EURANDOM PREPRINT SERIES

2021-004

May 31, 2021

Fork-join and redundancy systems with heavy-tailed job sizes

Y. Raaijmakers, S. Borst, O. Boxma
ISSN 1389-2355

Fork-join and redundancy systems with heavy-tailed job sizes

Youri Raaijmakers[a,*], Sem Borst[a], Onno Boxma[a]

[a]*Eindhoven University of Technology*
Department of Mathematics and Computer Science

Abstract

We investigate the tail asymptotics of the response time distribution for the cancel-on-start (c.o.s.) and cancel-on-completion (c.o.c.) variants of redundancy-d scheduling and the fork-join model with heavy-tailed job sizes. We present bounds, which only differ in the pre-factor, for the tail probability of the response time in the case of the first-come first-served (FCFS) discipline. For the c.o.s. variant we restrict ourselves to redundancy-d scheduling, which is a special case of the fork-join model. In particular, for regularly varying job sizes with tail index $-\nu$ the tail index of the response time for the c.o.s. variant of redundancy-d equals $-\min\{d_{\text{cap}}(\nu-1), \nu\}$, where $d_{\text{cap}} = \min\{d, N-k\}$, N is the number of servers and k is the integer part of the load. This result indicates that for $d_{\text{cap}} < \frac{\nu}{\nu-1}$ the waiting time component is dominant, whereas for $d_{\text{cap}} > \frac{\nu}{\nu-1}$ the job size component is dominant. Thus, having $d = \lceil \min\{\frac{\nu}{\nu-1}, N-k\} \rceil$ replicas is sufficient to achieve the optimal asymptotic tail behavior of the response time. For the c.o.c. variant of the fork-join($n_{\text{F}}, n_{\text{J}}$) model the tail index of the response time, under some assumptions on the load, equals $1-\nu$ and $1-(n_{\text{F}}+1-n_{\text{J}})\nu$, for identical and i.i.d. replicas, respectively; here the waiting time component is always dominant.

Keywords: Parallel-server systems, fork-join, redundancy, heavy-tailed distributions, response time asymptotics

---

*Corresponding author
Email address: `y.raaijmakers@tue.nl` (Youri Raaijmakers)

1. Introduction

In recent years, the fork-join model has attracted strong interest. This model is a theoretical abstraction of the popular MapReduce framework [8]. MapReduce is a programming model for processing and generating big data sets with parallel algorithms on clusters. In MapReduce every job is divided into tasks which can be processed in parallel in any order. For completion of the job the completed tasks need to be joined together.

Fork-join model

In the fork-join($n_{\mathrm{F}}, n_{\mathrm{J}}$) model tasks of a job are assigned to n_{F} servers selected uniformly at random. Redundant tasks are abandoned as soon as n_{J} of the n_{F} tasks either enter service ('cancel-on-start', c.o.s.) or finish service ('cancel-on-completion', c.o.c.). The job is completed when all these n_{J} tasks complete service.

Analytical results for the fork-join model are unfortunately scarce. Tight characterizations of the response time are only known in the special case of $n_{\mathrm{F}} = n_{\mathrm{J}} = 2$, see [10]. For a survey on results in other special cases we refer to [28]. Results for the expectation of the response time are established when $n_{\mathrm{F}} = n_{\mathrm{J}} \to \infty$, see for example [3, 22]. For a more detailed overview of the results and applications we refer to [17].

Redundancy scheduling

Redundancy-d scheduling is a special case within the fork-join model. In redundancy-d scheduling replicas of a job are assigned to d servers selected uniformly at random. Redundant replicas are abandoned as soon as one of the d replicas either enters service or finishes service. Thus redundancy-d scheduling is equivalent to the fork-join model with $n_{\mathrm{F}} = d$ and $n_{\mathrm{J}} = 1$. Observe that the c.o.s. variant of redundancy-d is equivalent to the Join-the-Shortest-Workload-d (JSW-d) policy, which assigns each job to the server with the smallest workload among d servers selected uniformly at random, see [2]. The c.o.c. variant of redundancy-d shares similarities with a strategy that assigns the job to the server that provides the minimum response time among d servers selected uniformly at random, but involves possibly concurrent service of multiple replicas.

It has been empirically shown that redundancy scheduling can improve performance in parallel-server systems [30], especially in case of highly variable job sizes. More specifically, for large-scale applications such as Google search, the ability of redundancy scheduling to reduce the expectation and the tail of the response time has been demonstrated [7]. Our understanding of redundancy scheduling is growing, and especially the stability condition for c.o.c. redundancy policies has received considerable attention, however, expressions for performance metrics such as the expectation or the distribution of the response time remain scarce. In [16] analytical expressions for the expected response time are obtained for exponential job sizes and independent and identically distributed (i.i.d.) replicas. Under the assumption of asymptotic independence a fixed-point equation characterizing the response time distribution for identical and i.i.d. replicas is derived in [19].

In this paper we examine the tail behavior of the response time when job sizes are heavy-tailed, which is one of the most relevant scenarios in redundancy scheduling and the fork-join model. Indeed, heavy tails in parallel processing are encountered in conjunction with the MapReduce framework developed at Google and its Hadoop open source implementation [9]. Moreover, measurement studies show that workload characteristics such as file sizes, CPU times, and session lengths tend to be heavy-tailed, see [17, 23, 31] and the references therein. The tail behavior of the waiting time distribution of the single server queue is well known, see for example [29] or [31, Chapter 2]. Let W_{FCFS} denote the waiting time for the single-server queue with the FCFS discipline, for subexponential (see Definition 2 in Appendix A) residual job sizes B^{res},

$$\mathbb{P}(W_{\text{FCFS}} > x) \sim \frac{\tilde{\rho}}{1-\tilde{\rho}}\mathbb{P}(B^{\text{res}} > x) \quad \text{as } x \to \infty, \tag{1}$$

where $\tilde{\rho} := \frac{\mathbb{E}[B]}{\mathbb{E}[A]}$ denotes the load with A the interarrival time and B the job size, and

$$\mathbb{P}(B^{\text{res}} > x) = \frac{1}{\mathbb{E}[B]} \int_{y=x}^{\infty} \mathbb{P}(B > y) \mathrm{d}y.$$

In particular for regularly varying (see Definition 6 in Appendix A) job size distributions with index $-\nu$, i.e., $\mathbb{P}(B > x) = x^{-\nu} L(x)$ with $L(\cdot)$ a slowly varying function at infinity,

$$\mathbb{P}(W_{\text{FCFS}} > x) \sim \frac{\tilde{\rho}}{1 - \tilde{\rho}} \frac{1}{(\nu - 1)\mathbb{E}[B]} L(x) x^{1-\nu} \quad \text{as } x \to \infty. \tag{2}$$

One way to understand the tail index $1 - \nu$ is the following. The workload (and waiting time) in an $M/G/1$ queue is distributed as a geometric($\tilde{\rho}$) sum of residual job sizes B^{res}. According to the theory of regular variation [4], loosely speaking, regular variation is preserved under integration, and asymptotically one can integrate as if $L(y)$ is kept outside the integral; so

$$\mathbb{P}(B^{\text{res}} > x) = \frac{1}{\mathbb{E}[B]} \int_{y=x}^{\infty} L(y) y^{-\nu} \mathrm{d}y \sim \frac{1}{(\nu - 1)\mathbb{E}[B]} L(x) x^{1-\nu} \quad \text{as } x \to \infty, \tag{3}$$

which implies that if B is regularly varying with index $-\nu$, then B^{res} is regularly varying with index $1 - \nu$.

The tail behavior in the single-server queue has also been studied for other service disciplines. For regularly varying job sizes, the random order of service (ROS) discipline has the same tail index as the FCFS discipline, but with a smaller pre-factor [6]. For the last-come first-served with preemptive resume (LCFS-PR) discipline and the processor sharing (PS) discipline the tail index of the response time for regularly varying job sizes is the same as the tail index of the job size, see [32] and [33], respectively. Thus, from a tail perspective, these service disciplines perform better than the FCFS discipline.

Closer related to the c.o.s. variant of redundancy scheduling are the results for the tail behavior of the waiting time for the Join-the-Shortest-Workload (JSW) policy or equivalently the $GI/G/N$ queue, see [12] and [13]. The key idea in [12, 13] to first consider deterministic interarrival times made the derivation of the tail behavior substantially more tractable. In [13] it is shown that for long-tailed residual job sizes and $\tilde{\rho} > k$, where $k := \lfloor \tilde{\rho} \rfloor$ is the integer part of the load,

$$\mathbb{P}(W_{\text{JSW}} > x) \geq \frac{\tilde{\rho}^{N-k} + o(1)}{(N-k)!} \mathbb{P}\left(B^{\text{res}} > \frac{\tilde{\rho}+\delta}{\tilde{\rho}-k} x\right)^{N-k} \quad \text{as } x \to \infty, \tag{4}$$

for any $\delta > 0$. For subexponential residual job sizes and $\tilde{\rho} < k+1$ it is shown that

$$\mathbb{P}(W_{\text{JSW}} > x) \leq \binom{N}{k} \left(\frac{(k+1)\tilde{\rho}}{(k+1)-\tilde{\rho}} + o(1)\right)^{N-k} \mathbb{P}\left(B^{\text{res}} > x(1-\delta)\right)^{N-k} \quad \text{as } x \to \infty. \tag{5}$$

A heuristic explanation for the exponent $N-k$ is as follows. After the arrival of $N-k$ big jobs, $N-k$ servers will be working on these big jobs for a very long time. The other k servers form an unstable $GI/G/k$ system, which implies that the workload drifts linearly to infinity. Thus eventually the workload at all N servers will exceed level x, causing the waiting time of an arriving job to be larger than x.

In this paper we investigate the tail behavior of the response time for both the c.o.s. and c.o.c. variants of redundancy scheduling and the fork-join model when job sizes are heavy-tailed. Throughout the paper we assume that the system under consideration is in steady state. For regularly varying job sizes with tail index $-\nu$ and the FCFS discipline it is shown that the response time for the c.o.s. variant of redundancy-d has tail index $-\min\{d_{\text{cap}}(\nu-1), \nu\}$, where $d_{\text{cap}} = \min\{d, N-k\}$ and $k = \lfloor \tilde{\rho} \rfloor$. For small loads, this result indicates that for $d < \frac{\nu}{\nu-1}$ the waiting time component is dominant, whereas for $d > \frac{\nu}{\nu-1}$ the job size component is dominant. Thus, having $d = \lceil \min\{\frac{\nu}{\nu-1}\} \rceil$ replicas already achieves the optimal asymptotic tail behavior of the response time and creating even more replicas yields no improvements in terms of response time tail asymptotics. For high loads, the results indicate that creating many replicas yields no benefits for the tail index of the response time. For the c.o.c. variant of the more general fork-join$(n_{\text{F}}, n_{\text{J}})$ model with identical and i.i.d. replicas the tail index of the response time is $1-\nu$ and $1-(n_{\text{F}}+1-n_{\text{J}})\nu$, respectively, and the waiting time component is always dominant. Note that in this case

Table 1: Overview of the tail index for the c.o.s. and c.o.c. varaint of redundancy scheduling with various service disciplines where the job size is regularly varying with tail index $-\nu$. The star indicates that for this scenario we obtained results for the more general fork-join($n_{\mathrm{F}}, n_{\mathrm{J}}$) model.

	c.o.s.		c.o.c.	
	$GI/G/N$ (Red-N)	Red-d	Red-d (identical)	Red-d (i.i.d.)
FCFS	$-\min\{(N-k)(\nu-1), \nu\}$ [12, 13]	$-\min\{d_{\text{cap}}(\nu-1), \nu\}$	$1-\nu$ $(*)$	$1-d\nu$ $(*)$
LCFS-PR		$-\nu$	$-\nu$ $(*)$	$-d\nu$ $(*)$
PS		$-\nu$	$-\nu$ [26]	$-d\nu$ [26]

the tail index is independent of the load of the system and for identical replicas even independent of the number of replicas. In the special case of redundancy-d scheduling with identical and i.i.d. replicas it follows that the tail index of the response time is $1-\nu$ and $1-d\nu$, respectively. All these results for the c.o.c. variant rely on the fact that the upper bound system, which is used in the proof, is stable. The stability condition of this system does not necessarily coincide with the stability condition of the original fork-join model.

For the LCFS-PR discipline in the fork-join model we show that the response time tail is just as heavy as the job size tail, implying that for the c.o.c. variant this discipline achieves better tail asymptotics than the FCFS discipline. For the c.o.s. variant the LCFS-PR discipline has better tail asymptotics than the FCFS discipline for scenarios with low load and a small number of replicas; in all other scenarios both service disciplines have similar tail asymptotics. In [26] it is shown that for the c.o.c. variant of redundancy-d scheduling with the PS discipline the tail index of the response time is $-\nu$ for identical replicas and $-d\nu$ for i.i.d. replicas. Table 1 provides an overview of the tail index for the various models and service disciplines.

The remainder of the paper is organized as follows. In Section 2 we provide a model description and state preliminary results. In Section 3 we characterize the tail behavior of the response time for the c.o.s. variant of redundancy scheduling and the c.o.c. variant of the more general fork-join model with the FCFS discipline, with some proofs deferred to the appendix. In Section 4 we discuss the tail behavior in the fork-join model with the LCFS-PR discipline. Section 5 provides numerical results on the tail behavior of the response time in redundancy scheduling with Pareto distributed job sizes. Section 6 contains conclusions and some suggestions for further research. The paper ends with two appendices. Appendix A collects various definitions and results for heavy-tailed random variables, which will be used in the paper. Appendix B provides the proof for one of the main theorems in this paper.

2. Model description and preliminaries

Consider a system of N parallel unit-speed servers. Jobs arrive at the epochs of an arbitrary renewal process, with successive interarrival times A_i, $i \geq 1$, each distributed as a generic random variable A. When a job arrives, a dispatcher assigns replicas of the job to n_{F} servers chosen uniformly at random (without replacement), where $1 \leq n_{\mathrm{F}} \leq N$. We consider two possible variants where redundant replicas are abandoned as soon as n_{J} of the n_{F} replicas either enter service (c.o.s.) or finish service (c.o.c.). If in the c.o.s. variant multiple replicas enter service at exactly the same time, then one of these replicas is chosen uniformly at random and starts service. A special case of the fork-join model is redundancy-d scheduling, where $n_{\mathrm{F}} = d$ and $n_{\mathrm{J}} = 1$. Note that in the c.o.s. variant of redundancy-d the dependency structure between the replicas does not play a role, since at all times there is only one replica of the job in service. In contrast, in the c.o.c. variant of redundancy-d, and also in the fork-join model, several replicas of the same job may be in service at the same time, and hence the dependency structure does matter. We thus allow the replica sizes $B_1, \ldots, B_{n_{\mathrm{F}}}$ of a job to be governed by some joint distribution function $F_B(b_1, \ldots, b_{n_{\mathrm{F}}})$, where B_i, $i = 1, \ldots, n_{\mathrm{F}}$, are each distributed as some random variable B, but not necessarily independent. Special cases of the dependency structure are: i) perfect dependency between the variables, so-called identical replicas, where the job size is preserved for all replicas, i.e., $B_i = B$, for all $i = 1, \dots, n_{\mathrm{F}}$, ii) no dependency at all, so-called i.i.d. replicas.

Finally, let us denote the steady-state waiting times of the replicas at their n_{F} servers (the time until their service starts if they are still in the system) by $W_1, \dots, W_{n_{\mathrm{F}}}$ and the steady-state response time by R. Let $X_{(n_{\mathrm{J}})}$ denote the n_{J}th order statistic of a set of random variables $X_1, \dots, X_N$.

3. FCFS discipline

In this section we analyze the tail asymptotics of the response time with the FCFS discipline. For the c.o.s. variant (Section 3.1) we restrict ourselves to redundancy-d scheduling, whereas for the c.o.c. variant (Section 3.2) we allow for the more general fork-join model.

3.1. Cancel-on-start

Observe that the steady-state response time in the c.o.s. variant of redundancy-d is given by

$$R = \min\{W_1, \dots, W_d\} + B. \tag{6}$$

We refer to the time between the arrival of a job and the moment the first replica goes into service as the waiting time $W_{\min} = \min\{W_1, \dots, W_d\}$ of a job. As mentioned earlier, the c.o.s. variant of redundancy-d is equivalent to the Join-the-Shortest-Workload-d (JSW-d) policy, which assigns each job to the server with the smallest workload among d servers selected uniformly at random.

For general interarrival times and job sizes the stability condition for the system with the JSW-d policy and FCFS is given by $\tilde{\rho} = \frac{\mathbb{E}[B]}{\mathbb{E}[A]} < N$, see [11].

In [13, Theorem 1.6] lower and upper bounds are derived for the tail probability of the waiting time for the JSW policy. The same methodology can be used to find lower and upper bounds for JSW-d, and hence for the c.o.s. variant of redundancy scheduling with $1 \leq d \leq N$ replicas, resulting in Theorem 1. The two derived lower bounds in this theorem hold for every value of $\tilde{\rho}$, but they are asymptotically dominant for different regions of $\tilde{\rho}$, as explained after the theorem. Note that for $d = N$ Theorem 1 recovers the results of [13] as captured in (4) and (5), whereas for $d = 1$ the system is equivalent to a $GI/G/1$ queue for which the tail behavior is given by (2).

Theorem 1. *Consider the c.o.s. variant of redundancy-d scheduling with the FCFS discipline. Let* $k = \lfloor \tilde{\rho} \rfloor \in \{0, 1, \ldots, N-1\}$ *be the integer part of the load and* $\delta > 0$.

i) If the residual job size B^{res} *is long-tailed, then*

$$\mathbb{P}(W_{\min} > x) \geq \frac{1}{\binom{N}{d}} \frac{\tilde{\rho}^d + o(1)}{d!} \left(\bar{B}^{\text{res}} \left((1+\delta)\, x \right) \right)^d. \tag{7}$$

ii) If $\tilde{\rho} < N - d$ *and the residual job size* B^{res} *is subexponential, then*

$$\mathbb{P}(W_{\min} > x) \leq \binom{N}{d} \left(\frac{(k+1)\tilde{\rho}}{k+1-\tilde{\rho}} + o(1) \right)^d \left(\bar{B}^{\text{res}} \left(\frac{x(1-\delta)}{k+1} \right) \right)^d. \tag{8}$$

iii) If the residual job size B^{res} *is long-tailed, then*

$$\mathbb{P}(W_{\min} > x) \geq \frac{\tilde{\rho}^{N-k} + o(1)}{(N-k)!} \left(\bar{B}^{\text{res}} \left(\frac{\tilde{\rho} + \delta}{\tilde{\rho} - k} x \right) \right)^{N-k}. \tag{9}$$

iv) If $\tilde{\rho} > N - d$ *and the residual job size* B^{res} *is subexponential, then*

$$\mathbb{P}(W_{\min} > x) \leq \binom{N}{k} \left(\frac{(k+1)\tilde{\rho}}{k+1-\tilde{\rho}} + o(1) \right)^{N-k} \left(\bar{B}^{\text{res}} \left(\frac{(k+1-N+d)x(1-\delta)}{k+1} \right) \right)^{N-k}. \tag{10}$$

Proof. Let $\boldsymbol{V} = (V_1, \ldots, V_N)$ denote the vector of residual workloads of the servers. Recall that $V_{(i)}$ denotes the ith order statistic of the set $V_1, \ldots, V_N$. The proof of i) follows from the inequality

$$\mathbb{P}(W_{\min} > x) \geq \frac{1}{\binom{N}{d}} \mathbb{P}(V_{(1)} > x, \ldots, V_{(d)} > x),$$

with $\frac{1}{\binom{N}{d}}$ corresponding to the probability that the replicas of an arbitrary job are assigned to the servers with the d largest workloads, and where

$$\mathbb{P}(V_{(1)} > x, \ldots, V_{(d)} > x) \geq \frac{\tilde{\rho}^d + o(1)}{d!} \left(\bar{B}^{\text{res}} \left((1+\delta)\, x \right) \right)^d,$$

by similar arguments as in the proof of Lemma 3.1 in [13]. The proof of iii) follows from the inequality

$$\mathbb{P}(W_{\min} > x) \geq \mathbb{P}(V_1 > x, \ldots, V_N > x),$$

where

$$\mathbb{P}(V_1 > x, \ldots, V_N > x) \geq \frac{\tilde{\rho}^{N-k} + o(1)}{(N-k)!} \left(\bar{B}^{\text{res}} \left(\frac{\tilde{\rho} + \delta}{\tilde{\rho} - k} x \right) \right)^{N-k},$$

by similar arguments as in the proof of Theorem 5.1 in [13]. The proof of ii) and iv) can be found in Appendix B. □

As reflected in the proof sketches, the asymptotic lower bounds in (7) and (9) correspond to two different scenarios for a large value of $W_{\min}$ to occur. Scenario 1 involves the arrival of d jobs of size x or larger 'overlapping in time'. In the JSW-d system these jobs will be assigned to d different servers with overwhelming probability for large x, and thus the workload at these d servers will exceed x. A newly arriving job that is so unfortunate as to sample exactly these d servers (which happens with probability $1/\binom{N}{d}$) will experience a waiting time larger than x. Scenario 2 involves the arrival of $N-k$ sufficiently large jobs 'overlapping in time', which instantaneously causes the workloads at $N-k$ servers to become large as described above, assuming $N-k \leq d$. This will also result in subsequent jobs all being assigned to one of the other k servers and hence create overload, so that the workloads at these servers will gradually start growing. Thus, eventually the workloads at all servers will be large, and every arriving job will experience a large waiting time. Observe that this scenario corresponds to that in the GI/G/N queue discussed in [13], as illustrated by the match with Equation (4).

Scenarios 1 and 2 are asymptotically dominant in case $d \leq N-k$ and $d \geq N-k$, respectively, reflecting that a large waiting time is most likely due to a minimum number of $d_{\text{cap}} = \min\{d, N-k\}$ large jobs. Note that in Scenario 1 the workloads at all servers will in fact grow large as well when $d \geq N-k$, but that Scenario 2 dominates in that case.

Scenarios with large workloads at l servers, with $d < l < N$, do not asymptotically contribute to the probability of a large waiting time. This may be intuitively explained by observing the following. (i) If such scenarios involve strictly more than d large workloads without resulting in overload of all servers (so $d < l < N - k$) then they are asymptotically much less likely than Scenario 1. (ii) If such scenarios involve $l \geq N - k$ large workloads, this will quickly result in overload of all servers, just like in Scenario 2.

Corollary 1 (Analogous to Corollary 1.1 in [13])**.** *Let the residual job size B^{res} be long-tailed and dominated varying and $k < \tilde{\rho} < k + 1$, i.e., $\tilde{\rho}$ not an integer value. Then there exist constants c_1 and c_2 such that, for all x,*

$$c_1 \left(\bar{B}^{\text{res}}(x)\right)^{d_{\text{cap}}} \leq \mathbb{P}(W_{\min} > x) \leq c_2 \left(\bar{B}^{\text{res}}(x)\right)^{d_{\text{cap}}},$$

where $d_{\text{cap}} = \min\{d, N - k\}$.

Proof. The result follows directly from Theorem 1, the last inclusion in (A.3) and the definition of dominated variation (Definition 4 in Appendix A). □

Remark 1. *Note that in Corollary 1 we exclude integer values for the load. Most of the heavy-tail results focus on the case where the load is not an integer, since the integer case is significantly more delicate to analyze. For a detailed study on the integer case in the $GI/G/2$ queueing system we refer to [5].*

Corollary 2. *For the c.o.s. variant of redundancy-d scheduling with the FCFS discipline:*

i) if $B \in RV(-\nu)$, then $W_{\min} \in ORV(d_{\text{cap}}(1 - \nu))$,

ii) if $B \in RV(-\nu)$, then $R \in ORV(-\min\{d_{\text{cap}}(\nu - 1), \nu\})$.

Proof. It is well known that if $B \in RV(-\nu)$, then $B^{\text{res}} \in RV(1 - \nu)$, see (3). The proof of i) follows by applying this result to Corollary 1 together with the inclusion $RV \subset \mathcal{L} \cap \mathcal{D}$ from (A.3) and Lemma 6 in Appendix A. The proof of ii) follows by i), Equation (6) and Lemma 5 in Appendix A. □

From Corollary 2 we conclude that the waiting time component is dominant in the response time tail as long as $d_{\text{cap}} \leq \frac{\nu}{\nu-1}$, but otherwise the job size component is dominant. Better than that ($x^{-\nu}$ tail behavior) is, obviously, not possible for the response time. In other words, having more than $\frac{\nu}{\nu-1}$ replicas will not provide any improvement in the tail behavior. For example, consider a system with a sufficiently small load. If $\nu = 4/3$, then $d = 4$ already yields $R \in ORV(-\nu)$, and from a tail perspective choosing $d > 4$ yields no benefits. If $\nu = 3/2$, then it does not pay to take d larger than 3. If $\nu \geq 2$ (so B has a finite second moment), then it does not pay to take d larger than 2. For high loads, the results indicate that creating many replicas yields no benefits for the tail index of the response time.

3.2. Cancel-on-completion

In this section we analyze the tail behavior for the c.o.c. variant of the fork-join($n_{\text{F}}, n_{\text{J}}$) model. The steady-state response time is given by

$$R = (W + B)_{(n_{\text{J}})}. \tag{11}$$

Next, we provide an upper and lower bound for the waiting time and response time via the workload in an auxiliary single-server queue, which is similar to [25, Lemma 1].

Lemma 1. *The waiting time $W_{(n_{\text{J}})}$ and the response time R in the c.o.c. variant of the fork-join($n_{\text{F}}, n_{\text{J}}$) model with the FCFS discipline are stochastically bounded from above by the waiting time W_{U} and response time R_{U}, respectively, in a $GI/G/1/FCFS$ queue with interarrival time A and job size $B_{(n_{\text{J}})}$.*

Proof. Consider an auxiliary system in which all jobs are assigned to the same n_{F} servers and where servers wait with serving a new job until all the n_{J} replicas are finished. This system is equivalent to the $GI/G/1/FCFS$ queue as defined in the lemma. Let ω_i be the workload at server i, where we define the workload as the amount of work a server needs to complete to become idle in the absence of any arrivals. By induction it can be shown that ω_i is bounded from above by the workload ω_{U} in the auxiliary $GI/G/1/FCFS$ system at all times. Assume that $\omega_{(N)} \leq \omega_{\mathrm{U}}$ after the m-th arrival. Then, after the $(m+1)$-th arrival the new workload is

$$\omega_{s_l^{m+1}} = \max\{(\omega_{\boldsymbol{s}^{m+1}} + b)_{(n_{\mathrm{J}})}, \omega_{s_l}\} \leq \max\{(\omega_{(N)} + b)_{(n_{\mathrm{J}})}, \omega_{(N)}\} = \omega_{(N)} + b_{(n_{\mathrm{J}})},$$

for $l = 1, \ldots, n$, since $\omega_i \leq \omega_{(N)}$ for all $i = 1, \ldots, N$. Thus the increase in maximum workload is bounded by $b_{(n_{\mathrm{J}})}$, which is exactly the increase in workload in the corresponding $GI/G/1/FCFS$ queue. Observe that the bound for the workload implies $W_i \leq W_{\mathrm{U}}$ for all $i = 1, \ldots, n_{\mathrm{F}}$, from which it follows that $W_{(n_{\mathrm{J}})} \leq W_{\mathrm{U}}$ and

$$R = (W + B)_{(n_{\mathrm{J}})} \leq W_{\mathrm{U}} + B_{(n_{\mathrm{J}})} = R_{\mathrm{U}}.$$

$\square$

Lemma 2. *The waiting time $W_{(n_{\mathrm{J}})}$ and the response time R in the c.o.c. variant of the fork-join($n_{\mathrm{F}}, n_{\mathrm{J}}$) model with the FCFS discipline are stochastically bounded from below by the waiting time W_{L} and response time R_{L}, respectively, in a $GI/G/1/FCFS$ queue with a random selection of the arrivals based on Bernoulli experiments with probability $1/K$, i.e., mean interarrival time $K\mathbb{E}[A]$, and with a job size $B_{(n_{\mathrm{J}})}$, where $K = \binom{N}{n_{\mathrm{F}}}\frac{n_{\mathrm{F}}!}{(n_{\mathrm{F}}-n_{\mathrm{J}}+1)!}$.*

Proof. Consider an auxiliary system in which we only allow arrivals to n_{F} specific servers, say 1 up and until n_{F}, and for which the first n_{J} job sizes are the smallest and in increasing order. Thus, the smallest job is assigned to server 1, the second smallest job to server 2, etcetera up and until server n_{J}. Note that the other arrivals are only deleted in the auxiliary system and not in the original system, thus the amount of work cannot be lower. The auxiliary system is equivalent to the $GI/G/1/FCFS$ queue as defined in the lemma, since server n_{J} always finishes the n_{J}-th replica last. Similarly to Lemma 1, it can be shown that the workload ω_i at server i in the original system, is bounded from below by the workload ω_{L} in the auxiliary $GI/G/1/FCFS$ system. Observe that the bound for the workload implies $W_i \geq W_{\mathrm{L}}$ for all $i = 1, \ldots, n_{\mathrm{F}}$, from which it follows that $W_{(n_\mathrm{J})} \geq W_\mathrm{L}$ and

$$R = (W + B)_{(n_\mathrm{J})} \geq W_\mathrm{L} + B_{(n_\mathrm{J})} = R_\mathrm{L}.$$

□

A sufficient stability condition for general interarrival times and job sizes is $\rho_\mathrm{U} := \frac{\mathbb{E}[B_{(n_\mathrm{J})}]}{\mathbb{E}[A]} < 1$, which can be proved via the upper bound system given in Lemma 1. The exact stability condition for the c.o.c. variant of the fork-join$(n_\mathrm{F}, n_\mathrm{J})$ model, and also redundancy-d scheduling, with the FCFS discipline in such a general setting is still unknown.

Theorem 2. *If $\rho_\mathrm{U} < 1$ and the residual job size $B^{\mathrm{res}}_{(n_\mathrm{J})}$ is subexponential, then for the c.o.c. variant of the fork-join$(n_\mathrm{F}, n_\mathrm{J})$ model scheduling with the FCFS discipline:*

$$\frac{\rho_\mathrm{L}}{1-\rho_\mathrm{L}} \bar{B}^{\mathrm{res}}_{(n_\mathrm{J})}(x) \leq \mathbb{P}(W_{(n_\mathrm{J})} > x) \leq \frac{\rho_\mathrm{U}}{1-\rho_\mathrm{U}} \bar{B}^{\mathrm{res}}_{(n_\mathrm{J})}(x) \quad \text{as } x \to \infty,$$

where $\rho_\mathrm{L} = \frac{\mathbb{E}[B_{(n_\mathrm{J})}]}{K\mathbb{E}[A]}$ with $K = \binom{N}{d}\frac{n_\mathrm{F}!}{(n_\mathrm{F}-n_\mathrm{J}+1)!}$ and $\rho_\mathrm{U} = \frac{\mathbb{E}[B_{(n_\mathrm{J})}]}{\mathbb{E}[A]}$.

Proof. Upper bound: By Lemma 1 the waiting time of a job is bounded from above by the waiting time W_U in a $GI/G/1/FCFS$ queue with interarrival time A and job size $B_{(n_\mathrm{J})}$. Thus, by the subexponentiality of $B^{\mathrm{res}}_{(n_\mathrm{J})}$, we can apply known results for the single-server queue, see (1), and obtain

$$\mathbb{P}(W_\mathrm{U} > x) \sim \frac{\rho_\mathrm{U}}{1-\rho_\mathrm{U}} \bar{B}^{\mathrm{res}}_{(n_\mathrm{J})}(x) \quad \text{as } x \to \infty. \tag{12}$$

Lower bound: By Lemma 2 the waiting time of a job is bounded from below by the waiting time W_L in a $GI/G/1/FCFS$ queue with a random selection of the arrivals based on Bernoulli experiments with probability $1/K$, i.e., mean interarrival time $K\mathbb{E}[A]$, and job size $B_{(n_\mathrm{J})}$. Again, by the subexponentiality of $B^{\mathrm{res}}_{(n_\mathrm{J})}$, by applying known results for the single-server queue we obtain

$$\mathbb{P}(W_\mathrm{L} > x) \sim \frac{\rho_\mathrm{L}}{1-\rho_\mathrm{L}} \bar{B}^{\mathrm{res}}_{(n_\mathrm{J})}(x) \quad \text{as } x \to \infty. \tag{13}$$

By combining (12) and (13) we get the desired statement. □

The next corollary provides insight in the tail behavior when the distribution of the n_{J}th order statistic of the job size is regularly varying, i.e., $B_{(n_{\mathrm{J}})} \in RV(-\tilde{\nu})$. Observe that, in the special case of identical replicas $B_{(n_{\mathrm{J}})} \in RV(-\nu)$ when $B \in RV(-\nu)$, thus in this case $\tilde{\nu} = \nu$, whereas for i.i.d. replicas $B_{(n_{\mathrm{J}})} \in RV(-(n_{\mathrm{F}} + 1 - n_{\mathrm{J}})\nu)$ when $B \in RV(-\nu)$ (see [20]), thus in this case $\tilde{\nu} = (n_{\mathrm{F}} + 1 - n_{\mathrm{J}})\nu$.

Corollary 3. *For the c.o.c. variant of the fork-join($n_{\mathrm{F}}, n_{\mathrm{J}}$) model with the FCFS discipline and $\rho_{\mathrm{U}} < 1$:*

i) if $B_{(n_{\mathrm{J}})} \in RV(-\tilde{\nu})$, then $W_{(n_{\mathrm{J}})} \in ORV(1-\tilde{\nu})$,

ii) if $B_{(n_{\mathrm{J}})} \in RV(-\tilde{\nu})$, then $R \in ORV(1-\tilde{\nu})$.

Proof. For regularly varying residual job sizes we know that

$$\mathbb{P}(B^{\mathrm{res}}_{(n_{\mathrm{J}})} > x) \sim \frac{1}{(\tilde{\nu}-1)\mathbb{E}[B_{(n_{\mathrm{J}})}]} L(x) x^{1-\tilde{\nu}} \quad \text{as } x \to \infty,$$

see (3). The proof of i) follows by Theorem 2 and Lemma 6. For the response time we can again use Lemmas 1 and 2 as in Theorem 2. Using the known result for the tail behavior in the single-server queue, see (2), together with Lemma 4 we obtain that

$$\mathbb{P}(R > x) \geq \mathbb{P}(R_{\mathrm{L}} > x) = \mathbb{P}(W_{\mathrm{L}} + B_{(n_{\mathrm{J}})} > x) \sim \frac{\rho_{\mathrm{L}}}{1-\rho_{\mathrm{L}}} \frac{L(x)x^{1-\tilde{\nu}}}{(\tilde{\nu}-1)\mathbb{E}[B_{(n_{\mathrm{J}})}]} \quad \text{as } x \to \infty,$$

and

$$\mathbb{P}(R > x) \leq \mathbb{P}(R_{\mathrm{U}} > x) = \mathbb{P}(W_{\mathrm{U}} + B_{(n_{\mathrm{J}})} > x) \sim \frac{\rho_{\mathrm{U}}}{1-\rho_{\mathrm{U}}} \frac{L(x)x^{1-\tilde{\nu}}}{(\tilde{\nu}-1)\mathbb{E}[B_{(n_{\mathrm{J}})}]} \quad \text{as } x \to \infty.$$

Now we can apply Lemma 6 in Appendix A and obtain the desired result. □

Remark 2. *For identical replicas we can even find a better upper bound in Theorem 2. Indeed, consider the system in which all replicas are completely served. This system is equivalent to a $GI/G/1/FCFS$ queue with a random selection of the arrivals based on Bernoulli experiments with probability n_{F}/N, i.e., mean interarrival time $\frac{N\mathbb{E}[A]}{n_{\mathrm{F}}}$, and job size B, which is equal to $B_{(n_{\mathrm{J}})}$ in the case of identical replicas.*

Observe that all the results for the tail index rely on the fact that the upper bound system is stable. The stability condition of this system does not necessarily coincide with the stability condition of the original fork-join model. We conjecture that these tail index results are valid whenever the original fork-join model is stable. However, note that constructing a tractable upper bound system with the same stability condition as the original fork-join model is hard, because this stability condition is unknown.

Interestingly, in contrast to the c.o.s. variant of redundancy, we observe that the tail index in the c.o.c. variant of the fork-join model does not depend on the load of the system. The main difference between the two variants is that for the c.o.s. variant we need multiple big jobs for a large value of $W_{\min}$ to occur, whereas for the c.o.c. variant we only need one big job. Moreover, note that a big job means that at least $n_{\mathrm{F}} + 1 - n_{\mathrm{J}}$ replica sizes should be big since we cancel the redundant replicas as soon as the first n_{J} replicas complete service. This is the reason why for i.i.d. replicas we get the tail index $1 - (n_{\mathrm{F}} + 1 - n_{\mathrm{J}})\nu$ and for identical replicas $1 - \nu$.

In the remainder of this subsection we focus on two special cases of the dependency structure, namely identical and i.i.d. replicas.

For the special case of identical replicas in the c.o.c. variant of the fork-join model with the FCFS discipline we have concluded: if $B \in RV(-\nu)$, then $R \in ORV(1 - \nu)$ which is independent of the number of replicas. We may conclude that the tail index is the same as for the single-server queue, see (2). Moreover, if we compare the tail index of the c.o.s. and c.o.c. variants of redundancy scheduling with identical replicas it follows that the c.o.s. variant always performs better from a tail perspective.

For the special case of i.i.d. replicas in the c.o.c. variant of the fork-join model with the FCFS discipline we have concluded: if $B \in RV(-\nu)$, then $R \in ORV(1 - (n_{\mathrm{F}} + 1 - n_{\mathrm{J}})\nu)$. If $n_{\mathrm{F}} = n_{\mathrm{J}} = 1$, then $R \in ORV(1 - \nu)$ which is consistent with the case of identical replicas. Moreover, if we compare the tail index of the c.o.s. and c.o.c. variants of redundancy scheduling with i.i.d. replicas it follows that the c.o.c. variant always performs better from a tail perspective.

Observe that this statement is in contrast with the case for identical replicas.

We studied two special structures for the dependency between replicas. The general case, with a vector $(B_1, \dots, B_{n_{\mathrm{F}}})$ of possibly dependent and multivariate regularly varying job sizes, will be more involved. For further information on multivariate regular variation we refer to [27] or [4, Appendix A1.5] and the references therein.

We determined the tail behavior for the c.o.s. variant of redundancy scheduling and the c.o.c. variant of the more general fork-join model. It can be concluded that the analysis of the c.o.s. variant is much more challenging than of the c.o.c. variant. One of the reasons is that for the c.o.s. variant multiple big jobs might be needed to have a large waiting time while for the c.o.c. variant only one big job is needed. In some sense this is remarkable, since for the stability condition it is the other way around: The stability condition for the c.o.s. variant of redundancy scheduling is known, whereas for the c.o.c. variant of the fork-join model, and also redundancy-d scheduling, it is still an open problem for non-exponential job size distributions.

4. LCFS-PR discipline

In this section we study the tail behavior of the response time in the fork-join model with the LCFS-PR discipline. First, we discuss known results for the single-server queue and in Sections 4.1 and 4.2 the tail behavior for the c.o.s. and c.o.c. variants of the fork-join model is discussed, respectively.

For the $GI/G/1$ queue with regularly varying job sizes the tail behavior of the response time distribution is known

$$\mathbb{P}(R_{\text{LCFS-PR}} > x) \sim \mathbb{E}[N_{\text{bp}}](1-\tilde{\rho})^{-\nu} L(x) x^{-\nu} \quad \text{as } x \to \infty, \tag{14}$$

where N_{bp} denotes the number of jobs completed during a busy period, see [32]. One way to understand (14) is the following. First observe that for the LCFS-PR discipline

$$R_{\text{LCFS-PR}} \stackrel{d}{=} P,$$

where P is the busy period of a $GI/G/1$ queue. Let $V(t)$ be the amount of work in the system at time t and assume that the first job arrives in an empty system at time 0. The busy period P is then defined as

$$P := \inf\{t > 0 : V(t) = 0\}.$$

Let the cycle maximum $C_{\max}$ be defined by

$$C_{\max} := \sup\{V(t), 0 \leq t \leq P\}.$$

It is shown, see for example [18, Corollary 2.2], that subexponentiality of B implies that $\mathbb{P}(C_{\max} > x) \sim \mathbb{P}(W_{\max} > x)$, where $W_{\max}$ is the maximum waiting time during a busy period, and from [1] we know that,

$$\mathbb{P}(W_{\max} > x) \sim \mathbb{E}[N_{\mathrm{bp}}]\mathbb{P}(B > x) \quad \text{as } x \to \infty.$$

Combining both relations gives

$$\mathbb{P}(C_{\max} > x) \sim \mathbb{E}[N_{\mathrm{bp}}]\mathbb{P}(B > x) \quad \text{as } x \to \infty.$$

A large maximum waiting time is most likely due to one large job. After this large job, the system behaves normally and the workload goes to zero with negative drift $-(1-\tilde{\rho})$. Hence if $C_{\max}$ is large, then one would expect that

$$P \approx \frac{C_{\max}}{1-\tilde{\rho}},$$

from which it follows that

$$\mathbb{P}(P > x) \sim \mathbb{E}[N_{\mathrm{bp}}](1-\tilde{\rho})^{-\nu}L(x)x^{-\nu} \quad \text{as } x \to \infty.$$

Observing that the busy period coincides with the response time of a job for the LCFS-PR discipline gives the desired result in (14).

4.1. Cancel-on-start

Note that for the LCFS-PR discipline the c.o.s. variant of the fork-join$(n_{\mathrm{F}}, n_{\mathrm{J}})$ model is equivalent to the system where replicas of each job are assigned to n_{J}

servers chosen uniformly at random (without replacement), since all replicas immediately go into service. Thus, each queue is equivalent with a $GI/G/1/LCFS\text{-}PR$ queue with a random selection of the arrivals based on Bernoulli experiments with probability n_{J}/N, i.e., mean interarrival time $N\mathbb{E}[A]/n_{\mathrm{J}}$ and mean job size $\mathbb{E}[B]$. Hence the stability condition is $\tilde{\rho} < \frac{N}{n_{\mathrm{J}}}$. For regularly varying job sizes the tail behavior of the response time is given by

$$\mathbb{P}(R > x) = \mathbb{P}(\max_{i=1,\dots,n_{\mathrm{J}}} R_{\mathrm{LCFS-PR}} > x) \sim n_{\mathrm{J}}\mathbb{P}(R_{\mathrm{LCFS-PR}} > x) \quad \text{as } x \to \infty,$$

see for example [21].

Observe that a similar reasoning is applicable for any service discipline in which all replicas immediately go into service. Another example is the processor-sharing (PS) discipline for which the tail behavior of the response time for the single-server queue with regularly varying job sizes with index $-\nu$ is given by

$$\mathbb{P}(R_{\mathrm{PS}} > x) \sim (1-\tilde{\rho})^{-\nu} L(x) x^{-\nu} \quad \text{as } x \to \infty,$$

see for example [31, Chapter 3] or [33].

4.2. Cancel-on-completion

In this section we analyze the tail asymptotics for the c.o.c. variant of the fork-join($n_{\mathrm{F}}, n_{\mathrm{J}}$) model with the LCFS-PR discipline. Similarly to the FCFS discipline in Section 3.2 we first state a lemma that provides an upper bound for the response time.

Lemma 3. *The response time in the fork-join($n_{\mathrm{F}}, n_{\mathrm{J}}$) model with the LCFS-PR discipline is stochastically bounded from above by the response time in a $GI/G/1/LCFS\text{-}PR$ queue with interarrival time A and job size $B_{(n_{\mathrm{J}})}$.*

Proof. Consider an auxiliary system in which all jobs are assigned to the same n_{F} servers and where servers wait with serving a new job until all the n_{J} replicas are finished. This system is equivalent to the $GI/G/1/LCFS\text{-}PR$ queue as defined in the lemma. Let ω_i be the workload at server i, where we define the workload as the amount of work a server needs to complete to become idle in the absence of any arrivals. It can be shown, by taking similar steps as in Lemma 1, that ω_i is bounded from above by the workload ω_{U} in the auxiliary $GI/G/1/LCFS\text{-}PR$ system at all times. Observe that the bound for the workload implies $R \leq R_{\text{U}}$. □

A sufficient stability condition for general interarrival times and job sizes is $\rho_{\text{U}} = \frac{\mathbb{E}[B_{(n_{\text{J}})}]}{\mathbb{E}[A]} < 1$, which can be proved via the upper bound system given in Lemma 3. The exact stability condition of the c.o.c. variant of the fork-join($n_{\text{F}}, n_{\text{J}}$) model, and also redundancy-d scheduling, with the LCFS-PR discipline in such a general setting is still unknown.

The next theorem provides insight in the tail behavior when the distribution of the n_{J}th order statistic of the job size is regularly varying, i.e., $B_{(n_{\text{J}})} \in RV(-\tilde{\nu})$. Similarly to the FCFS discipline it includes the special cases of identical replicas ($\tilde{\nu} = \nu$) and i.i.d. replicas ($\tilde{\nu} = (n_{\text{F}} + 1 - n_{\text{J}})\nu$).

Theorem 3. *For the c.o.c. variant of the fork-join($n_{\text{F}}, n_{\text{J}}$) model with the LCFS-PR discipline and $\rho_{\text{U}} < 1$: if $B_{(n_{\text{J}})} \in RV(-\tilde{\nu})$, then $R \in ORV(-\tilde{\nu})$.*

Proof. Upper bound: By Lemma 3 the response time is bounded from above by the response time in a $GI/G/1/LCFS\text{-}PR$ queue with interarrival time A and job size $B_{(n_{\text{J}})}$. Let R_{U} denote the response time in this upper bound system. Since $B_{(n_{\text{J}})}$ is regularly varying, we can apply known results for the single-server queue, see (14), and obtain

$$\mathbb{P}(R_{\text{U}} > x) \sim \mathbb{E}[N_{\text{bp}}](1 - \rho_{\text{U}})^{-\tilde{\nu}} L(x) x^{-\tilde{\nu}} \quad \text{as } x \to \infty,$$

where $\rho_{\text{U}} = \frac{\mathbb{E}[B_{(n_{\text{J}})}]}{\mathbb{E}[A]}$.

Lower bound: One could argue that R cannot have a heavier tail than R_{U}, but also not a lighter tail, since

$$\mathbb{P}(R > x) \geq \mathbb{P}(B_{(n_{\text{J}})} > x) = L(x) x^{-\tilde{\nu}}, \quad x > 0.$$

The proof follows by Lemma 6 in Appendix A. □

Remark 3. *For identical replicas we can even find a better upper bound in Theorem 3. Indeed, consider the system in which all replicas are completely* *served. This system is equivalent to a $GI/G/1/LCFS$-PR queue with a random selection of the arrivals based on Bernoulli experiments with probability n_{F}/N, i.e., mean interarrival time $\frac{N\mathbb{E}[A]}{n_{\mathrm{F}}}$, and job size B, which is equal to $B_{(n_{\mathrm{J}})}$ in the case of identical replicas.*

Theorem 3 indicates that for the LCFS-PR discipline the tail of the response time is just as heavy as the tail of the job size. Comparing the tail behavior in redundancy-d scheduling with the LCFS-PR discipline and with the FCFS discipline we can conclude that, for the c.o.s. variant, the LCFS-PR discipline has better tail behavior than (or equally good as) the FCFS discipline. Loosely speaking, the tail behavior of the LCFS-PR discipline is better in scenarios with small load and a small number of replicas d and the tail behavior of the two service disciplines is similar in all other scenarios. For the c.o.c. variant of the fork-join model the LCFS-PR discipline always has better tail behavior than the FCFS discipline for all dependency structures between the replicas.

5. Numerical results

In the previous sections we determined the tail behavior of the response time for heavy-tailed job sizes. In this section we provide simulation results for redundancy-d scheduling that illustrate this tail behavior in various scenarios. All the simulation experiments are conducted with 10^9 number of jobs. The figures are in log-log scale and we consider Pareto distributed job sizes with shape value $\nu = 1.5$, which means that $B \in RV(-1.5)$. Note that in the simulation $\mathbb{P}(R > x) = 0$ for x big enough, which explains the steep drop in all the figures.

In Figure 1 the tail behavior of the response time for the c.o.s. variant of redundancy is depicted, see Corollary 2 for the corresponding asymptotic bound. It can be seen that especially the lines for $d = 2$ and $d = N = 3$ are following the line representing tail index -0.5 quite well. For $d = 1$ it can be seen that at first it diverges, but after $x > 10$ it also runs parallel to the line representing tail index -0.5.

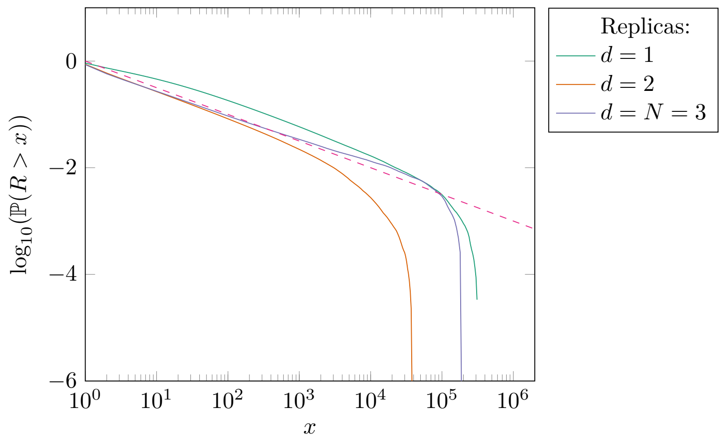

Figure 1: Tail behavior for the response time in the c.o.s. variant of redundancy-d scheduling with Pareto($\nu = 1.5, x_m = 1/3$) job sizes, $\mathbb{E}[B] = 1$, $N = 3$, $\tilde{\rho} = 2.5$ and the FCFS discipline. The dashed line depicts the function $y = x^{-0.5}$.

Figure 2 shows the tail behavior for the response time in the c.o.c. variant of redundancy with identical Pareto job sizes, see Corollary 3 for the asymptotic bound. It can be seen that for every number of replicas the tail index is equivalent to the value identified in Corollary 3. Interestingly, this figure shows that for $d = 2$ the asymptotic lower bound represents the exact tail behavior better than the upper bound.

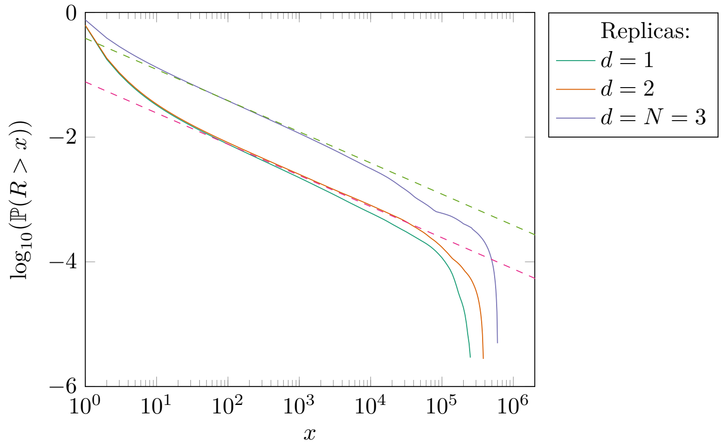

Figure 2: Tail behavior for the response time in the c.o.c. variant of redundancy-d scheduling with identical Pareto($\nu = 1.5, x_m = 1/3$) job sizes, $\mathbb{E}[B] = 1$, $N = 3$, $\tilde{\rho} = 0.5$ and the FCFS discipline. The dashed lines depict the tail behavior for the response time in the lower bound ($\mathbb{P}(R_{\mathrm{L}} > x)$) and in the upper bound ($\mathbb{P}(R_{\mathrm{U}} > x)$) given in Corollary 3. Note that the system with $d = 1$ and $d = 3 = N$ is equivalent to the lower and upper bound system, respectively.

Figure 3 depicts the tail behavior for the response time in the c.o.c. variant of redundancy with i.i.d. Pareto job sizes. Note that according to Corollary 3 the tail index is given by $1 - d\nu$. To get the same tail behavior for all the numbers of replicas in Figure 3 we scaled the job size with d.

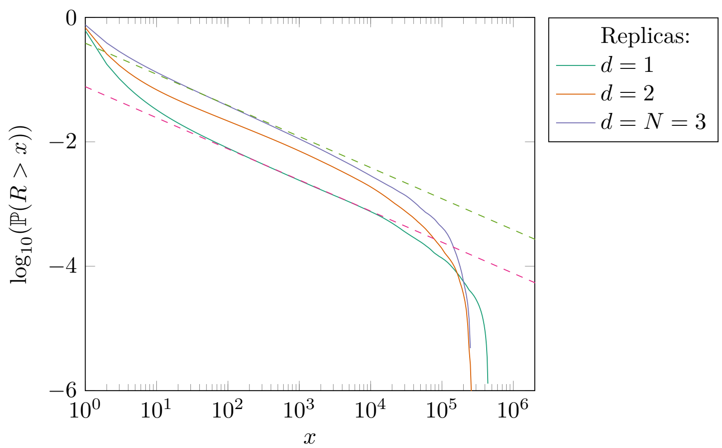

Figure 3: Tail behavior for the c.o.c. variant of redundancy-d scheduling with i.i.d. Pareto($\nu = 1.5/d, x_m = 1/3$) job sizes, $\mathbb{E}[B_{\min}] = 1$, $N = 3$, $\tilde{\rho} = 0.5$ and the FCFS discipline. The dashed lines depict the tail behavior for the response time in the lower bound ($\mathbb{P}(R_{\mathrm{L}} > x)$) and in the upper bound ($\mathbb{P}(R_{\mathrm{U}} > x)$) given in Corollary 3. Note that the system with $d = 1$ and $d = 3 = N$ is equivalent to the lower and upper bound system, respectively.

So far we only considered the FCFS discipline. Figure 4 shows the tail behavior of the response time for the c.o.c. variant of redundancy with the LCFS discipline.

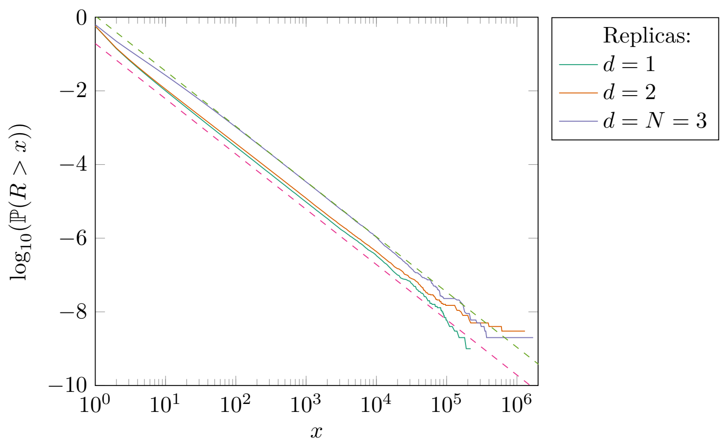

Figure 4: Tail behavior for the response time in the c.o.c. variant of redundancy-d scheduling with identical Pareto($\nu = 1.5, x_m = 1/3$) job sizes, $\mathbb{E}[B] = 1$, $N = 3$, $\tilde{\rho} = 0.5$ and the LCFS discipline. The dashed lines depict the tail behavior for the response time in the lower bound ($\mathbb{P}(R_{\mathrm{L}} > x)$) and in the upper bound ($\mathbb{P}(R_{\mathrm{U}} > x)$) given in Theorem 3. Note that the system with $d = 1$ and $d = 3 = N$ is equivalent to the lower and upper bound system, respectively.

6. Conclusion and suggestions for further research

In this paper we studied the tail behavior of the response time in redundancy-d scheduling and the fork-join model for heavy-tailed job sizes. In particular, for the c.o.s. variant of redundancy-d with the FCFS discipline and subexponential job sizes we determined the tail behavior of the response time and showed that it depends on the load of the system. For the c.o.c. variant of the fork-join model we observed that the tail behavior of the response time depends on the dependency structure between the replicas. For job sizes $B \in RV(-\nu)$, our results indicate that for the c.o.s. variant of redundancy scheduling in the scenario of sufficiently small load having $d = \lceil \frac{\nu}{\nu-1} \rceil$ replicas already achieves the optimal asymptotic tail behavior of the response time. For high loads, the results indicate that creating many replicas yields no benefits for the tail index of the response time. For the c.o.c. variant of the fork-join($n_{\mathrm{F}}, n_{\mathrm{J}}$) model with identical and i.i.d. replicas the tail index of the response time is $1 - \nu$ and $1 - (n_{\mathrm{F}} + 1 - n_{\mathrm{J}})\nu$, respectively. Thus, the tail index is independent of the load of the system and for identical replicas even independent of the number of replicas.

Observe that all the results on the tail index for the c.o.c. variant of the fork-join model rely on the fact that the upper bound system is stable. The stability condition of this system does not necessarily coincide with the stability condition of the original fork-join model. For further research one could study the tail index for these values for the load, i.e., whenever the original fork-join model is stable but the upper bound system is unstable.

A natural topic for further research would be to extend our analysis to heterogeneous servers or even more generally to job types that can have different speeds at the various servers, see for example the model in [24].

Another extension would be to analyze the tail behavior of the response time for the ROS service discipline. As mentioned in the introduction, for the single-server queue this discipline has the same tail index as the FCFS discipline. Simulation experiments (not included in this paper) suggest that this statement extends to redundancy-d scheduling.

Acknowledgments

The work in this paper is supported by the Netherlands Organisation for Scientific Research (NWO) through Gravitation grant NETWORKS 024.002.003. The authors would like to thank Sergey Foss for his helpful suggestions and careful reading.

References

[1] S. Asmussen. Subexponential asymptotics for stochastic processes: Extremal behaviour, stationary distributions and first passage times. *Annals of Applied Probability*, 8:354–374, 1997.

[2] U. Ayesta, T. Bodas, and I.M. Verloop. On redundancy-d with cancel-on-start a.k.a. join-shortest-work(d). *ACM SIGMETRICS Performance Evaluation Review*, 46(2):24–26, 2019.

[3] F. Baccelli, A.M. Makowski, and A. Shwartz. The fork-join queue and related systems with synchronization constraints: Stochastic ordering and computable bounds. *Advances in Applied Probability*, 21(3):629–660, 1989.

[4] N.H. Bingham, C.M. Goldie, and J.L. Teugels. *Regular Variation.* Cambridge University Press, 1987.

[5] J. Blanchet and K.R.A. Murthy. Tail asymptotics for delay in a half-loaded GI/GI/2 queue with heavy-tailed job sizes. *Queueing Systems*, 81:301–340, 2015.

[6] O.J. Boxma, S.G. Foss, J.M. Lasgouttes, and R. Núñez Queija. Waiting time asymptotics in the single server queue with service in random order. *Queueing Systems*, 46:35–73, 2004.

[7] J. Dean and L.Z. Barroso. The tail at scale. *Communications of the ACM*, 56(2):74–80, 2013.

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

[9] J. Dean and S. Ghemawat. MapReduce: A flexible data processing tool. *Communications of the ACM*, 53(1):72–77, 2010.

[10] L. Flatto and S. Hahn. Two parallel queues created by arrivals with two demands. *SIAM Journal on Applied Mathematics*, 44(5):1041–1053, 1984.

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

[12] S. Foss and D. Korshunov. Heavy tails in multi-server queues. *Queueing Systems*, 52(1):31–48, 2006.

[13] S. Foss and D. Korshunov. On large delays in multi-server queues with heavy tails. *Mathematics of Operations Research*, 37(2):201–218, 2012.

[14] S. Foss, D. Korshunov, and S. Zachary. *An Introduction to Heavy-Tailed and Subexponential Distributions.* Springer, New York, 2013.

[15] S. Foss and A. Richards. On sums of conditionally independent subexponential random variables. *Mathematics of Operations Research*, 35(1):102–119, 2010.

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

[17] M. Harchol-Balter. Open problems in queueing theory inspired by datacenter computing. *Queueing Systems*, 97:3–37, 2021.

[18] D. Heath, S. Resnick, and G. Samorodnitsky. Patterns of buffer overflow in a class of queues with long memory in the input stream. *Annals of Applied Probability*, 7:1021–1057, 1997.

[19] T. Hellemans and B. Van Houdt. Performance of redundancy-d with identical/independent replicas. *ACM Transactions on Modeling and Performance Evaluation of Computing Systems*, 4(2):1–28, 2019.

[20] A.H. Jessen and T. Mikosch. Regularly varying functions. *Publications de l'Institut Mathématique*, 80(94):171–192, 2006.

[21] T. Mikosch. Regular variation, subexponentiality and their applications in probability theory. Lecture notes, University of Groningen, 1999. `https://www.eurandom.tue.nl/pre-prints/` 1999-013.

[22] R.D. Nelson and A.N. Tantawi. Approximate analysis of fork/join synchronization in parallel queues. *IEEE Transactions on Computers*, 37(6):739–743, 1988.

[23] K. Park and W. Willinger. *Self-Similar Network Traffic and Performance Evaluation.* John Wiley and Sons, Inc., New York, 2000.

[24] Y. Raaijmakers, S.C. Borst, and O.J. Boxma. Delta probing policies for redundancy. *Performance Evaluation*, 127-128:21–35, 2018.

[25] Y. Raaijmakers, S.C. Borst, and O.J. Boxma. Redundancy scheduling with scaled Bernoulli service requirements. *Queueing Systems*, 93(1-2):67–82, 2019.

[26] Y. Raaijmakers, S.C. Borst, and O.J. Boxma. Stability and tail behavior of redundancy systems with processor sharing. *Performance Evaluation, to appear*, 147:1–40, 2021.

[27] S. Resnick. Heavy tailed analysis. Lecture notes, Cornell University, 2005. `https://www.eurandom.tue.nl/pre-prints/` 2005-024.

[28] A. Thomasian. Analysis of fork/join and related queueing systems. *ACM Computing Surveys*, 47(2):1–71, 2014.

[29] N. Veraverbeke. Asymptotic behaviour of Wiener-Hopf factors of a random walk. *Stochastic Processes and their Applications*, 5(1):27–37, 1977.

[30] A. Vulimiria, P.B. Godfrey, R. Mittal, J. Sherry, S. Ratnasamy, and S. Shenker. Low latency via redundancy. *CoNEXT '13: Proceedings of the ninth ACM Conference on Emerging Networking Experiments and Technologies*, pages 283–294, 2013.

[31] A.P. Zwart. *Queueing Systems with Heavy Tails.* PhD thesis, Eindhoven University of Technology, 2001. `https://research.tue.nl/nl/publications/queueing-systems-with-heavy-tails`.

[32] A.P. Zwart. Tail asymptotics for the busy period in the $GI/G/1$ queue. *Mathematics of Operations Research*, 26(3):485–493, 2001.

[33] A.P. Zwart and O.J. Boxma. Sojourn time asymptotics in the $M/G/1$ processor sharing queue. *Queueing Systems*, 35:141–166, 2000.

Appendix A. Preliminary results

In this appendix we introduce several classes of heavy-tailed distributions that are considered for the job size in this paper, see also [4, 14]. Let the complementary cumulative distribution function be defined as $\bar{F}_B(x) := 1 - F_B(x) = \mathbb{P}(B > x)$.

Definition 1. *B is heavy-tailed if, for all $\epsilon > 0$,*

$$\mathbb{E}[e^{\epsilon B}] = \infty,$$

or equivalently (see for example [14, Theorem 2.6]), if for all $\epsilon > 0$,

$$\mathbb{P}(B > x)e^{\epsilon x} \to \infty \text{ as } x \to \infty$$

Let $F_B^{n*}(x)$ be the n-fold convolution of $F_B(x)$ for $n = 2, 3, \ldots$, with $F_B^{1*}(x) \equiv F_B(x)$.

Definition 2. *B is subexponential, denoted by $B \in \mathcal{S}$, if*

$$\frac{\bar{F}_B^{2*}(x)}{\bar{F}_B(x)} = \frac{\mathbb{P}(B_1 + B_2 > x)}{\mathbb{P}(B > x)} \to 2 \quad \text{as } x \to \infty.$$

Examples of well-known subexponential distributions are Pareto, Lognormal and Weibull with a shape parameter between 0 and 1.

Definition 3. *B is long-tailed, denoted by $B \in \mathcal{L}$, if $\bar{F}_B(x+1) \sim \bar{F}_B(x)$ as $x \to \infty$.*

Here $f(x) \sim g(x)$ means $\frac{f(x)}{g(x)} \to 1$ as $x \to \infty$.

Definition 4. *B is dominated varying, denoted by $B \in \mathcal{D}$, if $\bar{F}_B(2x) \geq c\bar{F}_B(x)$ for some $c > 0$ and for all x.*

Definition 5. *B is $\mathcal{O}$-regularly varying, denoted by $B \in ORV$, if*

$$0 < \liminf_{x\to\infty} \frac{\bar{F}_B(\alpha x)}{\bar{F}_B(x)} \leq \limsup_{x\to\infty} \frac{\bar{F}_B(\alpha x)}{\bar{F}_B(x)} < \infty, \quad \forall \alpha \geq 1.$$

Furthermore, $B \in ORV(-\nu)$ if

$$c_1\alpha^{-\nu} < \liminf_{x\to\infty} \frac{\bar{F}_B(\alpha x)}{\bar{F}_B(x)} \leq \limsup_{x\to\infty} \frac{\bar{F}_B(\alpha x)}{\bar{F}_B(x)} < c_2\alpha^{-\nu}, \quad \forall \alpha \geq 1, \tag{A.1}$$

with positive constants c_1 and c_2.

Definition 6. *B is regularly varying of index $-\nu$, denoted by $B \in RV(-\nu)$, if*

$$\bar{F}_B(x) = L(x)x^{-\nu}, \quad x > 0, \tag{A.2}$$

with $L(x)$ a slowly varying function, i.e., $L(\alpha x)/L(x) \to 1$ for any $\alpha > 0$.

Observe that we have the following relations, see for example [4, Theorem 2.1.8] or [31, Chapter 2],

$$RV \subset \mathcal{D} \subset ORV \quad \text{and} \quad RV \subset \mathcal{L} \cap \mathcal{D} \subset \mathcal{S}. \tag{A.3}$$

We will analyze the tail asymptotics of the response time of an arbitrary job in steady state. For this we need some preliminary results that are stated in the lemmas below. The next lemma states that the minimum and the sum of two independent regularly varying random variables is again regularly varying.

Lemma 4. *Let X and Y be two independent regularly varying random variables with $\mathbb{P}(X > x) = L_1(x)x^{-\nu_1}$ and $\mathbb{P}(Y > x) = L_2(x)x^{-\nu_2}$. Then,*

i) $\min(X, Y) \in RV(-(\nu_1 + \nu_2))$,

ii) $X + Y \in RV(-\min\{\nu_1, \nu_2\})$.

Proof. See [4, Proposition 1.5.7]. □

A similar lemma as Lemma 4 can be proved for $\mathcal{O}$-regularly varying random variables.

Lemma 5. *Let X and Y be two independent $\mathcal{O}$-regularly varying random variables with index $-\nu_1$ and $-\nu_2$, respectively, then*

i) $\min(X, Y) \in ORV(-(\nu_1 + \nu_2))$,

ii) $X + Y \in ORV(-\min\{\nu_1, \nu_2\})$.

Proof. By definition of $ORV(-\nu)$ there exist c_i, $i = 1, \ldots, 4$, such that

$$c_1\alpha^{-\nu_1} < \liminf_{x\to\infty} \frac{\bar{F}_X(\alpha x)}{\bar{F}_X(x)} \leq \limsup_{x\to\infty} \frac{\bar{F}_X(\alpha x)}{\bar{F}_X(x)} < c_2\alpha^{-\nu_1}, \quad \forall \alpha \geq 1,$$

and

$$c_3\alpha^{-\nu_2} < \liminf_{x\to\infty} \frac{\bar{F}_Y(\alpha x)}{\bar{F}_Y(x)} \leq \limsup_{x\to\infty} \frac{\bar{F}_Y(\alpha x)}{\bar{F}_Y(x)} < c_4\alpha^{-\nu_2}, \quad \forall \alpha \geq 1.$$

Observe that by independence we have

$$\mathbb{P}(\min(X,Y) > x) = \mathbb{P}(X > x)\mathbb{P}(Y > x),$$

and therefore

$$c_1c_3\alpha^{-(\nu_1+\nu_2)} < \liminf_{x\to\infty} \frac{\bar{F}_{\min(X,Y)}(\alpha x)}{\bar{F}_{\min(X,Y)}(x)} \leq \limsup_{x\to\infty} \frac{\bar{F}_{\min(X,Y)}(\alpha x)}{\bar{F}_{\min(X,Y)}(x)} < c_2c_4\alpha^{-(\nu_1+\nu_2)}, \quad \forall \alpha \geq 1.$$

The proof of i) follows by the definition of $ORV(-\nu)$.

The proof of ii) is similar to the proof of the convolution closure for regularly varying distributions, see for example [21]. Since $\{X+Y > x\} \supset \{X > x\} \cup \{Y > x\}$ it follows that

$$\mathbb{P}(X+Y > x) \geq \mathbb{P}(X > x) + \mathbb{P}(Y > x) - \mathbb{P}(X > x)\mathbb{P}(Y > x).$$

For $0 < \delta < \frac{1}{2}$, we have that

$$\{X+Y > x\} \subset \{X > (1-\delta)x\} \cup \{Y > (1-\delta)x\} \cup \{X > \delta x, Y > \delta x\},$$

and therefore

$$\mathbb{P}(X+Y > x) \leq \mathbb{P}(X > (1-\delta)x) + \mathbb{P}(Y > (1-\delta)x) + \mathbb{P}(X > \delta x)\mathbb{P}(Y > \delta x).$$

Now if $\nu_1 = \min\{v_1, v_2\}$,

$$\frac{\mathbb{P}(X+Y > \alpha x)}{\mathbb{P}(X+Y > x)} \geq \frac{c_1}{c_2}(1-\delta)^{\nu_1}\alpha^{-\nu_1}(1+o(1)),$$

and

$$\frac{\mathbb{P}(X+Y > \alpha x)}{\mathbb{P}(X+Y > x)} \leq \frac{c_2}{c_1}(1-\delta)^{-\nu_1}\alpha^{-\nu_1}(1+o(1)).$$

The case for $\nu_2 = \min\{v_1, v_2\}$ follows by an analogous argument. By definition, see (A.1), we get that $X+Y \in ORV(-\min\{\nu_1, \nu_2\})$. □

Observe that these results could also be obtained by applying the principle of a single big jump, see for example [15].

Next we give an auxiliary lemma which states that a random variable is $\mathcal{O}$-regularly varying with index $-\nu$ under the condition $c_1 L(x)x^{-\nu} \leq \mathbb{P}(X > x) \leq c_2 L(x)x^{-\nu}$.

Lemma 6. *If* $c_1 L(x)x^{-\nu} \leq \mathbb{P}(X > x) \leq c_2 L(x)x^{-\nu}$*, then* $X \in ORV(-\nu)$*.*

Proof. From $c_1 L(x)x^{-\nu} \leq \mathbb{P}(X > x) \leq c_2 L(x)x^{-\nu}$ we get

$$\frac{c_1}{c_2}\alpha^{-\nu} < \liminf_{x\to\infty} \frac{\bar{F}_X(\alpha x)}{\bar{F}_X(x)} \leq \limsup_{x\to\infty} \frac{\bar{F}_X(\alpha x)}{\bar{F}_X(x)} < \frac{c_2}{c_1}\alpha^{-\nu}, \quad \forall \alpha \geq 1.$$

The proof follows by definition of $ORV(-\nu)$. □

Appendix B. Proof of the upper bounds in Theorem 1

In this appendix we will prove the upper bounds (8) and (10) on the tail of the waiting time for the c.o.s. variant of redundancy-d with the FCFS discipline. Our proof is based on the proof in [13] for the $GI/G/N$ queue, which corresponds to a system of N queues with the JSW-N policy. While the JSW-d policy with $1 \leq d \leq N$ requires essential and sometimes subtle adaptations, overall we follow the main line of reasoning of [13] and indicate for each lemma and theorem its counterpart in [13].

In most heavy-tail results in queueing theory, the interarrival time distribution does not have an effect on the waiting time tail behavior. With this in mind, similar to [12, 13], we first consider deterministic interarrival times, making the derivations more tractable, and thereafter prove Lemma 10 which allows us to extend the proof for deterministic interarrival times to generally distributed interarrival times. The idea behind the proof of the upper bounds is that we compare the system with the JSW-d policy to N auxiliary single-server queueing systems which work in parallel.

Let $\boldsymbol{s}^n$ denote the vector of d servers which are sampled for the nth job. For $n = 1, 2, \ldots$, let $\boldsymbol{V}_n = (V_{n1}, \ldots, V_{nN})$ be the vector of residual workloads at the arrival epoch of the nth job. The waiting time that the nth job experiences is $W_{\min,n} := \min\{V_{nj}, j \in \boldsymbol{s}^n\}$ and it joins server i_n, where $i_n = \min\{i \in \boldsymbol{s}^n : V_{ni} = W_{\min,n}\}$. Also,

$$V_{n+1,i} = \begin{cases} (V_{ni} + b_n - a_{n+1})^+ & \text{if } i = i_n, \\ (V_{ni} - a_{n+1})^+ & \text{if } i \neq i_n, \end{cases}$$

with job sizes b_n and interarrival times a_n Let $R(\boldsymbol{x}) = (R_1(\boldsymbol{x}), \ldots, R_N(\boldsymbol{x}))$ be the operator on $\mathbb{R}^N$ that orders the coordinates of $\boldsymbol{x} \in \mathbb{R}^N$ in nondecreasing order, i.e., $R_1(\boldsymbol{x}) \leq \cdots \leq R_N(\boldsymbol{x})$. Moreover, let $f_R : \mathbb{R} \to \mathbb{R}$ be the function that maps the server number to the number ordered by workload as in the operator $R(\cdot)$. For $n = 1, 2, \ldots$, put $\boldsymbol{D}_n = R(\boldsymbol{V}_n)$. Then $W_{\min,n} = D_{ni}$, where $i = f_R(i_n)$ and similar to the Kiefer-Wolfowitz recursion for the JSW policy, we get

$$\boldsymbol{D}_{n+1} = R((D_{n1} - a_{n+1})^+, \ldots, (D_{ni} + b_n - a_{n+1})^+, \ldots, (D_{nN} - a_{n+1})^+). \tag{B.1}$$

Observe that the operator $R(\cdot)$ is monotone, thus the sequence $\boldsymbol{D}_n$ satisfying Equation (B.1) satisfies the two monotonicity properties of Lemma 4.1 in [13] as well.

Hereafter we continue to assume deterministic interarrival times $a' \equiv \mathbb{E}[A]$ for both the original system and the auxiliary systems introduced below. Consider N auxiliary $D/G/1$ queueing systems which work in parallel. At every time instant T_n, $n = 1, 2, \ldots$, a batch of N jobs arrives, one job per server. Denote by U_{ni}, $i = 1, \ldots, N$, the waiting times in the ith $D/G/1$ queue and let b_{ni}, $n \geq 1$ and $i = 1, \ldots, N$, be independent random variables with common distribution B. We couple the job sizes of the $D/G/N$ redundancy-d system with job sizes at the N auxiliary $D/G/1$ queues: we let $b_n = b_{n,i_n}$, where $i_n = \min\{i \in \boldsymbol{s}^n : V_{ni} = W_{\min,n}\}$ as defined earlier. The deterministic interarrival times are $T_n = n(k+1)(a' - h)$, with

$$\frac{k}{k+1}\left(a' - \frac{\mathbb{E}[B]}{k+1}\right) < h < a' - \frac{\mathbb{E}[B]}{k+1}, \tag{B.2}$$

so that the auxiliary queueing systems are stable.

In Lemma 8 we upper bound the sum of waiting times by the sum of waiting times in the auxiliary $D/G/1$ queues and a light-tailed random variable. The proof of Lemma 8 uses the auxiliary Lemma 7 that provides an upper bound on the expected difference of the total workload at all the servers seen by the first and $(s+1)$th job when the workload at one of the servers is large. Note that the choice of this large workload is different from that used in [13].

Lemma 7 (Counterpart of Lemma 4.3 in [13])**.** *Consider a system with $k+1$ servers and assume $\mathbb{E}[B] > ka'$. For any $\epsilon > 0$, there exist $V_{\text{large}} < \infty$ and an integer $s \geq 1$ such that, for any initial value $\boldsymbol{D}_1$ with $D_{1,k+1} \geq V_{\text{large}}$,*

$$\mathbb{E}\left[\sum_{j=1}^{k+1} D_{1+s,j} - \sum_{j=1}^{k+1} D_{1j}\right] \leq s(\mathbb{E}[B] - (k+1)a' + \epsilon).$$

Proof. By property (2) of Lemma 4.1 in [13], it is enough to prove the result for initial values $D_{11} = \cdots = D_{1k} = 0$, $D_{1,k+1} = V_{\text{large}}$ only. Choose C such that $\mathbb{E}[\min\{a', C\}] \geq a' - \epsilon/2$. By property (1) of Lemma 4.1 in [13], we may prove the lemma with interarrival times $\min\{a', C\}$ instead of a'.

For $d \geq k+1$ the proof follows from Lemma 4.3 in [13], since the JSW-d and JSW policy are equivalent in the system with $k+1$ servers. For $d < k+1$, consider an auxiliary unstable $GI/G/k$ system with initial value $\hat{\boldsymbol{D}}_1 = 0$ and find s such that $\mathbb{E}\left[\sum_{i=1}^{k} \hat{D}_{1+s,i}\right] \leq s(\mathbb{E}[B] - ka' + \epsilon/2)$. Note that this system samples $d-1$ servers with probability $\frac{d}{k+1}$, i.e., the probability that server $k+1$ is sampled in the original system, and d servers with probability $1 - \frac{d}{k+1}$, i.e., the probability that server $k+1$ is not sampled in the original system. For an unstable system with workload vector $\hat{\boldsymbol{D}}_n$ we have that $\hat{D}_{ni} \to \infty$ as $n \to \infty$ for $i = 1, \ldots, k$.

Take $V_{\text{large}} = \max\{(s+1)C, V^*_{\text{large}}\}$, where V^*_{large} is defined as follows. Consider the system with initial values $D_{11} = \cdots = D_{1k} = 0$, $D_{1,k+1} = V^*_{\text{large}}$ and let the n^*th job be the first job that is assigned to the $(k+1)$th server, i.e., $n^* := \min\{n \geq 1 : i_n = k+1\}$ which clearly depends on the initial workload

V^*_{large}. Then take V^*_{large} such that

$$\min_{i=1,\ldots,k+1} D_{n^*i} \geq (s+1-n^*)\, C.$$

Note that such V^*_{large} exists, since increasing V^*_{large} leads, loosely speaking, to increasing workloads at the other k servers as well (because they are unstable). This definition of V^*_{large} ensures that the first time a job is allocated to server $k+1$ the workload at the other servers is large enough so that, without any additional work, these servers are not empty before the $(s+1)$th job. We cannot simply take $V_{\text{large}} = (s+1)C$ as in [13], because this does not guarantee that $D_{ni} > 0$ and $\hat{D}_{ni} > 0$, for all $n \in [\min\{s+1, n^*\}, s+1]$ and $i = 1, \ldots, k$, which is needed in the proof. Indeed, without additional constraints on V_{large} it may be that the job is allocated to the $(k+1)$th server, which has the smallest workload out of the d sampled servers, while at least one of the other $k+1-d$ servers is empty.

By the exact same steps as in Lemma 4.3 in [13] we can prove that

$$\begin{aligned}\sum_{j=1}^{k+1} D_{1+s,j} - \sum_{j=1}^{k+1} D_{1,j} &= \sum_{j=1}^{k} \hat{D}_{1+s,j} - \sum_{j=1}^{s} \min\{a', C\} \\ &\leq s(\mathbb{E}[B] - ka' + \frac{\epsilon}{2}) - s(a' - \frac{\epsilon}{2}) \quad \text{a.s.,}\end{aligned} \tag{B.3}$$

and the result follows. □

Lemma 8 (Counterpart of Lemma 6.2 in [13]). *There exists $\beta > 0$ such that, for any set of $k+1$ indices $I = \{i(1), \ldots, i(k+1)\}$, there is a random variable η_I such that $\mathbb{E}[e^{\beta \eta_I}] < \infty$ and, for any n, with probability 1,*

$$\sum_{i \in I} V_{ni} \leq \sum_{i \in I} U_{ni} + \eta_I. \tag{B.4}$$

Proof. Fix some $i^* \in I$. Observe that for $d \geq k+1$ the proof directly follows from Lemma 6.2 in [13], since the JSW-d and JSW policy are equivalent in the system with $k+1$ servers. For $d < k+1$, consider an auxiliary $GI/G/(k+1)$ redundancy-d system as in Lemma 7 with workloads $V^*_n = (V^*_{ni}, i \in I)$ with the same interarrival times equal to a', but whose service times b^*_n are chosen in a special manner. At any time n, if $i_n \in I$, then put $b_n^* = b_{n,i_n}$ and $i_n^* = i_n$. If $i_n^* \notin I$, then put $b_n^* = b_{n,i^*}$ and $i_n^* = i^*$. Applying property (1) of Lemma 4.1 in [13], we get that $R(V_{ni}, i \in I) \leq R(\boldsymbol{V}_n^*)$ coordinate-wise, for any n. Therefore,

$$\sum_{i\in I} V_{ni} \leq \sum_{i\in I} V_{ni}^*.$$

By the exact same steps as in Lemma 6.2 in [13] using Lemma 7 (the counterpart of Lemma 4.3 in [13]) we can prove Equation (B.4). □

Just like a crucial step in [13], Lemma 8 can be used to upper bound the waiting time in the N-server system by the waiting time in the corresponding system with *deterministic* interarrival times minus a negligible term. Note that the upper bounds are not as sharp as in [13] since, unlike [13], $W_{\min,n} \nleq \frac{1}{k+1}\sum_{i\in I} V_{ni}$ for every collection I.

Lemma 9 (Counterpart of Lemma 6.1 in [13])**.** *There exists a number $\beta > 0$ and a random variable η such that $\mathbb{E}[e^{\beta\eta}] < \infty$ and, for all n, with probability 1*
i) if $k \geq N - d$,

$$W_{\min,n} \leq \frac{k+1}{k+1-N+d} U_{n,(k+1)} + \eta,$$

where $U_{n,(k+1)}$ is the $(k+1)$th order statistic of vector $(U_{n1}, \dots, U_{nN})$,
ii) if $k \leq N - d$,

$$W_{\min,n} \leq (k+1)U_{n,(N-d+1)} + \eta.$$

Proof. i) For $k \geq N - d$ we have for every collection I of $k+1$ coordinates,

$$W_{\min,n} \leq \frac{1}{k+1-N+d} \sum_{i\in I} V_{ni}, \tag{B.5}$$

since $W_{\min,n}$ is no larger than the $(N-d+1)$th smallest value of V_{ni}, $i \in I$. Then it follows from Lemma 8 that

$$W_{\min,n} \leq \frac{1}{k+1-N+d} \sum_{i\in I} U_{ni} + \eta, \tag{B.6}$$

where $\eta := \max_{I:|I|=k+1} \eta_I$. Take I such that $\{U_{ni}, i \in I\}$ are the $k+1$ smallest coordinates of the vector $(U_{n1}, \dots, U_{nN})$. Then $U_{ni} \leq U_{n,(k+1)}$ for every $i \in I$.

ii) For $k \leq N-d$ we take the collection I of $k+1$ coordinates such that $\tilde{i}_n = \arg\min_i\{U_{ni} : i \in \boldsymbol{s}^n\} \in I$. Hence, $I \cap \boldsymbol{s}^n \neq \emptyset$ and again Equations (B.5) and (B.6) hold. Take the remaining coordinates of I such that $\{U_{ni}, i \in I \setminus \tilde{i}_n\}$ are the k smallest coordinates of the vector $(U_{n1}, \ldots, U_{n,\tilde{i}_n-1}, U_{n,\tilde{i}_n+1}, \ldots, U_{nN})$. Then $U_{ni} \leq U_{n,(N-d+1)}$ for every $i \in I$. Indeed, in the worst case $\{U_{ni} : i \in \boldsymbol{s}^n\}$ are the d largest coordinates of the vector $(U_{n1}, \ldots, U_{nN})$, but $\tilde{i}_n$ is defined as the argument that achieves the minimum of the set $\{U_{ni} : i \in \boldsymbol{s}^n\}$ from which it follows that $U_{n\tilde{i}_n} \leq U_{n,(N-d+1)}$. □

Theorem 4 (Analogous to Theorem 7.1 in [13])**.** *Let $\tilde{\rho} < k+1$ for some $k \in \{0, \ldots, N-1\}$. Then for any fixed h satisfying Equation* (B.2) *there exists $\beta > 0$ such that*
i) if $k \geq N-d$,

$$\mathbb{P}(W_{\min} > x+y) \leq \binom{N}{k}\left(\bar{F}_{M_{\mathrm{rw}}}\left(\frac{(k+1-N+d)x}{k+1}\right)\right)^{N-k} + const \cdot e^{-\beta y},$$

for all $x, y > 0$,
ii) if $k \leq N-d$,

$$\mathbb{P}(W_{\min} > x+y) \leq \binom{N}{d}\left(\bar{F}_{M_{\mathrm{rw}}}\left(\frac{x}{k+1}\right)\right)^{d} + const \cdot e^{-\beta y},$$

for all $x, y > 0$, where $F_{M_{\mathrm{rw}}}$ is the cumulative distribution function of the random variable

$$M_{\mathrm{rw}} := \sup_{n\geq 1}\left\{0, \sum_{j=1}^{n}(b_j - (k+1)(a'-h))\right\}.$$

Proof. Similarly to [13, Theorem 7.1], this proof relies on Lemma 9 which upper bounds the waiting time of the nth job in the two cases $k \geq N-d$ and $k \leq N-d$.

Consider deterministic arrival times $T_n = n(k+1)(a'-h)$ where h satisfies (B.2). For $k \geq N-d$, by Lemma 9 i),

$$\mathbb{P}(W_{\min,n} > x+y) \leq \mathbb{P}\left(U_{n,(k+1)} > \frac{(k+1-N+d)x}{k+1}\right) + \mathbb{P}(\eta > y),$$

and taking into account the independence of the auxiliary queueing systems, we obtain

$$\mathbb{P}(W_{\min,n} > x + y) \leq \binom{N}{k} \left(\mathbb{P}\left(U_{n1} > \frac{(k+1-N+d)x}{k+1}\right)\right)^{N-k} + \mathbb{P}(\eta > y).$$

Similarly, for $k \leq N - d$, by Lemma 9 ii),

$$\mathbb{P}(W_{\min,n} > x + y) \leq \mathbb{P}\left(U_{n,(N-d+1)} > \frac{x}{k+1}\right) + \mathbb{P}(\eta > y),$$

and taking into account the independence of the auxiliary queueing systems, we obtain

$$\mathbb{P}(W_{\min,n} > x + y) \leq \binom{N}{d} \left(\mathbb{P}\left(U_{n1} > \frac{x}{k+1}\right)\right)^{d} + \mathbb{P}(\eta > y).$$

The proof is completed by observing (1) that M_{rw} has the distribution of the maximum of a random walk, which is also the distribution of the steady-state waiting time U in any of the auxiliary $D/G/1$ queues, and (2) according to Lemma 9, $\mathbb{P}(\eta > y)$ is exponentially bounded. □

It is well known (cf. [14, Theorem 5.2]) that if the residual job size B^{res} is subexponential, then

$$\bar{F}_{M_{\text{rw}}}(x) \sim \frac{\mathbb{E}[B]}{(k+1)(a'-h) - \mathbb{E}[B]} \bar{B}^{\text{res}}(x) \quad \text{as } x \to \infty.$$

Taking h close to its minimal value, we arrive at the following estimate

$$\bar{F}_{M_{\text{rw}}}(x) \sim \left(\frac{(k+1)\tilde{\rho}}{(k+1) - \tilde{\rho} - \frac{(k+1)^2\epsilon}{a'}}\right) \bar{B}^{\text{res}}(x) \quad \text{as } x \to \infty, \tag{B.7}$$

where $\epsilon > 0$. We now have the ingredients to prove the upper bounds (8) and (10) in Theorem 1. Replacing x by $(1-\delta)x$ and y by δx in Theorem 4 results in

$$\mathbb{P}(W_{\min} > x) \leq \binom{N}{k} \left(\bar{F}_{M_{\text{rw}}}\left(\frac{(k+1-N+d)x(1-\delta)}{k+1}\right)\right)^{N-k} + const \cdot e^{-\beta\delta x},$$

and

$$\mathbb{P}(W_{\min} > x) \leq \binom{N}{d} \left(\bar{F}_{M_{\text{rw}}}\left(\frac{x(1-\delta)}{k+1}\right)\right)^{d} + const \cdot e^{-\beta\delta x},$$

for $k \geq N-d$ and $k \leq N-d$, respectively. Combined with Equation (B.7) this yields the upper bound.

So far we assumed deterministic interarrival times; the following lemma allows us to extend the proof to the case of generally distributed interarrival times. For clarity we highlight the metrics that correspond to the system with deterministic interarrival times by an apostrophe.

Lemma 10 (Counterpart of Lemma 1 in [12]). *If $\mathbb{P}(W'_{\min} > x) \leq \bar{G}(x)$ for some long-tailed distribution G, where $W'_{\min}$ denotes the waiting time in the system with deterministic interarrival times $a' \equiv \mathbb{E}[A]$, then*

$$\limsup_{x\to\infty} \frac{\mathbb{P}(W_{\min} > x)}{\bar{G}(x)} \leq 1.$$

Proof. Denote $\xi_n = a' - a_n$. Put $M_0 = 0$ and, for $n \geq 1$,

$$M_n = \max\{0, \xi_n, \xi_n + \xi_{n-1}, \ldots, \xi_n + \cdots + \xi_1\} = (\xi_n + M_{n-1})^+.$$

Similar to Lemma 1 in [12] we use induction to prove the inequality

$$W_{\min,n} \leq W'_{\min,n} + iM_n \quad \text{a.s.} \tag{B.8}$$

For $n = 1$ we have $0 \leq 0 + iM_1$. Assume the inequality is proved for some n, then

$$\begin{aligned} W_{\min,n+1} &= R(W_{\min,n} + e_{i_n} b_n - ia_{n+1})^+ \\ &\leq R(W'_{\min,n} + iM_n + e_{i_n} b_n - ia_{n+1})^+ \\ &= R(W'_{\min,n} + e_{i_n} b_n - ia' + i(M_n + \xi_{n+1}))^+. \end{aligned}$$

Since $(u+v)^+ \leq u^+ + v^+$,

$$W_{\min,n+1} \leq R(W'_{\min,n} + e_{i_n} b_n - ia')^+ + i(M_n + \xi_{n+1})^+ \equiv W'_{\min,n+1} + iM_{n+1},$$

and the proof of (B.8) is complete. The remainder of the proof follows by the exact same steps as in Lemma 1 in [12]. ☐