Queueing Syst (2018) 88:243–278
https://doi.org/10.1007/s11134-017-9559-5

Asymptotics of insensitive load balancing and blocking phases

Matthieu Jonckheere[1] · **Balakrishna J. Prabhu**[2]

Received: 11 September 2016 / Revised: 24 April 2017 / Published online: 24 November 2017

Abstract We study a single class of traffic acting on a symmetric set of processor-sharing queues with finite buffers, and we consider the case where the load scales with the number of servers. We address the problem of giving robust performance bounds based on the study of the asymptotic behaviour of the insensitive load balancing schemes, which have the desirable property that the stationary distribution of the resulting stochastic network depends on the distribution of job-sizes only through its mean. It was shown for small systems with losses that they give good estimates of performance indicators, generalizing henceforth Erlang formula, whereas optimal policies are already theoretically and computationally out of reach for networks of moderate size. We characterize the response of symmetric systems under those schemes at different scales and show that three amplitudes of deviations can be identified according to whether $\rho < 1$, $\rho = 1$, or $\rho > 1$. A central limit scaling takes place for a sub-critical load; for $\rho = 1$, the number of free servers scales like $n^{\frac{\theta}{\theta+1}}$ (θ being the buffer depth and n being the number of servers) and is of order 1 for super-critical loads. This further implies the existence of different phases for the blocking probability. Before a (refined) critical load $\rho_c(n) = 1 - an^{-\frac{\theta}{\theta+1}}$, the blocking is exponentially small and becomes of order $n^{-\frac{\theta}{\theta+1}}$ at $\rho_c(n)$. This generalizes the well-known quality-and-

A preliminary version appeared in the proceeding of Sigmetrics 2016.

✉ Balakrishna J. Prabhu
balakrishna.prabhu@laas.fr

Matthieu Jonckheere
mjonckhe@dm.uba.ar

[1] Instituto de cálculo, Universidad de Buenos Aires and Conicet, Buenos Aires, Argentina

[2] LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France

efficiency-driven regime, or Halfin—Whitt regime, for a one-dimensional queue and leads to a generalized staffing rule for a given target blocking probability.

Keywords Insensitive load balancing · Blocking phases · Mean-field scalings · QED-Jagerman–Halfin–Whitt regime

Mathematics Subject Classification 60K25 · 60F05 · 60F10

1 Introduction

Load balancing is a critical component in multi-servers systems such as call centres, server farms, as well as in distributed systems with many applications running on different servers (as a single example, see the load balancing needs of the CERN network [16]). Despite its intensive use, there are few efficient rules of thumb for dynamic load balancing schemes, i.e. when the decision of the dispatcher depends on the instantaneous load (for example, number of jobs) at each server, for which performance evaluation is within reach using currently known techniques. This fact becomes even more evident for large systems with asymmetric server speeds and blocking, where both the precise structure of optimal policies (for specific traffic descriptions) and their performance elude current knowledge and techniques. Moreover, there is currently a great need for more decentralized (and hence practically implementable) schemes, allowing us to deal with the massive amount of servers involved in the architecture of Web service companies. The performance trade-off between efficiency and level of information needed to achieve this level is far from being well understood.

Two types of ideas were historically employed to make progress on this front: considering large-scale networks and obtaining asymptotic results using propagation of chaos (asymptotic independence, which is essential for mean-field approximations) on the one hand; restricting the load balancing schemes to obtain more tractable reversible processes on the other hand. We aim here at combining both techniques and show that it may lead to very precise results which are typically out of reach with other methods. In turn, these results allow us to give universal (i.e. valid for all job-size distributions) lower bounds on performance. In particular, different asymptotic scalings for the blocking probability can be identified: an exponentially small blocking probability for sub-critical loads; a polynomial order in the critical regime; and a constant level for super-critical loads. Before describing our contribution more precisely, let us recall the vast effort of research in the two mentioned directions.

A first way of overcoming the analytical difficulty of load balancing problems for processor-sharing systems (and, more generally, symmetric queues) with generic job-size distribution is to restrict the routing policies so that the stationary regime of the system becomes invariant to the job-size distribution (except for its mean), leading to insensitive load balancing (see [6] for more details). To understand the underlying principles, it is useful to come back to the properties of the Erlang formula, which was clearly a revolution for performance evaluation of telephone networks and arguably the true start of queuing theory. The Erlang formula, which gives the probability of loss for a set of telephone lines, based its lasting success on simplicity and robustness:

1. the only assumptions that are required to apply the formula are Poisson arrivals and independent call durations;
2. the formula is insensitive to the call duration distribution and depends on a unique parameter: the traffic intensity; and
3. it can be efficiently computed using a recursive formula.

At the mathematical level, the key property is the reversibility of the birth-and-death process that models the system under Markovian assumptions, which then implies the insensitivity property: the stationary measure of the system does not depend on the whole call distribution but only on its mean. In [13,15], insensitive bounds for the blocking probability of circuits networks with overflow were developed based on reversibility conditions for an auxiliary bounding process. Those same principles were translated to performance models of best effort and streaming traffic in [5,6] for models with balanced allocations and in [4,6] for models with dynamic insensitive load balancing. For multi-class networks with insensitive load balancing, Markov decision programming techniques were employed in [26], structural results were provided in [22], while extensive simulations were proposed in [33]. This chunk of research dealt with networks of fixed size and allowed drawing the following conclusions. For networks with a unique class of traffic, the insensitive load balancing compares very accurately to optimal policies for a given job-size distribution, while delay estimations are a bit less accurate [4,6]. The penalization imposed by reversibility is greater for multi-class networks, while the sensitivity (of optimal sensitive policies) also deteriorates [22,26]. Hence, a small to moderate price has to be paid for robustness and simplicity. It is perhaps counter-intuitive to notice that, for models with infinite buffers, this price becomes very high. It was indeed proved that if the state space is infinite and in the absence of blocking, the optimal insensitive load balancing (for any reasonable criterion) is static (i.e. does not depend on the queue lengths) and is hence much less efficient than a state-dependent sensitive load balancing [21]. For sensitive schemes like join-the-shortest-queue, there is no characterization of the stationary measures in a general setting (see, for example, the approximations in [24]).

On the other side of the spectrum, a great deal of attention has been given in the last decades to mean-field-type results for different types of networks with load balancing applications. When the number of servers is larger, asymptotic independence of the state of the servers allows us to obtain a limiting deterministic dynamical system known as the mean-field limit. The first results were obtained for exponential service times, starting with the seminal works of [29,38], which proved mean-field limits for schemes like join-the-shortest-of-d (JSQ(d)) among n queues, where n is large. Transient functional law of large numbers and propagation of chaos were obtained in [11] for FIFO scheduling. More recently, [37] obtains the mean-field limit for the join-the-idle-queue (JIQ) and [30] computes the diffusive limit in the Halfin–Whitt regime for a class of policies, of which the JIQ and JSQ(d) policies are a special case. Interestingly, they show that JIQ is optimal at this diffusive scale. For the JSQ policy, the large-server heavy-traffic limit was derived in [9].

As one might suspect, for general service time distributions, the results are scarcer. In [10], the convergence of the mean-field limit of JIQ in the stationary regime was proved under light-traffic conditions. For service time distributions with decreasing

hazard rate and FIFO service discipline, propagation of chaos properties and asymptotic behaviour of the number of occupied servers were obtained for the JSQ(d) policy in [7].

All the above results concern systems without blocking and are sensitive to the job-size distribution. For systems with blocking, the optimality of JSQ for general inter-arrival time distributions and exponential service times was proved in [14,36], while the optimal insensitive policy for Poisson arrivals was characterized in [4]. These works deal with the optimality per se without quantifying the relationship between the blocking probability and the load per server, especially in the various asymptotic regimes. Recently, [41] investigated this relationship in the mean-field limit for the Markovian JSQ(d) scheme with blocking and homogeneous server speeds. This work was generalized to heterogeneous speeds in [31], who also conjectured its insensitivity property.

It is hence natural and complementary to look at insensitive networks with a very large number of servers and given buffer depth in order to see if the results obtained for finite networks scale appropriately. As detailed below, this leads to very precise results which are qualitatively different from the case without blocking and are out of reach for sensitive policies with blocking. This, in turn, provides simple dimensioning rules.

Contributions

We study the asymptotics of a set of n processor-sharing servers, each with buffer size θ, fed by a Poisson process of intensity ρn, when n gets large, under the family of insensitive load balancing schemes shown to be optimal (in the class of insensitive load balancing) in [4]. A more detailed description of the model is given in the next section.

Building upon closed-form expressions for the stationary measure, we characterize precisely the asymptotics of the stationary measure and the blocking probability for various scalings of the load. Consequently, we provide universal benchmarks for achievable performance which have no known counterpart for sensitive policies.

We first obtain the stationary measure of the number of occupied servers and give its transient mean-field limit. Considering the symmetric version of the model, we show that the functional law of large numbers also holds for the stationary version of the system (limits in n and t commute). The existence and uniqueness of the limiting stationary probabilities are proved through a monotonicity argument involving the Erlang formula, while the stationary point is characterized through the Erlang formula. This implies simple conclusions on the asymptotic behaviour of the blocking probability: the blocking probability is asymptotically vanishing for the sub-critical ($\rho < 1$) case and is equal to $1-\rho^{-1}$ for the super-critical case $\rho > 1$. In both cases, this blocking probability corresponds to the optimal blocking probability achievable by any non-anticipating policy. Of course, this is far from being sufficiently informative and we are led to focus on a more detailed study of the stationary distribution for large n, establishing both large deviations principles for the sub- and super-critical cases and moderate deviations results. We show that, when $\rho < 1$ is fixed, the blocking

probability is exponentially small, and we characterize the most probable deviations from the mean-field limit. The large deviation cost is shown to be a sum of two terms: the "distance" to the stationary point from distributions with a given mean plus the cost of having a different mean from the true stationary mean. We also show that a central limit theorem is valid for the occupation numbers around the stationary point of the mean field in the sub-critical regime. For the critical case $\rho = 1$, the right scaling is no longer of order $\sqrt{n}$. Using local limit theorems and exploiting the characterization of deviations from the mean-field limits, we show that the number of free servers scales like $n^{\frac{\theta}{\theta+1}}$, the limiting distribution depending on θ and coinciding with the normal distribution only for $\theta = 1$. In a third step, we study the critical case at a finer scale and show that a qualitative phase transition occurs at the critical load $\rho_c(n) = 1 - an^{-\frac{\theta}{\theta+1}}$, where θ is the buffer depth. The blocking probability is exponentially small until $\rho_c(n)$ and of order $n^{-\frac{\theta}{\theta+1}}$ at this critical load. This generalizes the Halfin–Whitt regime established for the $M/M/n/n$ system (in that case, the correct scaling for the moderate deviations stayed of order $\sqrt{n}$) and shows that the popular staffing rule established for the $M/M/n/n$ system does actually change with the value of θ when load balancing is employed. The super-critical regime is simpler to characterize, the deviations being of order 1. We illustrate these findings in simple numerical experiments. In particular, we observe that the bounds obtained compare favourably with the performance of JSQ, which is known to be optimal for exponential service times and outperform JIQ. Finally, we comment on how these results can be used for performance planning. For instance, given the number of servers, how maximum delay can be traded for blocking, and how the level of information needed for a possible implementation can be reduced. We also give insights into possible future work.

2 Review of the optimal insensitive load balancing policy

Notation We use the following notation, common to all sections. For any vector space (the exact one under consideration will be clear from the context), let e_i be the point defined by $(e_i)_i = 1,\ (e_i)_j = 0,\ j \neq i$. We denote

$$|x| = \sum_{i=1}^{k} x_i \quad \text{and} \quad \binom{|x|}{x} = \frac{|x|!}{x_1! \dots x_k!}.$$

We denote by $\mathbf{1}_S$ the indicator function of S, that is, the map taking the value 1 inside S and 0 outside, and denote, respectively, by $\mathbb{R}_+$ and $\mathbb{R}_+^*$ the set of non-negative and positive reals.

This section is a review of the relevant definitions, merits and results known for the insensitive load balancing policy investigated in this paper. The narrative here is for a more general model than the one we shall analyse. Nonetheless, it gives a flavour of the possible generalizations, some of which are elaborated upon in Sect. 6.

Consider a dispatcher and a set of n processor-sharing servers with speed μ_i for $i = 1, \dots, n$. Jobs with i.i.d. sizes sampled from a generic distribution of mean 1 arrive to the dispatcher according to a Poisson process of intensity λ. The dispatcher

routes an incoming job to one of the servers according to the following insensitive load balancing rule. Let $\theta = (\theta_1, \ldots, \theta_n)$ be a vector of natural numbers, and let $\mathcal{X} \subset \mathbb{N}^n$ be a finite-coordinate convex set describing the constraints on the number of jobs in each server (for instance, $\mathcal{X} = \{x : x_i \leq \theta_i, \forall i = 1, \ldots, n\}$). Then, the incoming job is routed to server i with probability

$$a_i^{\theta}(x) = \frac{\theta_i - x_i}{\sum_{j=1}^{N} (\theta_j - x_j)} 1_{x+e_i \in \mathcal{X}}. \tag{1}$$

Note that with this rule, the number of jobs in server i is smaller than θ_i for all $i = 1, \ldots, n$. One can view θ_i as the buffer size of server i, but it could be a smaller number chosen to guarantee a certain rate of service. Also note that this rule depends on the speeds μ_i only through the vector θ. Nevertheless, this load balancing rule was proved to be optimal[1] in the set of insensitive load balancing (for a unique class of traffic) in [4], i.e., given the speeds μ_i there exists an optimal vector θ such that this rule is optimal among all insensitive load balancing.

Let X be the stochastic process valued in $\mathcal{X}$ describing the number of ongoing jobs in each server. Under Poisson arrivals and exponentially distributed job-sizes, X is a continuous-time jump Markov process, on the state space $\mathcal{X}$, with infinitesimal generator $Q = (q(x, y))_{x,y \in \mathcal{X}}$ given by, $\forall x \in \mathcal{X}$,

$$\begin{cases} q(x, x - e_i) = \mu_i & \text{if } x - e_i \in \mathcal{X}; \\ q(x, x + e_i) = \lambda a_i(x) & \text{if } x + e_i \in \mathcal{X}; \\ q(x, y) = 0 & \text{if } y \in \mathcal{X},\ y \neq x - e_i, x + e_i. \end{cases} \tag{2}$$

We recall that the family of insensitive load balancing corresponds to the routing rates $\lambda_i() = \lambda a(\cdot)$ such that there exists a balance function $\Lambda : \mathcal{X} \to \mathbb{R}_+^*$,

$$\forall i,\ \forall x \in \mathcal{X}, x + e_i \in \mathcal{X}, \quad \lambda_i(x) = \Lambda(x + e_i)/\Lambda(x), \tag{3}$$

which is equivalent to the detailed balance criterion. The relationship between this criterion and insensitivity was first formulated in [40]. Under condition (3), the process X is reversible and the stationary distribution is given by

$$\pi(x) = \frac{\Lambda(x)\Phi(x)}{\sum_{y \in \mathcal{X}} \Phi(y)\Lambda(y)}, \tag{4}$$

with

$$\Phi(x) = \prod_{i=1}^{n} \mu_i^{-x_i}.$$

[1] Optimal in the sense that it minimizes the blocking probability or any convex criterion.

For the optimal insensitive load balancing corresponding to the mentioned rates $\lambda a_i(\cdot)$, the routing balance function Λ takes the form

$$\Lambda(x) = \Lambda_\theta(x) = \binom{|\theta - x|}{\theta - x} \lambda^{|x|}, \tag{5}$$

where $\binom{|\theta-x|}{\theta-x} = \frac{|\theta-x|!}{\prod_{i=1}^{n}(\theta_i - x_i)!}$ are the multinomial coefficients.

The blocking probability, B_θ, of an arriving job can be determined using the PASTA property to be $\pi(\theta)$.

3 Model and preliminary results

In the rest of the paper, we shall assume that the servers are homogeneous, that is, they have the same speed and the same buffer size. (We shall comment later on the possibility of extending those results.) Without loss of generality, let the common speed be 1. The common buffer size will be taken to be θ. (From now on, θ is a natural number and not a vector as in the previous section.) For the asymptotic analysis we have in mind, it turns out to be more convenient to define the state as the number of servers processing a certain number of jobs instead of the number of jobs being processed in every server. Let $\mathcal{S} = \{s \in \{0, 1, \ldots, n\}^{\theta+1} : \sum_{i=0}^{\theta} s_i = n\}$ be the set of states, where s_i corresponds to the number of servers with i jobs. In state $s \in \mathcal{S}$, the insensitive load balancing rule described in Sect. 2 will route an incoming job to a server with i jobs at rate

$$\lambda_i(s) = \lambda \frac{(\theta - i)s_i}{n\theta - \bar{s}}, \tag{6}$$

where $\bar{s} = \sum_{i=0}^{\theta} i s_i$.

Let $\{S^{(n)}(t) \in \mathcal{S}\}_{t\geq 0}$ be a stochastic process denoting, at time t, the number of servers with i jobs, $i = 0, \ldots, \theta$. Under Poisson arrivals and exponentially distributed job-sizes, $S^{(n)}(t)$ is a continuous-time jump Markov process on the state space $\mathcal{S}$ with the following transition rates:

$$S^{(n)}(t) \to \begin{cases} S^{(n)}(t) + e_i - e_{i-1} & \text{at rate } \lambda_{i-1}(s),\ i \geq 1; \\ S^{(n)}(t) + e_i - e_{i+1} & \text{at rate } s_{i+1}, \end{cases} \tag{7}$$

assuming that the transitions take the process to a state within $\mathcal{S}$.

The reversibility property of X is preserved by S. More precisely,

Theorem 1 *If the job-size distribution is exponential, the process $S^{(n)}(t)$ is a reversible Markov process and its stationary distribution is given by*

$$\pi^{(n)}(s) = \pi_0^{(n)} \frac{(n\theta - \bar{s})!}{(n\theta)!} \binom{n}{s} \prod_{k=0}^{\theta} \left(\frac{\theta!}{(\theta - k)!} (n\rho)^k \right)^{s_k}, \tag{8}$$

where $\bar{s}$ is the total number of jobs in the system, $\rho = \lambda/n$ is the load per server, and $\pi_0^{(n)}$ corresponds to the probability of the state with all servers empty, that is $\bar{s} = 0$ and $s = (n, 0, \ldots, 0)$.

Proof A sufficient condition for a probability measure to be the stationary measure of a Markov chain is that it satisfies the local balance equations. Consider two states s and $s + e_i - e_{i-1}$, both within $\mathcal{S}$. From (8),

$$\frac{\pi^{(n)}(s + e_i - e_{i-1})}{\pi^{(n)}(s)} = \frac{\lambda(\theta - (i-1))s_{i-1}}{n\theta - \bar{s}} \frac{1}{(s_i + 1)\mu}, \tag{9}$$

$$= \frac{\lambda_{i-1}(s)}{(s_i + 1)}, \tag{10}$$

which are in the same proportion as the local transition rates between these two states as computed from (7). □

Corollary 1 *Using the PASTA property, the blocking probability is given by*

$$B_\theta^{(n)} = \pi_0^{(n)} \frac{(n\rho)^{n\theta}(\theta!)^n}{(n\theta)!}. \tag{11}$$

Instead of using $\pi_0^{(n)}$ as the normalizing constant, we can resort to $B^{(n)}$ for this purpose and rewrite (8) as

$$\pi^{(n)}(s) = B_\theta^{(n)}(n\theta - \bar{s})! \binom{n}{s} \prod_{k=0}^{\theta} \left(\frac{1}{(\theta - k)!} (n\rho)^{k-\theta} \right)^{s_k}. \tag{12}$$

A special case that will reappear throughout this paper is the one with $\theta = 1$, which corresponds to the classical $M/M/n/n$ queue or the Erlang loss system. Upon setting $\theta = 1$ in (8), we obtain

$$\pi^{(n)}(s_0) = \frac{(n\rho)^{(n-s_0)}}{(n - s_0)!} \pi_0^{(n)}, \tag{13}$$

where

$$\pi_0^{(n)} = \sum_{k \le n} \frac{(n\rho)^{n-k}}{(n-k)!} \tag{14}$$

and s_0 is the number of empty servers. We hence retrieve the formula corresponding to the $M/M/n/n$ queue, as expected.

We remind the reader that for the JSQ policy, the stationary measure is quite intricate to compute, even for the case of two servers [24], making it difficult to predict the performance of this policy.

Once the stationary measure is determined, the stationary performance measures such as the mean sojourn time and the blocking probability can be numerically computed for any given set of parameters such as n, θ or ρ. However, for large n or ρ close to 1, the relationship between the performance measures and the parameters can be obtained in a more palatable (and exploitable) form using asymptotic analysis. The following sections will follow this path, leading to a mean-field limit as well as the characterization of the large and moderate deviations.

4 A mean-field deterministic limit

In this section, we give the limiting transient and stationary behaviour in the case of exponentially distributed job-sizes when n diverges. This limit, called the mean-field limit, has become a classical asymptotic regime for the analysis of large queueing systems and particle systems with a large number of servers or particles [3,23,25,31]. In load balancing applications, this type of analysis has been used for several policies whose stationary measure is either unknown or known for relatively small values of the number of servers (for example, shorter of d choices [29], joining the shortest queue [31]). Indeed, for data centres that can have hundreds to thousands of servers, the mean-field limit can give a first-order approximation to the system behaviour both in the transient and in the stationary phase.

In the mean-field limit, dynamics for the fraction of servers containing a certain number of jobs are as follows.

Theorem 2 *Fix $\rho < 1$ and $\theta \geq 1$. For exponentially distributed job-sizes, for all fixed time, $S^{(n)}(t)/n \to y(y)(t)$, in probability, with y which is the solution of the following set of differential equations:*

$$\frac{dy_j(t)}{dt} = \rho \frac{\theta - (j-1)}{\theta - \sum_k k y_k(t)} y_{j-1}(t) + y_{j+1}(t) \tag{15}$$

$$- \rho \frac{\theta - j}{\theta - \sum_k k y_k(t)} y_j(t) - y_j(t), \ 0 < j < \theta, \tag{16}$$

$$\frac{dy_\theta(t)}{dt} = \rho \frac{1}{\theta - \sum_k k y_k(t)} y_{\theta-1}(t) - y_\theta(t), \tag{17}$$

$$\frac{dy_0(t)}{dt} = y_1(t) - \rho \frac{\theta}{\theta - \sum_k k y_k(t)} y_0(t), \tag{18}$$

with $y(0) = \lim_{n\to\infty} \frac{S^{(n)}(0)}{n}$.

Proof We first show the result for $\rho < 1$. Let us write $\bar{S}^{(n)}(t) = S^{(n)}(t)/n$ and suppose that $\bar{S}^{(n)}(0) = y(0)$. Using Dynkin's formula, we obtain that, for all $0 \leq i \leq \theta$,

$$\bar{S}_i^n(t) = y_i(0) + \frac{1}{n} \int_0^t \beta_i(S_s^{(n)}) ds + M_i^n(t),$$

where β is the drift function defined by

$$\beta_i(s) = \lambda_{i-1}(s) + s_{i+1} - s_i - \lambda_i(s)$$

and M_i^n is the (usual additive) Lévy martingale. Define $s_0 = (0, 0, \ldots, 1)$ and $G_\varepsilon = \{s : |s - s_0| \geq \varepsilon\}$. It is a matter of routine argument to prove that, for all $y(0) \in G_{2\varepsilon}$, there exists some ε such that $y(t) \in G_\varepsilon$ for all $t \leq T$. Note that $\beta_i(s)/n = b(s/n)$ and that b is a globally Lipschitz function on G_ε for all $\varepsilon > 0$. Indeed, if $y \in G_\varepsilon$,

$$\begin{aligned}\left|\frac{\partial b_j(y)}{\partial y_l}\right| = \Big| & l\rho \frac{\theta - (j-1)}{(\theta - \sum_k k y_k)^2} y_{j-1} \\ & + 1_{l=j-1}\rho \frac{\theta - (j-1)}{(\theta - \sum_k k y_k)^2} + 1_{l=j+1} - l\rho \frac{\theta - j}{\theta - \sum_k k y_k} y_j \\ & - 1_{l=j}\rho \frac{\theta - j}{\theta - \sum_k k y_k} - 1_{l=j} \Big| \leq L_\varepsilon.\end{aligned}$$

Define now α to be the infinitesimal quadratic variation of the process, i.e.,

$$\begin{aligned}\alpha(s) &= \sum_{1 \leq i \leq \theta} |e_i - e_{i-1}|^2 (\lambda_{i-1}(s) + s_i) \\ &= \sum_{0 \leq i \leq \theta} \frac{1}{n^2} 2(\lambda_i(s) + s_i) \\ &\leq \frac{4\theta^2}{n^2}.\end{aligned}$$

The Cauchy-Schwartz inequality and Doob's inequality lead to

$$\begin{aligned}\mathbb{E}\left(\sup_{s \leq t} |M_i^n(s)|\right) &\leq \sqrt{\mathbb{E}\left(\sup_{s \leq t} |M_i^n(s)|^2\right)} \leq 2\sqrt{\mathbb{E}\left(M_i^n(t)^2\right)} \\ &\leq 2\sqrt{\mathbb{E}\int_0^t \alpha\left(S_s^n\right) ds} = \frac{4\theta}{n}.\end{aligned}$$

Define $\tau = \inf\{t : X_t \notin G_\varepsilon\}$. Now, for $t \leq \tau$,

$$\begin{aligned}\sup_{s \leq t} |\bar{S}^n(t) - y(t)|_1 &\leq \sum_{0 \leq i \leq \theta} \Big(\int_0^t |\beta_i(\bar{S}^n(u)) - \beta_i(y(u)| du + \sup_{s \leq t} |M_i^n(s)| \Big), \\ &\leq L_\varepsilon \sum_{0 \leq i \leq \theta} \int_0^t |\bar{S}^n(u) - y(u)|_1 du + \sum_{0 \leq i \leq \theta} \sup_{s \leq t} |M^n(s)|.\end{aligned}$$

Then, a direct application of Gronwall's lemma gives that

$$\sup_{s\le t}|\bar{S}^n(t)-y(t)|_1 \le \exp(L_\varepsilon t)\sum_{0\le i\le\theta}\sup_{s\le t}|M_i^n(s)|.$$

Hence, $\exp(L_\varepsilon t)\sum_{0\le i\le\theta}\sup_{s\le t}|M^n(s)| \le \delta$ for δ sufficiently small implies that $\bar{S}^n(s)\in G_\varepsilon$ for all $s\le t$, which further implies $\tau > t$.

Hence, Doob's inequality leads to

$$\mathbb{P}\left(\sup_{s\le t}|\bar{S}^n(s)-y(s)|>\delta\right)\le\mathbb{P}\left(\sum_{0\le i\le\theta}\sup_{s\le t}|M^n(s)|>\delta\exp(-L_\varepsilon t)\right)\le\frac{4\theta}{\delta n}\exp(L_\varepsilon t),$$

which concludes the proof for $\rho < 1$.

For $\rho \ge 1$, we need to prove that the process is pushed to the frontier when started close to it, and stays at the frontier of the state space later on. This can be done using the following coupling argument: Let $Y^n(t)$ be a $M/M/n\theta/n\theta$ queue. We can easily couple X^n and Y^n such that, almost surely for all times t, $|X^n(t)| \ge Y^n(t)$. This implies in particular that $|X^n(t)|/(n\theta) \ge Y^n(t)/(n\theta)$. The limit in probability of $Y^n(t)/(n\theta)$ is $\max(1, \rho + (y_0 - 1)e^{-t})$ (see, for instance, [34]). This shows that, for any $x \notin G_\varepsilon$, there exists a vanishing function g such that $1 \ge |X_\theta^n(t)|/(n\theta) \ge 1 - g(\varepsilon)$, which concludes the proof. □

Remark 1 We expect such convergence results to hold under generic job-size distribution (though, of course, the transient limit will depend on the service time distribution), but the proof becomes much more technical as one has to work with measure-valued processes, and falls out of the scope of this paper. Results like asymptotic independence for randomized load balancing schemes such as join-the-shortest-of-d-queues with generic job-sizes have been proved in [7].

Theorem 3 *For $0 < \rho \le 1$, the unique steady-state solution of the system of equations (15)–(18) is given by*

$$\hat{p}_j = \left(\frac{\theta-\hat{c}}{\rho}\right)^{\theta-j}\frac{1}{(\theta-j)!}\hat{p}_\theta, \tag{19}$$

$$\text{with } \hat{p}_\theta = \frac{1}{\sum_{k=0}^{\theta}\left(\frac{\theta-\hat{c}}{\rho}\right)^k\frac{1}{k!}}, \tag{20}$$

where

$$\hat{c} = \theta - \rho\zeta_\theta^{-1}(1-\rho), \tag{21}$$

with ζ_θ^{-1} as the inverse function of the Erlang blocking viewed as a function of the traffic intensity for a fixed buffer depth θ.

If $\rho > 1$, the unique solution is $\hat{c} = \theta$, $\hat{p}_j = 0$, for $j \le \theta - 1$, and $\hat{p}_\theta = 1$.

Proof Suppose first that $\rho < 1$. It can be easily verified that, with

$$\hat{c} = \sum_{k=0}^{\theta} k \hat{p}_k, \tag{22}$$

(19) and (20) are the steady-state solutions of (15)–(18). We now show that $\hat{c}$ as defined in (22) verifies (21).

After some simple algebraic manipulations, it can be verified that the fixed point equation (22) is equivalent to the equation

$$(1-\rho) \sum_{k=0}^{\theta} \left(\frac{\theta - x}{\rho}\right)^k \frac{1}{k!} = \left(\frac{\theta - x}{\rho}\right)^{\theta} \frac{1}{\theta!} \tag{23}$$

in the set $[0, \theta]$. Thus, solving (22) boils down to finding a traffic intensity $a = \frac{\theta - x}{\rho}$ such that the Erlang blocking formula with intensity a gives $1 - \rho$, i.e.,

$$\zeta_\theta(a) = \frac{a^\theta \frac{1}{\theta!}}{\sum_{k=0}^{\theta} a^k \frac{1}{k!}} = (1-\rho).$$

By a simple sample path argument, the Erlang formula is an increasing function of a that is 0 in 0 and 1 in $+\infty$. Hence, it is invertible and there is a unique $a > 0$ such that $\zeta_\theta(a) = 1 - \rho$. Now observe that by a conservation argument (the traffic entering vs. traffic outgoing), we have that, for all a,

$$a(1 - \zeta_\theta(a)) \leq \theta,$$

which boils down (given the definition of a) to

$$a \leq \frac{\theta}{\rho},$$

which in turn gives a unique solution to

$$\frac{\theta - x}{\rho} = a.$$

Now consider $\rho > 1$. It is straightforward to verify that $\hat{c} = \theta$, $\hat{p}_j = 0$, for $j \leq \theta - 1$, and $\hat{p}_\theta = 1$ is a solution. If $\hat{c} < \theta$, then the drift of the differential equation (19) cannot be 0, which implies that the solution given is unique. □

Using the generic method developed in [25] to invert limits in n and t under the assumption of reversibility of the Markov process under study (which is indeed verified here), we can state

Proposition 1 *For $\rho < 1$, $\pi^{(n)}$ converges pointwise to $\hat{p}$ when n and t converge to infinity.*

Proof We apply Corollary 1 in [25], which states that if:

- the sequence of stationary measures $(\pi^{(n)})$ is tight,
- there is convergence over fixed time intervals to a flow y_t,
- the process is reversible for fixed n,
- there exists a unique stationary limiting point $\hat{p}$ for y_t,

then the sequence $\pi^{(n)}$ converges weakly to $\delta_{\hat{p}}$. We remark that in our case tightness is immediate since the state space is finite. Convergence on compact intervals has been verified in Theorem 2, the process is reversible by definition, and uniqueness of the stationary limit of y has been shown in Theorem 3. □

Remark 2 By insensitivity, taking the limit in time first defines a sequence of limiting distributions $\pi^{(n)}$ which do not depend on the specific job-size distribution and which converge towards $\hat{p}$.

4.1 Performance consequences

Let B_θ denote the stationary blocking probability in the mean-field limit, that is, when $n \to \infty$ and $t \to \infty$. Using the PASTA property for fixed n, the blocking probability of a job is the probability that it finds all the servers in their blocking state, i.e., upon arrival all the servers have θ tasks.

Before deriving B_θ, we first give a lower bound on the blocking probability that could be achieved by any non-anticipating and size-unaware load balancing policy.

Proposition 2 *For $\theta > 0$, the blocking probability of any non-anticipating and size-unaware load balancing policy is greater than* $\max(0, 1 - \rho^{-1})$.

Proof We give the argument for $\rho > 1$. The argument for $\rho \leq 1$ is similar. Consider the system in which the resources are pooled, that is, there is one server of service rate $n\mu$ and buffer size $n\theta$. By a pathwise argument for Markovian versions of the systems (implying, by insensitivity, the result for all service times in the stationary regime), the blocking probability of this system will be less than any system with a set of disjoints servers.

In the pooled system, the scaled number of tasks in the system, $\frac{X(t)}{n}$, will follow the differential equation:

$$\dot{x}(t) = \rho - 1, \ 0 < x(t) < \theta, \tag{24}$$

$$\rho(1 - \underline{\mathrm{B}}_\theta) = 1, \ x(t) = \theta. \tag{25}$$

When $x(t)$ is in the interior of the state space, all tasks will be accepted. On the boundary $x(t) = \theta$, the tasks which cause overflow will be blocked. Hence, the blocking probability will be

$$\underline{\mathrm{B}}_\theta = 1 - \rho^{-1}.$$

□

We now show that the insensitive load balancing policy achieves this lower bound, which is independent of θ.

Proposition 3 *The limiting blocking probability of the insensitive load balancing policy is given by*

$$B_\theta = \begin{cases} 0 & \text{if} \quad \rho < 1; \\ 1 - \rho^{-1} & \text{otherwise.} \end{cases} \tag{26}$$

Proof For $\rho < 1$, it can be seen that $\hat{p}_\theta < 1$, and therefore $B_\theta = 0$.

For $\rho \geq 1$, we shall first prove this result for $\theta = 1$. For $\theta = 1$, from (22),

$$c = p_1,$$

and from (19),

$$p_1 = \frac{\rho(1 - B_1)}{1 - p_1}(1 - p_1) = \rho(1 - B_1).$$

Since $p_1 \leq 1$, B_1 is such that $\rho(1 - B_1) = 1$. That is, $B_1 = 1 - \rho^{-1}$.

Using Proposition 3 in [4], $B_\theta \leq B_1$ for all $\theta \geq 1$. Now, using the lower bound in Proposition 2, this implies that $B_\theta = B_1$. □

The stationary blocking probability of the insensitive policy is thus minimal in the considered class of policies, and it is independent of θ. Hence, even a buffer of size 1 is sufficient to get the optimal stationary behaviour. We would like to point out that the optimality is only valid in the limit $n \to \infty$. In order to compute the blocking probability (or other performance measures) for values of n that are large but finite, one has to look at finer scales, which will be the objective of Sect. 5.

5 Finer scales and estimates

While the results of the previous section are interesting for some performance metrics like the mean number of customers (which gives the mean waiting time via Little's formula), they are too rough to be really informative in terms of blocking probabilities. Indeed, any reasonable dynamic load balancing may achieve the given bounds. To get useful and discriminative estimates, we hence need to investigate the process $S^{(n)}$ at finer scales. In particular, we aim at determining when blocking can be considered a large deviation event (with a probability exponentially small in n) and when it will be at another scale.

5.1 Large deviations

Let $\mathcal{P} = \{p \in \mathbb{R}_+^\theta : \sum_{i=0}^\theta p_i = 1\}$. For $c > 0$, denote $\mathcal{S}_c^{(n)} = \{s \in \mathcal{S} : \bar{s} = nc\}$ and $\mathcal{P}_c^{(n)} = \{q \in \mathcal{P} : nq \in \mathcal{S}_c^{(n)}\}$. Since $\bar{s} = \sum_k k s_k$, we have $\sum_k q_k = c$, $\forall q \in \mathcal{S}_c^{(n)}$.

Thus, $\mathcal{P}_c^{(n)}$ is the set of discrete probability distributions taking values on a lattice of unit size $1/n$ and having a first moment of c.

Define $p \in \mathcal{P}_c^{(n)}$ by

$$p_k(c) = \frac{1}{(\theta - k)!} \left(\frac{\theta - c}{\rho} \right)^{\theta - k} \frac{1}{\psi(c)}, \tag{27}$$

where

$$\psi(c) = \sum_{k=0}^{\theta} \frac{1}{k!} \left(\frac{\theta - c}{\rho} \right)^{k} \tag{28}$$

is a normalizing constant which ensures that p is a probability vector. There need not be a vector in $\mathcal{P}_c^{(n)}$ satisfying (27), in which case we define p to be the vector[2] in $\mathcal{P}_c^{(n)}$ which is closest (say in norm l^1) to satisfying (27). To simplify the notation, let $\pi^{(n)}(q; c) = \pi^{(n)}(nq)\mathbb{1}_{\{q \in \mathcal{P}_c\}}$ be the stationary probability of observing $q \in \mathcal{P}_c^{(n)}$.

Let

$$\bar{c} = \arg \max_{c \in [0,\theta]} e^c \psi(c). \tag{29}$$

We first characterize $\bar{c}$. As a consequence of the definition of ψ, simple algebraic computations show that $\bar{c}$ coincides (as it intuitively should) with the asymptotic mean value $\hat{c}$ found in Theorem 3, i.e.,

Proposition 4 $\bar{c} = \hat{c} = \theta - \rho \zeta_\theta^{-1}(1 - \rho)$.

The proof of the proposition is similar to the one for $\hat{c}$ in Theorem 3.

Let $\hat{p} = p(\hat{c})$. The large deviation cost is shown to be a sum of two terms: the "distance" to the stationary point from distributions with equal means plus the cost of having a different mean from the stationary mean. More precisely, we have the following large deviation estimates for $S^{(n)}$:

Theorem 4 *For* $\rho < 1$,

$$\lim_{n \to \infty} \frac{1}{n} \log \left(\frac{\pi^{(n)}(q; c)}{\pi^{(n)}(\hat{p}; \hat{c})} \right) = (c - \hat{c}) + \log \left(\tfrac{\psi(c)}{\psi(\hat{c})} \right) - D_{KL}(q(c) \| p(c)), \tag{30}$$

where D_{KL} *is the Kullback–Leibler divergence.*

Proof Applying Stirling's approximation in the term containing $n\theta$ in (12) and noting that $\sum_k k q_k = c$, we get

[2] When the value of c is clear from the context, we will use the notation p instead of $p(c)$.

$$\pi^{(n)}(q;c) \sim B_\theta^{(n)}(2\pi(n\theta - nc))^{1/2} e^{-n\theta+nc} \cdot \binom{n}{nq} \prod_{k=0}^{\theta} \left(\frac{1}{(\theta-k)!} \left(\frac{n\theta - nc}{n\rho} \right)^{\theta-k} \right)^{nq_k} \tag{31}$$

$$= B_\theta^{(n)}(2\pi(n\theta - nc))^{1/2} e^{-n\theta+nc} \psi^n \cdot \binom{n}{nq} \prod_{k=0}^{\theta} p_k^{nq_k}. \tag{32}$$

Thus, $\pi^{(n)}(q;c)$ is proportional to the multinomial distribution with p_k as the probability of success of the kth class. Using Stirling's approximation in (32), we get

$$\pi^{(n)}(q;c) \sim B_\theta^{(n)}(2\pi(n\theta - nc))^{1/2} e^{-n\theta+nc} \psi^n (2\pi n)^{-\theta/2} \cdot \prod_{k=0}^{\theta} \left(\frac{p_k}{q_k} \right)^{nq_k} \frac{1}{q_k^{1/2}}, \tag{33}$$

from which the desired result can be deduced. □

Corollary 2 *For* $\rho < 1$,

$$\lim_{n\to\infty} -\frac{1}{n} \log \left(\frac{\pi^{(n)}(q;c)}{\pi^{(n)}(p;c)} \right) = D_{KL}(q\|p). \tag{34}$$

Corollary 2 says that, conditioned on observing nc jobs in the system, the probability of observing a certain distribution of jobs over the servers concentrates around p. The probability of observing any other $q \in \mathcal{P}_c^{(n)}$ decreases exponentially with rate $nD_{KL}(q\|p)$, that is, the Kullback–Leibler divergence serves as the large-deviation rate function. This result is akin to Sanov's theorem in information theory [8].

Corollary 3 *For* $\rho < 1$,

$$\lim_{n\to\infty} \frac{1}{n} \log \left(\frac{\pi^{(n)}(p;c)}{\pi^{(n)}(p;\hat{c})} \right) = (c - \hat{c}) + \log \left(\frac{\psi(c)}{\psi(\hat{c})} \right). \tag{35}$$

Corollary 3 states that the scaled number of tasks in the system concentrates around $\hat{c}$ exponentially with rate $(c - \hat{c}) + \log(\psi(c)/\psi(\hat{c}))$.

5.2 Asymptotics for the blocking probability

In this section, we shall look at the asymptotics for the blocking probability, which is the main performance measure of interest for our system. Two different asymptotic regimes will be considered: (i) the number of servers, n, scales linearly with the arrival

rate, $n\rho$, and (ii) the Halfin–Whitt regime [12] which has a linear term as in (i) along with a sub-linear term that represents the safety margin.

The starting point for both these asymptotic regimes will be an integral characterization of B_θ which is derived from the generating function of the stationary measure (see Theorem 13 in Appendix A.5).

Theorem 5 *For $\rho \in (0, 1)$, the blocking B_θ has the asymptotic form*

$$\lim_{n\to\infty} B_\theta^{(n)} \exp(nR(\gamma_{\theta,\rho})) \left(\frac{2\pi n}{\alpha_{\theta,\rho}}\right)^{1/2} = 1, \tag{36}$$

where

$$R(t) = \log\left(\sum_{k=0}^{\theta} \frac{t^k}{k!}\right) - \rho t, \tag{37}$$

$$\gamma_{\theta,\rho} = \arg\max_{t\in(0,\infty)} R(t) = \frac{\theta - \hat{c}}{\rho}, \tag{38}$$

and

$$\alpha_{\theta,\rho} = \frac{(1-\rho)}{\rho}\left(\frac{\theta}{\rho\gamma_{\theta,\rho}} - 1\right). \tag{39}$$

Corollary 4 *For $\theta = 1$, $\gamma_{\theta,\rho} = \frac{1-\rho}{\rho}^{-1}$ and $\alpha_{\theta,\rho} = 1$. Thus,*

$$B_1^{(n)} \sim e^{n(1-\rho)}\rho^n(2\pi n)^{-1/2}. \tag{40}$$

For the proof of Theorem 5, we shall need the following result, whose proof follows along the same lines as that of Theorem 3.

Lemma 1 *Let $\gamma_{\theta,\rho} = \arg\max_{t\in(0,\infty)} R(t)$. Then, $\gamma_{\theta,\rho}$ is the unique solution of the equation*

$$(1-\rho)\sum_{k=0}^{\theta} \frac{x^k}{k!} = \frac{x^\theta}{\theta!}, \tag{41}$$

and $x = \zeta_\theta^{-1}(1-\rho)$.

Proof of Theorem 5 Using calculations on the generating functions for the stationary probability (see Appendix A.5), we obtain that

$$\lim_{n\to\infty} B_\theta^{(n)} n\rho \int_0^\infty \left(\sum_{k=0}^{\theta} \frac{1}{k!}t^k\right)^n e^{-tn\rho} dt = 1. \tag{42}$$

The asymptotic form of the integral can be determined using Laplace's method, which says that

$$\int_0^\infty e^{nf(t)}dt \approx e^{nf(t_0)}\left(\frac{2\pi}{n(-f''(t_0))}\right)^{1/2}, \tag{43}$$

where t_0 is the unique maximizer of f in $(0, \infty)$.

Define

$$R(t) = \log\left(\sum_{k=0}^{\theta}\frac{t^k}{k!}\right) - \rho t. \tag{44}$$

Then

$$\lim_{n\to\infty} B_\theta^{(n)} e^{nR(t_0)}\left(\frac{2\pi n\rho^2}{-R''(t_0)}\right)^{1/2} = 1, \tag{45}$$

where t_0 maximizes R and is characterized in Lemma 1. For Laplace's method to be applicable, one needs $R''(t_0) < 0$. This is shown in Lemma 2, which appears in Appendix A.4. We shall now compute $\alpha_{\theta,\rho} = -\rho^{-2}R''(\gamma_{\theta,\rho})$, and the main result then follows from (45).

Let $g_\theta(t) = \sum_{k=0}^{\theta}\frac{t^k}{k!}$. Since $\gamma_{\theta,\rho}$ is the maximizer of $R(t)$, we have $R'(\gamma_{\theta,\rho}) = 0$, which upon rearrangement gives

$$\frac{g_{\theta-1}(\gamma_{\theta,\rho})}{g_\theta(\gamma_{\theta,\rho})} = \rho. \tag{46}$$

From the definition of g_θ and Lemma 1, we have $g_\theta(\gamma_{\theta,\rho}) = \frac{1}{1-\rho}\frac{\gamma_{\theta,\rho}^\theta}{\theta!}$, and

$$g_{\theta-2}(\gamma_{\theta,\rho}) = g_{\theta-1}(\gamma_{\theta,\rho}) - \frac{\gamma_{\theta,\rho}^{\theta-1}}{(\theta-1)!} \tag{47}$$

$$= \frac{\gamma_{\theta,\rho}^\theta}{\theta!}\left(\frac{\rho}{1-\rho} - \frac{\theta}{\gamma_{\theta,\rho}}\right). \tag{48}$$

Thus,

$$R''(\gamma_{\theta,\rho}) = \frac{g_{\theta-2}(\gamma_{\theta,\rho})}{g_\theta(\gamma_{\theta,\rho})} - \left(\frac{g_{\theta-1}(\gamma_{\theta,\rho})}{g_\theta(\gamma_{\theta,\rho})}\right)^2 \tag{49}$$

$$= \rho - \frac{(1-\rho)\theta}{\gamma_{\theta,\rho}} - \rho^2, \tag{50}$$

from which the expression for $\alpha_{\theta,\rho}$ follows. □

For $\rho > 1$, we cannot directly use the technique that was used for $\rho < 1$ because the maximum of $R(t)$ in the interval $[0, \infty)$ occurs at $t = 0$, which is not an interior point of the support of R. So, we shall resort to a theorem due to Erdelyi that treats this case.

Theorem 6 *Let $\rho > 1$ and $n(\rho - 1)$ be bounded away from 0. As $n \to \infty$,*

$$B_\theta^{(n)} \sim 1 - \rho^{-1} + \frac{1}{(\rho-1)^\theta n^\theta} + o(n^{-\theta}). \tag{51}$$

Proof We shall apply Erdelyi's theorem (see Theorem 1.1 in the arXiv preprint of [32] for a precise statement with the notation relevant for our proof) with $f(t) = -R(t)$ and $[a, b) = [0, \infty)$. Let us verify that the four conditions of this theorem are satisfied by $-R(t)$.

For $\rho \geq 1$, the function $-R(t)$ is increasing in the interval $[0, \infty)$ with minimum at $t = 0$. To see this,

$$-R'(t) = \rho - \frac{\sum_{k=0}^{\theta-1} \frac{t^k}{k!}}{\sum_{k=0}^{\theta} \frac{t^k}{k!}} \geq 1 - \frac{\sum_{k=0}^{\theta-1} \frac{t^k}{k!}}{\sum_{k=0}^{\theta} \frac{t^k}{k!}} \geq 0.$$

From Lemma 4, in a neighbourhood of 0, $-R(t)$ is analytic with expansion

$$-R(t) = (\rho - 1)t + \frac{t^{\theta+1}}{(\theta+1)!} + o\left(t^{\theta+1}\right) \tag{52}$$

and $-R(t)$ is continuously differentiable with an analytic derivative. Finally, to show the absolute convergence of the integral, note that

$$\int_0^\infty e^{-n(-R(t))} dt \leq \int_0^\infty e^{-n(\rho-1)t} dt = \frac{1}{n(\rho-1)} < \infty, \tag{53}$$

as long as $n(\rho-1)$ is bounded away from 0. Thus, all the necessary conditions required by Erdelyi's theorem are satisfied.

The various parameters in the asymptotic expansion (1.5) in [32] are: $R(0) = 0$, $\alpha = 1$, $a_0 = \rho - 1$, $a_1 = \cdots = a_{\theta-1} = 0$, $a_\theta = \frac{1}{(\theta+1)!}$, which gives $\beta_0 = (\rho-1)^{-1}$, $\beta_1 = \cdots \beta_{\theta-1} = 0$ and $\beta_\theta = -\frac{(\rho-1)^{-(\theta+2)}}{\theta!}$, so that

$$\int_0^\infty e^{nR(t)} dt \sim \frac{1}{(\rho-1)n} - \frac{1}{(\rho-1)^{\theta+2} n^{\theta+1}} + o\left(n^{-(\theta+1)}\right). \tag{54}$$

Substituting the above asymptotic expansion in (42), we get the claimed result. □

Since $R(0) = 0$, there is no exponential decay of the blocking probability when $\rho > 1$.

Corollary 5 *Setting $\theta = 1$ in the above theorem, we get the corresponding result for $\theta = 1$ obtained in* [17] *(see Theorem* 13 *there).*

The previous theorems give the asymptotics of the blocking probability for a fixed load per server for large number of servers. For $\rho < 1$, the blocking probability goes to 0 exponentially quickly in n, while for $\rho > 1$ it goes to $1 - \rho^{-1}$, a strictly positive quantity. The next theorem looks at the scaling law that results in a polynomial blocking probability. For the Erlang C model, that is, a system without blocking, this regime has the following interpretation: if the cost of servers is high, Halfin and Whitt [12] observed that it could be beneficial to reduce the number of servers in such a way that the probability of waiting is strictly positive and less than 1, instead of going to 0 with the number of servers. This increase in the waiting probability has the benefit of requiring $\lambda + O(\lambda^{1/2})$ instead of $b\lambda$, $b > 1$, servers. Thus, one gains in the cost and the utilization of servers at the expense of the waiting probability. In order to evoke this trade-off between these two quantities, this scaling regime is also called the quality-and-efficiency-driven (QED) regime. We note that this asymptotic regime was already studied for the Erlang B system in Jagerman [17] (see Theorem 14), but the interpretation in terms of a trade-off is due to Halfin and Whitt [12] for systems without blocking and Whitt [39] for systems with blocking. These works were followed by more precise asymptotics in both the Erlang B and the Erlang C systems [18,20].

In addition to the classical asymptotic regimes, such as mean field and QED, one can define an intermediate regime known as the non-degenerate slowdown (NDS) regime (see [2]). The feature of the NDS regime is that the mean sojourn time is of the same (non-degenerate) order as the mean service time. For the insensitive load balancing policy that we are investigating, this relationship between the mean sojourn time and the mean service time is verified irrespective of the load because the buffer length is finite and there is a nonzero probability of routing an arrival to a non-empty queue. Thus, the NDS regime is rather trivial for this policy and is not investigated here.

The following theorem gives the QED scaling for the balanced load balancing policy and can be viewed as a generalization of the QED result for the Erlang loss model.

Theorem 7 *For $a \in (-\infty, \infty)$, let*

$$n\rho = n + an^{1/(\theta+1)}. \tag{55}$$

Then,

$$\lim_{n\to\infty} B_\theta^{(n)} n^{\theta/(\theta+1)} \int_0^\infty \exp\left(au - \frac{u^{(\theta+1)}}{(\theta+1)!}\right) du = 1. \tag{56}$$

Proof The proof follows similar reasoning as in the proof of Theorem 14 in [17]. From (42) and using Lemma 4,

$$1 = \lim_{n\to\infty} B_\theta^{(n)} n\rho \int_0^\infty \left(\sum_{k=0}^{\theta} \frac{1}{k!} t^k\right)^n e^{-tn\rho} dt \tag{57}$$

$$\sim \lim_{n\to\infty} B_\theta^{(n)} n\rho \int_0^\infty \exp\left(nt - n\frac{t^{(\theta+1)}}{(\theta+1)!} - tn\rho\right) dt \tag{58}$$

$$\sim \lim_{n\to\infty} B_\theta^{(n)} n\rho \int_0^\infty \exp\left(an^{1/(\theta+1)} t - n\frac{t^{(\theta+1)}}{(\theta+1)!}\right) dt. \tag{59}$$

Setting $u = tn^{1/(\theta+1)}$ in the integral gives

$$\sim \lim_{n\to\infty} B_\theta^{(n)} (n^{\theta/(\theta+1)} + a) \int_0^\infty \exp\left(au - \frac{u^{(\theta+1)}}{(\theta+1)!}\right) du. \tag{60}$$

□

Note that a can be positive or negative, which means that even with a total charge larger than the number of servers, the blocking probability can decay to 0 provided that (55) is satisfied asymptotically. For $a = 0$, using simple computations, Theorem 7 leads to the following corollary:

Corollary 6 *If* $\rho = 1$,

$$B_\theta^{(n)} \sim \frac{(\theta+1)!^{\frac{1}{\theta+1}}}{\theta+1} \Gamma\Big(\frac{1}{\theta+1}\Big) n^{-\theta/(\theta+1)}, \tag{61}$$

where Γ *is the Gamma function.*

Note that for $\theta = 1$, we retrieve that

$$B_1^{(n)} \sim (0.5\pi n)^{-1/2}. \tag{62}$$

Remark 3 It turns out that the scaling of (55) is the same as that in [28] (see display (4.9)) and differs only in the context in which it is obtained. The one in [28], which they call the QED-c regime, appears in the context of an Erlang C system with abandonments in which the arrival rate has an additive uncertainty. The parameter c, which in our model translates to $1/(\theta+1)$, is the exponent of the additive uncertainty in [28].

5.3 Moderate deviations

Using the previous estimates of the blocking probability (i.e., of the normalizing constant of the stationary distribution) we can now characterize the deviations around $\hat{p}$ of size smaller than $O(n)$ for a fixed value of ρ. Three amplitudes of deviations will be identified according to whether $\rho < 1$, $\rho = 1$, or $\rho > 1$. The proofs of the three results in this subsection appear in the Appendix.

The first result is for $\rho < 1$ and is a central-limit-theorem-type scaling when the deviations around the mean are of the order of $\sqrt{n}$.

Theorem 8 *For $\rho < 1$,*

$$\frac{1}{\sqrt{n}}\left(\left(S^{(n)}(\infty)\right)_{0\le i<\theta} - n(\hat{p})_{0\le i<\theta}\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \Sigma), \tag{63}$$

where

$$\begin{aligned}\Sigma^{-1} = {} & \psi(1, 1, \ldots, 1)\cdot(1, 1, \ldots, 1)^\top \\ & - \left(\frac{1}{\theta - \hat{c}}\right)(\theta, \theta-1, \ldots, 1)\cdot(\theta, \theta-1, \ldots, 1)^\top \\ & + \begin{pmatrix} 1/\hat{p}_0 & 0 & \ldots & 0 \\ 0 & 1/\hat{p}_1 & \ldots & 0 \\ \vdots & \ldots & \ddots & \vdots \\ 0 & 0 & \ldots & 1/\hat{p}_{\theta-1}\end{pmatrix}.\end{aligned} \tag{64}$$

Proof Please see Appendix A.1. □

Corollary 7 *For $\rho < 1$, $\theta = 1$, we have $\hat{c} = \rho$, $\hat{p}_0 = 1 - \rho$, and $\psi = \rho^{-1}$, leading to $\Sigma^{-1} = \rho^{-1}$ and*

$$\frac{S_0^{(n)}(\infty) - n(1-\rho)}{\sqrt{n\rho}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 1). \tag{65}$$

The next case corresponds to $\rho = 1$. For $a, z \in \mathbb{R}$ and $\theta \ge 1$, define

$$\widehat{\Phi}_\theta(z; a) = \int_z^\infty \exp\left(au - \frac{u^{(\theta+1)}}{(\theta+1)!}\right)du. \tag{66}$$

For $\theta = 1$ and $a = 0$, $(2\pi)^{-1/2}\widehat{\Phi}_\theta$ reduces to the complementary cumulative distribution function of the standard normal distribution.

Theorem 9 *For $\rho = 1$ and $z \in \mathbb{R}_+$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_{\theta-1}^{(n)}(\infty)}{n^{\theta/(\theta+1)}} > z\right) = \frac{\widehat{\Phi}_\theta(z; 0)}{\widehat{\Phi}_\theta(0; 0)}. \tag{67}$$

Proof Please see Appendix A.2. □

Corollary 8 *For $\rho = 1$, $\theta = 1$, and $z > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S^{(n)}(\infty)}{\sqrt{n}} > z\right) = 2(1 - \Phi(z)), \tag{68}$$

where Φ is the distribution function of the standard normal distribution.

Unlike in the $\rho < 1$ case, the deviations are now no longer of $O(\sqrt{n})$, but are of a higher order. On the other hand, the fluctuations take $S^{(n)}$ with high probability only to states with either $\theta - 1$ or θ jobs. All other configurations are on a scale lower than $n^{\theta/(\theta+1)}$. This is in contrast to the behaviour for $\rho < 1$, where the fluctuations can take the process to states with number of jobs ranging from 0 to θ. Thus, for $\rho = 1$, conditioned on being accepted, a customer has a high probability of being routed to a server of $\theta - 1$ jobs. This property has a direct consequence on the state information the dispatcher needs to take routing decisions. We shall elaborate upon this in Sect. 6.

Finally, for $\rho > 1$, the following result shows that the deviations around $n\theta$ are of $O(1)$ and are geometrically distributed. Moreover, the excursions take $S^{(n)}$ only to states with $S^{(n)}_{\theta-1} > 0$ and $S^{(n)}_i = 0$ for $i < \theta - 1$. That is, at a random time, there will be a geometrically distributed number of servers with $\theta - 1$ clients and there will be no servers with fewer than $\theta - 1$ clients.

Theorem 10 *For $\rho > 1$,*

$$S^{(n)}_{\theta-1}(\infty) \xrightarrow[n\to\infty]{d} \text{Geo}\left(\rho^{-1}\right), \tag{69}$$

and the blocking probability is

$$B^{(n)}_{\theta} \sim 1 - \rho^{-1}. \tag{70}$$

Proof See Appendix A.3. □

The previous theorems give a more precise characterization of the system state as well as its performance for a fixed value of ρ (not depending on n), both in terms of blocking and waiting time. In particular, for $\rho < 1$, the blocking is exponentially small in n, while the mean sojourn time is

$$\frac{1}{\rho} \sum_i \frac{\theta - i}{\theta - \hat{c}} i \hat{p}_i,$$

with a deviation of order $\frac{1}{\sqrt{n}}$.

5.4 Numerical experiments

In Fig. 1, we first provide a comparison of the blocking probability obtained by the insensitive policy analysed in this paper with that of two other policies, namely JSQ and JIQ. The results were obtained through simulations. There are 20 servers each with a buffer size of 10. Two different job-size distributions were used: (i) exponential distribution and (ii) discrete distribution, which we call custom, with point masses at 0.1 and 10. The probability of a job-size being 0.1 (resp. 10) was 9/9.9 (resp. 0.9/9.9).

The JSQ policy is known to be optimal for exponential job-size distributions and homogeneous server speeds, and thus gives a natural benchmark for comparison. The JIQ policy is an interesting policy from the practical point of view as it requires little

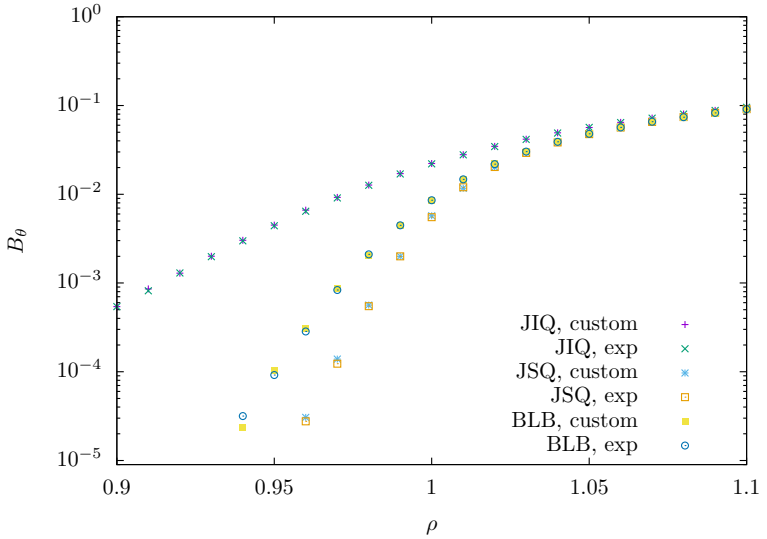

Fig. 1 Comparison of the blocking probability for different load balancing policies. The number of servers is 20. Buffer size is 10

state information compared to JSQ and the insensitive policy, while at the same time it is optimal in the mean-field limit, that is, its blocking probability goes to 0 when the number of servers goes to ∞. We have not included the JSQ(d) policy in our comparison as this policy is not optimal in the mean-field limit. We observed in the simulations that in the symmetric case and for high loads, the performance of JSQ and JIQ changes very little when changing the job-size distribution.

While JIQ requires less state information, it can be seen that, even for a load of 0.9, a few orders of magnitude of gains can be obtained by using the state information. The drawback of JIQ comes from the fact that, at high loads, it behaves more and more like Bernoulli routing since there are fewer empty servers available. Thus, while JIQ is optimal in the mean-field limit, the number of servers required to get close to this limit will be much higher than that of JSQ or the insensitive policy, motivating the asymptotic expressions for fixed n that we provided. For systems with a smaller number of servers in which state information can be obtained relatively cheaply, JSQ or balanced policies can give a considerable performance advantage over JIQ.

As mentioned in the introduction, in the case of symmetric speeds, our motivation for studying the insensitive policy comes mainly from the fact that precise asymptotic estimates can be obtained, which is obviously not the case with JSQ and JIQ. For asymmetric speed, insensitive policies might actually present performance gains over JSQ, but this falls out of the scope of this paper.

We now illustrate the relationship between the blocking probability and the various parameters such as θ, n and λ for the insensitive policy. For this, we evaluate the predictive abilities of some of the results obtained in this section by comparing them with the blocking probability obtained from simulating the Markov chain $S^{(n)}$. In Fig. 2, we plot the blocking probability for $n = 200$ servers and for different values of

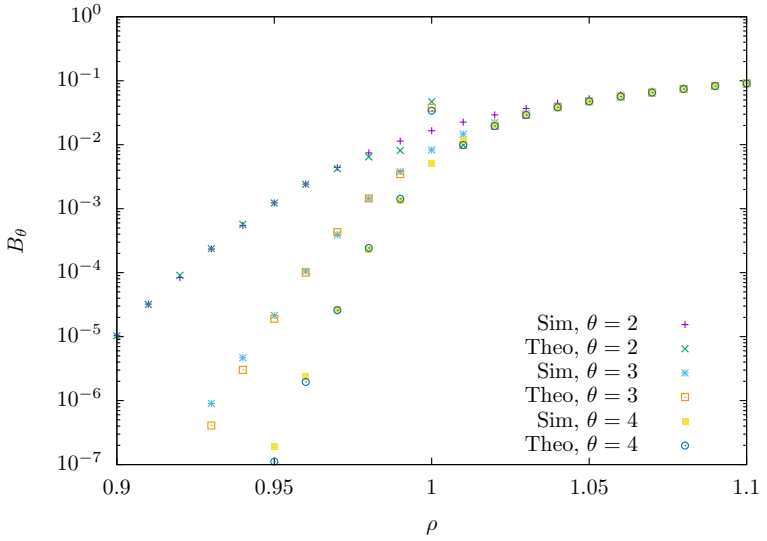

Fig. 2 Comparison of the blocking probability computed from Theorems 5 and 10 with that obtained from simulations. The number of servers is 200

ρ and θ. The theoretical values were calculated using Theorem 5 for $\rho < 1$, Theorem 9 for $\rho = 1$, and Theorem 10 for $\rho > 1$.

We observe that, even for $n = 200$, the prediction is already reasonably accurate for $\theta = 2$ and $\theta = 3$, except for loads very close to 1, where the accuracy is lower. (This comes from a singularity in the expression of the blocking probability at 1.)

6 Engineering insights and future work

6.1 Performance planning

The asymptotic analysis of insensitive load balancing gives a conservative planning tool for managing the performance relationship between the load ρ and the delay guarantees depending on θ, and the blocking guarantees depending both on n, θ. Indeed, in many applications, a given level of quality of service in terms of delay has to be reached and this can be done by fixing θ. For a given buffer depth θ, the mean delay of a job entering the system will be less than θ. (The server speed and the mean job-size are fixed to 1.) On the other hand, for a given θ and n, we have precisely characterized the asymptotics of the blocking probability, unveiling the critical load $\rho_c(n)$ as the frontier of the acceptable blocking probability for most applications. Hence, one can adapt the number of servers n to cope with a target blocking probability given the load, or adapt the load given the number of servers. Note that this planning is completely out of reach for specific sensitive policies.

Another way of looking at it is by considering the staffing rule which is the number of servers necessary to obtain a vanishing blocking probability in the limit when the total charge is large. In [17,39], the staffing rule for $\theta = 1$ was shown to be $\lambda + O(\lambda^{1/2})$,

that is, at least these many servers are required to get a vanishing blocking probability when λ is large.

Theorem 7 generalizes the known results for $\theta = 1$ to larger values of θ, leading to the following staffing rule:

Proposition 5 *For a fixed target blocking probability, the number of servers has to scale as* $\lambda + a\lambda^{1/(\theta+1)}$*, where* a *is determined by the target blocking probability and can be computed using* (56).

6.2 Practical schemes under the critical load

One of the major criticisms of state-dependent policies such as JSQ or the policy under study in this paper is that the dispatcher needs to know the state of every server in order to route an incoming job. The process of collecting state information can add significant delays and lead to lost revenue [27]. Practical policies such as JSQ(d) [29] or JIQ [27] play on the trade-off between information and optimality and aim to perform much better than state-independent policies while at the same time needing much less information than the whole set of servers. For example, JSQ(d), with the knowledge of the state of only d (which can be fixed number independent of n) servers, has a considerable gain at least in the case of exponentially distributed job-sizes and in the absence of blocking when $d = 2$ compared to $d = 1$.

While, at first glance, the insensitive load balancing policy seems to require full state information, Theorem 9 lends a helping hand in alleviating this need. Recall that this theorem has the following implication: for $\rho = 1$ and n large, most of the servers will have either θ or $\theta - 1$ jobs. One possible scheme to exploit this property is based on the idea first proposed for JIQ, which was motivated by the observation that collecting state information at arrival instants should be avoided in order to reduce delays for jobs. In JIQ, the servers inform the dispatcher (or leave information on a bulletin board) when they become idle. The dispatcher[3] then knows which servers are idle, and it routes an incoming packet to one of these servers, if there is one; otherwise, it routes based on no information. Thus, upon arrival a job can be routed immediately based on state information collected previously.

For the insensitive policy, one can conceive a scheme in which servers inform the dispatcher whether they have $\theta - 1$ or fewer than $\theta - 1$ jobs. (This scheme automatically implies that the dispatcher also knows which servers have θ jobs.) When a job arrives, the dispatcher will need to determine the state of only those servers with fewer than $\theta - 1$ jobs. Since this number is expected to be on a smaller scale than $n^{\theta/(\theta+1)}$ (thanks to Theorem 9), one can expect to reduce the information flow between the servers and dispatchers at arrival instants. One of our future works will be to characterize precisely the variations in the number of servers with fewer than $\theta - 1$ jobs. A back of the envelope calculation based upon the proof of Theorem 9 leads one to believe that there will be $O(n^{(k+1)/(\theta+1)})$ servers with k jobs and hence $O(n^{(\theta-1)/(\theta+1)})$ servers with fewer than $\theta - 1$ jobs, but this remains to be rigorously investigated.

[3] We are assuming a single dispatcher.

Of course, this reasoning is valid for a given blocking probability of order $n^{-\frac{\theta}{\theta+1}}$ and this could be significantly reduced for other blocking targets (and hence other loads).

6.3 Multi-speed servers

The above planning was simplified by the fact that we considered a symmetric system depending only on three possibly interdependent parameters (n, ρ, θ). In a future work, we aim at generalizing the analysis to servers with different speeds or even to servers with state-dependent speed. Though this generalization falls out of the scope of this paper, let us underline the possibility of this analysis by giving its first step, the expression of the stationary measure for the occupation of a multi-speed server farm.

Consider a server farm with n servers that are classified according to their speed into J different types. A server of type j has speed c_j and buffer size θ_j, and there are n_j servers of type j.

As for the symmetric system, it is convenient here to study the number of servers processing jobs instead of the number of jobs being processed. Let $\mathcal{S}_j = \{s \in \{0, 1, \ldots, n_j\}^{\theta_j} : \sum_{i=0}^{\theta_j - 1} s_i \leq n_j\}$, and let $\mathcal{S}^{(n)} = \prod_{j=1}^{J} \mathcal{S}_j$. Further, let $\{S^{(n)}(t) \in \mathcal{S}\}_{t \geq 0}$, where $n = \sum_{i=1}^{J} n_j$. Let $S^{(n)}(t)$ be a random process defined on $\mathcal{S}^{(n)}(t)$, where the component (i, j) of $\mathcal{S}^{(n)}(t)$ denotes the number of servers of type j with i customers at time t.

We shall use a boldface font to denote an element of $\mathcal{S}_j$ and use calligraphic font to denote an element of $\mathcal{S}^{(n)}$. So, $\mathbf{s}_j$ would be a vector in $\mathcal{S}_j$, and an element $\mathsf{s} \in \mathcal{S}^{(n)}$ can be written as $\mathsf{s} = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_J)$.

In state s, the arrival rate to servers of type j with i tasks is given by

$$\lambda_{i,j}(\mathsf{s}) = \frac{(\theta_j - i)s_{i,j}}{\sum_j \left(n_j \theta_j - \bar{s}_j\right)} n\rho, \tag{71}$$

where $\bar{s}_j = \sum_i i s_{i,j}$ is the total number of tasks in severs of type j.

Theorem 11 *If the job-size distribution is exponential, the process $\mathcal{S}(t)$ is a reversible Markov process and the stationary distribution of $\mathcal{S}^{(n)}(t)$ is given by*

$$\pi(\mathsf{s}) = \pi(\mathsf{0}) \frac{(n\theta - \bar{\mathsf{s}})!}{(n\theta)!} \prod_{j=1}^{J} \binom{n_j}{\mathbf{s}_j} \prod_{k=0}^{\theta} \left(\frac{\theta_j !}{(\theta_j - k)!} (n\rho_j)^k \right)^{s_{k,j}}, \tag{72}$$

where $\bar{\mathsf{s}} = \sum_j \bar{s}_j$ is the total number of tasks in the system and $\rho_j = \rho c_j^{-1}$.

Given this first result established, all the steps in the presented analysis might (and should) be considered. This would in particular allow us to characterize the optimal trunk reservation parameters (θ_i) for various trade-offs of loads, delays and blocking.

6.4 Future research directions

Other than the directions described in the previous subsection, several open questions deserve attention. A natural related model would be the generalization of the Erlang C model, that is, the model studied in this work but with a common waiting room where arrivals wait when all the servers are in the blocking phase. More fundamental questions that merit investigation are as follows:

- Can similar results be established for sensitive policies (like join-the-shortest-of-d-queues among n)? Are the meaningful scalings similar?
- Can we quantify the optimality gaps for specific families of jobs-size distributions?
- Can we obtain even finer estimates for the blocking probabilities in the QED regime, in the spirit of the body of work establishing precise asymptotics for the Erlang formula [19]?

Acknowledgements This work was partially supported by the Basque Center for Applied Mathematics BCAM and the Bizkaia Talent and European Commission through COFUND programme, under the project titled "High-dimensional stochastic networks and particles systems", awarded in the 2014 Aid Programme with request Reference Number AYD-000-273, and by the STIC-AmSud Project No. 14STIC03. We thank the referees for their insightful comments which have improved the quality of the paper.

Proofs of results in Sect. 5

A.1 Proof of Theorem 8

Proof We first prove a local convergence. Let $q = \hat{p} + \beta/\sqrt{n}$, and let $c = \sum_k kq_k$, $\bar{\beta} = \sum_i i\beta_i$. Since $\sum_k q_k = \sum_k \hat{p}_k = 1$, we have

$$\sum_k \beta_k = 0, \quad c = \hat{c} + \frac{\bar{\beta}}{\sqrt{n}}. \tag{73}$$

We remind the reader that in order to simplify notation, we shall use p instead of $p(c)$. Starting from (33),

$$\frac{\pi(q)}{\pi(\hat{p})} \sim \left(\frac{\psi(c)}{\psi(\hat{c})}\right)^n e^{\sqrt{n}\bar{\beta}} \prod_k \left(\frac{p_k}{q_k}\right)^{nq_k} \tag{74}$$

$$= \left(\frac{\psi(c)}{\psi(\hat{c})}\right)^n e^{\sqrt{n}\bar{\beta}} \prod_k \left(\frac{\hat{p}_k}{q_k}\frac{p_k}{\hat{p}_k}\right)^{nq_k} \tag{75}$$

$$= e^{\sqrt{n}\bar{\beta}} \prod_k \left(\frac{\hat{p}_k}{q_k}\frac{p_k\psi(c)^{-1}}{\hat{p}_k\psi(\hat{c})^{-1}}\right)^{nq_k} \tag{76}$$

$$= e^{\sqrt{n}\bar{\beta}} \prod_k \left(\frac{\hat{p}_k}{q_k}\right)^{nq_k} \prod_k \left(\frac{p_k\psi(c)^{-1}}{\hat{p}_k\psi(\hat{c})^{-1}}\right)^{nq_k}. \tag{77}$$

We shall compute the asymptotics of the two products separately. The first product gives

$$\prod_k \left(\frac{\hat{p}_k}{q_k}\right)^{nq_k} = \prod_k \left(1 + \frac{\beta_k}{\hat{p}_k\sqrt{n}}\right)^{-n\hat{p}_k - \sqrt{n}\beta_k} \tag{78}$$

$$\sim \prod_k \exp\left(-\sqrt{n}\beta_k - \frac{\beta_k^2}{2\hat{p}_k}\right) \tag{79}$$

$$= \prod_k \exp\left(-\frac{\beta_k^2}{2\hat{p}_k}\right), \tag{80}$$

where the last equality follows from (73). For the second product, from (27),

$$\log\left(\frac{p_k\psi(c)}{\hat{p}_k\psi(\hat{c})}\right) = \log\left(\left(\frac{\theta - \hat{c} - \bar{\beta}/\sqrt{n}}{\rho}\right)^{\theta-k}\left(\frac{\theta-\hat{c}}{\rho}\right)^{-(\theta-k)}\right) \tag{81}$$

$$\sim \log\left(1 - \frac{(\theta-k)\bar{\beta}}{(\theta-\hat{c})\sqrt{n}} + \frac{(\theta-k)(\theta-k-1)}{2}\frac{\bar{\beta}^2}{(\theta-\hat{c})^2 n}\right) \tag{82}$$

$$\sim \frac{-(\theta-k)\bar{\beta}}{(\theta-\hat{c})\sqrt{n}} - \frac{(\theta-k)}{2}\frac{\bar{\beta}^2}{(\theta-\hat{c})^2 n}. \tag{83}$$

Thus,

$$\prod_k \left(\frac{p_k\psi(c)^{-1}}{\hat{p}_k\psi(\hat{c})^{-1}}\right)^{nq_k} \sim \exp\left((n\hat{p}_k + \sqrt{n}\beta_k)\left(\frac{-(\theta-k)\bar{\beta}}{(\theta-\hat{c})\sqrt{n}} - \frac{(\theta-k)\bar{\beta}^2}{2(\theta-\hat{c})^2 n}\right)\right) \tag{84}$$

$$\sim -\sqrt{n}\bar{\beta} + \frac{\bar{\beta}^2}{2(\theta-\hat{c})}, \tag{85}$$

where the equalities (73), $\sum_k \beta_k = \bar{\beta}$ and $\sum_k \hat{p}_k = \hat{c}$ helped in the simplification.

Substituting the asymptotics of the two products in (77), we get

$$\frac{\pi(q)}{\pi(p)} = \exp\left(\frac{\bar{\beta}^2}{2(\theta-\hat{c})}\right)\prod_k \exp\left(-\frac{\beta_k^2}{2\hat{p}_k}\right). \tag{86}$$

Consider the exponent on the RHS. Since $\sum_i \beta_i = 0$, we have $\bar{\beta} = \sum_i i\beta_i = \sum_{i=0}^{\theta-1} i\beta_i - \theta\sum_{i=0}^{\theta-1}\beta_i = -\sum_i(\theta-i)\beta_i$. Therefore,

$$\frac{\bar{\beta}^2}{2(\theta-\hat{c})} - \frac{1}{2p_k}\sum_k \beta_k^2 = \frac{1}{2(\theta-\hat{c})}\left(\sum_{k=0}^{\theta-1}(\theta-k)\beta_k\right)^2 - \frac{1}{2\hat{p}_k}\sum_{k=0}^{\theta-1}\beta_k^2 - \frac{1}{2\hat{p}_\theta}\left(\sum_{i=0}^{\theta-1}\beta_i\right)^2 \tag{87}$$

$$= \frac{1}{2(\theta-\hat{c})}\left(\sum_{k=0}^{\theta-1}(\theta-k)\beta_k\right)^2 - \frac{1}{2\hat{p}_k}\sum_{k=0}^{\theta-1}\beta_k^2 - \frac{\psi}{2}\left(\sum_{i=0}^{\theta-1}\beta_i\right)^2. \tag{88}$$

Since the multivariate Gaussian distribution has exponent $-\frac{1}{2}\beta\Sigma^{-1}\beta$, we can deduce from the above equation the inverse of the covariance matrix to be that stated in the theorem and the local convergence of $\frac{\pi(q)}{\pi(p)}$ to the Gaussian density.

Using the approximation in (32), combined with the blocking probability estimates obtained in Theorem 5, it can be easily seen that

$$\pi(\hat{p}) \sim n^{-\theta/2},$$

which in turn implies that, for any $q = \hat{p} + \beta/\sqrt{n}$,

$$\pi(q)n^{-\theta/2} \to \exp\left(\frac{1}{2(\theta-\hat{c})}\left(\sum_{k=0}^{\theta-1}(\theta-k)\beta_k\right)^2 - \frac{1}{2\hat{p}_k}\sum_{k=0}^{\theta-1}\beta_k^2 - \frac{\psi}{2}\left(\sum_{i=0}^{\theta-1}\beta_i\right)^2\right).$$

Generalizing slightly the previous computations, the same would hold for any $q = \hat{p} + (\beta+\varepsilon_n)/\sqrt{n}$, with ε_n vanishing when n goes to infinity. Hence, to derive a global convergence result of the distribution function as stated in the theorem, we can now appeal to a variant of Scheffé's lemma (see, for instance, Theorem 1.29 in [35] with $\delta_i(n) = \frac{1}{\sqrt{n}}$, $i = 1, \ldots, k$ and $k = \theta$). □

A.2 Proof of Theorem 9

Proof Instead of defining q according to a predefined scaling as in the previous proof, we shall this time define it with an arbitrary scaling which shall be made precise later. Let $q = \hat{p} + \beta^{(n)}$, where again we have $\sum_k \beta_k^{(n)} = 0$. For $\rho = 1$, $\hat{p}_0 = \cdots = \hat{p}_{\theta-1} = 0$, $\hat{p}_\theta = 1$, so we shall assume that $\beta_k^{(n)} \geq 0$ for $k < \theta$. Also, for $\rho = 1$, we have $\hat{c} = \theta$ and $\psi(\hat{c}) = 1$ so that

$$c = \theta + \bar{\beta}^{(n)}, \quad \psi(c) = \sum_{j=0}^{\theta}\frac{(-\bar{\beta}^{(n)})^j}{j!}, \tag{89}$$

where $\bar{\beta}^{(n)} = \sum_k k\beta_k^{(n)} < 0$.

Our starting point is again (33), which for the present case reduces to

$$\frac{\pi(q)}{\pi(\hat{p})} \sim \psi(c)^n e^{n\bar{\beta}^{(n)}}\prod_k\left(\frac{p_k}{q_k}\right)^{nq_k} \tag{90}$$

$$= \left(\frac{\psi(c)}{q_\theta}\right)^{nq_\theta} e^{n\bar{\beta}^{(n)}}\prod_{k=0}^{\theta-1}\left(\frac{p_k\psi(c)}{\beta_k^{(n)}}\right)^{nq_k} \tag{91}$$

$$= \left(\frac{\psi(c)}{q_\theta}\right)^{nq_\theta} e^{n\bar{\beta}^{(n)}} \prod_{k=0}^{\theta-1} \left(\frac{1}{(\theta-k)!} \frac{(\theta - \hat{c} - \bar{\beta}^{(n)})^{\theta-k}}{\beta_k^{(n)}}\right)^{nq_k} \tag{92}$$

$$= \left(\frac{\psi(c)}{q_\theta}\right)^{nq_\theta} e^{n\bar{\beta}^{(n)}} \prod_{k=0}^{\theta-1} \left(\frac{1}{(\theta-k)!} \frac{(-\bar{\beta}^{(n)})^{\theta-k}}{\beta_k^{(n)}}\right)^{nq_k}. \tag{93}$$

Since $\beta^{(n)} \sim 0$ and $\bar{\beta}^{(n)} < 0$, the value of $k < \theta$ that makes the largest contribution is $\theta - 1$. For all other values of k, $\frac{(\bar{\beta}^{(n)})^{\theta-k}}{\beta_k^{(n)}} \to 0$ with respect to this fraction for $k = \theta - 1$. That is, fluctuations under this scaling will be visible only in $S_{\theta-1}^{(n)}$ and $S_\theta^{(n)}$ and not in lower values of k. In other words, given the number in the system, we can deduce the configuration to be $S_{\theta-1}^{(n)} = n\theta - nc$ and $S_\theta^{(n)} = n - S_{\theta-1}^{(n)}$. Therefore, there is only one vector p in the set $\mathcal{P}_c^{(n)}$. As a consequence, the only possible value of q in (90) is p, which then leads to

$$\frac{\pi(q)}{\pi(\hat{p})} \sim \psi(c)^n e^{n\bar{\beta}^{(n)}}. \tag{94}$$

Consider $q_{\theta-1} = \beta^{(n)} \geq 0$, where $\beta^{(n)}$ is a scalar from now on. Since $q_\theta = 1 - \beta^{(n)}$, we have $\bar{\beta}^{(n)} = -\beta^{(n)}$. Let us compute the asymptotics of the term with ψ:

$$n \log(\psi(c)) = n \log\left(\sum_{j=0}^{\theta} \frac{{\beta^{(n)}}^j}{j!}\right) \sim n\left(\beta^{(n)} - \frac{{\beta^{(n)}}^{\theta+1}}{(\theta+1)!}\right), \tag{95}$$

where the last asymptotic form is a consequence of Lemma 4. Substituting the above relation back in (94), we get

$$\frac{\pi(q)}{\pi(\hat{p})} \sim \exp\left(-n\frac{{\beta^{(n)}}^{\theta+1}}{(\theta+1)!}\right), \tag{96}$$

where we have used the identity $\bar{\beta}^{(n)} = -\beta^{(n)}$ which was noted previously.

Consequently, the right scaling for $\beta^{(n)}$ is $zn^{-1/(\theta+1)}$, for $z > 0$, which means that $S_{\theta-1}^{(n)} = n\beta^{(n)}$ lives on a scale of $n^{\theta/(\theta+1)}$. As for the proof of the central limit theorem, we can pass from local to global convergence by combining (96), the estimate of the blocking probabilities given in Theorem 7, and Theorem 1.29 in [35] with $\delta_1(n) = \frac{1}{\sqrt{n}}$ and $k = 1$. □

A.3 Proof of Theorem 10

Proof Following the same steps as in the proof of Theorem 9 until (93), we can arrive at the conclusion that $S_{\theta-1}^{(n)}$ and $S_\theta^{(n)}$ will be nonzero. Note that the only difference with the $\rho = 1$ case is that now

$$\psi(c) = \sum_{j=0}^{\theta} \frac{(-\bar{\beta}^{(n)})^j}{\rho^j j!}, \tag{97}$$

and p_k has a factor $\rho^{-(\theta-k)}$. Going further until (96) leads us to

$$\frac{\pi(q)}{\pi(\hat{p})} \sim \exp\left(-n\beta^{(n)} + n\frac{\beta^{(n)}}{\rho} - n\frac{{\beta^{(n)}}^{\theta+1}}{\rho^{\theta+1}(\theta+1)!}\right). \tag{98}$$

The only possible scaling is, thus, $\beta^{(n)} = zn^{-1}$, which means that the fluctuations of $S_\theta^{(n)}$ around $n\theta$ are $O(1)$.

We cannot carry on from this stage along the same line as that in the proof of Theorem 10, because to arrive at (94) we had assumed that the nonzero fluctuations were increasing with n. (This was needed to apply Stirling's approximation.) So, we shall work directly with the stationary distribution. From (12),

$$\mathbb{P}(S_{\theta-1}^{(n)} = s) = B_\theta^{(n)} s! \frac{n!}{s!(n-s)!}(n\rho)^{-s} \tag{99}$$

$$= \mathbb{P}(S_{\theta-1}^{(n)} = 0)\frac{n!}{(n-s)!}(n\rho)^{-s} \tag{100}$$

$$\sim \mathbb{P}(S_{\theta-1}^{(n)} = 0)\rho^{-s}, \tag{101}$$

which is a consequence of Stirling's approximation. □

A.4 Concavity of R

Lemma 2 *The function $R : \mathbb{R}_+ \to \mathbb{R}$ defined by*

$$R(t) = \log\left(\sum_{k=0}^{\theta} \frac{t^k}{k!}\right) - \rho t \tag{102}$$

is concave.

Proof Recall that $g_\theta(t) = \sum_{k=0}^{\theta} \frac{t^k}{k!}$. Rewrite $g_\theta(t)$ in terms of the incomplete gamma function using the following steps:

$$g_\theta(t) = \frac{1}{\Gamma(\theta+1,0)} \int_0^\infty (t+u)^\theta e^{-u} du \tag{103}$$

$$= e^t \tilde{\Gamma}(\theta+1, t), \tag{104}$$

where $\tilde{\Gamma}$ is the normalized incomplete gamma function, that is, $\tilde{\Gamma}(m,x) = \frac{\Gamma(m,x)}{\Gamma(m,0)}$.

To show the concavity of R, we shall show that its second derivative is negative. Note that $g'_\theta(t) = g_{\theta-1}(t)$ so that

$$R'(t) = \frac{g_{\theta-1}(t)}{g_\theta(t)} - \rho \tag{105}$$

and

$$R''(t) = \frac{g_\theta(t)g_{\theta-2}(t) - g_{\theta-1}(t)^2}{g_\theta(t)^2} \tag{106}$$

$$= \frac{\tilde{\Gamma}(\theta+1,t)\tilde{\Gamma}(\theta-1,t) - \tilde{\Gamma}(\theta,t)^2}{\tilde{\Gamma}(\theta+1,t)^2}. \tag{107}$$

It is shown in [1] that $\tilde{\Gamma}$, viewed as a function of θ, is log-concave for all $t > 0$. We can thus infer that R is concave in $(0, \infty)$. □

A.5 Generating functions

Let

$$\mathcal{M}^{(n)}(\mathbf{z}) = \sum_s \pi(s) \prod_{k=0}^{\theta} z_k^{s_k} \tag{108}$$

be the moment generating function of $S^{(n)}$.

Theorem 12

$$\mathcal{M}^{(n)}(\mathbf{z}) = B_\theta n\rho \int_0^\infty \left(\sum_{k=0}^{\theta} \frac{1}{k!} t^k z_{\theta-k} \right)^n e^{-tn\rho} dt. \tag{109}$$

Proof From (12) and using the facts that $x! = \int_0^\infty t^x e^{-t} dt$ and $(n\theta - \bar{s}) = \sum_k (\theta - k)s_k$,

$$\bar{\mathcal{M}}^{(n)}(z) = B_\theta \sum_s \int_0^\infty t^{\sum_k (\theta-k)s_k} e^{-t} dt \binom{n}{s} \prod_{k=0}^{\theta} \left(\frac{(n\rho)^{-(\theta-k)} z_k}{(\theta-k)!} \right)^{s_k} \tag{110}$$

$$= B_\theta \sum_s \binom{n}{s} \int_0^\infty \prod_{k=0}^{\theta} \left(\frac{((n\rho)^{-1}t)^{(\theta-k)} z_k}{(\theta-k)!} \right)^{s_k} e^{-t} dt \tag{111}$$

$$= B_\theta \int_0^\infty \sum_s \binom{n}{s} \prod_{k=0}^{\theta} \left(\frac{((n\rho)^{-1}t)^{(\theta-k)} z_k}{(\theta-k)!} \right)^{s_k} e^{-t} dt \tag{112}$$

$$= B_\theta \int_0^\infty \left(\sum_{k=0}^{\theta} \frac{1}{k!} t^k (n\rho)^{-k} z_{\theta-k} \right)^n e^{-t} dt, \tag{113}$$

where the last identity is a consequence of the multinomial theorem followed by a relabelling of the index inside the sum. Finally, making the transformation $t \mapsto tn\rho$ inside the integral completes the proof. □

Let

$$\bar{\mathcal{M}}^{(n)}(z) = \sum_j \left(\sum_{s:\sum_k k s_k = j} \pi(s) \right) z^j \tag{114}$$

be the moment generating function of the number of tasks in steady state.

Lemma 3

$$\bar{\mathcal{M}}^{(n)}(z) = \mathcal{M}^{(n)}\left(z^0, z^1, \ldots, z^\theta\right). \tag{115}$$

Proof From its definition,

$$\begin{aligned}
\bar{\mathcal{M}}^{(n)}(z) &= \sum_j \left(\sum_{s:\sum_k k s_k = j} \pi(s) \right) z^j && (116)\\
&= \sum_j \left(\sum_{s:\sum_k k s_k = j} \pi(s) z^{k s_k} \right) && (117)\\
&= \sum_j \left(\sum_{s:\sum_k k s_k = j} \pi(s) (z^k)^{s_k} \right) && (118)\\
&= \sum_s \pi(s) (z^k)^{s_k} && (119)\\
&= \mathcal{M}^{(n)}\left(z^0, z^1, \ldots, z^\theta\right). && (120)
\end{aligned}$$

□

Theorem 13

$$\bar{\mathcal{M}}^{(n)}(z) = B_\theta n\rho \int_0^\infty \left(\sum_{k=0}^{\theta} \frac{1}{k!} t^k z^{\theta-k} \right)^n e^{-tn\rho} dt. \tag{121}$$

Proof The result is a direct consequence of Theorem 12 and Lemma 3. □

A.6 Miscellaneous results

Lemma 4 *For* $\theta \geq 1$,

$$\log\left(\sum_{i=0}^{\theta} \frac{t^i}{i!} \right) = t - \frac{1}{(\theta+1)!} t^{\theta+1} + o\left(t^{\theta+1}\right), \quad \text{as} \quad t \to 0. \tag{122}$$

Proof Let $h_\theta(t) = \log\left(\sum_{i=0}^{\theta} \frac{t^i}{i!}\right)$. The proof is based on computing the coefficients in the Taylor series expansion of h around 0. The first derivative of h is

$$h_\theta^{(1)}(t) = 1 - \frac{t^\theta}{\theta!} g_\theta(t)^{-1}, \tag{123}$$

where $g_\theta(t) = \sum_{i=0}^{\theta} \frac{t^i}{i!}$, which gives the coefficient of t as 1.

For $k \leq \theta$, taking the kth derivative of $h^{(1)}$ and evaluating it using the product rule for higher-order derivatives, we obtain the $(k+1)$th derivative of h as

$$h_\theta^{(k+1)}(t) = -\sum_{j=0}^{k} \binom{k}{j} \frac{t^{\theta-j}}{(\theta-j)!} \left(g_\theta(t)^{-1}\right)^{(k-j)}, \tag{124}$$

where $(g_\theta(t)^{-1})^{(k-j)}$ is the $(k-j)$th derivative of $g_\theta(t)^{-1}$. Assuming that the derivatives of $g_\theta(t)^{-1}$ do not go to ∞ at $t = 0$ (which can be seen to be true), at $t = 0$, the only nonzero derivative is obtained for $k = \theta$ and $j = k$. That is,

$$h_\theta^{(k+1)}(0) = \begin{cases} 0 & 1 \leq k < \theta; \\ -1 & k = \theta. \end{cases} \tag{125}$$

□

References

1. Alzer, H., Baricz, A.: Functional inequalities for the incomplete gamma function. J. Math. Anal. Appl. **385**(1), 167–178 (2012)
2. Atar, R.: A diffusion regime with nondegenerate slowdown. Oper. Res. **60**(2), 490–500 (2012)
3. Benäim, M.: Dynamics of stochastic approximation algorithms. Lecture Notes in Math. Springer, Berlin, Séminaire de Probabilités XXXIII (1999)
4. Bonald, T., Jonckheere, M., Proutière, A.: Insensitive load balancing. SIGMETRICS Perform. Eval. Rev. **32**(1), 367–377 (2004)
5. Bonald, T., Massoulié, L., Proutière, A., Virtamo, J.: A queueing analysis of max-min fairness, proportional fairness and balanced fairness. Queueing Syst. Theory Appl. **53**(1–2), 65–84 (2006)
6. Bonald, T., Proutière, A.: Insensitive bandwidth sharing in data networks. Queueing Syst. Theory Appl. **44**(1), 69–100 (2003)
7. Bramson, M., Lu, Y., Prabakhar, B.: Asymptotic independence of queues under randomized load balancing. Queueing Syst. **71**, 247–292 (2012)
8. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, Hoboken (2006)
9. Eschenfeldt, P., Gamarnik, D.: Join the shortest queue with many servers. In: The Heavy Traffic Asymptotics (2015)
10. Foss, S., Stolyar, A.: Large-scale join-idle-queue system with general service times. ArXiv e-prints May (2016)
11. Graham, C.: Chaoticity on path space for a queueing network with selection of the shortest queue among several. J. Appl. Probab. **37**(1), 198–211 (2000)
12. Halfin, S., Whitt, W.: Heavy-traffic limits for queues with many exponential servers. Oper. Res. **29**(3), 567–588 (1981)
13. Hordijk, A.: Insensitive bounds for performance measures. In: Proceedings of the 12th International Teletraffic Congress (ITC 12), Torino (1988)

14. Hordijk, A., Koole, G.: On the optimality of the generalized shortest queue policy. Probab. Eng. Inf. Sci. **4**(4), 477–487 (1990)
15. Hordijk, A., Ridder, A.: Insensitive bounds for the stationary distribution of non-reversible Markov chains. J. Appl. Probab. **25**(1), 9–20 (1988)
16. http://information-technology.web.cern.ch/services/load-balancing-services
17. Jagerman, D.L.: Some properties of the Erlang loss function. Bell Syst. Tech. J. **53**(3), 525–551 (1974)
18. Janssen, A.J.E.M., Van Leeuwaarden, J.S.H., Zwart, B.: Gaussian expansions and bounds for the Poisson distribution applied to the Erlang B formula. Adv. Appl. Probab. **40**(1), 122–143 (2008)
19. Janssen, A.J.E.M., van Leeuwaarden, J.S.H., Zwart, A.P.: Corrected asymptotics for a multi-server queue in the Halfin–Whitt regime. Queueing Syst. Theory Appl. **58**(4), 261–301 (2008)
20. Janssen, A.J.E.M., van Leeuwaarden, J.S.H., Zwart, A.P.: Refining square-root safety staffing by expanding Erlang C. Oper. Res. **59**(6), 1512–1522 (2011)
21. Jonckheere, M.: Insensitive versus efficient dynamic load balancing in networks without blocking. Queueing Syst. **54**(3), 193–202 (2006)
22. Jonckheere, M., Mairesse, J.: Towards an Erlang formula for multiclass networks. Queueing Syst. **66**(1), 53–78 (2010)
23. Kipnis, C., Landim, C.: Scaling Limits of Interacting Particle Systems. Grundlehren der Mathematischen Wissenschaften. Springer, Berlin (2013)
24. Knessl, C., Yao, H.: On the finite capacity shortest queue problem. Prog. Appl. Math. **2**(1), 1–34 (2011)
25. Le Boudec, J.Y.: The stationary behaviour of fluid limits of reversible processes is concentrated on stationary points. Netw. Hetegnereous Med. **8**(2), 1529–540 (2013)
26. Leino, J., Virtamo, J.: Insensitive load balancing in data networks. Comput. Netw. **50**(8), 1059–1068 (2006)
27. Lu, Y., Xie, Q., Kliot, G., Geller, A., Larus, J.R., Greenberg, A.: Join-idle-queue: a novel load balancing algorithm for dynamically scalable web services. Perform. Eval. **68**(11), 1056–1071 (2011)
28. Maman, S.: Uncertainty in the demand for service: the case of call centers and emergency departments. Master's Thesis, Technion, April (2009)
29. Mitzenmacher, M.: The power of two choices in randomized load balancing. Ph.D. Thesis (1996)
30. Mukherjee, D., Borst, S.C., van Leeuwaarden, J.S.H., Whiting, P.A.: Universality of load balancing schemes on the diffusion scale. J. Appl. Probab. **53**(4), 1111–1124 (2016). 12
31. Mukhopadhyay, A., Karthik, A., Mazumdar, R.R., Guillemin, F.: Mean field and propagation of chaos in multi-class heterogeneous loss models. Perform. Eval. **91**, 117–131 (2015). (Special issue: performance 2015)
32. Nemes, G.: An explicit formula for the coefficients in Laplace's method. Constr. Approx. **38**(3), 471–487 (2013). arXiv:1207.5222
33. Pla, V., Virtamo, J., Martinez-Bauset, J.: Optimal robust policies for bandwidth allocation and admission control in wireless networks. Comput. Netw. **52**(17), 3258–3272 (2008)
34. Robert, P.: Stochastic Networks and Queues. Springer, Berlin (2003)
35. Sagitov, S.: Weak Convergence of Probability Measures. http://www.math.chalmers.se/~serik/WeakConv/C-space.pdf (2013)
36. Sparaggis, P.D., Towsley, D., Cassandras, C.G.: Extremal properties of the shortest/longest non-full queue policies in finite-capacity systems with state-dependent service rates. J. Appl. Probab. **30**(1), 223–236 (1993)
37. Stolyar, A.: Pull-based load distribution among heterogeneous parallel servers: the case of multiple routers. ArXiv e-prints (2015)
38. Vvedenskaya, N.D., Dobrushin, R.L., Karpelevich, F.I.: Queueing system with selection of the shortest of two queues: an asymptotic approach. Probl. Inf. Transm. **32**(1), 15–27 (1996)
39. Whitt, W.: Heavy traffic approximations for service systems with blocking. Bell Syst. Tech. J. **63**(4), 689–708 (1984)
40. Whittle, P.: Partial balance and insensitivity. J. Appl. Probab. **22**(1), 168–176 (1985)
41. Xie, Q., Dong, X., Lu, Y., Srikant, R.: Power of d choices for large-scale bin packing: a loss model. In: Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, SIGMETRICS'15, pp. 321–334. ACM, New York (2015)