marked sum. The replica with the minimal marked sum is marked; all the others remain unmarked. On the event of a tie, the replica with the lower index is marked.

It is clear from the construction that marked replicas are precisely those corresponding to that a LW(d) policy would select to route to a server.

On the other hand, we shall prove that under R(d), the tasks that eventually make it to the server are exactly the marked ones (whereas unmarked tasks are those to be canceled). As a consequence, the claim will follow.

For job k, let $M_k \in \{1, \ldots, N\}$ denote the corresponding marked replica (named by the server assigned to it) and let $S_k \in \{1, \ldots, N\}$ denote the task that eventually makes it to the server (also

named by the server assigned to it). It thus suffices to prove that for all k, $M_k = S_k$. This is proved by induction on k. For $k = 1$, recall that the system starts empty. The claim is clear because both the marking scheme and R(d) act according to the same tie breaking rule.

Next, assuming that $M_j = S_j$ for all $j < k$, we show that $M_k = S_k$. Arguing by contradiction, assume that for job k, the marked replica, M_k, is distinct from the replica that makes it to the server, S_k. Let s_k be the time of arrival of this job and t_k the time it is accepted to service. Let $W(M_k)$ and $W(S_k)$ be the marked sum that M_k and, respectively, S_k see ahead of them upon arrival. Since M_k is the marked replica, $W(M_k) \leq W(S_k)$ (and on the event of equality, $M_k < S_k$). Note that the aforementioned marked sum corresponds to replicas that arrived earlier than job k, because they appear ahead of one of the two replicas is the respective queues. Hence the induction hypothesis applies to them. That is, (a) these marked replicas make it to their respective server. Similarly, (b) all unmarked replicas ahead of replicas M_k and S_k are to be canceled before reaching a server.

Now, during $[s_k, t_k)$, the two replicas M_k and S_k of job k are present in the queues, hence the respective servers are necessarily continuously busy throughout this time interval. Hence by (a) and (b) above, the marked sum ahead of M_k and S_k at time t_k is given by $W(M_k) - (t_k - s_k)$ and $W(M_k) - (t_k - s_k)$, respectively (where this conclusion is valid even in the special case $s_k = t_k$).

By definition of S_k and t_k, we have $W(S_k) - (t_k - s_k) = 0$. Hence $W(M_k) - (t_k - s_k) \leq 0$, and since this quantity expresses service duration it must be zero. It follows that at time t_k- replica M_k has zero marked sum ahead. Moreover, there can be no unmarked replica ahead of M_k in service at that time, by (b) above, and there can be no unmarked replica ahead of M_k in the queue either, because such a replica would be entering service at t_k, again contradicting (b).

Thus, both M_k and S_k see no replicas ahead of them at time t_k- (whether marked or unmarked), and it is up to the tie breaking rule to select which enters service. Since, by definition, S_k enters service, it must be that $S_k < M_k$. On the other hand, we also have $W(M_k) = W(S_k)$, and so tie breaking has also been applied at the time of arrival to determine which is to be marked, hence $M_k < S_k$. This is a contradiction. It follows that $M_k = S_k$, and the induction argument is complete. This completes the proof. □

3.3 Proofs for LQF

As before, denote $\hat{Q}_i^n = n^{-1/2} Q_i^n$, and $\hat{S}_i^n(t) = n^{-1/2}(S_i(nt) - nt)$, and similarly, denote

$$\hat{A}_i^n(t) = n^{-1/2}(A_i(nt) - nt).$$

The processes A_i^n and S_i^n are given as $A_i \circ \mathfrak{I}\lambda_i^n$ and $S_i \circ \mathfrak{I}\mu_i^n$, respectively. Therefore

$$n^{-1/2}(A_i^n(t) - \mathfrak{I}\lambda_i^n(t)) = n^{-1/2}(A_i(n\mathfrak{I}\bar{\lambda}_i^n(t)) - n\mathfrak{I}\bar{\lambda}_i^n(t)) = (\hat{A}_i^n \circ \mathfrak{I}\bar{\lambda}_i^n)(t). \tag{84}$$

Similarly,

$$n^{-1/2}(S_i^n(t) - \mathfrak{I}\mu_i^n(t)) = (\hat{S}_i^n \circ \mathfrak{I}\bar{\mu}_i^n)(t). \tag{85}$$

Proof of Theorem 2.11. Fix $\varepsilon > 0$ and $T > 0$. Denote the average of a vector X^n with N components as $X_M^n = N^{-1}\sum_i X_i^n$, where $X^n = A^n, D^n$ or Q^n as well as $\hat{A}^n, \hat{D}^n, \hat{Q}^n$. Define

$$\tau = \tau^n = \inf\{t : \text{there exists } m \in \{1, ..., N\} : \hat{Q}_M^n(t) - \hat{Q}_m^n(t) > 2\varepsilon\}, \tag{86}$$

and let $\Omega^n = \{\tau < T\}$. We prove that $\mathbb{P}(\Omega^n) \to 0$ as $n \to \infty$. Define

$$\sigma = \sigma^n = \sup\{t < \tau : \hat{Q}^n_M(t) - \hat{Q}^n_m(t) \le \varepsilon\}.$$

Since the initial condition for $\hat{Q}^n$ is zero, $\sigma \in [0, \tau)$. Denote $J = [\sigma, \tau]$. Then on the event Ω^n we have $\hat{Q}^n_M(\tau) - \hat{Q}^n_m(\tau) > 2\varepsilon$, $\hat{Q}^n_M(\sigma) - \hat{Q}^n_m(\sigma) \le \varepsilon$. Therefore, recalling the convention $X[J] = X(\tau) - X(\sigma)$,

$$\hat{Q}^n_M[J] - \hat{Q}^n_m[J] \ge \varepsilon, \text{ and } \hat{Q}^n_M(u) > \hat{Q}^n_m(u) \text{ for all } u \in J = [\sigma, \tau]. \tag{87}$$

We may work with a deterministic index in place of the random m, by appealing to the union bound. That is,

$$\begin{aligned}
&\mathbb{P}(\tau < T) \le \sum_{j=1}^{N} \mathbb{P}(\Omega^n(j)), \quad \text{where} \\
&\Omega^n(j) = \Big\{ \text{there exist } s, t \in [0, T],\ s < t, \text{ such that, with } J = [s, t],\ \hat{Q}^n_M[J] - \hat{Q}^n_j[J] \ge \varepsilon, \\
&\qquad\qquad \text{and } \hat{Q}^n_M(u) > \hat{Q}^n_j(u) \text{ for all } u \in J \Big\}.
\end{aligned} \tag{88}$$

The result will follow once we show that each $\mathbb{P}(\Omega^n(j)) \to 0$. We thus fix a deterministic index j, and, without loss of generality assume that $j = 1$. The corresponding time interval is denoted by $J = [s, t]$, while $\Omega^n(1)$ is abbreviated as Ω^n.

On Ω^n, using (87) and the balance equation (32), we have

$$A^n_M[J] - D^n_M[J] - A^n_1[J] + D^n_1[J] \ge \varepsilon n^{1/2}. \tag{89}$$

The cumulative amount of time the server has worked on class-i jobs during J is given by $B^n_i[J] = B^n_i(t) - B^n_i(s)$. Denote $J_i = [B^n_i(s), B^n_i(t)]$. Then by (32), $D^n_i[J] = S^n_i[J_j]$. By (87), $LQ^n(u) \ne 1$ for every $u \in J$. Therefore class 1 receives no service, $D^n_1[J] = 0$ and $D^n_M[J] = N^{-1} \sum_{i \ne 1} D^n_i[J]$. Using these facts in (89), we obtain

$$A^n_M[J] - N^{-1} \sum_{i \ne 1} S^n_i[J_i] - A^n_1[J] \ge \varepsilon n^{1/2}. \tag{90}$$

Dividing by $n^{1/2}$, and using (84) and (85) in the above yields

$$N^{-1} \sum_i \hat{A}^n_i \circ \mathfrak{I} \bar{\lambda}^n_i[J] - N^{-1} \sum_{i \ne 1} \hat{S}^n_i \circ \mathfrak{I} \bar{\mu}^n_i[J_i] - \hat{A}^n_1 \circ \mathfrak{I} \bar{\lambda}^n_1[J] + Z^n \ge \varepsilon, \tag{91}$$

where

$$Z^n = n^{1/2} \Big\{ N^{-1} \sum_i \mathfrak{I} \bar{\lambda}^n_i[J] - N^{-1} \sum_{i \ne 1} \mathfrak{I} \bar{\mu}^n_i[J_i] - \mathfrak{I} \bar{\lambda}^n_1[J] \Big\}. \tag{92}$$

For a bound on Z^n, note that

$$N^{-1} \sum_i \mathfrak{I} \bar{\lambda}^n_i[J] - \mathfrak{I} \bar{\lambda}^n_1[J] \le \Big(\max_i \sup_{t \in [0,T]} \bar{\lambda}^n_i(t) - \min_i \inf_{t \in [0,T]} \bar{\lambda}^n_i(t) \Big)(t - s), \tag{93}$$

and

$$\sum_{i \ne 1} \mathfrak{I} \bar{\mu}^n_i[J_i] \ge \min_i \inf_{t \in [0,T]} \bar{\mu}^n_i(t) \sum_{i \ne 1} |J_i| = \Big(\min_i \inf_{t \in [0,T]} \bar{\mu}^n_i(t) \Big)(t - s), \tag{94}$$

where $|J_i|$ denotes the length of the interval J_i, and we used (87) to determine that the server is non-idle during $[s,t]$ and does not work on class-i jobs, by which $\sum_{i\neq 1}|J_i| = t-s$. Using the two inequalities above and the assumed condition (39), we conclude that

$$Z^n \leq -Cn^{1/2}(t-s),$$

for some positive constant C. Therefore, by using (91) and the bound on Z^n, we have that on Ω^n,

$$N^{-1}\sum_i w_T(\hat{A}_i^n \circ \mathfrak{I}\bar{\lambda}_i^n, |J|) + N^{-1}\sum_{i\neq 1} w_T(\hat{S}_i^n \circ \mathfrak{I}\bar{\mu}_i^n, |J_i|) + w_T(\hat{A}_1^n \circ \mathfrak{I}\bar{\lambda}_1^n, |J|) - Cn^{1/2}(t-s) \geq \varepsilon.$$

The fact that $\mathbb{P}(\Omega^n) \to 0$ now follows by the $\mathcal{C}$-tightness of $\hat{A}_i^n$ and $\hat{S}_i^n$ and the boundedness of $\bar{\lambda}_i^n$ and $\bar{\mu}_i^n$ by an argument similar to the one used in the proof of Theorem 2.1 (this is the argument starting at equation (60)). Since $\varepsilon > 0$ is arbitrary, the result follows. □

Proof of Proposition 2.12. Given the results of Theorem 2.11, this proposition can be proved by a technique very similar to that used for proving Proposition 2.6, hence the proof is omitted. □

Acknowledgment. The authors are grateful to the three referees for their most valuable feedback.

References

[1] Ananthanarayanan, G., Ghodsi, A., Shenker, S., and Stoica, I. (2013). Effective straggler mitigation- Attack of the clones. *USENIX NSDI* (pp. 185-198).

[2] Ananthanarayanan, G., Kandula, S., Greenberg, A. G., Stoica, I., Lu, Y., Saha, B., and Harris, E. (2010). Reining in the Outliers in Map Reduce Clusters using Mantri. *OSDI* (Vol. 10, No. 1, p. 24).

[3] Andrews, M., Kumaran, K., Ramanan, K., Stolyar, A., Vijayakumar, R., and Whiting, P. (2004). Scheduling in a queuing system with asynchronously varying service rates. *Probability in the Engineering and Informational Sciences*, 18(2), 191-217.

[4] Atar, R., and Dupuis, P. (2002). A differential game with constrained dynamics and viscosity solutions of a related HJB Equation. *Nonlinear Analysis: Theory, Methods and Applications*, 51(7), 1105-1130.

[5] Atar, R., and Saha, S. (2016). An ϵ-Nash Equilibrium with High Probability for Strategic Customers in Heavy Traffic. *Mathematics of Operations Research.*

[6] Baharian, G., and Tezcan, T. (2011). Stability analysis of parallel server systems under longest queue first. *Mathematical Methods of Operations Research*, 74(2), 257-279.

[7] Bertsimas, D., Paschalidis, I. C., and Tsitsiklis, J. N. (1998). Asymptotic buffer overflow probabilities in multiclass multiplexers: An optimal control approach. *IEEE Transactions on Automatic Control*, 43(3), 315-335.

[8] Billingsley, P. (2013). *Convergence of probability measures.* Wiley.

[9] van der Boor, M., Borst, S. C., van Leeuwaarden, J. S., and Mukherjee, D. (2018). Scalable load balancing in networked systems: A survey of recent advances. arXiv preprint arXiv:1806.05444.

[10] Borst, S., Boxma, O., Groote, J. F., and Mauw, S. (2003). Task allocation in a multi server system. *Journal of Scheduling*, 6(5), 423-436.

[11] Bramson, M. (1998). State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems*, 30(1-2), 89-140.

[12] Bramson, M., Lu, Y., and Prabhakar, B. (2013). Decay of tails at equilibrium for FIFO join the shortest queue networks. *The Annals of Applied Probability*, 23(5), 1841-1878.

[13] Brutlag, J. (2009). Speed matters for Google web search. *Google*.

[14] Bramson, M. (2011). Stability of join the shortest queue networks. *The Annals of Applied Probability*, 1568-1625.

[15] Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American statistical association*, 100(469), 36-50.

[16] Chen, H. and Mandelbaum, A. (1991). Leontief systems, RBVs and RBMs. In *Applied Stochastic Analysis*.

[17] Chen, H., and Yao, D. D. (2013). *Fundamentals of queueing networks: Performance, asymptotics, and optimization* (Vol. 46). Springer Science and Business Media.

[18] Chen, H., and Ye, H. Q. (2012). Asymptotic optimality of balanced routing. *Operations research*, 60(1), 163-179.

[19] Čudina, Milica, and Kavita Ramanan. (2011). Asymptotically optimal controls for time-inhomogeneous networks. *SIAM Journal on Control and Optimization*, 49.2, 611-645.

[20] Dean, J., and Barroso, L. A. (2013). The tail at scale. *Communications of the ACM*, 56(2), 74-80.

[21] Dean, J., and Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1), 107-113.

[22] Dupuis, P. (2003). Explicit solution to a robust queueing control problem. *SIAM journal on control and optimization*, 42(5), 1854-1875.

[23] Durrett, R. (1999). *Essentials of stochastic processes* (Vol. 1). New York: Springer.

[24] Eschenfeldt, P., and Gamarnik, D. (2016). Supermarket queueing system in the heavy traffic regime. Short queue dynamics. *arXiv preprint*, arXiv:1610.03522.

[25] Foschini, G. J. (1977). On heavy traffic diffusion analysis and dynamic routing in packet switched networks. *Computer Performance*, 499-513.

[26] Foschini, G., and Salz, J. (1978). A basic dynamic routing problem and diffusion. *IEEE Transactions on Communications*, 26(3), 320-327.

[27] Foss, S., and Chernova, N. (1998). On the stability of a partially accessible multi station queue with state dependent routing. *Queueing Systems*, 29(1), 55-73.

[28] Gardner, K., Harchol-Balter, M., and Scheller-Wolf, A. (2016). A better model for job redundancy: Decoupling server slowdown and job size. In *IEEE MASCOTS* (pp. 1-10).

[29] Gardner, K., Zbarsky, S., Doroudi, S., Harchol-Balter, M., and Hyytia, E. (2015). Reducing latency via redundant requests: Exact analysis. *ACM SIGMETRICS Performance Evaluation Review*, 43(1), 347-360.

[30] Gardner, K., Harchol-Balter, M., Scheller-Wolf, A., Velednitsky, M., and Zbarsky, S. (2017). Redundancy-d: The power of d choices for redundancy. *Operations Research*.

[31] Hampshire, R. C., Harchol-Balter, M., and Massey, W. A. (2006). Fluid and diffusion limits for transient sojourn times of processor sharing queues with time varying rates. *Queueing Systems*, 53(1-2), 19-30.

[32] He, Y. T., and Down, D. G. (2008). Limited choice and locality considerations for load balancing. *Performance Evaluation*, 65(9), 670-687.

[33] Honnappa, H., Jain, R., and Ward, A. R. (2014). On transitory queueing. *arXiv preprint*, arXiv:1412.2321.

[34] Honnappa, H., Jain, R., and Ward, A. R. (2015). A queueing model with independent arrivals, and its fluid and diffusion limits. *Queueing Systems*, 80(1-2), 71-103.

[35] Joshi, G., Liu, Y., and Soljanin, E. (2012). Coding for fast content download. *IEEE Allerton* (pp. 326-333).

[36] Joshi, G., Liu, Y., and Soljanin, E. (2014). On the delay storage trade off in content download from coded distributed storage systems. *IEEE Journal on Selected Areas in Communications*, 32(5), 989-997.

[37] Kandula, S., Sengupta, S., Greenberg, A., Patel, P., and Chaiken, R. (2009). The nature of data center traffic: measurements and analysis. *ACM Conference on Internet Measurement* (pp. 202-208).

[38] Keller, J. B. (1982). Time-dependent queues. *SIAM Review*, 24(4), 401-412.

[39] Koole, G., and Righter, R. (2008). Resource allocation in grid computing. *Journal of Scheduling*, 11(3), 163-173.

[40] Le, L. B., Modiano, E., Joo, C., and Shroff, N. B. (2010). Longest-queue-first scheduling under SINR interference model. *ACM MobiHoc* (pp. 41-50).

[41] Li, B., Boyaci, C., and Xia, Y. (2012). Performance guarantee under longest-queue-first schedule in wireless networks. *IEEE Transactions on Information Theory*, 58(9), 5878-5889.

[42] Lipshutz, D., and Ramanan, K. (2016). On directional derivatives of Skorokhod maps in convex polyhedral domains. *arXiv preprint*, arXiv:1602.01860.

[43] Maguluri, S. T., Hajek, B., and Srikant, R. (2014). The stability of longest-queue-first scheduling with variable packet sizes. *IEEE Transactions on Automatic Control*, 59(8), 2295-2300.

[44] Mandelbaum, A., and Ramanan, K. (2010). Directional derivatives of oblique reflection maps. *Mathematics of Operations Research*, 35(3), 527-558.

[45] Mandelbaum, A., W. A. Massey. (1995). Strong approximations for time-dependent queues. *Math. Oper. Res.* 20(1).

[46] Massey, W. A. (1981). Non-stationary queues. Ph.D. thesis, Stanford University.

[47] Massey, W. A., and W. Whitt. (1998). Uniform acceleration expansions for markov chains with time-varying rates. *Ann. Appl. Probab.* 8(4) 1130-1155.

[48] Massey, W. A. (1985). Asymptotic analysis of the time dependent M/M/1 queue. *Math. Oper. Res.* 305-327.

[49] McDonald, D. R., and Turner, S. R. E. (2000). Comparing load balancing algorithms for distributed queueing networks. *Analysis of Communication Networks: Call Centres, Traffic, and Performance*, 28, 105.

[50] Newell, G. F. (1968). Queues with time-dependent arrival rates I—the transition through saturation. *Journal of Applied Probability*, 5(2), 436-451.

[51] Newell, G. F. (1968). Queues with time-dependent arrival rates II—The maximum queue and the return to equilibrium. *Journal of Applied Probability*, 5(3), 579-590.

[52] Newell, G. F. (1968). Queues with time-dependent arrival rates III—A mild rush hour. *Journal of Applied Probability*, 5(3), 591-606.

[53] Ousterhout, K., Wendell, P., Zaharia, M., and Stoica, I. (2013). Sparrow: distributed, low latency scheduling. *ACM SOSP* (pp. 69-84).

[54] Özkan, E., and Ward, A. R. (2015). On the control of fork-join networks. arXiv preprint, arXiv: 1505.04470.

[55] Pedarsani, R., and Walrand, J. (2012). Stability of Lu-Kumar networks under Longest-Queue and Longest-Dominating-Queue scheduling. *Preprint.*

[56] Reiman, M. I. (1982). The heavy traffic diffusion approximation for sojourn times in Jackson networks. In *Applied probability computer science: the interface* (pp. 409-421). Birkhäuser Boston.

[57] Reiman, M. I. (1984). Some diffusion approximations with state space collapse. In *Modelling and performance evaluation methodology* (pp. 207-240). Springer Berlin Heidelberg.

[58] Shah, N. B., Lee, K., and Ramchandran, K. (2012). *The MDS queue: Analyzing latency performance of codes and redundant requests.* Technical Report.

[59] Shah, N. B., Lee, K., and Ramchandran, K. (2016). When do redundant requests reduce latency?. *IEEE Transactions on Communications*, 64(2), 715-722.

[60] Souders, S. (2009). Velocity and the bottom line.

[61] Stolyar, A. L., and Ramanan, K. (2001). Largest weighted delay first scheduling: Large deviations and optimality. *Annals of Applied Probability*, 1-48.

[62] Torrieri, D. (2015). *Principles of spread-spectrum communication systems.* Springer. ISO 690.

[63] Vulimiri, A., Godfrey, P. B., Mittal, R., Sherry, J., Ratnasamy, S., and Shenker, S. (2013). Low latency via redundancy. *ACM CoNEXT* (pp. 283-294).

[64] Vvedenskaya, N. D., Dobrushin, R. L. V., and Karpelevich, F. I. (1996). Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Informatsii*, 32(1), 20-34.

[65] Wan, P. J., Xu, X., Wang, Z., Tang, S., and Wan, Z. (2012). Stability analyses of longest-queue-first link scheduling in MC-MR wireless networks. *ACM MobiHoc* (pp. 45-54).

[66] Wang, D., Joshi, G., and Wornell, G. (2014). Efficient task replication for fast response times in parallel computation. In *ACM SIGMETRICS Performance Evaluation Review* (Vol. 42, No. 1, pp. 599-600).

[67] Whitt, W. (2002). *Stochastic-Process Limits: an introduction to stochastic-process limits and their application to queues*. Springer Science and Business Media.

[68] Whitt, W. (2016). Heavy-traffic limits for a single-server queue leading up to a critical point. *Operations Research Letters*, 44(6), 796-800.

[69] Williams, R. J. (1998). Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing systems*, 30(1), 27-88.

[70] Xu, H., and Li, B. (2014). Repflow: Minimizing flow completion times with replicated flows in data centers. *IEEE INFOCOM* (pp. 1581-1589).

[71] Ying, L., Srikant, R., Eryilmaz, A., and Dullerud, G. E. (2006). A large deviations analysis of scheduling in wireless networks. *IEEE Transactions on Information Theory*, 52(11), 5088-5098.

[72] Zaharia, M., Konwinski, A., Joseph, A. D., Katz, R. H., and Stoica, I. (2008). Improving MapReduce Performance in Heterogeneous Environments. In *OSDI* (Vol. 8, No. 4, p. 7).

[73] Zhang, H., Hsu, G. H., and Wang, R. (1995). Heavy traffic limit theorems for a sequence of shortest queueing systems. *Queueing Systems*, 21(1), 217-238.