Power-of-two sampling in redundancy systems

Operations Research Letters 50 (2022) 699–706

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Power-of-two sampling in redundancy systems: The impact of assignment constraints

Ellen Cardinaels [a,*], Sem Borst [a], Johan S.H. van Leeuwaarden [b]

[a] Eindhoven University of Technology, the Netherlands
[b] Tilburg University, the Netherlands

ARTICLE INFO

Article history:
Received 11 November 2021
Received in revised form 15 July 2022
Accepted 12 October 2022
Available online 17 October 2022

Keywords:
Power-of-two
Parallel-server systems
Load balancing
Redundancy scheduling
Light traffic
Stochastic comparison

ABSTRACT

A classical sampling strategy for load balancing policies is power-of-two, where any server pair is sampled with equal probability. This does not cover practical settings with assignment constraints which force non-uniform sampling. While intuition suggests that non-uniform sampling adversely impacts performance, this was only supported through simulations, and rigorous statements have remained elusive. Building on product-form distributions for redundancy systems, we prove the stochastic dominance of uniform sampling for a four-server system as well as arbitrary-size systems in light traffic.

1. Introduction

Load balancing policies applying a power-of-d sampling strategy assign a job to the 'best' among a randomly selected subset of $d \geq 2$ parallel servers. These policies were originally investigated in a balls-and-bins context where it was shown that the maximum bin occupancy is exponentially reduced when instead of purely random assignment the least loaded bin among the d selected bins is chosen [4]. This concept was also shown to be highly effective in queueing contexts, in particular in the many-server regime where the above policy results in a doubly exponential improvement in terms of the queue length distribution per server compared to purely random assignment [19,26]. Moreover, power-of-d policies only involve a low implementation overhead and hence provide scalability, which is critical in large-scale systems such as data centers. More recently, power-of-d sampling policies have also been proposed in the context of redundancy scheduling where replicas of each job are dispatched to a randomly selected subset of d servers [10]. We refer to [5] for further background on scalable load balancing algorithms.

In the classical power-of-d setup, it is implicitly assumed that the d servers are selected uniformly at random, with or without replacement. This is a natural assumption when all jobs and servers are mutually exchangeable, and also mathematically convenient (see the discussion of related literature below for further details). However, this assumption excludes situations with assignment constraints, such as locality constraints or compatibility relations between jobs and servers, which force non-uniform server sampling. In the special case that $d = 2$ such assignment constraints can conveniently be visualized in a graph consisting of N nodes, one for each server. When a non-negative weight is associated with each edge, edges or server pairs can be sampled according to these weights. When server pairs are sampled uniformly at random, all edges receive a weight equal to $1/\binom{N}{2}$. This raises the following interesting question: *How do non-uniform edge weights affect the system performance, as compared to uniform edge weights?* One intuitively expects performance to benefit from more flexibility (i.e., more non-zero weights) and homogeneity (i.e., uniform edge weights). This intuition was supported through heuristic arguments and simulations for the Join-the-Shortest-Queue policy in specific topological settings, see for instance the thesis of Mitzenmacher [18], the seminal paper of Turner [25] and the more recent work of Gast [13]. However to the best of our knowledge, rigorous statements on the performance impact of assignment constraints in the power-of-choice setting have remained elusive so far.

In the present paper we establish stochastic comparison results which corroborate the above-mentioned 'common wisdom' in redundancy systems, and prove in some specific settings that the classical uniform power-of-two policy outperforms *any* power-of-two policy with assignment constraints. To establish these results

* Corresponding author.
E-mail address: e.cardinaels@tue.nl (E. Cardinaels).

https://doi.org/10.1016/j.orl.2022.10.006

we employ the product-form expressions for the stationary occupancy distribution obtained by Gardner *et al.* [12]. Unfortunately, the detailed job-level state description yields expressions that do not give immediate insight into the system performance. Careful inspection and further manipulation of the detailed product-form expressions, however, allows us to derive stochastic comparison results.

We first establish closed-form expressions for the stationary distribution of the total number of jobs for four-server systems, which we then use to show a stochastic ordering result for a ring graph compared to a complete graph, confirming the above intuition. For systems of arbitrary size, closed-form expressions for the stationary distribution of the total number of jobs in the system seem out of reach. However, focusing on a light-traffic scenario allows us to extract the essential information to compare the stationary distributions of systems with different edge selection probabilities. This comparison gives rise to an optimization problem in terms of the edge selection probabilities for which the classical uniform power-of-two policy arises as the optimal solution. Moreover, the light-traffic comparison can be interpreted as a design guideline for an efficient weighted power-of-two policy in the presence of assignment constraints.

The literature focusing on the classical uniform power-of-d sampling policy is extensive and vibrant as the inherent symmetry of these policies lends itself well to asymptotic analysis in a many-server regime. Seminal results in such settings were obtained by Mitzenmacher [19] and Vvedenskaya *et al.* [26] using fluid-limit techniques, and later closely related mean-field concepts were studied in [10,15–17].

As in [1,9,28], these techniques can also be used for the analysis of particular asymmetric assignment constraints where mutually exchangeable servers are clustered in a finite number of pools. However, fluid-limit and mean-field techniques are usually not well-suited to scenarios with asymmetric assignment constraints corresponding to a graph as mentioned above. Indeed, Gast [13] and Turner [25] use an approximation scheme and simulations to demonstrate that the classical Join-the-Shortest-Queue(2) (JSQ(2)) policy outperforms a restricted JSQ(2) policy where the assignment is governed by a ring graph. In contrast, the results in [7,20,23,27] establish conditions in terms of the assignment constraints that yield performance comparable to the classical uniform power-of-d policies in a many-server regime.

Non-uniform selection of subsets of servers in a JSQ context has also been considered by He and Down [14] and Sloothaak *et al.* [24], where it is shown that the diffusion scaled queue length process coincides with that of a fully pooled system in a heavy-traffic regime. In [8] conditions in terms of the assignment constraints are established to draw a similar conclusion for redundancy policies. Rather than imposing conditions on the assignment constraints, our results aim to connect and compare the performance of systems operating under the classical uniform power-of-two sampling policies and those with assignment constraints. We will focus on systems with a fixed number of servers in moderate or light traffic.

The remainder of this paper is organized as follows. In Section 2 we present a detailed model description and discuss some broader context and preliminaries. In Subsection 3.1 we set out to prove a stochastic comparison between two small systems. Next we consider systems of arbitrary size in Subsection 3.2, and establish a light-traffic comparison between the classical uniform power-of-two policy and weighted power-of-two policies. A discussion of the results and some pointers for further research are provided in Section 4.

2. Model description and preliminaries

2.1. Model description

Before elaborating on redundancy scheduling, we first define the power-of-two policies to sample server pairs subject to the assignment constraints. Jobs arrive according to a Poisson process with rate $N\lambda$, with N the total number of parallel servers. When a job arrives, the server pair available for its assignment is $\{i, j\}$ with probability $p_{\{i,j\}}$, $i, j, \in \{1, \ldots, N\}$ and $i \neq j$. For simplicity we refer to such a job as a type-$\{i, j\}$ job. Due to the properties of the Poisson process, one can equivalently take the view that type-$\{i, j\}$ jobs arrive according to a Poisson process with rate $N\lambda p_{\{i,j\}}$.

As alluded to in the introduction, one can think of an underlying simple graph structure with N nodes and (non-negative) edge weight $p_{\{i,j\}}$ for the edge $\{i, j\}$. Server pairs are then sampled proportionally to these edge weights for each arriving job. Let $\mathcal{E} := \{\{i, j\} \mid p_{\{i,j\}} > 0, i, j = 1, \ldots, N, i \neq j\}$ denote the set of all edges with a non-zero weight, or alternatively, all possible job types that can occur in the system. Without loss of generality we assume that $\sum_{e\in\mathcal{E}} p_e = 1$, and refer to p_e as the selection probability of the edge e. An example where this underlying graph is given by a ring graph is depicted in Fig. 1a.

The setting where $p_{\{i,j\}} \equiv 1/E$ for all $i \neq j$, with $E = \binom{N}{2}$ the total number of different server pairs, corresponds to the typical power-of-two setting with uniform sampling. We will henceforth refer to this setting as the *classical* power-of-two policy, while any setting with non-uniform sampling is referred to as a *weighted* power-of-two policy.

Remark 1. When jobs can be assigned to $d \geq 2$ servers, one can think of p_e as the selection probability of hyper-edge e in a hypergraph with N nodes where each hyper-edge is incident to d distinct nodes. In the remainder of this paper we will focus on the case where $d = 2$.

For an arriving type-$\{i, j\}$ job, under the redundancy policy, replicas are assigned to both servers i and j, with $i, j, \in \{1, \ldots, N\}$ and $i \neq j$. The service requirements of the two replicas are independent and exponentially distributed with unit mean. Each server has speed $\mu > 0$ and handles the assigned jobs in a First-Come-First-Served manner. Once the first replica finishes service, the remaining replica will be discarded instantaneously.

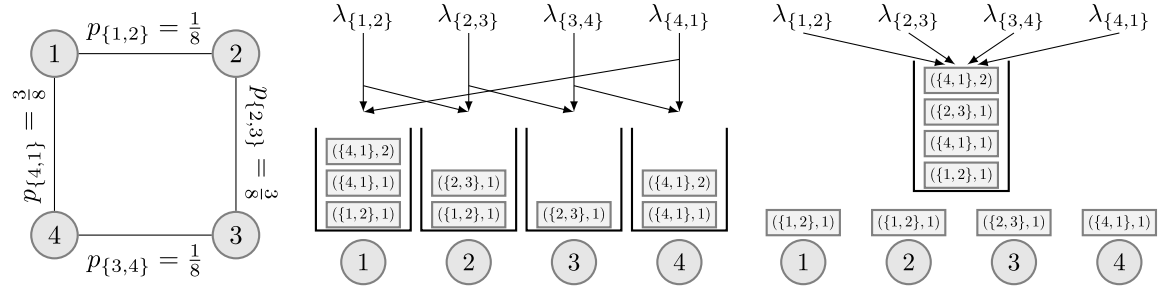
As discussed in the introduction, it is intuitively plausible that uniform sampling outperforms non-uniform sampling. This intuitive notion is formalized in the following conjecture.

Conjecture 1. Let Q^* and $Q(P)$ denote the total number of jobs in stationarity in a redundancy system with N servers operating according to the classical power-of-two policy and a weighted power-of-two policy with edge selection probabilities $P = (p_{\{i,j\}})_{i,j}$, respectively. Then, Q^* is stochastically smaller than $Q(P)$, i.e.,

$$Q^* \leq_{\text{st}} Q(P).$$

We will establish stochastic comparison results and light-traffic limits to support the above conjecture, building on the product-form expressions for redundancy systems that will be outlined in the next subsection.

Remark 2. Conjecture 1 implicitly assumes both Q^* and $Q(P)$ to exist. As shown in [12] this is the case for Q^* if and only if $\lambda < \mu$. It can be further deduced from [12] that the latter condition is also necessary for $Q(P)$ to exist. The sufficient condition requires the



(a) A ring as an underlying graph structure. (b) Representation with one queue per server and replicas. (c) Central queue representation.

Fig. 1. With the underlying graph structure as indicated in Panel (a) with $N = 4$ servers, Panels (b) and (c) give two different representations of the state $\boldsymbol{c} = (\{1, 2\}, \{4, 1\}, \{2, 3\}, \{4, 1\})$. The notation $(\{i, j\}, k)$ stands for the kth arrival of a type-$\{i, j\}$ job.

aggregate arrival rate of any subset of edges to be strictly smaller than the aggregate service rate of the servers at its endpoints, i.e., for all $S \subseteq \mathcal{E}$ it must hold that $N\lambda \sum_{\{i,j\}\in S} p_{\{i,j\}} < \mu|\cup_{\{i,j\}\in S}\{i, j\}|$.

Remark 3. Note that Conjecture 1 contrasts with the universality result in [8, Theorem 1] in a *heavy-traffic* regime. In particular, it is shown in [8] that for a broad range of weighted policies the system in heavy traffic achieves complete resource pooling and exhibits state space collapse.

2.2. Product-form expressions

The system occupancy at time t of a system operating under the redundancy policy may be represented in terms of a vector $(c_1, \ldots, c_{Q(t)})$, with $Q(t)$ denoting the total number of jobs in the system at time t and $c_q \in \mathcal{E}$. One can think of the occupancy vector as a central queue where $c_q \in \mathcal{E}$ indicates the type of the qth oldest job in the system at time t. It is easily verified that the system occupancy evolves as a Markov process by virtue of the exponential traffic assumptions.

From a modeling perspective, the state description yields a system that is equivalent to a system where replicas are positioned in two dedicated queues in front of the two selected servers. To illustrate this, consider the graph structure depicted in Fig. 1a with $N = 4$ servers and assume that the state of the system is given by $\boldsymbol{c} = (\{1, 2\}, \{4, 1\}, \{2, 3\}, \{4, 1\})$. Hence, the oldest job in the system is replicated to servers 1 and 2, the second oldest job is replicated to servers 1 and 4, etc. Fig. 1b represents the system with dedicated queues at each of the servers, while Fig. 1c represents the system with a centralized queue. In the latter case a server that becomes idle scans this central queue and will initiate service of a replica of the first job it is compatible with.

It was shown in [12] that, under the stability conditions mentioned in Remark 2, the stationary distribution of the system occupancy is

$$\pi(c_1, \ldots, c_Q) = C \prod_{i=1}^{Q} \frac{N\lambda p_{c_i}}{\mu(c_1, \ldots, c_i)}, \tag{1}$$

with C a normalization constant and

$$\mu(c_1, \ldots, c_i) = \mu|\bigcup_{j=1}^{i}\{c_j\}| \tag{2}$$

the aggregate service rate of the system in the state $(c_1, \ldots, c_i)$. For instance, the stationary probability of state $\boldsymbol{c}$ depicted in Fig. 1 is given by $\pi(\boldsymbol{c}) = C(\frac{4\lambda}{\mu})^4 \cdot \frac{p_{\{1,2\}}}{2}\frac{p_{\{4,1\}}}{3}\frac{p_{\{2,3\}}}{4}\frac{p_{\{4,1\}}}{4}$.

3. Main results

We now use the product-form expressions to assess the performance of various systems with respect to their assignment constraints. However, the detailed job-level state description ingrained in the product-form expressions does not provide much insight into the overall performance and does not allow a direct comparison. In order to make a meaningful comparison, we therefore consider the stationary distribution of the total number of jobs in the system, Q. This distribution may be expressed in terms of the detailed product-form expressions as

$$\mathbb{P}\{Q = q\} = \sum_{\boldsymbol{c}\in\mathcal{E}^q} \pi(\boldsymbol{c}), \tag{3}$$

with $q \geq 0$. Hence, all $|\mathcal{E}|^q$ states $\boldsymbol{c} = (c_1, c_2, \ldots, c_q)$ with $c_i \in \mathcal{E}$ for all $i = 1, \ldots, q$ must be aggregated to determine $\mathbb{P}\{Q = q\}$. Besides the fact that there are exponentially many terms, the various terms are also highly different. The difference between two terms is caused by the various job types that could occur but mainly by the order in which they appear.

However, for small systems we can enumerate all possible server rate sequences $(|c_1|, |c_1 \cup c_2|, \ldots, |c_1 \cup \cdots \cup c_q|)$, which results in stationary distributions with particular underlying structures amenable for comparison as we will show in Subsection 3.1. Unfortunately, this enumeration strategy does not lead to tractable expressions for larger systems. Therefore we consider larger systems in a light-traffic regime to partially suppress the complexity (captured in the normalization constant), and reveal the essential dependence of the stationary distribution on the edge selection probabilities. In particular, the stationary probability in (3) reduces in a light-traffic regime to a polynomial of degree q in function of the selection probabilities. This again allows to compare systems with different underlying structures as will be demonstrated in Subsection 3.2.

3.1. Four-server systems

We will derive closed-form expressions for the summation in (3) for small systems, for which the computations are already quite tedious. The focus will be on the classical power-of-two policy and a particular subset of weighted policies, namely those policies governed by ring graphs. Let $\epsilon \in (0, 1)$ and N be even, and define the edge selection probabilities as

$$p_{\{i,i+1\}} = \begin{cases} \epsilon\frac{2}{N}, & \text{if } i \text{ is even} \\ (1-\epsilon)\frac{2}{N}, & \text{if } i \text{ is odd} \end{cases} \tag{4}$$

with $i = 1, \ldots N$ and $p_{\{1,N\}} = p_{\{N,N+1\}}$. Note that the example in Fig. 1a is a special case of this setting with $N = 4$ and $\epsilon = 3/4$.

When $\epsilon = 1/2$, all probabilities are equal to N^{-1}. The average arrival rate across all edges is given by λ and therefore this graph is referred to as the homogeneous ring. In all other cases we refer to this underlying graph as the heterogeneous ring.

Lemma 1. *The stationary distribution of the total number of jobs in a system with the uniform complete graph structure on $N = 4$ servers is given by*

$$\mathbb{P}\{Q_4^* = q\} = \frac{1}{9}(1-\rho)(3-\rho)(3-2\rho) \times \left\{-4\left(\frac{2\rho}{3}\right)^q + \frac{1}{2}\left(\frac{\rho}{3}\right)^q + \frac{9}{2}\rho^q\right\},$$

with $q \geq 0$ and $\rho := \frac{\lambda}{\mu} < 1$.

Lemma 2. *The stationary distribution of the total number of jobs in a system with the heterogeneous ring structure on $N = 4$ servers is given by*

$$\mathbb{P}\{Q_4^{\text{het}}(\epsilon) = q\} = \frac{(1-\rho)(1-(1-\epsilon)\rho)(1-\epsilon\rho)(3-2\rho)}{3-2\rho+(1-\epsilon)\epsilon\rho^2} \cdot \left\{\frac{6\epsilon(1-\epsilon)}{2-9\epsilon(1-\epsilon)}\left(\frac{2\rho}{3}\right)^q + \frac{(1-\epsilon)^2}{\epsilon(2-3(1-\epsilon))}((1-\epsilon)\rho)^q + \frac{\epsilon^2}{(1-\epsilon)(2-3\epsilon)}(\epsilon\rho)^q + \frac{1+\epsilon(1-\epsilon)}{\epsilon(1-\epsilon)}\rho^q\right\}, \tag{5}$$

with $q \geq 0$, $\epsilon \in (0,1)$ and $\rho := \frac{\lambda}{\mu} < 1$.

Note that the stationary distribution (5) is symmetric around $\epsilon = 1/2$, reflecting the symmetry in the edge selection probabilities. The derivations of the results in Lemmas 1 and 2 are deferred to the online appendix.

The next proposition proves a partial version of Conjecture 1 for systems with $N = 4$ servers and weighted policies that correspond to homogeneous ring graphs.

Proposition 1. *Let Q_4^* and Q_4^{hom} denote the total number of jobs in stationarity in a system with a uniform complete graph structure and a homogeneous ring, respectively, with $N = 4$ servers and $\lambda < \mu$. Then, Q_4^* is stochastically smaller than Q_4^{hom}, i.e.,*

$$Q_4^* \leq_{\text{st}} Q_4^{\text{hom}}. \tag{6}$$

The proof of Proposition 1 uses Lemmas 1 and 2 and can be found in the online appendix. A numerical comparison of the above derived stationary distributions is depicted in Fig. 2. Although the figure clearly supports the result in Proposition 1, it also suggests that the absolute differences between the various distributions are fairly small. Furthermore, the above-described settings are compared to a setting where the ring structure is disconnected by choosing $\epsilon = 0$ or $\epsilon = 1$. This system is equivalent to two independent single-server queues with arrival rate 2λ and service rate 2μ. Hence, the total number of jobs in the system is determined by a sum of two independent and geometrically distributed random variables with parameter $\rho := \lambda/\mu$, resulting in a negative binomial distribution. Alternatively, it can be seen that (5) indeed converges to $(q+1)(1-\rho)^2\rho^q$ when $\epsilon \downarrow 0$ or $\epsilon \uparrow 1$.

Moreover, from (5) it can be deduced that the probability of an empty system, i.e., $\mathbb{P}\{Q_4^{\text{het}}(\epsilon) = 0\}$, decreases the more ϵ deviates from $1/2$ (the online appendix). While considering Fig. 2, it can be observed that $\mathbb{P}\{Q_4^{\text{het}}(\epsilon) \geq q\}$ increases for any fixed q the more ϵ deviates from $1/2$, revealing a degradation of the system performance the more the selection probabilities of the ring structure differ from the uniform probabilities.

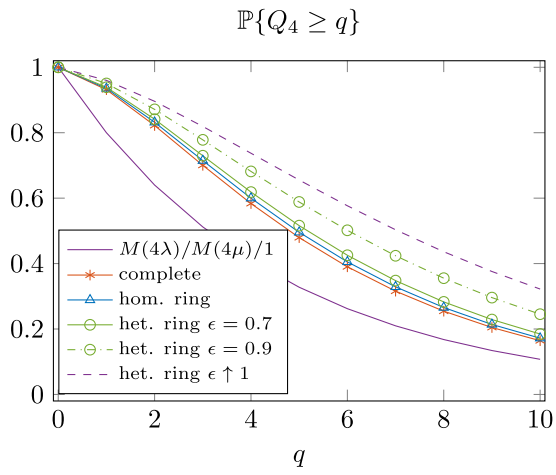


Fig. 2. The stationary distributions of the total number of jobs compared for various power-of-d policies with $N = 4$ servers and $\rho = 0.8$: classical power-of-4 policy, classical power-of-2 policy, weighted power-of-2 policies governed by a homogeneous ring and heterogeneous rings with $\epsilon = 0.7$, $\epsilon = 0.9$ and $\epsilon \uparrow 1$. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

In addition, a setting where a job can be replicated to all $N = 4$ servers is considered, so $d = 4$ instead of $d = 2$. This fully pooled scenario is equivalent in performance to a single-server queue with arrival rate 4λ and service rate 4μ, and yields the stochastically smallest total number of jobs in the system. Indeed, the service rate in this system is always equal to 4μ when there are jobs present, while in all other cases the service rate is at most equal to 4μ.

The challenges with the detailed product-form expressions mentioned at the beginning of this section could be overcome by judicious state aggregation for scenarios with a small number of servers and weighted policies that correspond to ring graphs, although the expressions were already rather unwieldy in this case. The stationary distributions (and their normalization constants) in Lemmas 1 and 2 for larger values of N become more intricate, let alone their comparison for different edge selection probabilities. This makes it complicated to extend Proposition 1 for larger system sizes, though it is assumed the hold for any fixed N as illustrated by means of simulation in the online appendix. Alternatively, the stationary distribution of the total number of jobs in a system with N servers operating under the classical power-of-d policy could be derived using a similar approach as outlined in [10, Section 4.2.2], implicitly relying on the corresponding generating function. For any $q \geq 1$,

$$\mathbb{P}\{Q_N^* = q\} = C\frac{\rho^q}{\binom{N-1}{d-1}^q} \sum_{\boldsymbol{n} \in R(q)} (-1)^{|\boldsymbol{n}|+q}(|\boldsymbol{n}|)! \prod_{j=1}^{q} \frac{F_j(N,d)^{n_j}}{n_j!}.$$

In the above formula, we define $R(q) := \{\boldsymbol{n} \in \mathbb{N}^q : 1 \cdot n_1 + 2 \cdot n_2 + \cdots + q \cdot n_q = q\}$, $|\boldsymbol{n}| := n_1 + n_2 + \cdots + n_q$ and C as in [10, Theorem 2]. Moreover,

$$F_j(N,d) = \sum_{\boldsymbol{k} \in \tilde{R}_j(N,d)} \prod_{i=d}^{N} \binom{i-1}{d-1}^{k_{i-d+1}},$$

where $\tilde{R}_j(N,d) \subset \{0,1\}^{N-d+1}$ contains all binary vectors with precisely j entries equal to 1. Although the above expressions are presented in closed form, they are unwieldy. This would be exacerbated in case the occupancy probabilities of the weighted power-of-two policy would be derived using the explicit expression for

the generating function obtained in [8, Proposition 1], hence a general comparison between the stationary distributions seems out of reach.

Coupling arguments are commonly used as an alternative method to establish stochastic dominance properties, when the actual distributions are not tractable. However, this approach seems out of reach because the systems under consideration do not necessarily have the same number of job types or equal arrival probabilities for mutual job types. Also, coupling arguments would yield a stronger stochastic comparison result for the entire process over time, which may in fact not hold. This implies that a coupling approach might just be doomed to fail regardless.

3.2. Light-traffic results

As discussed in the above paragraph, stochastic dominance results in full generality do not seem within reach. As we now proceed to demonstrate however, this degree of generality can be tackled if we consider a light-traffic regime.

Let $Q_\lambda(P)$ be a random variable with the stationary distribution of the total number of jobs in the system with an underlying graph structure with edge selection probabilities $P = (p_{\{i,j\}})_{i,j}$. With C the normalization constant in (1) equal to $\mathbb{P}\{Q_\lambda(P) = 0\}$ it can easily be seen that

$$\mathbb{P}\{Q_\lambda(P) = q\} = \mathbb{P}\{Q_\lambda(P) = 0\} \cdot \alpha_q(P) \cdot \left(\frac{N\lambda}{\mu}\right)^q, \tag{7}$$

with $q \geq 1$. Define

$$\alpha_q(P) \coloneqq \sum_{\boldsymbol{c} \in \mathcal{E}^q} \prod_{i=1}^{q} \frac{p_{c_i}}{\left|\bigcup_{j=1}^{i} \{c_j\}\right|}, \tag{8}$$

which only depends on the edge selection probabilities and not on the total arrival rate $N\lambda$ or the server speed μ. Obviously, the probability of an empty system, $\mathbb{P}\{Q_\lambda(P) = 0\}$, tends to one when λ approaches zero. Note that $\alpha_1(P)$ is always equal to $1/2$ since $|\{c\}| = 2$ for all $c \in \mathcal{E}$. The value $\alpha_q(P)$ for any $q \geq 2$ depends on the edge selection probabilities, for instance, $|\{c_1\} \cup \{c_2\}|$ is equal to 2, 3 or 4 whenever the edges c_1 and c_2 are either the same, have one common endpoint, or no common endpoints, respectively. Therefore, $\alpha_q(P)$ will determine how various underlying structures will perform compared to each other when λ approaches zero as formalized in the following theorem.

Theorem 1. *For any $q \geq 1$, if $P' = (p'_{\{i,j\}})_{i,j}$ and $P = (p_{\{i,j\}})_{i,j}$ are two sets of edge selection probabilities such that $\alpha_q(P') \leq \alpha_q(P)$, then*

$$\lim_{\lambda \downarrow 0} \frac{\mathbb{P}\{Q_\lambda(P') \geq q\}}{\mathbb{P}\{Q_\lambda(P) \geq q\}} \leq 1.$$

The following lemma allows us to establish Theorem 1.

Lemma 3. *Let $P' = (p'_{\{i,j\}})_{i,j}$ and $P = (p_{\{i,j\}})_{i,j}$ be two sets of edge selection probabilities, then for any $q \geq 0$*

$$\frac{\mathbb{P}\{Q_\lambda(P') \geq q\}}{\mathbb{P}\{Q_\lambda(P) \geq q\}} = \frac{\alpha_q(P') + o(1)}{\alpha_q(P) + o(1)} \tag{9}$$

as $\lambda \downarrow 0$.

Proof. The Taylor expansion of $\mathbb{P}\{Q_\lambda(P) = 0\}$ near λ equal to zero yields

$$\mathbb{P}\{Q_\lambda(P) = 0\} = \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{N\lambda}{\mu}\right)^q \frac{\mathrm{d}^k}{\mathrm{d}x^k} \mathbb{P}\{Q_\lambda(P) = 0\}|_{\lambda \downarrow 0}, \tag{10}$$

with $x \coloneqq N\lambda/\mu$. Combining (10) with the observations in (7) and (8) gives $\mathbb{P}\{Q_\lambda(P) = q\} = \alpha_q(P)(N\lambda/\mu)^q + o(\lambda^q)$ from which (9) follows. □

From Theorem 1 it can be deduced that Conjecture 1 holds in a light-traffic regime once we are able to establish an inequality relation for the $\alpha_q(P)$ values involved. More precisely, none of the weighted power-of-two policies achieves better performance than the classical power-of-two policy when it can be shown that the uniform edge selection probabilities yield the smallest values of $\alpha_q(P)$ for all $q \geq 1$. So, proving Conjecture 1 in a light-traffic regime boils down to an optimization problem in terms of $\alpha_q(P)$ as a function of the edge selection probabilities $P = (p_{\{i,j\}})_{i,j}$. Note that $\alpha_q(\cdot)$ is a multivariate polynomial of degree q, hence it is continuous. Moreover, the set of edge selection probabilities is compact in $\mathbb{R}^E$, with $E = \binom{N}{2}$, implying that $\alpha_q(\cdot)$ must attain its global minimum. With the above observations in mind, we now present the following conjecture.

Conjecture 2. Let $\alpha_q^* \coloneqq \alpha_q(P')$ with $P' = (p'_{\{i,j\}})_{i,j}$ such that $p'_{\{i,j\}} \equiv 1/\binom{N}{2}$ for all i and j, $i \neq j$, then

$$\alpha_q^* = \min\left\{\alpha_q(P) \mid P = \left(p_{\{i,j\}}\right)_{i,j}\right\},$$

for all $q \geq 0$.

The above conjecture was shared in personal communication with Brosch, Laurent and Steenkamp, who showed that $\alpha_q(P)$, as a function of $P = (p_{\{i,j\}})_{i,j}$, is a convex polynomial for $q = 2$ and 3. Hence, choosing all edge selection probabilities to be uniform will minimize $\alpha_q(\cdot)$ [6, Theorem 2]. Polak later extended this convexity result to $q \leq 9$ [21, Theorem 1.1]. In [21] convexity is established once the Hessian matrix of α_q is positive semidefinite via a symmetry reduction. Proving semidefiniteness of the obtained lower-dimensional matrices increases in complexity as it becomes computationally harder to obtain the matrix coefficients for larger values of q. Combining [21, Theorem 1.1] with Theorem 1 results in the following corollary.

Corollary 1. *Let Q_λ^* and $Q_\lambda(P)$ denote the total number of jobs in stationarity in a system with N servers operating according to the classical power-of-two policy and a weighted power-of-two policy with edge selection probabilities $P = (p_{\{i,j\}})_{i,j}$, respectively. Then, for $q \leq 9$,*

$$\lim_{\lambda \downarrow 0} \frac{\mathbb{P}\{Q_\lambda^* \geq q\}}{\mathbb{P}\{Q_\lambda(P) \geq q\}} \leq 1.$$

Table 1 gives a comparison between α_q^* and $\alpha_q(P)$ for several values of q and when the underlying structure is a ring. We observe that for a fixed number of servers N, in analogy to the observations in Subsection 3.1, the performance of the system governed by the homogeneous ring is closer to the performance of the uniform case than the heterogeneous rings as $\alpha_q^*/\alpha_q(P)$ in this case is closer to 1.

Computing α_q for a given set of edge selection probabilities $P = (p_{\{i,j\}})_{i,j}$ is time-consuming as it requires summation over $|\mathcal{E}|^q$ terms, and also formed the bottleneck to prove the convexity results in [21] for values of $q \geq 10$. However, the above numerical results support the statement in Conjecture 2.

Remark 4. The condition $\alpha_q(P') \leq \alpha_q(P)$ in Theorem 1 is not sufficient to establish the stochastic dominance result in Conjecture 1

Table 1
The fraction of $\alpha_q^*/\alpha_q(P)$ for various values of q when the edge selection probabilities P correspond to a ring structure. Lemma 3 implies that this fraction corresponds to $\mathbb{P}\{Q_\lambda^* \geq q\}/\mathbb{P}\{Q_\lambda(P) \geq q\}$ when $\lambda \downarrow 0$.

	$N=4$				$N=8$		
	$q=2$	$q=4$	$q=10$	$q=16$	$q=2$	$q=4$	$q=10$
hom. ring	0.9804	0.9432	0.9046	0.9004	0.9754	0.8947	0.6586
het. ring $\epsilon=0.7$	0.9713	0.9055	0.8957	0.7850	0.9700	0.8707	0.5831
het. ring $\epsilon=0.9$	0.9448	0.8063	0.5481	0.4509	0.9543	0.8051	0.4095

for any fixed value of λ, even when this inequality could be shown to hold for all $q \geq 1$, which is due to the behavior of the normalization constant. However, a sufficient condition would be

$$\frac{\alpha_{q-1}(P')}{\alpha_q(P')} \geq \frac{\alpha_{q-1}(P)}{\alpha_q(P)} \tag{11}$$

for all $q \geq 1$, which also implies the condition in Theorem 1. The fact that $Q(P') \leq_{\text{st}} Q(P)$ once condition (11) is fulfilled for all $q \geq 1$ follows from a direct comparison of the respective stationary distributions in (7). Details of the proof are deferred to the online appendix.

Remark 5. We used the product-form distributions to establish the light-traffic result in Theorem 1, while usually a light-traffic approach is only considered when explicit formulas are lacking, and then based on the powerful framework developed by Reiman and Simon [22]. The latter framework outlines an approach to determine the coefficients of the Taylor expansion in (10). To derive these coefficients, one has to take into account the arrival and departure times of individual jobs, as well as the exact service rate at each moment in time, which is complicated by the fact that multiple servers can be processing a replica of the same job. Hence, it is notationally and computationally more convenient to leverage the product-form expressions which directly furnish the desired coefficients in terms of (8).

4. Discussion

4.1. Design implications

In Section 3.2 we proved a partial version of Conjecture 1 implying that non-uniform edge selection probabilities cannot yield better performance than uniform ones in a light-traffic regime. In many situations however, strictly uniform edge selection probabilities may simply not be feasible because of assignment constraints. Theorem 1, in conjunction with Lemma 3, then provides a specific guideline for the design of an efficient assignment policy subject to these constraints as we will now illustrate.

Assume that there are K different job types with arrival rates $\lambda_1, \ldots, \lambda_K$ and $\sum_{k=1}^K \lambda_k = N\lambda$. In the system design one has to choose (once and for all) for each job type $k = 1, \ldots, K$ which server pair (or edge) $e \in \mathcal{E}$ is eligible for assignment. For example, the various job types may correspond to requests for different data objects, which each can only be stored at two servers. The assignment policy can thus be represented in terms of binary decision variables $(x_{e,k})_{e,k}$, which are equal to 1 if job type k can be assigned to the servers at the endpoints of edge e, and 0 otherwise. The aim is to find values for the variables $\boldsymbol{x} = (x_{e,k})_{e,k}$ yielding edge selection probabilities $P(\boldsymbol{x}) = (p_e(\boldsymbol{x}))_e$ that stochastically minimize the number of jobs in the system. The edge selection probabilities may be expressed in terms of the decision variables as

$$p_e(\boldsymbol{x}) = \frac{1}{N\lambda} \sum_{k=1}^K \lambda_k x_{e,k} \quad \text{for all } e \in \mathcal{E}. \tag{12}$$

Recalling that $\mathbb{P}\{Q(P(\boldsymbol{x})) \geq q\} = \alpha_q(P(\boldsymbol{x})) \cdot (N\lambda/\mu)^q + o(\lambda^q)$ for $q \geq 1$ and $\alpha_1(P(\boldsymbol{x})) \equiv 1/2$, Lemma 3 and Theorem 1 suggest that the following minimization problem should be solved in order to obtain the ideal distribution of the various types as captured by (12):

$$\begin{aligned}
\min \quad & \alpha_2(P(\boldsymbol{x})) = \sum_{\boldsymbol{c} \in \mathcal{E}^2} \frac{p_{c_1}(\boldsymbol{x})}{2} \frac{p_{c_2}(\boldsymbol{x})}{|\{c_1\} \cup \{c_2\}|} \\
\text{s.t.} \quad & \sum_{e \in \mathcal{E}} x_{e,k} = 1 \\
& \text{for all } k = 1, \ldots, K, && \text{(13a)} \\
& N\lambda \sum_{e \in \mathcal{I}} p_e(\boldsymbol{x}) = \sum_{k=1}^K \lambda_k \sum_{e \in \mathcal{I}} x_{e,k} < \mu(\mathcal{I}) \\
& \text{for all } \mathcal{I} \subseteq \mathcal{E}, && \text{(13b)} \\
& x_{e,k} \in \{0,1\} \quad \text{for all } (e,k) \in |\mathcal{E}| \times K.
\end{aligned}$$

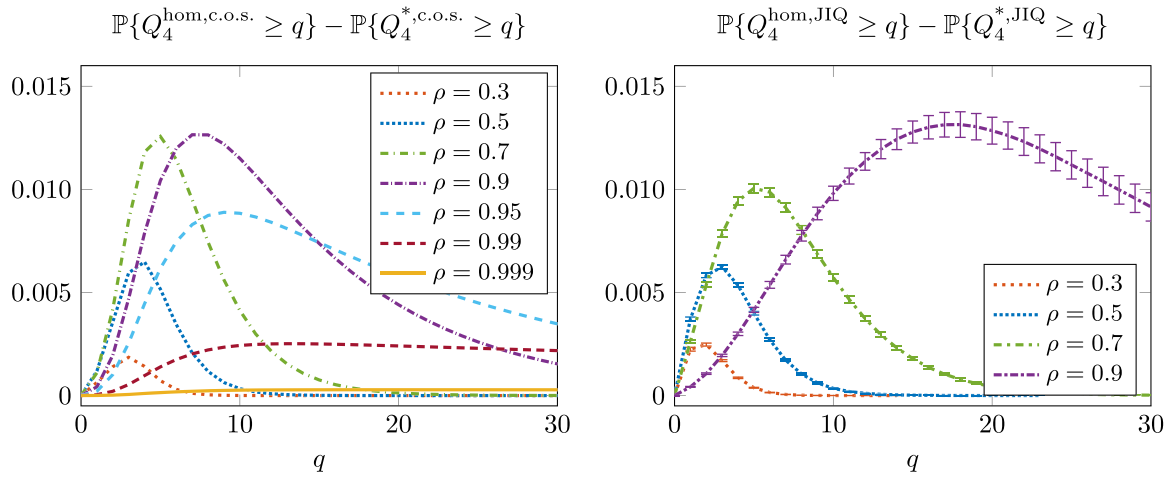
In the above optimization problem, (13a) guarantees that each job type k gets assigned to precisely one server pair or edge. Moreover, (13b) ensures that the system is stable, with $\mu(\mathcal{I})$ as defined in (2), in accordance with the stability conditions identified in [12].

Remark 6. A feasible solution $\boldsymbol{x} = (x_{e,k})_{e,k} \in \{0,1\}^{|\mathcal{E}| \times K}$ cannot be constructed for all arrival rate vectors $\{\lambda_k\}_k$. This is for instance the case when the sufficient conditions to guarantee stability, stated in Remark 2, are not satisfied, even if the necessary condition $N\lambda = \sum_{k=1}^K \lambda_k < N\mu$ is met. A simple counter example can be constructed in a system with $N = 4$ servers and $K = 2$ job types with arrival rates $\lambda_1 = (2+\delta)\mu$ and $\lambda_2 = (2-2\delta)\mu$ for any $\delta \in (0,1)$. Even though $\lambda < \mu$, there exists no allocation of the two job types that yields a stable system as $\lambda_1 \geq 2\mu$.

4.2. Other load balancing policies

The broader theme of the present paper is comparing the performance of weighted power-of-d policies with that of the classical power-of-d policy. As mentioned earlier, the notion that the performance of the latter policy serves as an upper bound for the performance of the former policies was supported through heuristic arguments and simulations by Gast [13], Mitzenmacher [18] and Turner [25] in a JSQ context. We proved that this property indeed holds for redundancy policies both in small systems and in systems of arbitrary size in the light-traffic regime. We focused on redundancy policies in view of the explicit product-form distributions, but we expect that the stochastic comparison results extend to load balancing policies beyond redundancy policies. This introduces interesting directions for further research as the above methods cannot directly be applied to analyze these alternative policies.

The redundancy policy described in Section 2 is often referred to as the *redundancy cancel-on-completion* (c.o.c.) policy. A natural policy to investigate as well is the *redundancy cancel-on-start* (c.o.s.) policy. Instead of discarding the redundant replicas once one of them finished service, redundant replicas are now discarded



(a) Redundancy c.o.s., based on Lemmas 4 and 5. (b) JIQ, based on simulations.

Fig. 3. A comparison between the stationary distributions of the total number of jobs for four-server systems with a uniform complete graph and a homogeneous ring as underlying structures.

once one of them starts service. Hence, the job is served at the server where its replica encountered the smallest workload, yielding an alternative implementation of the *Join-the-Smallest-Workload* (JSW) policy [2,3]. Whenever several replicas find idle servers upon arrival, the job will undergo service at the server that has been idle for the longest time, referred to as Assign-to-Longest-Idle-Server (ALIS).

The system occupancy at time t is given by $(c_1, \ldots, c_{Q'(t)}; u_1, \ldots, u_{L(t)})$ with $Q'(t)$ the total number of *waiting* jobs in the system at time t and $c_q \in \mathcal{E}$ denoting the qth oldest waiting job in the system. There are $L(t)$ idle servers at time t and server $u_l \in \{1, \ldots, N\}$ is the lth longest idle server in the system. Note that it is not feasible to have simultaneously waiting type-$\{i, j\}$ jobs and either server i or j idle. It was shown in [2,3] that, under suitable stability conditions, the stationary distribution of the system occupancy is

$$\pi^{\text{c.o.s.}}(c_1, \ldots, c_{Q'}; u_1, \ldots, u_L) = C' \prod_{i=1}^{Q'} \frac{N\lambda p_{c_i}}{\mu(c_1, \ldots, c_i)} \prod_{l=1}^{L} \frac{\mu}{\lambda_{\mathcal{C}(u_1,\ldots,u_l)}}, \tag{14}$$

with C' the normalization constant, $\mu(c_1, \ldots, c_i)$ as defined in (2) and

$$\lambda_{\mathcal{C}(u_1,\ldots,u_l)} = N\lambda \sum_{e \in \mathcal{E}\,:\, e \cap \{u_1,\ldots,u_l\} \neq \emptyset} p_e$$

the total arrival rate of jobs that can be served by the idle servers $\{u_1, \ldots, u_l\}$. Comparing this product-form expression with (1) for the redundancy c.o.c. policy reveals that obtaining stationary probabilities at an aggregate level is now also affected by the servers that are idle and the relative times they became idle.

For small systems it is possible to obtain the stationary distribution of the total number of jobs in the system from (14). The next two lemmas mirror the results in Subsection 3.1.

Lemma 4. *The stationary distribution of the total number of jobs in a system with the uniform complete graph structure on $N = 4$ servers operating under the redundancy c.o.s. policy is given by*

$$\mathbb{P}\{Q_4^{*,\text{c.o.s.}} = q\} = C_4^{*,\text{c.o.s.}} \left\{ 20\rho^q + \frac{4}{3^3}\left(\frac{\rho}{3}\right)^q - \frac{5 \cdot 2^5}{3^3}\left(\frac{2\rho}{3}\right)^q \right\}, \tag{15}$$

with $q \geq 1$, $\rho := \frac{\lambda}{\mu} < 1$ and

$$C_4^{*,\text{c.o.s.}} = \frac{(1-\rho)(3-\rho)(3-2\rho)}{(1+\rho)(3+\rho)(3+2\rho)}.$$

Lemma 5. *The stationary distribution of the total number of jobs in a system with the homogeneous ring structure on $N = 4$ servers operating under the redundancy c.o.s. policy is given by*

$$\mathbb{P}\{Q_4^{\text{hom,c.o.s.}} = q\} = C_4^{\text{hom,c.o.s.}} \left\{ 5\rho^q + \frac{1}{3}\left(\frac{\rho}{2}\right)^q - 2\left(\frac{2\rho}{3}\right)^q \right\}, \tag{16}$$

with $q \geq 1$, $\rho := \frac{\lambda}{\mu} < 1$ and

$$C_4^{\text{hom,c.o.s.}} = \frac{48(1-\rho)(2-\rho)(3-2\rho)}{-2\rho^3 + 55\rho^2 + 121\rho + 66}.$$

The proofs of Lemmas 4 and 5 are given in the online appendix. A comparison between the two stationary distributions in Lemmas 4 and 5 for various values of ρ can be found in Fig. 3a, which suggests that an equivalent result as in Proposition 1 holds for the redundancy c.o.s. policy, namely, $Q_4^{*,\text{c.o.s.}} \leq_{\text{st}} Q_4^{\text{hom,c.o.s.}}$.

Remark 7. In Fig. 3a it can be seen that the difference between the cumulative distributions of $Q_4^{\text{hom,c.o.s.}}$ and $Q_4^{*,\text{c.o.s.}}$, though still positive, narrows for values of λ approaching μ. This observation is in line with the heavy-traffic results in [8, Theorem 1] for both redundancy c.o.c. and c.o.s. policies, showing that $(1 - \lambda/\mu)Q(P)$ converges in distribution to an exponentially distributed random variable with unit mean for any set of edge selection probabilities that do not create local bottlenecks when $\lambda \uparrow \mu$.

A crucial difference between the product-form distributions for redundancy c.o.c. and c.o.s. is the fact that the normalization constant of the former corresponds to the probability that the system is completely idle, while for the latter it corresponds to the probability that there are no waiting jobs in the system and all servers are occupied. From this it immediately follows that the normalization constant will not tend to 1 in a light-traffic regime, implying

that a direct generalization of the reasoning in Subsection 3.2 is not applicable. It is worthwhile to note that there exist alternative state descriptors for which the normalization constant does coincide with the probability that the system is completely idle, see for instance [11, Theorem 3.10]. However, the corresponding product-form stationary distribution is inherently more complex than the one in (14), which would yield additional challenges when proving an equivalent formulation of Theorem 1 for the redundancy c.o.s. policy.

As mentioned earlier, the notion that non-uniform sampling cannot yield better performance than uniform sampling is a quite natural one and expected to apply more broadly for load balancing policies beyond JSQ and redundancy strategies.

We will now numerically illustrate this for the so-called Join-the-Idle-Queue (JIQ) policy, which has attracted significant attention in the load balancing literature recently. The JIQ policy assigns an arriving job to an idle (compatible) server, if any. Otherwise, the job is assigned to a randomly selected (compatible) server. Since no expressions are available for the stationary distribution, we used simulations to compare the empirical distributions of systems with a homogeneous ring and a uniform complete graph for $N = 4$ servers for various values of ρ. From Fig. 3b it can again be observed that the stochastic ordering result holds, i.e., $Q_4^{*,\mathrm{JIQ}} \leq_{\mathrm{st}} Q_4^{\mathrm{hom,\,JIQ}}$. The comparison in Fig. 3b is based on 50 simulation runs per value of ρ, each run consisting of 10 000 000 events. Besides the average difference between the cumulative distributions of $Q_4^{\mathrm{hom,JIQ}}$ and $Q_4^{*\mathrm{JIQ}}$, also its 95% confidence intervals are plotted.

Data availability

No data was used for the research described in the article.

Acknowledgements

The work of S. Borst was partly supported by the Dutch Research Council (NWO) through Gravitation grant NETWORKS-024.002.003. The work of J.S.H. van Leeuwaarden was partly supported by VICI grant 202.068.

Appendix. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.orl.2022.10.006.

References

[1] J. Abdul Jaleel, S. Doroudi, K. Gardner, A. Wickeham, A general "power-of-d" dispatching framework for heterogeneous systems, Queueing Syst. (2022), https://doi.org/10.1007/s11134-022-09736-z, in press.
[2] I. Adan, I. Kleiner, R. Righter, G. Weiss, FCFS parallel service systems and matching models, Perform. Eval. 127–128 (2018) 253–272.
[3] U. Ayesta, T. Bodas, I.M. Verloop, On a unifying product form framework for redundancy models, Perform. Eval. 127–128 (2018) 93–119.
[4] Y. Azar, A.Z. Broder, A.R. Karlin, E. Upfal, Balanced allocations, SIAM J. Comput. 29 (1) (1999) 180–200.
[5] M. van der Boor, S.C. Borst, J.S.H. van Leeuwaarden, D. Mukherjee, Scalable load balancing in networked systems: universality properties and stochastic coupling methods, in: Proc. ICM '18, 2018.
[6] D. Brosch, M. Laurent, A. Steenkamp, Optimizing hypergraph-based polynomials modeling job-occupancy in queuing with redundancy scheduling, SIAM J. Optim. 31 (3) (2021) 2227–2254.
[7] A. Budhiraja, D. Mukherjee, R. Wu, Supermarket model on graphs, Ann. Appl. Probab. 29 (3) (2019) 1740–1777.
[8] E. Cardinaels, S.C. Borst, J.S.H. van Leeuwaarden, Heavy-traffic universality of redundancy systems with assignment constraints, Preprint, accepted at Oper. Res.
[9] K. Gardner, J. Abdul Jaleel, A. Wickeham, S. Doroudi, Scalable load balancing in the presence of heterogeneous servers, Perform. Eval. 145 (2021) 102–151.
[10] K. Gardner, M. Harchol-Balter, A. Scheller-Wolf, M. Velednitsky, S. Zbarsky, Redundancy-d: the power of d choices for redundancy, Oper. Res. 65 (4) (2017) 1078–1094.
[11] K. Gardner, R. Righter, Product forms for FCFS queueing models with arbitrary server-job compatibilities: an overview, Queueing Syst. 96 (1) (Oct. 2020) 3–51.
[12] K. Gardner, S. Zbarsky, S. Doroudi, M. Harchol-Balter, E. Hyytiä, A. Scheller-Wolf, Queueing with redundant requests: exact analysis, Queueing Syst. 83 (3–4) (2016) 227–259.
[13] N. Gast, The power of two choices on graphs: the pair-approximation is accurate?, ACM SIGMETRICS Perform. Eval. Rev. 43 (2) (2015) 69–71.
[14] Y.T. He, D.G. Down, Limited choice and locality considerations for load balancing, Perform. Eval. 65 (9) (2008) 670–687.
[15] T. Hellemans, T. Bodas, B. Van Houdt, Performance analysis of workload dependent load balancing policies, Proc. ACM Meas. Anal. Comput. Syst. 3 (2) (2019) 1–35.
[16] T. Hellemans, B. Van Houdt, On the power-of-d-choices with least loaded server selection, Proc. ACM Meas. Anal. Comput. Syst. 2 (2) (2018) 1–22.
[17] T. Hellemans, B. Van Houdt, Mean waiting time in large-scale and critically loaded power of d load balancing systems, Proc. ACM Meas. Anal. Comput. Syst. 5 (2) (2021).
[18] M. Mitzenmacher, The power of two choices in randomized load balancing, PhD thesis, University of California, Berkeley, 1996.
[19] M. Mitzenmacher, The power of two choices in randomized load balancing, IEEE Trans. Parallel Distrib. Syst. 12 (10) (2001) 1094–1104.
[20] D. Mukherjee, S.C. Borst, J.S.H. van Leeuwaarden, Asymptotically optimal load balancing topologies, Proc. ACM Meas. Anal. Comput. Syst. 2 (1) (2018).
[21] S.C. Polak, Symmetry reduction to optimize a graph-based polynomial from queueing theory, SIAM J. Appl. Algebra Geom. 6 (2) (2022) 243–266.
[22] M.I. Reiman, B. Simon, Open queueing systems in light traffic, Math. Oper. Res. 14 (1) (1989) 26–59.
[23] D. Rutten, D. Mukherjee, Load balancing under strict compatibility constraints, Math. Oper. Res. (2022), https://doi.org/10.1287/moor.2022.1258, in press.
[24] F. Sloothaak, J.R. Cruise, S. Shneer, M. Vlasiou, B. Zwart, Complete resource pooling of a load-balancing policy for a network of battery swapping stations, Queueing Syst. 99 (2021) 65–120.
[25] S.R.E. Turner, The effect of increasing routing choice on resource pooling, Probab. Eng. Inf. Sci. 12 (1998) 109–124.
[26] N.D. Vvedenskaya, R.L. Dobrushin, F.I. Karpelevich, Queueing system with selection of the shortest of two queues: an asymptotic approach, Probl. Inf. Transm. 32 (1) (1996) 20–34.
[27] W. Weng, X. Zhou, R. Srikant, Optimal load balancing with locality constraints, Proc. ACM Meas. Anal. Comput. Syst. 4 (3) (2020).
[28] D. Zhan, G. Weiss, Many-server scaling of the N-system under FCFS-ALIS, Queueing Syst. 88 (1) (2018) 27–71.