Control of parallel non-observable queues: asymptotic equivalence and optimality of periodic policies*

Jonatha Anselmi†, Bruno Gaujal‡ and Tommaso Nesti§

Abstract

We consider a queueing system composed of a dispatcher that routes deterministically jobs to a set of non-observable queues working in parallel. In this setting, the fundamental problem is which policy should the dispatcher implement to minimize the stationary mean waiting time of the incoming jobs. We present a structural property that holds in the classic scaling of the system where the network demand (arrival rate of jobs) grows proportionally with the number of queues. Assuming that each queue of type r is replicated k times, we consider a set of policies that are periodic with period $k\sum_r p_r$ and such that exactly p_r jobs are sent in a period to each queue of type r. When $k \to \infty$, our main result shows that all the policies in this set are *equivalent*, in the sense that they yield the same mean stationary waiting time, and *optimal*, in the sense that no other policy having the same aggregate arrival rate to *all* queues of a given type can do better in minimizing the stationary mean waiting time. This property holds in a strong probabilistic sense. Furthermore, the limiting mean waiting time achieved by our policies is a convex function of the arrival rate in each queue, which facilitates the development of a further optimization aimed at solving the fundamental problem above for large systems.

1 Introduction

In computer and communication networks, the access of jobs to resources (web servers, network links, etc.) is usually regulated by a dispatcher. A fundamental problem is which algorithm should the dispatcher implement to minimize the mean delay experienced by jobs. There is a vast literature on this subject and the structure of the optimal algorithm strongly depends on *i)* the information available to the dispatcher, *ii)* the topology of the network and *iii)* how jobs are processed by resources. We are interested in a scenario where:

- The dispatcher has *static* information of the system;
- The network topology is parallel;
- Resources process jobs according to the first-come-first-served discipline.

Static information means that the dispatcher knows the probability distributions of job sizes and inter-arrival times but cannot observe the dynamic state of resources (e.g., the current number of jobs in their queues). This is motivated by the fact that real networks are composed of hundreds or thousands of resources, and allowing for dynamic information implies a non-negligible communication overhead and the problem of delayed information [29]. On the other hand, inferring statistics of inter-arrival times and job sizes in advance is a much easier task. The scenario above is of interest in volunteer computing, cloud computing, web server farms, etc.; see, e.g., [23, 24, 19] respectively.

*Research partially supported by grant SA-2012/00331 of the Departm ent of Industry, Innovation, Trade and Tourism (Basque Government) and grant MTM2010-17405 (Ministerio de Ciencia e Innovación, Spain) which sponsored an internship of Tommaso Nesti at BCAM.

†Basque Center for Applied Mathematics (BCAM), Al. de Mazarredo 14, 48009 Bilbao, Basque Country, Spain; INRIA Bordeaux Sud Ouest, 200 av. de la Vieille Tour, 33405 Talence, France. Email: jonatha.anselmi@inria.fr

‡INRIA and LIG, Zirst 51, Av. J. Kuntzmann, MontBonnot Saint-Martin, 38330, France. Email: bruno.gaujal@inria.fr

§Basque Center for Applied Mathematics (BCAM), Al. de Mazarredo 14, 48009 Bilbao, Basque Country, Spain; Dipartimento di Matematica, University of Pisa, Largo B. Pontecorvo 5, 56127 Pisa, Italy. Email: nesti@mail.dm.unipi.it

In this framework, the problems of finding a policy that minimizes the mean stationary delay and of determining the minimum mean stationary delay are both considered difficult; see, e.g., [2, 1] for an overview. A policy can be defined as a function that maps a natural number n, corresponding to the n-th job arriving to the dispatcher, to a probability mass function P_n over the set of resources. When the n-th job arrives, the dispatcher sends it to resource i with probability $P_n(i)$ independently of any other event in the past, present or future. Unfortunately, the problem of finding an optimal policy is intractable and for this reason two extreme families of policies have received particular attention in the literature: *probabilistic* policies, obtained when P_n is constant (in n), and *deterministic* policies, obtained when P_n puts the whole mass on a single resource.

When dealing with probabilistic policies, the difficulty of the problem is simplified by the fact that the arrival process at each resource is a renewal process, provided that the same holds for the arrival process at the dispatcher. This allows one to decompose the problem and, using the theory of the mean waiting time of the single GI/GI/1 queue, to immediately reduce it to a relatively simple optimization problem. This problem is usually convex and there exist efficient numerical procedures for their solution; e.g., [30, 34, 32, 12, 13, 8].

Contrariwise, when dealing with deterministic policies, one of the main difficulties is that the arrival process at each resource is hardly ever a renewal process. This prevents one from decomposing the problem and directly using the classic theory of the single queue as it has been done for probabilistic policies. Given this difficulty, researchers divided this problem in two subproblems:

i) In the first subproblem, the optimal deterministic policy is searched among all the deterministic policies ensuring that the long-term fractions of jobs to be sent to each resource is kept fixed (denote such fractions by vector p);

ii) In the second subproblem, the output of the first subproblem is employed to develop a further optimization over p.

In this paper, we focus on the first subproblem and, under some system scaling, we identify a set of policies that are optimal. This result is used to reduce the second subproblem to the solution of a convex optimization.

One of the folk theorem of queueing theory says that determinism in the inter-arrival times minimizes the waiting time of the single queue [17, 22]. In view of this classic insight and fixing fractions p, it is not surprising that an optimal policy tries to make the arrival process at each resource as *regular* (or less variable) as possible. Thus, our stochastic scheduling problem can be converted into a problem in word combinatorics. If the dispatcher must ensure fractions p, the main result known in the literature is that *balanced sequences* are optimal admission sequences [18, 1]. However, only very few vectors of rates p are balanceable and to decide whether or not p is balanceable is itself a hard combinatorial problem; see, e.g., [2, Chapter 2], which also contains an overview of which vectors are balanceable, and [36]. Matter of fact, the problem of finding an optimal deterministic policy is still considered difficult [10, 3, 37, 21, 2, 7, 20]. The only exceptions are when resources are stochastically equivalent, where round-robin[1] is known to be optimal in a strong sense [26], or when the dispatcher routes jobs to two resources, where balanced sequences can be always constructed no matter the value of rates p [18]. In presence of more than two queues, we stress that balanced sequences with given rates p do not exist in general. This non-existence makes the problem difficult and one still wonders which structure should an optimal policy have when p is not balanceable. When the routing is performed to two resources, jobs join the dispatcher following a Poisson process and service times have an exponential distribution, the optimal rates p as function of the inter-arrival and service times have a fractal structure, see [15, Figure 8]. This puts further light on the complexity of the problem.

While deterministic policies are believed to be more difficult to study than probabilistic policies (deterministic and probabilistic in the sense described above), they can achieve a significantly better performance [3]. This holds also for the variance of the waiting time because, as discussed above, the arrival process at each resource is much more regular in the deterministic case, especially if there are several resources as we show in this paper.

1.1 Contribution

In the framework described above, we are interested in deriving structural properties of deterministic policies when the system size is large. We study a scaling of the system where the arrival rate of jobs, λk, grows to

[1]Round-robin sends the n-th job to resource $(n \bmod R) + 1$, where R is the total number of resources.

infinity proportionally with the number of resources (queues in the following), Rk, while keeping the network load (or utilization) fixed. This scaling is often used in the queueing community. Specifically, there are R types of queues, and k is the number of queues in each type, i.e., the parameter that we will let grow to infinity. Beyond issues related to the tractability of the problem, this type of scaling is motivated by the fact that the size of real systems is large and that replication of resources is commonly used to increase system reliability.

First, with respect to a class of periodic policies, we define the random variable of the waiting time of each incoming job. This is done using Lindley's equation [25] and a suitable initial randomization. Using such randomization, we can adapt the framework developed by Loynes in [27] to our setting where jobs are sent to a set of parallel queues, and in our first result, Theorem 1, we show the monotone convergence in distribution of the waiting time of each incoming job. To the best of our knowledge, this is the first paper that characterizes the stationary distribution of the waiting time of incoming jobs in parallel FCFS queues when policies are deterministic.

Then, with respect to a given vector $p \in \mathbb{N}^R$, we define a certain subset of policies that are periodic with period $k\sum_r p_r$ and such that exactly p_r jobs are sent in a period to each queue of type r. While further details will be developed in Section 3.1, this set is meant to imply that queues of a given type are visited in a round-robin manner and that arrivals are "well distributed" among the different queue types. When $k \to \infty$, our main result states that all the policies in this set are *equivalent*, in the sense that they yield the same mean stationary waiting time, and *optimal*, in the sense that no other policy having the same aggregate arrival rate to *all* queues of a given type can do better in minimizing the mean stationary waiting time. In particular, we show that the stationary waiting time converges both in distribution and in expectation to the stationary waiting time of a system of independent D/GI/1 queues whose parameters only depend on p, λ and the distribution of the service times. This is shown in Theorems 2 and 3, respectively.

The main idea underlying our proof stands in analyzing the sequence of stationary waiting times along appropriate subsequences. Along these subsequences, it is possible to extract a powerful pattern for the arrival process of each queue that is common to all members of the subsequence. Such pattern is exploited to establish monotonicity properties in the language of stochastic orderings. These are properties holding for the considered subsequences only: they do not hold true along an arbitrary subsequence and counterexamples can be given. These properties will imply the uniform integrability of the sequence of stationary waiting times and will allow us to work on expected values.

Summarizing, fixing the proportions p of jobs to send to each queue type and given k large, our results state that all the periodic policies have almost the same performance and this lets us conclude that our limit is robust for predicting performance. Using known properties of the D/GI/1 queue, we also observe that the stationary mean waiting time obtained in our limit is a convex function of p. This facilitates the solution of subproblem *ii)* above, which is thus captured by the solution of a convex optimization problem.

This paper is organized as follows. Section 2 introduces the model under investigation and provides a characterization of the stationary waiting time (Theorem 1); Section 3 introduces a class of policies and presents our main results (Theorems 2 and 3); Section 4 is devoted to proofs; finally, Section 5 draws the conclusions of this paper.

2 Parallel queueing model

We consider a queueing system composed of R types of queues (or resources, servers) working in parallel. Each queue of type r is replicated k times, for all $r = 1, \ldots, R$, so there are kR queues in total. Parameter k is a scaling factor and we will let it grow to infinity. The service discipline of each queue is first-come-first-served (FCFS) and the buffer size of each queue is infinite. A stream of jobs (or customers) joins the queues through a dispatcher. The dispatcher routes each incoming job to a queue according to some policy and instantaneously. Figure 1 illustrates the structure of the queueing model under investigation. In the following, indices r, κ, n will be implicitely assumed to range from 1 to R, from 1 to k, in $\mathbb{N}$, respectively.

All the random variables that follow will be considered belonging to a fixed underlying probability triple $(\Omega, \mathcal{F}, \Pr)$.

Let $(T_n^{(k)})_{n\in\mathbb{N}}$ and $(S_{n,\kappa,r}^{(k)})_{n\in\mathbb{N}}$ be given sequences of i.i.d. random variables in $\mathbb{R}_+$[2]. These sequences are

[2]For any $E \subseteq \mathbb{R}$, we let $\mathbb{E}_+ \stackrel{\text{def}}{=} \{x \in E : x > 0\}$.

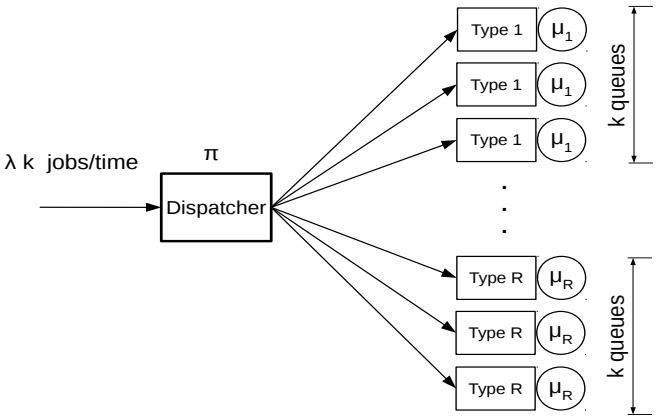


Figure 1: Structure of the parallel queueing model under investigation.

all assumed to be independent each other. Quantity $T_n^{(k)}$ is interpreted as the inter-arrival time between the n-th and the $(n+1)$-th jobs arriving to the dispatcher. Quantity $S_{n,\kappa,r}^{(k)}$ is interpreted as the service times of the n-th job arriving at the κ-th queue of type r. We assume that $S_{1,\kappa,r}^{(k)} =_{st} S_{1,1,r}^{(k)}$ and that $\mathbb{E}S_{n,\kappa,r}^{(k)} = \mu_r^{-1}$. For the arrival process at the dispatcher, we will refer to the following cases.

Case 1. *The process* $(T_n^{(k)})_{n\in\mathbb{N}}$ *is a renewal process with rate* λk *and such that* $\operatorname{Var} T_n^{(k)} = o(1/k)$.

Case 2. *The process* $(T_n^{(k)})_{n\in\mathbb{N}}$ *is a Poisson process with rate* λk.

Case 3. *The process* $(T_n^{(k)})_{n\in\mathbb{N}}$ *is constant with rate* λk, *i.e.,* $T_n^{(k)} = (\lambda k)^{-1}$.

It is clear that Cases 2 and 3 are both more restrictive than Case 1.

Let $\|\cdot\|$ denote the L_1-norm.

Let $q \stackrel{\text{def}}{=} (q_{r,\kappa})_{\forall r,\kappa} \in \mathbb{Q}_+^R \times \mathbb{Q}_+^k$ be such that $\|q\| = 1$. Quantity $q_{r,\kappa}$ will be interpreted as the proportion of jobs sent to queue (r,κ) in a period, or equivalently in the long term.

Let $n^* \stackrel{\text{def}}{=} n^*(q) \stackrel{\text{def}}{=} \min\{n \in \mathbb{Z}_+ : n\,q \in \mathbb{Z}_+^R \times \mathbb{Z}_+^k\}$. Since q is a vector of rational numbers, $n^* < \infty$.

Let V be a discrete random variable with values in $\{1,\ldots,n^*\}$ such that $\Pr(V = i) = 1/n^*$, for all $i = 1,\ldots,n^*$. We assume that V is independent of any other random variable.

Let $\mathcal{A}_q(k)$ be the set of all functions $\pi : \mathbb{N} \to \{1,\ldots,R\} \times \{1,\ldots,k\}$ such that for all r and κ

$$q_{r,\kappa} = \frac{1}{n^*} \sum_{n=1}^{n^*} \mathbf{1}_{\{\pi(n)=(r,\kappa)\}}, \qquad \pi(n) = \pi(n+n^*) \tag{1}$$

for all n, where $\mathbf{1}_E$ denotes the indicator function of event E. Thus, these functions are periodic with period n^* and $n^* q_{r,\kappa}$ is the number of jobs sent in a period to queue (r,κ). We refer to each element $\pi \in \mathcal{A}_q(k)$ as a *policy* (or a q-policy) operated by the dispatcher, and it is interpreted as follows: $\pi(n) = (r,\kappa)$ means that the n-th job arriving to the dispatcher is sent to the k-th queue of type r if $n \geq V$, otherwise it means that the n-th job is discarded. Thus, the outcome of random variable V gives the index of the first job that is actually served by some queue. In other words, $\pi(V)$ is the first queue that serves some job.

Let $(T_{n,\kappa,r}^{(k)}(\pi))_{n\in\mathbb{N}}$ be the sequence of inter-arrival times that are induced by policy π at the κ-th queue of type r (under any of the cases above). By construction, $T_{n,\kappa,r}^{(k)}(\pi)$ is the sum of a deterministic number of inter-arrival times seen at the dispatcher. The arrival process $(T_{n,\kappa,r}^{(k)}(\pi))_{n\in\mathbb{N}}$ can be made stationary if it is allowed a shift in time and a suitable randomization for the inter-arrival time of the first arrival of each queue. We assume for now that this has been done (details will be given at the beginning of Section 4). As done in

[27] and according to [14, p. 456], this implies that we can extend the stationary process $(T^{(k)}_{n,\kappa,r}(\pi))_{n\in\mathbb{N}}$ to form a stationary process $(T^{(k)}_{n,\kappa,r}(\pi))_{n\in\mathbb{Z}}$ (clearly, the same holds for the process $(S^{(k)}_{n,\kappa,r})_{n\in\mathbb{N}}$).

The waiting time of the n-th job arriving to the κ-th queue of type r induced by a policy $\pi \in \mathcal{A}(k)$ is denoted by $W^{(k)}_{n,r,\kappa}(\pi)$. It is the time between its arrival at the dispatcher (or equivalently at the queue) and the start of its service, and it is defined as follows: for $n = 0$, $W^{(k)}_{n,r,\kappa}(\pi) = 0$ and for $n > 0$,

$$W^{(k)}_{n,r,\kappa}(\pi) \stackrel{\text{def}}{=} \left(W^{(k)}_{n-1,r,\kappa}(\pi) + S^{(k)}_{n,\kappa,r} - T^{(k)}_{n,\kappa,r}(\pi)\right)^+, \tag{2}$$

where $x^+ \stackrel{\text{def}}{=} \max\{x, 0\}$. Equation (2) is known as Lindley's recursion [25]. The assumption that $W^{(k)}_{0,r,\kappa}(\pi) = 0$ serves to avoid technicalities[3]. It is known that the sequence of random variables $(W^{(k)}_{n,r,\kappa}(\pi))_{n\in\mathbb{N}}$ converges in distribution to the random variable

$$W^{(k)}_{r,\kappa}(\pi) \stackrel{\text{def}}{=} \left(\sup_{n\geq 0} \sum_{n'=0}^{n} S^{(k)}_{-n',\kappa,r} - T^{(k)}_{-n',\kappa,r}(\pi)\right)^+ \tag{3}$$

and that $W^{(k)}_{n,r,\kappa}(\pi) \leq_{st} W^{(k)}_{n+1,r,\kappa}(\pi)$, where $\leq_{st}$ denote the usual stochastic order; see [27]. We refer to $W^{(k)}_{r,\kappa}(\pi)$ as the *stationary* waiting time of jobs at the κ-th queue of type r.

Given π, let $f : \mathbb{N} \to \mathbb{N}\times\{1,\ldots,R\}\times\{1,\ldots,k\}$ be a mapping with the following meaning: $f(n) = (n', r, \kappa)$ means that the n-th job arriving to the dispatcher is the n'-th customer joining queue (r, κ). If $n < V$, no job is sent to any queue and thus we assume $f(n) = 0$. Note that $f(n)$ is a deterministic function of random variable V. Since V is uniform over $\{1,\ldots,n^*\}$, for all $n \geq n^*$ we have

$$\Pr(f_2(n) = r, f_3(n) = \kappa) = q_{r,\kappa}, \tag{4}$$

where f_j refers to the j-th component of f. With this notation, quantity $W^{(k)}_{f(n)}(\pi)$ is the waiting time of the n-th job arriving to the dispatcher induced by a policy $\pi \in \mathcal{A}_q(k)$, for all $n \geq n^*$.

Now, let $(Q_{r,\kappa})_{\forall r,\kappa}$ denote a partition of set $\{1,\ldots,n^*\}$. The intervals $Q_{r,\kappa}$, for all r and k, are thus disjoint and we further require that the number of points in $Q_{r,\kappa}$ is $n^*\, q_{r,\kappa}$. Let

$$W^{(k)}(\pi) \stackrel{\text{def}}{=} \sum_{r=1}^{R}\sum_{\kappa=1}^{k} \mathbf{1}_{\{V\in Q_{r,\kappa}\}} W^{(k)}_{r,\kappa}(\pi). \tag{5}$$

Since V is independent of the $W^{(k)}_{r,\kappa}(\pi)$'s, the distribution of $W^{(k)}(\pi)$ is the finite mixture of the $W^{(k)}_{r,\kappa}(\pi)$'s with weights q. Next theorem says that $W^{(k)}(\pi)$ can be interpreted as the right random variable describing the stationary waiting time of jobs achieved with policy $\pi \in \mathcal{A}_q(k)$. It is proven by adapting the framework developed by Loynes in [27]. In the remainder of the paper, convergence in distribution (in probability) is denoted by $\stackrel{d}{\to}$ (respectively, $\stackrel{\Pr}{\longrightarrow}$).

Theorem 1. *Let Case 1 hold. Let q be such that $q_{r,\kappa}\lambda < \mu_r$ for all r, κ and $\pi \in \mathcal{A}_q(k)$. Then,*

$$W^{(k)}_{f(n)}(\pi) \leq_{st} W^{(k)}_{f(n+1)}(\pi) \tag{6}$$

and

$$W^{(k)}_{f(n)}(\pi) \xrightarrow[n\to\infty]{d} W^{(k)}(\pi). \tag{7}$$

By using the monotone convergence theorem, Theorem 1 implies that also the moments of $W^{(k)}_{f(n)}(\pi)$ converge to the moments of $W^{(k)}(\pi)$, provided that they are finite.

Finally, for a given $p \in \mathbb{R}^R_+$, we define the auxiliary random variable $W_r(p)$, which corresponds to the stationary waiting time of a D/GI/1 queue with inter-arrival times $(T_{n,r}(p))_{n\in\mathbb{N}}$ where $T_{n,r}(p) = \|p\|/(p_r\lambda)$

[3]Using a standard coupling argument and that each queue will empty in finite time almost surely, what follows can be generalized easily to the case where $W^{(k)}_{0,r,\kappa}(\pi) \geq 0$.

and service times $(S_{n,1,r})_{n\in\mathbb{N}}$, for all r. Let also $\overline{V} \stackrel{\text{def}}{=} \overline{V}(p)$ be a random variable with values in $\{1,\ldots,R\}$ independent of any other random variable and such that $\Pr(\overline{V}=r) = p_r/\|p\|$, for all r, and let

$$W(p) \stackrel{\text{def}}{=} \sum_{r=1}^{R} \mathbf{1}_{\{\overline{V}=r\}} W_r(p). \tag{8}$$

Note that $\mathbb{E}W(p) = \sum_r \frac{p_r}{\|p\|}\mathbb{E}W_r(p)$ can be interpreted as the waiting time of R independent D/GI/1 queues averaged over weights $p/\|p\|$.

We will be interested in establishing convergence results for $W^{(k)}$ when $k\to\infty$. With respect to some policies to be defined, we will show forms of convergence to $W(p)$.

2.1 Discussion

Some remarks about the model above and Theorem 1 follow:

- To the best of our knowledge, the characterization of the *distribution* of the stationary waiting time $W^{(k)}$ given in Theorem 1 is new in the literature, though its proof is based on classical arguments. While we believe that such characterization has its own interest, in our paper we need to formally define the distribution of the stationary waiting time (rather than only its expected value as existing works do) because we can only prove our main result, Theorem 3, through a distributional convergence argument, Theorem 2.
- Though several works focused on finding policies that minimize the expected value $\mathbb{E}W^{(k)}$ (see the Introduction), the analysis of $\mathbb{E}W^{(k)}$ in our scaling where $k\to\infty$ seems new.
- We assume that q is a vector of rational numbers. From a practical standpoint, this is not a loss of generality for obvious reasons. Our approach needs this structure to prove (7). In the case where $\pi(n)$ is not periodic, a case that we do not consider here, and the limit $\lim_{n\to\infty}\frac{1}{n}\sum_{n'=1}^{n}\mathbf{1}_{\{\pi(n')=(r,\kappa)\}}$ does not exist, it would be interesting to know whether some convergence in distribution of $W^{(k)}_{f(n)}(\pi)$ occurs.
- The fact that in Case 1 we require that $\operatorname{Var} T_n^{(k)} = o(k)$ is not a loss of generality for Theorem 1, as we do not let $k\to\infty$ there. Case 1 covers the case where $T_n^{(k)}$ has a (Gk,α)-phase-type distribution where both G and α are not functions of k.
- We may have assumed that each queue of type r was replicated kz_r times, $z_r\in\mathbb{N}$, instead of just k times. This is equivalent to our setting, as we can assume that there are z_r queues of type r when $k=1$.

3 Main results

For $p\in\mathbb{Z}_+^R$, let

$$\mathcal{P}_p(k) \stackrel{\text{def}}{=} \left\{ q\in\mathbb{Q}_+^R\times\mathbb{Q}_+^k : \sum_{\kappa=1}^{k}\frac{q_{r,\kappa}}{k} = \frac{p_r}{\|p\|},\ \forall r \right\} \tag{9}$$

and

$$\overline{\mathcal{A}}_p(k) \stackrel{\text{def}}{=} \bigcup_{q\in\mathcal{P}_p(k)} \mathcal{A}_q(k). \tag{10}$$

The set $\overline{\mathcal{A}}_p(k)$ is interpreted as the set of all periodic policies for which $\lambda\frac{p_r}{\|p\|}$ is the mean arrival rate of jobs to each queue of type r. We consider the following problem.

Problem 1. *Let $p\in\mathbb{Z}_+^R$ be given. Determine the optimizers and the optimal objective function value of*

$$\min_{\pi\in\overline{\mathcal{A}}_p(k)} \mathbb{E}W^{(k)}(\pi). \tag{11}$$

As discussed in the Introduction, this is considered as a difficult problem. We are interested in establishing structural properties of Problem 1 when k is large. In the following, we define a certain class of policies and then present our main results.

3.1 A class of periodic policies

For $p \in \mathbb{Z}_+^R$, we define $\mathcal{C}_p^{(k)}$ as the subset of all policies $\pi \in \overline{\mathcal{A}}_p(k)$ that satisfy the following properties:

- Among every $k\|p\|$ consecutively-arriving jobs, exactly p_r jobs are sent to each queue of type r, for each r. That is,
$$\sum_{n'=n+1}^{k\|p\|+n} \mathbf{1}_{\{\pi(n')=(r,\kappa)\}} = p_r, \quad \forall n.$$
- Among every $\|p\|$ consecutively-arriving jobs, exactly p_r jobs of type r arrive. That is,
$$\sum_{\kappa=1}^{k} \sum_{n'=n+1}^{\|p\|+n} \mathbf{1}_{\{\pi(n')=(r,\kappa)\}} = p_r, \quad \forall n.$$
- Let $n_1, n_2, \ldots$ be the subsequence of all jobs that are sent to queues of type r. Then, $\pi_2(n_1) = 1$, and $\pi(n_{j+1}) = \pi(n_j) + (0,1)$ if $\pi_2(n_j) < k$ and $\pi(n_{j+1,r}) = (r,1)$ otherwise.

The first property says that π is periodic with period $k\|p\|$ and that in a period exactly p_r jobs are sent to each queue of type r. The second property says that arrivals remain "well distributed" among the different types when k increases. Finally, the third property says that queues of type r are accessed by jobs in a round-robin (or cyclic) order starting from queue 1, and we need it to ensure that the cardinality of $\mathcal{C}_p^{(k)}$, i.e., $|\mathcal{C}_p^{(k)}|$, does not vary with k. In fact, one can verify that

$$|\mathcal{C}_p^{(k)}| = \frac{\|p\|!}{\prod_r (p_r!)}. \tag{12}$$

These policies can be implemented in a distributed manner. More precisely, one can think that there are two tiers of dispatchers: the dispatcher in the first tier schedules jobs *inter*-group, while the dispatchers in the second tier, R in total, schedule jobs *intra*-group and implement round-robin.

With respect to a sequence of policies $(\pi^{(k)})_{k\in\mathbb{N}}$, where $\pi^{(k)} \in \mathcal{C}_p^{(k)}$, we will show (Lemma 2)

$$T_{n,\kappa,r}^{(k)}(\pi^{(k)}) \xrightarrow[k\to\infty]{\Pr} \frac{\|p\|}{p_r \lambda}, \qquad \forall n. \tag{13}$$

This means that the finite dimensional distributions of the arrival process at each queue will be 'close' to the deterministic process, when k is large, which implies that some form of convergence to $W(p)$ should occur in view of the continuity of the stationary waiting time [11].

3.2 Asymptotic equivalence and optimality

Next theorem proves a first form of convergence of $W^{(k)}(\pi^{(k)})$ to $W(p)$.

Theorem 2. *Let Case 1 hold. Let $p \in \mathbb{Z}_+^R$ be such that $\lambda \frac{p_r}{\|p\|} < \mu_r$ for all r. Let also an arbitrary sequence $(\pi^{(k)})_{k\in\mathbb{N}}$ be given where $\pi^{(k)} \in \mathcal{C}_p^{(k)}$ for all k. Then*

$$W^{(k)}(\pi^{(k)}) \xrightarrow[k\to\infty]{d} W(p). \tag{14}$$

Unfortunately, Theorem 2 is not enough to claim that $\mathbb{E}W^{(k)}(\pi^{(k)})$ converges as well. Convergence of the first moment is important from an operational standpoint, as in practice one desires to optimize over $\mathbb{E}W^{(k)}$ or $\operatorname{Var} W^{(k)}$. Under some additional assumptions, next theorem states that also the expected value and the

variance of $W^{(k)}$ converge. Furthermore, it states that all the policies in set $\mathcal{C}_p^{(k)}$ are both asymptotically equivalent and asymptotically optimal, with respect to the criterion in Problem 1.

Theorem 3. *Let Case 2 or 3 hold. Let $p \in \mathbb{Z}_+^R$ be such that $\lambda \frac{p_r}{\|p\|} < \mu_r$ for all r. Let also an arbitrary sequence $(\pi^{(k)})_{k\in\mathbb{N}}$ be given where $\pi^{(k)} \in \mathcal{C}_p^{(k)}$ for all k. If $\mathbb{E}[(S_{1,\kappa,r}^{(k)})^3] < \infty$, then*

$$\liminf_{k\to\infty} \min_{\pi\in\overline{\mathcal{A}}_p(k)} \mathbb{E}W^{(k)}(\pi) = \lim_{k\to\infty} \mathbb{E}W^{(k)}(\pi^{(k)}) \tag{15}$$

$$\lim_{k\to\infty} \mathbb{E}W^{(k)}(\pi^{(k)}) = \mathbb{E}W(p). \tag{16}$$

Furthermore, if $\mathbb{E}[(S_{1,\kappa,r}^{(k)})^5] < \infty$, then

$$\lim_{k\to\infty} \operatorname{Var} W^{(k)}(\pi^{(k)}) = \sum_r \frac{p_r}{\|p\|} \left(\operatorname{Var} W_r(p) + (\mathbb{E}W_r(p) - \mathbb{E}W(p))^2 \right). \tag{17}$$

Provided that k is large, thus, no other policy in $\overline{\mathcal{A}}_p(k) \setminus \mathcal{C}_p^{(k)}$ can do better than any $\pi^{(k)} \in \mathcal{C}_p^{(k)}$ to minimize $\mathbb{E}W^{(k)}$. It also explicits the limiting value of $\mathbb{E}W^{(k)}(\pi^{(k)})$, which is $\mathbb{E}W(p)$. It is known that $\mathbb{E}W(p)$ is a convex function in p (e.g., [30]).

To prove Theorem 3, we use the well-known fact that [5]

$$W_r(p) \leq_{icx} W_{r,\kappa}^{(k)}(\pi) \tag{18}$$

for any $\pi \in \overline{\mathcal{A}}_p(k)$. Using this lower bound first and then that the waiting time of the D/GI/1 queue is convex increasing in its arrival rate, it is not difficult to show that

$$\mathbb{E}W(p) \leq \mathbb{E}W^{(k)}(\pi). \tag{19}$$

Then, we prove that the sequence $\mathbb{E}W^{(k)}(\pi^{(k)})$ is upper bounded by a sequence that converges to the lower bound in (19). An observation here is that the lower bound (18) holds under conditions that are weaker than those assumed in this paper; see [22]. For instance, it is possible to extend (19) (and thus Theorem 3) to the case where i) policies are not periodic, ii) the fractions of jobs to send in each queue are not necessarily rational numbers, and iii) policies are randomized [31], that is the case where $\pi(n)$ is any probability mass function over the set of queues. We do not investigate these extensions in further detail.

4 Proofs

In this section, we develop proofs for Theorems 1, 2 and 3. Before doing this, we fix some additional notation and show how it is possible to make the arrival process at each queue stationary (as assumed in Section 2).

Let us consider the κ-th queue of type r and its arrival process $(T_{n,\kappa,r}^{(k)})_{n\in\mathbb{N}}$. Each inter-arrival time depends on the policy $\pi^{(k)}$ implemented by the dispatcher, i.e., $T_{n,\kappa,r}^{(k)} = T_{n,\kappa,r}^{(k)}(\pi^{(k)})$, though in the following we drop such dependency for notational simplicity. Since $\pi^{(k)}$ is periodic by construction with period $k\|p\|$ and p_r jobs have to be sent within a cycle to each type-r queue, the sequence $(T_{n,\kappa,r}^{(k)})_{n\in\mathbb{N}}$ is composed of a repeated pattern of p_r inter-arrival times, that we can write as

$$A_{1,\kappa,r}^{(k)}, A_{2,\kappa,r}^{(k)}, \ldots, A_{p_r,\kappa,r}^{(k)}, \tag{20}$$

where each quantity $A_{j,\kappa,r}^{(k)}$, $j = 1, \ldots, p_r$, is the sum of a deterministic number (that depends on $\pi^{(k)}$) of inter-arrival times to the dispatcher. We denote such number by $a_{j,r}^{(k)}$ and notice that it does not vary with κ because by symmetry the arrival processes of all queues of a given type are equal, in distribution, up to a shift in time. Thus,

$$A_{j,\kappa,r}^{(k)} =_{st} \sum_{n=1}^{a_{j,r}^{(k)}} T_n^{(k)}, \tag{21}$$

where $=_{st}$ denotes equality in distribution. It is also clear that

$$\sum_{j=1}^{p_r} a_{j,r}^{(k)} = k\|p\|. \tag{22}$$

We want to make $(T_{n,\kappa,r}^{(k)})_{n\in\mathbb{N}}$ stationary. This can be done as follows by randomizing over the first inter-arrival time of queue (r,κ). Now, let us consider the auxiliary random variables $U_{r,\kappa}$, for all r and κ, which we assume independent each other and of any other random variable and having a uniform distribution in $[0,1]$. Then, we take

$$T_{1,r,\kappa}^{(k)} \stackrel{\text{def}}{=} \sum_{j=1}^{p_r} A_{j,r,\kappa}^{(k)} \mathbf{1}_{\left\{U_{r,\kappa}\in\left[\frac{j-1}{p_r},\frac{j}{p_r}\right]\right\}} \tag{23}$$

Therefore, if $T_{1,r,\kappa}^{(k)} = A_{j,r,\kappa}^{(k)}$, for some $j < p_r$, then $T_{2,r,\kappa}^{(k)} = A_{j+1,r,\kappa}^{(k)}$ and so forth according to the pattern (20). Defined in this manner, one can see that $(T_{n,r,\kappa}^{(k)})_{n\in\mathbb{N}}$ is stationary. Furthermore, using (22)

$$\mathbb{E}T_{n,\kappa,r}^{(k)} = \tfrac{1}{p_r}\sum_{j=1}^{p_r} a_{j,r}^{(k)}\, \mathbb{E}T_n^{(k)} = \tfrac{1}{p_r}\sum_{j=1}^{p_r} \tfrac{a_{j,r}^{(k)}}{k\lambda} = \tfrac{\|p\|}{p_r\lambda}. \tag{24}$$

At this point, the issue is the following. Consider two queues, say (r,κ) and (r',κ'), and suppose that $T_{1,r,\kappa}^{(k)} = A_{j,r,\kappa}^{(k)}$ and $T_{1,r',\kappa'}^{(k)} = A_{j',r',\kappa'}^{(k)}$. We should check whether policy $\pi^{(k)}$ is actually able to induce an arrival process equal (samplepath-wise) to the one built above. One can easily see that this can be done with a possible shift of time for the arrival process at the queues and possibly discarding a finite number of jobs. This is allowed because these operations do not change the stationary behavior.

In the remainder, we will use stochastic orderings. We will denote by $\leq_{st}$, $\leq_{cx}$ and $\leq_{icx}$, the *usual stochastic order*, the *convex order* and the *increasing convex order*, respectively; we point to, e.g., [33, 35] for their definition.

We will also refer to the following lemma, which can be easily proven by using a diagonalization argument.

Lemma 1. *Let $(D_k)_{k\in\mathbb{N}}$ be a sequence of ordered finite sets having $N > 1$ elements each. For $n \in D_k$, let $\#n \in \mathbb{N}$ denote the position of n in D_k (e.g., if $D_k = \{c,a,b\}$ then $\#a = 2$). Let $(n_k)_{k\in\mathbb{N}}$ be a sequence such that $n_k \in D_k$, for all k. Let $(f_k)_{k\in\mathbb{N}}$ be a sequence of functions where $f_k : D_k \to \mathbb{R}$. If $\lim_{k\to\infty} f_k(n_k) = c$ for all sequences $(n_k)_{k\in\mathbb{N}}$ such that $\#n_k = \#n_1$, for all $n_1 \in D_1$, then $\lim_{k\to\infty} f_k(n_k) = c$ for all sequences $(n_k)_{k\in\mathbb{N}}$.*

4.1 Proof of Theorem 1

Let random variables V_1 and V_2 be given such that $V_1 =_{st} V_2 =_{st} V$. We prove (6) through Strassen's theorem building a coupling $(\tilde{W}_{f(n)}^{(k)}(\pi), \tilde{W}_{f(n+1)}^{(k)}(\pi))$ of $W_{f(n)}^{(k)}(\pi)$ and $W_{f(n+1)}^{(k)}(\pi)$ through V_1 and V_2 ensuring that $\tilde{W}_{f(n)}^{(k)}(\pi) \leq \tilde{W}_{f(n+1)}^{(k)}(\pi)$. This is done as follows. First, let $\tilde{W}_{f(n)}^{(k)}(\pi)$ and $\tilde{W}_{f(n+1)}^{(k)}(\pi)$ be the Loynes waiting times (see [27, p. 501]) obtained when the first queue to serve a job is given by the outcome of V_1 and V_2, respectively; thus, $\tilde{W}_{f(n)}^{(k)}(\pi)$ is the waiting time at time 0 with n' jobs in the past at queue (r,κ), provided that $f(n) = (n',r,\kappa)$. Then, we take $V_2 = V_1 - 1$ if $V_1 > 1$ otherwise $V_2 = n^*$, and let the (Loynes) waiting times be driven by the same realizations of the random inter-arrival and service times.

We now prove (7). Since π is periodic with period n^*, we first observe that π returns the same queue along subsequence $(n\,n^* + i)_{n\in\mathbb{N}}$, for all $i = 1,\ldots,n^*$. Similarly, also the second and third component of $f(n\,n^* + i)$ do not change along these subsequences, though they are not known in advance because they depend on the outcome of random variable V, see (4). Thus, for all $i = n^* + 1,\ldots,2n^*$, by construction we have

$$\Pr(f_2(n\,n^* + i) = r, f_3(n\,n^* + i) = \kappa) = \Pr(f_2(i) = r, f_3(i) = \kappa) = q_{r,\kappa}, \tag{25}$$

and we get

$$\lim_{n\to\infty} \Pr(W_{f(nn^*+i)}^{(k)}(\pi) \leq t) \tag{26a}$$

$$= \lim_{n\to\infty} \sum_{r,\kappa} q_{r,\kappa} \Pr(W^{(k)}_{n,r,\kappa}(\pi) \le t \,|\, (f_2(i), f_3(i)) = (r,\kappa)) \tag{26b}$$

$$= \sum_{r,\kappa} q_{r,\kappa} \Pr(W^{(k)}_{r,\kappa}(\pi) \le t) \tag{26c}$$

$$= \Pr(W^{(k)}(\pi) \le t). \tag{26d}$$

In (26b), we have conditioned on $f(i)$. In (26c), we have used that $W^{(k)}_{n,r,\kappa}(\pi)$ converges in distribution to $W^{(k)}_{r,\kappa}(\pi)$; see [27]. In (26d), we have used the definition of $W^{(k)}(\pi)$. Now, since the limit in (26d) does not depend on i, the proof is concluded by applying Lemma 1.

4.2 Proof of Theorem 2

We first observe that we can prove this theorem under some assumption on the sequence $(\pi^{(k)})_{k\in\mathbb{N}}$. Given $\pi^{(1)} \in \mathcal{C}_p(1)$, we require that for all k:

$$\pi^{(k)}_1(n) = \pi^{(1)}_1(n), \qquad \forall n. \tag{27}$$

These sequences will be assumed along this proof. If this theorem holds for these sequences, then it also holds for all the sequences in view of Lemma 1 and of the fact that the cardinality of $\mathcal{C}_p(k)$ does not vary with k.

For $m \in \mathbb{N}$, let

$$k_m \stackrel{\text{def}}{=} m \,\mathrm{lcm}(p), \tag{28}$$

where $\mathrm{lcm}(p)$ denotes the least common multiple of $p_1, \ldots, p_R$. The subsequences $(k_m + i)_{m\in\mathbb{N}}$, for all $i = 1, \ldots, \mathrm{lcm}(p)$, play a key role in our proof and will be especially used also in the proof of Theorem 3. Along these subsequences, next fact holds true and follows by construction of the policies in set $\mathcal{C}^{(k)}_p$: it is a direct consequence of the fact that queues of the same type are visited in a round-robin manner.

Fact 1. *For $m > 1$, $j = 1, \ldots, p_r$ and $i = 1, \ldots, \mathrm{lcm}(p)$, $a^{(p_r m+i)}_{j,r} = a^{(p_r(m-1)+i)}_{j,r} + \|p\|$.*

Unfolding the recursion in Fact 1, for $m \in \mathbb{N}$, $i = 1, \ldots, \mathrm{lcm}(p)$, we get

$$a^{(k_m+i)}_{j,r} = a^{(k_{m-1}+i)}_{j,r} + \tfrac{\|p\|}{p_r}\mathrm{lcm}(p) = a^{(i)}_{j,r} + m\tfrac{\|p\|}{p_r}\mathrm{lcm}(p) \tag{29}$$

and therefore

$$\frac{a^{(k_m+i)}_{j,r}}{k_m + i} = \frac{a^{(i)}_{j,r} + m\frac{\|p\|}{p_r}\mathrm{lcm}(p)}{m\cdot\mathrm{lcm}(p) + i} \xrightarrow[m\to\infty]{} \frac{\|p\|}{p_r}. \tag{30}$$

Since (30) holds for all $i = 1, \ldots, \mathrm{lcm}(p)$, by using Lemma 1 we obtain

$$\lim_{k\to\infty} \tfrac{a^{(k)}_{j,r}}{k} = \tfrac{\|p\|}{p_r}. \tag{31}$$

Lemma 2. *Under the hypotheses of Theorem 2, $T^{(k)}_{n,r,\kappa} \to \frac{\|p\|}{\lambda p_r}$ in probability, as $k \to \infty$.*

Proof. For all $\epsilon > 0$,

$$\Pr\left(|T^{(k)}_{n,r,\kappa} - \tfrac{\|p\|}{\lambda p_r}| \ge \epsilon\right) = \Pr\left(|T^{(k)}_{n,r,\kappa} - \mathbb{E}T^{(k)}_{n,r,\kappa}| \ge \epsilon\right) \tag{32a}$$

$$\le \frac{1}{\epsilon^2}\mathrm{Var}\, T^{(k)}_{n,r,\kappa} \tag{32b}$$

$$= \frac{1}{\epsilon^2}\left(\mathbb{E}(\mathrm{Var}\, T^{(k)}_{n,r,\kappa}|U_{r,\kappa}) + \mathrm{Var}\,\mathbb{E}(T^{(k)}_{n,r,\kappa}|U_{r,\kappa})\right) \tag{32c}$$

$$= \frac{1}{\epsilon^2}\left(\frac{1}{p_r}\sum_{j=1}^{p_r} \mathrm{Var}\, A^{(k)}_{j,r,\kappa} + (\mathbb{E}A^{(k)}_{j,r,\kappa} - \mathbb{E}T^{(k)}_{n,r,\kappa})^2\right) \tag{32d}$$

$$= \frac{1}{\epsilon^2}\left(\frac{1}{p_r}\sum_{j=1}^{p_r} a_{j,r}^{(k)} \operatorname{Var} T_1^{(k)} + \left(\frac{a_{j,r}^{(k)}}{\lambda k} - \frac{\|p\|}{\lambda p_r}\right)^2\right) \tag{32e}$$

$$\xrightarrow[k\to\infty]{} 0. \tag{32f}$$

In (32b), (32c) and (32e), we have used Chebyshev's inequality, the law of total variance and that the $T_n^{(k)}$'s are i.i.d., respectively. In (32f), we have used (31) and that $a_{j,r}^{(k)} \operatorname{Var} T_1^{(k)} \to 0$ because $\operatorname{Var} T_1^{(k)} = o(k)$ and $a_{j,r}^{(k)} \le k\|p\|$. □

Since $T_{n,r,\kappa}^{(k)}$ converge in probability for each n, also the finite dimensional distributions of the process $(T_{n,r,\kappa}^{(k)})_{n\in\mathbb{N}}$ converge to the one of the constant process with rate $\frac{\lambda p_r}{\|p\|}$. Together with the fact that $\mathbb{E}T_{n,r,\kappa}^{(k)} = \lim_{k\to\infty} \mathbb{E}T_{n,r,\kappa}^{(k)} = \frac{\|p\|}{\lambda p_r}$, we can use the continuity of the stationary waiting time (see [11, Theorem 22]) to establish that

$$W_{r,\kappa}^{(k)}(\pi^{(k)}) \xrightarrow[k\to\infty]{d} W_r(p) \tag{33}$$

holds. Using (33) and that Cesaro sums converge if each addend converges, we obtain

$$\lim_{k\to\infty} \Pr(W^{(k)}(\pi^{(k)}) \le t) = \lim_{k\to\infty} \sum_{r=1}^{R} \frac{p_r}{\|p\|k} \sum_{\kappa=1}^{k} \Pr(W_{r,\kappa}^{(k)}(\pi^{(k)}) \le t) \tag{34a}$$

$$= \sum_{r=1}^{R} \frac{p_r}{\|p\|} \Pr(W_r(p) \le t) = \Pr(W(p) \le t) \tag{34b}$$

as desired.

4.3 Proof of Theorem 3

Proof of (15) *and* (16). Given that $\mathcal{C}_p^{(k)} \subseteq \overline{\mathcal{A}}_p(k)$, (15) and (16) hold true if

$$\mathbb{E}W(p) \le \mathbb{E}W^{(k)}(\pi) \tag{35}$$

for all $\pi \in \overline{\mathcal{A}}_p(k)$ and

$$\lim_{k\to\infty} \mathbb{E}W^{(k)}(\pi^{(k)}) \le \mathbb{E}W(p). \tag{36}$$

Let $\mathbb{E}W_{r,\kappa}(x)$ be the mean waiting time of a D/GI/1 queue with arrival rate λx and i.i.d. service times having the same distribution of $S_{1,r,\kappa}$. Inequality (35) is a fairly direct application of known results: for all $\pi \in \overline{\mathcal{A}}_p(k)$,

$$\mathbb{E}W^{(k)}(\pi) = \sum_{r,\kappa} q_{r,\kappa} \mathbb{E}W_{r,\kappa}^{(k)}(\pi) \tag{37a}$$

$$\ge \sum_{r,\kappa} q_{r,\kappa} \mathbb{E}W_{r,\kappa}(kq_{r,\kappa}) \tag{37b}$$

$$\ge \sum_{r,\kappa} \frac{p_r}{k\|p\|} \mathbb{E}W_{r,\kappa}(\tfrac{p_r}{\|p\|}) \tag{37c}$$

$$= \sum_{r} \frac{p_r}{\|p\|} \mathbb{E}W_r(p) = \mathbb{E}W(p). \tag{37d}$$

In (37b), we have used the lower bound in [22]. In (37c), we have used Karamata's inequality once noticing that i) $\mathbb{E}W_{r,\kappa}(x) = \mathbb{E}W_{r,1}(x)$, ii) the majorization $(\frac{p_r}{\|p\|}, \dots, \frac{p_r}{\|p\|}) \prec (kq_{r,1}, \dots, kq_{r,k})$ holds, and iii) the mean waiting time of a D/GI/1 queue is convex increasing in the arrival rate (see, e.g., [30, Theorem 5], [16]), which means that $q_{r,\kappa}\mathbb{E}W_{r,\kappa}(kq_{r,\kappa})$ is convex in $q_{r,\kappa}$.

We now prove (36). As in the proof of Theorem 2, this can be done assuming that (27) holds. Thus, the sequences (27) will be assumed along this proof.

Associated to each queue of type r, we define an auxiliary random variable, $\overline{T}_r^{(k)}$, such that

$$\overline{T}_r^{(k)} =_{st} \sum_{n=1}^{\min_{j=1,\dots,p_r} a_{j,r}^{(k)}} T_n^{(k)}. \tag{38}$$

Next lemma provides properties satisfied by $\overline{T}_r^{(k)}$ that will be used later. We recall that $k_m = m\,\mathrm{lcm}(p)$, see (28).

Lemma 3. *Under the hypotheses of Theorem 3, the following properties hold:*

i) We have

$$\lim_{k\to\infty} \mathbb{E}\overline{T}_r^{(k)} = \lim_{k\to\infty} \min_{j=1,\dots,p_r} \tfrac{a_{j,r}^{(k)}}{\lambda k} = \tfrac{\|p\|}{\lambda p_r}. \tag{39}$$

ii) $\overline{T}_r^{(k)} \to \frac{\|p\|}{\lambda p_r}$ *in probability, as* $k \to \infty$.

iii) For all $i = 1, \dots, \mathrm{lcm}(p)$,

$$-\overline{T}_r^{(k_{m+1}+i)} \leq_{icx} -\overline{T}_r^{(k_m+i)}. \tag{40}$$

Proof. Proof of i). This is an immediate consequence of (31).

Proof of ii). For all $i = 1, \dots, \mathrm{lcm}(p)$, let $j_i^* \in \arg\min_{j=1,\dots,p_r} \frac{a_{j,r}^{(i)}}{i}$. Then, from (30), we get

$$j_i^* \in \arg\min_{j=1,\dots,p_r} \frac{a_{j,r}^{(k_m+i)}}{k_m+i}, \quad \forall m > 1. \tag{41}$$

In view of Lemma 1, the convergence in ii) holds if we can show that $\overline{T}_r^{(k_m+i)} \xrightarrow[m\to\infty]{\Pr} \frac{\|p\|}{\lambda p_r}$, for all $i = 1, \dots, \mathrm{lcm}(p)$. Given (41), this amounts to show that

$$\sum_{n=1}^{a_{j_i^*,r}^{(k_m+i)}} T_n^{(k_m+i)} \xrightarrow[m\to\infty]{\Pr} \tfrac{\|p\|}{\lambda p_r}, \quad \forall i = 1, \dots, \mathrm{lcm}(p). \tag{42}$$

We prove the former by showing (the stronger statement) that $\sum_{n=1}^{a_{j,r}^{(k)}} T_n^{(k)} \xrightarrow[k\to\infty]{\Pr} \frac{\|p\|}{\lambda p_r}$, for all j. Now, using that $\{|X-c|>2\epsilon\} \subseteq \{|X-\mathbb{E}X|>\epsilon\} \cup \{|\mathbb{E}X - c|>\epsilon\}$ for a random variable X, we have

$$\begin{aligned}\Pr\Big(|\sum_{n=1}^{a_{j,r}^{(k)}} T_n^{(k)} - \tfrac{\|p\|}{p_r\lambda}| \geq \epsilon\Big) \leq &\Pr\Big(|\sum_{n=1}^{a_{j,r}^{(k)}} T_n^{(k)} - \mathbb{E}\sum_{n=1}^{a_{j,r}^{(k)}} T_n^{(k)}| \geq \epsilon\Big) + \\ &\Pr\Big(|\tfrac{\|p\|}{p_r\lambda} - \mathbb{E}\sum_{n=1}^{a_{j,r}^{(k)}} T_n^{(k)}| \geq \epsilon\Big).\end{aligned}$$

The second term in the right-hand side of former inequality tends to zero as $k \to \infty$ by (31). The following shows that also the first term goes to zero:

$$\Pr\Big(|\sum_{n=1}^{a_{j,r}^{(k)}} T_n^{(k)} - \mathbb{E}\sum_{n=1}^{a_{j,r}^{(k)}} T_n^{(k)}| \geq \epsilon\Big) \leq \frac{1}{\epsilon^2}\sum_{n=1}^{a_{j,r}^{(k)}} \mathrm{Var}\, T_n^{(k)} \tag{44a}$$

$$\leq \frac{1}{\epsilon^2} a_{j,r}^{(k)} o(1/k) \xrightarrow[k\to\infty]{} 0. \tag{44b}$$

In (44a), we have used Chebyshev's inequality and that the $T_n^{(k)}$'s are independent. In (44b), we have used that $a_{j,r}^{(k)} \leq k\|p\|$.

Proof of iii). We use that $X \leq_{icx} Y$ if and only if there exists an other random variable Z such that $X \leq_{st} Z$ and $Z \leq_{cx} Y$ [28].

Using (29) and that $\frac{a_{j_i^*,r}^{(k_m+i)}}{(k_m+i)} \leq \frac{\|p\|}{p_r}$ for all m (by (30)), the first observation is that

$$a_{j_i^*,r}^{(k_{m+1}+i)} \geq a_{j_i^*,r}^{(k_m+i)} + \operatorname{lcm}(p)\frac{a_{j_i^*,r}^{(k_m+i)}}{k_m+i}, \tag{45}$$

where j_i^* is defined above in point i). Thus, we have

$$-\overline{T}_r^{(k_{m+1}+i)} =_{st} -\sum_{n=1}^{a_{j_i^*,r}^{(k_{m+1}+i)}} T_n^{(k_{m+1}+i)} \tag{46a}$$

$$\leq_{st} -\sum_{n=1}^{a_{j_i^*,r}^{(k_m+i)}+\operatorname{lcm}(p)\frac{a_{j_i^*,r}^{(k_m+i)}}{k_m+i}} T_n^{(k_{m+1}+i)} \stackrel{\text{def}}{=} -Z. \tag{46b}$$

Now, it remains to show that $-Z \leq_{cx} -\overline{T}_r^{(k_m+i)}$, which is equivalent to show that

$$Z \leq_{cx} \overline{T}_r^{(k_m+i)}, \tag{47}$$

see [33, Theorem 3.A.12]. Since

$$\mathbb{E}Z = \frac{1}{\lambda}\frac{a_{j_i^*,r}^{(k_m+i)} + \operatorname{lcm}(p)\frac{a_{j_i^*,r}^{(k_m+i)}}{k_m+i}}{k_m+i+\operatorname{lcm}(p)} = \frac{1}{\lambda}\frac{a_{j_i^*,r}^{(k_m+i)}}{k_m+i} = \mathbb{E}\overline{T}_r^{(k_m+i)}, \tag{48}$$

(47) holds trivially under Case 3. Now, let Case 2 hold. Noticing that both $\overline{T}_r^{(k_m+i)}$ and Z have Erlang distributions with the same mean, to prove (47) is enough to show that $\operatorname{Var} Z \leq \operatorname{Var}\overline{T}_r^{(k_m+i)}$; see [35, p. 14]. We have

$$\lambda^2 \operatorname{Var} Z = \frac{a_{j_i^*,r}^{(k_m+i)} + \operatorname{lcm}(p)\frac{a_{j_i^*,r}^{(k_m+i)}}{k_m+i}}{(k_{m+1}+i)^2} \tag{49a}$$

$$= \frac{a_{j_i^*,r}^{(k_m+i)}}{(k_m+i)^2}\frac{(k_m+i)(k_m+i+\operatorname{lcm}(p))}{(k_m+i+\operatorname{lcm}(p))^2} \tag{49b}$$

$$\leq \frac{a_{j_i^*,r}^{(k_m+i)}}{(k_m+i)^2} = \lambda^2 \operatorname{Var}\overline{T}_r^{(k_m+i)} \tag{49c}$$

as desired. □

We now present an argument that allows us to uniformly bound the second moment of $W_{r,\kappa}^{(k)}$.

Let $\delta_r \stackrel{\text{def}}{=} \frac{1}{2}\left(\frac{\|p\|}{p_r\lambda} - \frac{1}{\mu_r}\right)$ and

$$k^* \stackrel{\text{def}}{=} \min\left\{k > 0 : 0 \leq \frac{\|p\|}{p_r\lambda} - \min_{j=1,\ldots,p_r}\frac{a_{j,r}^{(k')}}{k'\lambda} \leq \delta_r, \quad \forall k' \geq k\right\}. \tag{50}$$

Lemma 4. $k^* < \infty$.

Proof. This is immediate because $\delta_r > 0$ by hypothesis, $\lim_{k' \to \infty} \min_{j=1,\ldots,p_r} \frac{a_{j,r}^{(k')}}{k'\lambda} = \frac{\|p\|}{p_r \lambda}$ (see (39)), and $\frac{\|p\|}{p_r\lambda} \geq \min_{j=1,\ldots,p_r} \frac{a_{j,r}^{(k)}}{k\lambda}$ for all k. □

Let $\overline{W}_r^{(k)}$ denote the stationary waiting time of a GI/GI/1 queue with (i.i.d.) interarrival times $(\overline{T}_{n,r}^{(k)})_{n\in\mathbb{N}}$ where $\overline{T}_{n,r}^{(k)} =_{st} \overline{T}_r^{(k)}$ (see (38)) and service times $(S_{n,\kappa,r}^{(k)})_{n\in\mathbb{N}}$. By coupling the $\overline{T}_{n,r}^{(k)}$'s and the $T_{n,\kappa,r}^{(k)}$'s in the obvious manner, one can easily see that $(\overline{T}_{1,r}^{(k)}, \ldots, \overline{T}_{n,r}^{(k)}) \leq (T_{1,\kappa,r}^{(k)}, \ldots, T_{n,\kappa,r}^{(k)})$[4] and therefore we have $(\overline{T}_{1,r}^{(k)}, \ldots, \overline{T}_{n,r}^{(k)}) \leq_{st} (T_{1,\kappa,r}^{(k)}, \ldots, T_{n,\kappa,r}^{(k)})$. Using, e.g., [6, pp. 217, 220], this implies

$$W_{r,\kappa}^{(k)} \leq_{st} \overline{W}_r^{(k)}. \tag{51}$$

Furthermore, given that $-\overline{T}_{n,r}^{(k_{m+1}+i)} \leq_{icx} -\overline{T}_{n,r}^{(k_m+i)}$ for all $i = 1, \ldots, \mathrm{lcm}(p)$ (by Lemma 3) and that the $(\overline{T}_{n,r}^{(k)})_{n\in\mathbb{N}}$ are independent, we can use [5, p. 337] to establish that

$$\overline{W}_r^{(k_{m+1}+i)} \leq_{icx} \overline{W}_r^{(k_m+i)}, \tag{52}$$

for all $i = 1, \ldots, \mathrm{lcm}(p)$. Therefore, given $\mathbb{E}\left((W_{r,\kappa}^{(k)})^2\right) < \infty$ for all k and Lemma 4, we can uniformly bound the second moment of $W_{r,\kappa}^{(k)}$ as follows

$$\begin{aligned}
\sup_{k\geq k^*} \mathbb{E}\left((W_{r,\kappa}^{(k)})^2\right) &\leq \sup_{k\geq k^*} \mathbb{E}\left((\overline{W}_r^{(k)})^2\right) && \text{(53a)}\\
&= \max_{i=1,\ldots,\mathrm{lcm}(p)} \sup_{m: k_m+i\geq k^*} \mathbb{E}\left((\overline{W}_r^{(k_m+i)})^2\right) && \text{(53b)}\\
&= \max_{i=1,\ldots,\mathrm{lcm}(p)} \mathbb{E}\left((\overline{W}_r^{(k_{m_i^*}+i)})^2\right) && \text{(53c)}\\
&< \infty, && \text{(53d)}
\end{aligned}$$

where $m_i^* \stackrel{\text{def}}{=} \min\{m : k_m + i \geq k^*\}$. In (53a) and (53c), we have used (51) and (52), respectively. In (53d), we have used that

$$\mathbb{E}\overline{T}_{n,r}^{(k)} = \min_{j=1,\ldots,p_r} \frac{a_{j,r}^{(k)}}{k\lambda} \geq \frac{\|p\|}{p_r\lambda} - \delta_r = \frac{1}{2}\frac{\|p\|}{p_r\lambda} + \frac{1}{2}\frac{1}{\mu_r} > \frac{1}{\mu_r}, \quad \forall k \geq k^*, \tag{54}$$

i.e. the ergodicity condition, and that the third moment of service times is finite, which imply that the second moment of $\overline{W}_r^{(k)}$ is finite [5, pg. 270].

Now, using the continuity of the stationary waiting time of GI/GI/1 queues [5, Corollary X.6.4] and part i) and ii) of Lemma 3, we have $\overline{W}_r^{(k)} \xrightarrow[k\to\infty]{d} W_r(p)$, and given the uniform integrability (53) we have that also the expected values converge, i.e.,

$$\lim_{k\to\infty} \mathbb{E}\overline{W}_r^{(k)} = \mathbb{E}W_r(p). \tag{55}$$

With the above relations, we can conclude the proof of (36)

$$\begin{aligned}
\lim_{k\to\infty} \mathbb{E}W^{(k)}(\pi^{(k)}) &= \lim_{k\to\infty} \sum_r \sum_\kappa \frac{p_r}{\|p\|k} \mathbb{E}W_{r,\kappa}^{(k)}(\pi^{(k)}) && \text{(56a)}\\
&\leq \lim_{k\to\infty} \sum_r \frac{p_r}{\|p\|} \mathbb{E}\overline{W}_r^{(k)} && \text{(56b)}
\end{aligned}$$

[4] Given $x, y \in \mathbb{R}^d$, here $x \leq y$ means $x_i \leq y_i$ for all $i = 1, \ldots, d$.

$$= \sum_r \frac{p_r}{\|p\|} \mathbb{E}W_r(p). \tag{56c}$$

Proof of (17). Let Q be a discrete random variable with values in $\{1,\dots,R\}\times\{1,\dots,k\}$ such that $\Pr(Q=(r,\kappa)) = \frac{p_r}{\|p\|k}$. We assume that this random variable is independent of any other random variable. By definition of $W^{(k)}(\pi^{(k)})$ and using the law of total variance, we obtain

$$\operatorname{Var} W^{(k)}(\pi^{(k)}) = \mathbb{E}(\operatorname{Var} W^{(k)}(\pi^{(k)})|Q) + \operatorname{Var}\mathbb{E}(W^{(k)}(\pi^{(k)})|Q) \tag{57a}$$

$$= \sum_{r,\kappa} \tfrac{p_r}{\|p\|k} \left(\operatorname{Var} W^{(k)}_{r,\kappa}(\pi^{(k)}) + \left(\mathbb{E}W^{(k)}_{r,\kappa}(\pi^{(k)}) - \mathbb{E}W^{(k)}(\pi^{(k)})\right)^2 \right). \tag{57b}$$

When $k\to\infty$, we have already established that $\mathbb{E}W^{(k)}(\pi^{(k)}) \to \mathbb{E}W(p)$ and that $\mathbb{E}W_r(p) \le \mathbb{E}W^{(k)}_{r,\kappa}(\pi^{(k)}) \le \mathbb{E}\overline{W}^{(k)}_{r,\kappa} \to \mathbb{E}W_r(p)$. Therefore, it only remains to show that the second moment of $W^{(k)}_{r,\kappa}(\pi^{(k)})$ converge to $\mathbb{E}[W_r(p)^2]$. This is done by using the same argument above for the convergence of the first moment. Hence, using the continuity of the waiting time and of the square function [5, Corollary X.6.4] and part i) and ii) of Lemma 3, we obtain $(\overline{W}^{(k)}_r)^2 \xrightarrow{d} W_r(p)^2$, as $k\to\infty$. Furthermore, the second moment of $(\overline{W}^{(k)}_r)^2$ is finite because the fifth moment of the service times is finite [5, pg. 270] and (52) ensures that the sequence $(\overline{W}^{(k)}_r)^2$ is uniformly integrable because it is non-increasing along subsequences $(k_m+i)_{m\in\mathbb{N}}$, for all $i=1,\dots,\operatorname{lcm}(p)$. Thus, $\mathbb{E}[(\overline{W}^{(k)}_r)^2] \to \mathbb{E}[W_r(p)^2]$. Together with (18), as desired we obtain

$$\lim_{k\to\infty} \mathbb{E}[(W^{(k)}_{r,\kappa}(\pi^{(k)}))^2] = \mathbb{E}[W_r(p)^2]. \tag{58}$$

5 Conclusions

We have derived structural properties concerning Problem 1. Fixing the proportion of jobs to send on each queue, p, we have shown that periodic policies are asymptotically equivalent and optimal. The limiting mean waiting time, $\mathbb{E}W(p)$ (see (8)), is expressed in terms of independent D/GI/1 queues and has the convenient property of being convex in p. We believe that these structural properties provide researchers and practitioners with new means about the considered problem. For instance, one consequence of these results is that the problem of computing the optimal proportions of jobs to send to each queue, which is considered a difficult problem (see the Introduction), can be captured by the solution of an optimization problem of the form:

$$\min \mathbb{E}W(p) \quad \text{s.t.:} \quad p\in\mathcal{S},$$

for $\mathcal{S}$ compact and convex, and we stress that $\mathbb{E}W(p)$ is a convex function of p. In the case where service times have an exponential distribution, $\mathbb{E}W(p)$ admits a simple characterization because it is the weighted mean waiting time of R D/M/1 queues, which can be computed efficiently [9]; see also [4].

As discussed in Section 3.1, the policies in $\mathcal{C}^{(k)}_p$ can be implemented in a distributed manner in a two-tier hierarchy. Here, our results imply that the dispatcher in the first tier has a negligible influence on jobs waiting times, which means that it does not need to have full information about the statistics of network resources.

References

[1] E. Altman, B. Gaujal, and A. Hordijk. Balanced sequences and optimal routing. *J. ACM*, 47(4):752–775, July 2000.

[2] E. Altman, B. Gaujal, and A. Hordijk. Multimodularity, convexity, and optimization properties. *Math. Oper. Res.*, 25(2):324–347, 2000.

[3] J. Anselmi and B. Gaujal. Optimal routing in parallel, non-observable queues and the price of anarchy revisited. In *International Teletraffic Congress*, pages 1–8, 2010.

[4] J. Anselmi and B. Gaujal. The price of forgetting in parallel and non-observable queues. *Perform. Eval.*, 68(12):1291–1311, Dec. 2011.

[5] S. Asmussen. *Applied Probability and Queues*. Wiley, 1987.

[6] F. Baccelli and P. Brémaud. *Elements of queueing theory: Palm Martingale calculus and stochastic recurrences*, volume 26. springer, 2003.

[7] A. Bar-Noy, R. Bhatia, J. Naor, and B. Schieber. Minimizing service and operation costs of periodic scheduling (extended abstract). In H. J. Karloff, editor, *SODA*, pages 11–20. ACM/SIAM, 1998.

[8] C. H. Bell and S. Stidham. Individual versus social optimization in the allocation of customers to alternative servers. *Management Science*, 29(7):831–839, 1983.

[9] U. N. Bhat. *An Introduction to Queueing Theory: Modeling and Analysis in Applications*. Birkhauser Verlag, 2008.

[10] S. Bhulai, T. Farenhorst-Yuan, B. Heidergott, and D. van der Laan. Optimal balanced control for call centers. *Annals OR*, 201(1):39–62, 2012.

[11] A. Borovkov. *Stochastic processes in queueing theory*. Applications of mathematics. Springer-Verlag, 1976.

[12] S. C. Borst. Optimal probabilistic allocation of customer types to servers. ACM SIGMETRICS '95/PERFORMANCE '95, pages 116–125, New York, NY, USA, 1995. ACM.

[13] M. B. Combé and O. J. Boxma. Optimization of static traffic allocation policies. *Theor. Comput. Sci.*, 125(1):17–43, 1994.

[14] J. Doob. *Stochastic Processes*. Wiley Publications in Statistics. John Wiley & Sons, 1953.

[15] B. Gaujal, E. Hyon, and A. Jean-Marie. Optimal routing in two parallel queues with exponential service times. *Discrete Event Dynamic Systems*, 16:71–107, January 2006.

[16] L. Gun, A. Jean-Marie, A. M. Makowski, and T. Tedijanto. Convexity results for parallel queues with bernoulli routing. Technical Report, TR 1990-52. University of Maryland, USA, 1990.

[17] B. Hajek. The proof of a folk theorem on queuing delay with applications to routing in networks. *J. ACM*, 30(4):834–851, Oct. 1983.

[18] B. Hajek. Extremal splitting of point processes. *Math. Oper. Res.*, 10:543–556, 1986.

[19] M. Harchol-Balter, A. Scheller-Wolf, and A. R. Young. Surprising results on task assignment in server farms with high-variability workloads. In *SIGMETRICS/Performance*, pages 287–298. ACM, 2009.

[20] A. Hordijk, G. M. Koole, and J. A. Loeve. Analysis of a customer assignment model with no state information. In *Probability in the Engineering and Informational Sciences*, volume 8, pages 419–429, 1994.

[21] A. Hordijk and D. van der Laan. Periodic routing to parallel queues and billiard sequences. *Mathematical Methods of Operations Research*, 59(2):173–192, 2004.

[22] P. Humblet, M. I. of Technology. Laboratory for Information, and D. Systems. *Determinism Minimizes Waiting Time in Queues*. LIDS-P-. Laboratory for Information and Decision Systems, 1982.

[23] B. Javadi, D. Kondo, J.-M. Vincent, and D. P. Anderson. Discovering statistical models of availability in large distributed systems: An empirical study of seti@home. *IEEE Trans. Parallel Distrib. Syst.*, 22(11):1896–1903, 2011.

[24] B. Javadi, P. Thulasiraman, and R. Buyya. Cloud resource provisioning to extend the capacity of local resources in the presence of failures. In *HPCC-ICESS*, pages 311–319. IEEE Computer Society, 2012.

[25] D. V. Lindley. The theory of queues with a single server. *Mathematical Proceedings of the Cambridge Philosophical Society*, 48:277–289, 1952.

[26] Z. Liu and R. Righter. Optimal load balancing on distributed homogeneous unreliable processors. *Journal of Operations Research*, 46(4):563–573, 1998.

[27] R. Loynes. The stability of a queue with nonindependent interarrival and service times. *Mathematical Proceedings of the Cambridge Philosophical Society*, 58:497–520, 1962.

[28] A. Makowski. On an elementary characterization of the increasing convex ordering, with an application. *Journal of Applied Probability*, 31:834–840, 1994.

[29] M. Mitzenmacher. How useful is old information? *IEEE Trans. Parallel Distrib. Syst.*, 11(1):6–20, 2000.

[30] M. J. Neely and E. Modiano. Convexity in queues with general inputs. *IEEE Transactions on Information Theory*, 51(2):706–714, 2005.

[31] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994.

[32] J. Sethuraman and M. S. Squillante. Optimal stochastic scheduling in multiclass parallel queues. SIGMETRICS '99, pages 93–102, New York, NY, USA, 1999. ACM.

[33] M. Shaked and J. G. Shanthikumar. *Stochastic orders and their applications*. Academic Pr, 1994.

[34] J. G. Shanthikumar and S. H. Xu. Asymptotically optimal routing and service rate allocation in a multiserver queueing system. *Operations Research*, 45:464–469, 1997.

[35] D. Stoyan and D. Daley. *Comparison methods for queues and other stochastic models*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1983.

[36] R. Tijdeman. Fraenkel's conjecture for six sequences. *Discrete Mathematics*, 222(1-3):223–234, 2000.

[37] D. van der Laan. Routing jobs to servers with deterministic service times. *Math. Oper. Res.*, 30(1):195–224, 2005.