EURANDOM PREPRINT SERIES

2019-014

December 2, 2019

Stability of Redundancy Systems with Processor Sharing

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

Stability of Redundancy Systems with Processor Sharing

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

[a] *Department of Mathematics and Computer Science, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands*

Abstract

We investigate the stability condition for redundancy-d systems where each of the servers follows a processor-sharing (PS) discipline. We allow for generally distributed job sizes, with possible dependence among the d replica sizes being governed by an arbitrary joint distribution. We establish that the stability condition is characterized by the expectation of the minimum of d replica sizes being less than the mean interarrival time per server. In the special case of identical replicas, the stability condition is insensitive to the job size distribution given its mean, and the stability condition is inversely proportional to the number of replicas. In the special case of i.i.d. replicas, the stability threshold decreases (increases) in the number of replicas for job size distributions that are NBU (NWU). We also discuss extensions to scenarios with heterogeneous servers.

Keywords: Parallel-server system, redundancy, stability, processor-sharing

1. Introduction

The interest in redundancy systems has strongly grown in recent years, fueled by empirical evidence that redundancy improves performance in applications with many servers such as web page downloads and Google search queries [1, 2]. The most important feature of redundancy is the replication of each incoming job. The replicas are instantaneously allocated to, say, d different servers, chosen uniformly at random (without replacement), and abandoned as soon as either the first of these d replicas starts service ('cancel-on-start' c.o.s.) or the first of these d replicas finishes service ('cancel-on-completion' c.o.c.).

As alluded to above, allocating replicas of the same job to multiple servers has the potential to improve delay performance. Indeed, adding replicas increases the chance for one of the replicas to find a short queue, and the c.o.s. version is in fact equivalent to a Join-the-Smallest-Workload (JSW) like policy. On the other hand, it may also result in potential vulnerabilities and could cause instability in the c.o.c. version since the same job may be in service at multiple servers, potentially wasting service capacity. The above trade-off already suggests that establishing the stability condition is not straightforward. Indeed, despite the numerous studies on redundancy [3, 4, 5, 6, 7, 8, 9], the stability condition is only known in a few (special) cases.

Gardner et al. [6] obtained an analytical expression for the expected latency. They also proved that, in the scenario of c.o.c. redundancy with i.i.d. replicas, exponential job sizes and

*Corresponding author

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

the FCFS scheduling discipline, the stability condition is given by $\rho < 1$, where $\rho := \frac{\lambda \mathbb{E}[X]}{N}$ denotes the load of the system, λ the arrival rate of the jobs, N the number of servers and $\mathbb{E}[X]$ the expected job size. Note that the stability condition is independent of the number of replicas.

Anton et al. [3] proved the stability condition for various scheduling disciplines, both for i.i.d. and identical replicas, in the case of exponential job sizes. They showed that for i.i.d. replicas the stability conditions for FCFS, PS and Random Order of Service (ROS) are the same. For identical replicas the stability condition for PS is given by $\rho < 1/d$, and is thus inversely proportional to the number of replicas. The proofs rely on scaling limits of appropriate lower- and upperbound systems. A summary of the stability condition results, except the trivial cases $d = 1$ and $d = N$, is provided in Table 1.

For general job size distributions, the stability condition remains an open problem. For i.i.d. replicas and a specific non-exponential distribution, the scaled Bernoulli distribution, the stability condition is asymptotically given by $\rho N < K^{d-1}$, where K is the scale of the job size [9]. Observe that the stability condition is independent of the number of servers, but depends on the scale parameter K. For identical replicas there are, to the best of our knowledge, no explicit expressions for the stability condition for any scheduling discipline.

The present paper focuses on the stability condition for the PS discipline and extends the results from [3] to general job size distributions with possible dependence among the replicas. This covers the extreme scenarios of perfect dependence (identical replicas) and no dependence at all (i.i.d. replicas), as previously considered in the literature, as special cases. The PS discipline is highly relevant as an idealization of Round-Robin scheduling policies that are widely implemented in time-shared computer systems for fairness reasons. To prove the stability condition we consider carefully chosen lower- and upperbound systems. The lowerbound system is similar to the lowerbound system proposed in [3]. An alternative novel representation of this lowerbound system allows us to prove a necessary stability condition for general job size distributions, via the use of fluid limits. The upperbound system gives an indication for the sufficient stability condition, again via the use of fluid limits.

The key property of these fluid limits is that all the components of the fluid limit in the lower- and upperbound systems remain equal at all times when starting with the same initial conditions. More importantly, even the fluid limits of the lower- and upperbound systems coincide when starting with the same initial conditions. Hence the stability conditions for the fluid limits of the lower- and upperbound systems coincide. The above two properties reflect that the queue lengths in the original system have the tendency to remain equal when starting from large initial values. This implies that each of the d replicas of a given job will be served at *approximately* the same rate, and hence the total amount of service capacity consumed by an arbitrary job is $d\mathbb{E}[\min\{X_1, \ldots, X_d\}]$, yielding the stability condition $\tilde{\rho} := \frac{d\lambda \mathbb{E}[\min\{X_1,\ldots,X_d\}]}{N} < 1$.

We establish that the stability condition is characterized by the expectation of the minimum of d replica sizes being less than the mean interarrival time per server. Consequently, for identical replicas the stability condition turns out to be insensitive to the job size distribution given its mean, and the stability threshold is inversely proportional to the number of replicas. In [3] it is concluded that in this case PS causes the worst possible reduction among all work-conserving service disciplines for the stability condition, since the stability condition is equal to that in a system where all replicas are fully served. We show that this conclusion for identical replicas extends to general job sizes. In the special case of i.i.d. replicas, the stability threshold decreases in the number of replicas for job size distributions that are New-Better-than-Used (NBU) and increases for distributions that are New-Worse-than-Used (NWU). Comparing the stability criteria to the known sufficient and asymptotically necessary stability condition for FCFS with scaled

Table 1: Summary of the stability condition results. The X indicates a scenario for which we obtain results.

	Exponential		Bernoulli	General	
	i.i.d.	identical	i.i.d.	i.i.d.	identical
FCFS	$\rho < 1$ [6]		$\rho N < K^{d-1}$ [9]		
PS	$\rho < 1$ [3]	$\rho < \frac{1}{d}$ [3]		X	X
ROS	$\rho < 1$ [3]	$\rho < 1$ [3]			

Bernoulli job sizes, we observe that these differ, in contrast to the case of exponential job sizes considered in [3]. In particular, for the scaled Bernoulli job sizes the stability threshold is larger for PS than for FCFS.

The remainder of the paper is organized as follows. In Section 2 we present a detailed description of the redundancy-d model with the PS discipline. The stability conditions for both the lower- and upperbound systems, and therefore also of the original stochastic model, are proved in Section 3. Section 4 contains numerical experiments that show the accuracy of the lower- and upperbound systems used to prove the stability condition. Section 5 contains conclusions and some suggestions for further research.

2. Model description

Consider a system with N parallel servers where jobs arrive as a Poisson process of rate λ. Each of the N servers follows a processor-sharing (PS) discipline. In this paper we focus on the case of homogeneous server speeds, but some of the statements can be extended to heterogeneous server speeds as further discussed in Section 5. When a job arrives, the dispatcher immediately assigns replicas to $d \leq N$ servers selected uniformly at random (without replacement). As soon as the first of these d replicas finishes service the remaining ones are abandoned. We allow the replica sizes $X_1, \ldots, X_d$ of a job to be governed by some joint distribution $F(x_1, \ldots, x_d)$, where X_i, $i = 1, \ldots, d$, are each distributed as a generic random variable X, but not necessarily independent. By Sklar's Theorem, see e.g., [10], we know that any joint distribution can be written in terms of marginal distribution functions and a copula $C : [0, 1]^d \to [0, 1]$ which describes the dependency structure between the variables. Note that we do not allow the dependency structure to depend on the state of the system. Special cases of the dependency structure are i) perfect dependency, so-called identical replicas, where the job size is preserved for all replicas, i.e., $X_i = X$, $i = 1, \ldots, d$, ii) no dependency at all, so-called i.i.d. replicas.

Observe that in this paper the load of the system is defined by $\tilde{\rho} = \frac{d\lambda \mathbb{E}[\min\{X_1,\ldots,X_d\}]}{N}$, as opposed to the load defined in [3, 6] and used in Table 1.

3. Stability analysis

In this section we analyze the stability condition for redundancy-d systems with PS. In Sections 3.1 and 3.2 carefully chosen lower- and upperbound systems are introduced. In Section 3.3 we consider the corresponding fluid limit models of these lower- and upperbound systems to prove the stability condition. Figure 1 depicts an overview of this section.

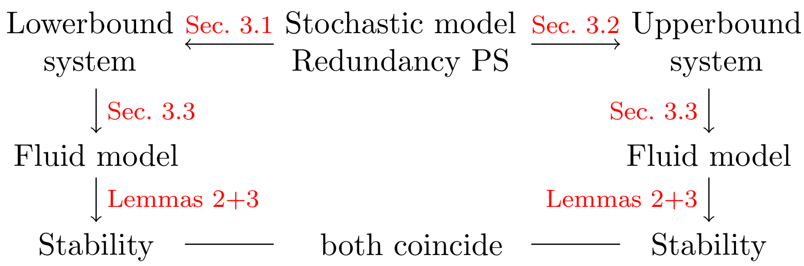

Figure 1: Overview of the various systems.

3.1. Lowerbound

In this section we introduce a lowerbound system which is the same as our original system except for two differences. In the lowerbound system replicas always receive a service rate equal to the reciprocal of the minimum queue length, at that time, of all the sampled servers, instead of a service rate equal to the reciprocal of the queue length at that specific server. The second difference is that in the lowerbound system the sizes of all the replicas are equal to $X_{\min} := \min\{X_1, \ldots, X_d\}$, the minimum size of the d replicas in the original system. However, this latter difference actually just follows from the first difference. Namely, if all replicas receive service at equal rate it immediately follows that the smallest replica will always complete service first, after which the other replicas are abandoned.

The above-described system provides a lowerbound in the sense that the residual sizes of the replicas are always smaller than or equal to the residual sizes of the replicas in the original system. In particular, the sojourn time of a job in the lowerbound system is always smaller than or equal to that in the original system. The ordering of the residual sizes of the replicas can be formally proved with sample-path arguments for so-called 'monotonic' processor-sharing networks, see Lemma 1 in [11].

Next, we give an example to further clarify the operation of the lowerbound system.

Example 1. *Assume that the queue lengths at the* $d = 2$ *sampled servers are* $(4, 5)$*. Then, in the original system the replicas receive service at rates* $1/4$ *and* $1/5$*, respectively. In the lowerbound system both replicas receive service at rate* $1/4 = 1/\min\{4, 5\} = \max\{1/4, 1/5\}$.

Remark 1. *For* $d = 1$ *and* $d = N$ *this lowerbound system is exactly the same as the original system and in the special case of identical replicas this lowerbound system is exactly the same as the lowerbound system introduced in [3].*

We now provide an alternative way of viewing the lowerbound system, without any redundant replicas.

Consider $M = \binom{N}{d}$ (fictitious) job classes, where each job class corresponds to one of the $\binom{N}{d}$ possible combinations of d servers. Let $\boldsymbol{s}^i \subseteq \{1, \ldots, N\}$ denote the set of servers corresponding to the i-th job class. Without loss of generality, we suppose that job class 1 corresponds to the set of servers $\boldsymbol{s}^1 = \{1, \ldots, d\}$, job class 2 to the set $\boldsymbol{s}^2 = \{1, \ldots, d-1, d+1\}$ and finally job class M to the set $\boldsymbol{s}^M = \{N-d+1, \ldots, N\}$. All the jobs of a certain class receive service at the same rate, which depends on the number of jobs present of other classes, namely those classes that correspond to sets of servers that have a server in common. Thus, we can also see this as M individual (virtual) queues, one for each job class, that follow a PS discipline with a rate that depends on the number of jobs present at the other virtual queues. Note that only jobs from class i arrive at virtual queue i and these jobs arrive according to a Poisson process with rate $\lambda_i = \lambda/M$, see also Figure 2.

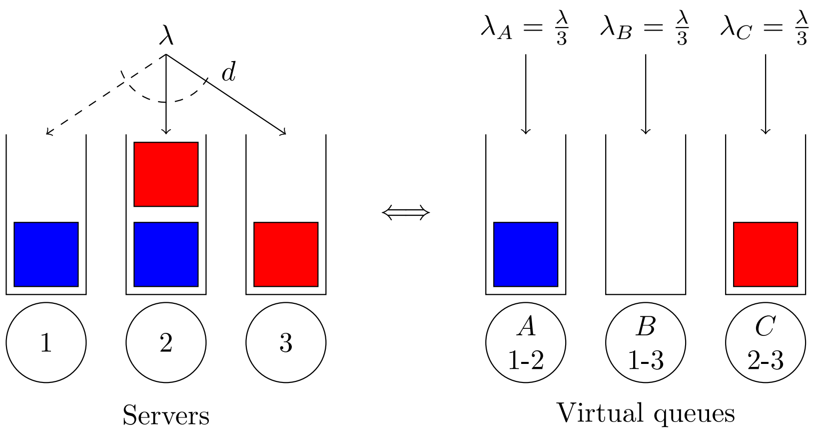

Figure 2: Visualization of the lowerbound systems.

Example 2. *Consider the example visualized in Figure 2 with $N = 3$ servers and $d = 2$ replicas. In the lowerbound system we have $M = 3$ separate $M/X_{min}/1/PS$ virtual queues, for clarity denoted by A, B and C. Moreover $\mathbf{s}^A = \{1, 2\}$, $\mathbf{s}^B = \{1, 3\}$ and $\mathbf{s}^C = \{2, 3\}$. Note that in general the number of servers in the original system and the number of virtual queues may differ. Let Q^*_j denote the number of jobs at server j, $j \in \{1, 2, 3\}$, and Q_i the number of jobs at virtual queue i, $i \in \{A, B, C\}$. Observe that at server 2 on the left hand side two types of jobs can arrive, i.e., jobs which have replicas at servers 1 and 2 (this corresponds to job class A on the right hand side) and jobs which have replicas at servers 2 and 3 (this corresponds to job class C on the right hand side). Thus, the queue length at server 2 is equal to the sum of the queue lengths at virtual queues A and C, i.e., $Q^*_2 = Q_A + Q_C$. Therefore, the blue job receives service at rate $1 = 1/\min\{Q^*_1, Q^*_2\} = 1/\min\{Q_A + Q_B, Q_A + Q_C\}$.*

In this example virtual queue A is working at rate $\frac{Q_A}{\min\{Q_A+Q_B, Q_A+Q_C\}}$. Equivalently, virtual queues B and C are working at rates $\frac{Q_B}{\min\{Q_A+Q_B, Q_B+Q_C\}}$ and $\frac{Q_C}{\min\{Q_A+Q_C, Q_B+Q_C\}}$, respectively.

In general, the service rate, in the lowerbound system, at virtual queue i is

$$\frac{Q_i}{\min\{\sum_{i_1 \in \mathcal{S}^i_1} Q_{i_1}, \dots, \sum_{i_d \in \mathcal{S}^i_d} Q_{i_d}\}},$$

where $\mathcal{S}^i_j = \{k : s^i_j \in \mathbf{s}^k\}$ denotes the set of job classes which share server s^i_j in their corresponding server set and Q_i the number of jobs at virtual queue i. Thus $\sum_{i_1 \in \mathcal{S}^i_1} Q_{i_1}$ is the number of jobs at server s^i_1. In Example 2 we have $\mathcal{S}^A_1 = \mathcal{S}^B_1 = \{A, B\}$, $\mathcal{S}^A_2 = \mathcal{S}^C_1 = \{A, C\}$ and $\mathcal{S}^B_2 = \mathcal{S}^C_2 = \{B, C\}$.

Again, note that this alternative lowerbound system, explained in Example 2, behaves exactly the same as the lowerbound system described at the beginning of this subsection, in the sense that a job at virtual queue i receives exactly the same service rate as the d replicas at the servers in the original system that belong to $\mathbf{s}^i$. However, in the alternative lowerbound system there is no replication of jobs, i.e., the dependency is captured in the service rate, which makes the alternative lowerbound system easier to analyze.

3.2. Upperbound

In this section we introduce an upperbound system which is the same as our original system except for two differences. In the upperbound system replicas always receive a service rate equal to the reciprocal of the maximum queue length, at that time, of all the sampled servers, instead of a service rate equal to the reciprocal of the queue length at that specific server. The second difference is that in the upperbound system the sizes of all the d replicas are equal to $X_{\min} = \min\{X_1, \ldots, X_d\}$, the minimum size of the d replicas in the original system. Again, similar to the lowerbound system, note that this latter difference is a direct implication of the first difference since all the d replicas receive service at equal rate. Therefore the smallest replica will always complete service first.

The above-described system provides an upperbound in the sense that the residual sizes of the replicas are always larger than or equal to the residual sizes of the replicas in the original system. Once again, this ordering can be proved with sample-path arguments for so called monotonic processor-sharing networks.

For the upperbound system we give an example similar to Example 1.

Example 3. *Assume that the queue lengths at the $d = 2$ sampled servers are $(4, 5)$. Then, in the original system the replicas receive service at rates $1/4$ and $1/5$, respectively. In the upperbound system both replicas receive service at a rate $1/5 = 1/\max\{4, 5\} = \min\{1/4, 1/5\}$.*

Remark 2. *For $d = 1$ and $d = N$ this upperbound system is exactly the same as the original system.*

Just like for the lowerbound system, we now provide an alternative way of viewing the upperbound system. This is accomplished by taking the same set of $M = \binom{N}{d}$ virtual queues. The only difference in Example 2 is that now the blue job receives service at rate $1 = 1/\max\{Q_1^*, Q_2^*\} = 1/\max\{Q_A + Q_B, Q_A + Q_C\}$.

In general, the service rate, in the upperbound system, at virtual queue i is

$$\frac{Q_i}{\max\{\sum_{i_1 \in \mathcal{S}_1^i} Q_{i_1}, \ldots, \sum_{i_d \in \mathcal{S}_d^i} Q_{i_d}\}}.$$

Because of the similarities of the lower- and upperbound systems, the fluid limits are very similar as well. Therefore we discuss both fluid limits simultaneously in the next section.

3.3. Fluid limit

Let $Q_i(t)$ denote the number of jobs at virtual queue i at time t with $\boldsymbol{Q}(t) = (Q_1(t), \ldots, Q_M(t))$. The time until virtual queue i first becomes empty is denoted by $\tau_i = \inf\{t | Q_i(t) = 0\}$. The service rate per job at virtual queue i at time s is given by

$$\varphi_i(\boldsymbol{Q}(s)) = \frac{1}{\min\{\sum_{i_1 \in \mathcal{S}_1^i} Q_{i_1}(s), \ldots, \sum_{i_d \in \mathcal{S}_d^i} Q_{i_d}(s)\}}, \tag{1}$$

for the lowerbound system and

$$\varphi_i(\boldsymbol{Q}(s)) = \frac{1}{\max\{\sum_{i_1 \in \mathcal{S}_1^i} Q_{i_1}(s), \ldots, \sum_{i_d \in \mathcal{S}_d^i} Q_{i_d}(s)\}}, \tag{2}$$

for the upperbound system whenever $Q_i(s) > 0$ and $\varphi_i(\boldsymbol{Q}(s)) = 0$ for $Q_i(s) = 0$. The attained service process, i.e., the cumulative amount of processing time per job allocated to virtual queue i during the interval $[s, t]$, is

$$\eta_i(s, t) = \int_{u=s}^{t} \varphi_i(\boldsymbol{Q}(u))\mathrm{d}u.$$

For $0 \le t \le \tau_i$, $\eta_i(0, t)$ is continuous and strictly increasing from an initial value of 0, and for $t > \tau_i$ it is (Lipschitz) continuous nondecreasing in the second argument.

Let U_{ik} be the arrival time of the kth job at virtual queue i. Define $A_i(t) = \max\{k : U_{ik} < t\}$ as the number of jobs that arrive at virtual queue i during the interval $(0, t]$. The virtual queue length process satisfies

$$Q_i(t) = \sum_{l=1}^{Q_i(0)} \mathbb{1}_{\{v'_{il} > \eta_i(t)\}} + \sum_{k=1}^{A_i(t)} \mathbb{1}_{\{v_{ik} > \eta_i(U_{ik}, t)\}}, \tag{3}$$

where v'_{il} denotes the residual job size of the lth initial job at virtual queue i and v_{ik} the job size of the kth arriving job at virtual queue i.

We consider the behavior of the system on a fluid scale and define the scaled processes

$$\begin{aligned}
&\overline{Q}_i^n(t) = Q_i(nt)/n,\\
&\overline{\eta}_i^n(s, t) = \int_{u=s}^{t} \varphi_i(\overline{\boldsymbol{Q}}^n(u))\mathrm{d}u = \int_{u=s}^{t} n\varphi_i(\boldsymbol{Q}(nu))\mathrm{d}u = \eta_i(ns, nt),\\
&\overline{\tau}_i^n = \inf\{t|\overline{Q}_i^n(t) = 0\} = \tau_i/n.
\end{aligned}$$

The scaled attained service time is continuous and strictly increasing for $0 \le t < \overline{\tau}_i^n$, and continuous non-decreasing for $t \ge \overline{\tau}_i^n$ in the second argument.

Definition 1. *A non-negative continuous function $\overline{Q}_i(\cdot)$ is a fluid-model solution if it satisfies the functional equation*

$$\overline{Q}_i(t) = \overline{Q}_i(0)(1 - G(\overline{\eta}_i(0, t))) + \frac{\lambda}{M}\int_{s=0}^{t} (1 - F_{X_{min}}(\overline{\eta}_i(s, t)))ds, \tag{4}$$

where $G(\cdot)$ is the service time distribution of initial jobs, $F_{X_{min}}(\cdot)$ the service time distribution of arriving jobs and

$$\overline{\eta}_i(s, t) = \int_{u=s}^{t} \varphi_i(\overline{\boldsymbol{Q}}(u))du.$$

Theorem 1. *The limit point of any convergent subsequence of $(\overline{Q}_i^n(t); t \ge 0)$ is almost surely a solution of the fluid-model Equation* (4).

Proof. The proof is presented in Appendix A.1. □

Next, we prove a monotonicity result for the fluid limit given by Equation (4).

Definition 2. *We say that the service rate per job is monotonic if*

$$\frac{\varphi_i(\overline{\boldsymbol{Q}}^a(t))}{\overline{Q}_i^a(t)} \ge \frac{\varphi_i(\overline{\boldsymbol{Q}}^b(t))}{\overline{Q}_i^b(t)}, \quad \forall \overline{\boldsymbol{Q}}^a(t) \le \overline{\boldsymbol{Q}}^b(t) : \overline{Q}_i^a(t) > 0,$$

see also [11].

Lemma 1. *Let $\varphi_i(\cdot)$ be monotonic for all $i = 1\ldots, M$ and consider the fluid limits with initial conditions $\overline{\boldsymbol{Q}}^a(0)$ and $\overline{\boldsymbol{Q}}^b(0)$, where $\overline{\boldsymbol{Q}}^a(0) \leq \overline{\boldsymbol{Q}}^b(0)$, then $\varphi_i(\overline{\boldsymbol{Q}}^a(t)) \geq \varphi_i(\overline{\boldsymbol{Q}}^b(t))$ for all $i = 1\ldots, M$ and $t \geq 0$ for which $\overline{Q}_i^a(t) > 0$.*

Proof. Assume that there exists an $i \in \{1, \ldots, M\}$ such that t_0 is the first time for which $\varphi_i(\overline{\boldsymbol{Q}}^a(t_0)) < \varphi_i(\overline{\boldsymbol{Q}}^b(t_0))$ with $\overline{Q}_i^a(t) > 0$. Then, there must exist an $i^* \in \{1, \ldots, M\}$ such that $\overline{Q}_{i^*}^a(t) > \overline{Q}_{i^*}^b(t)$ for some $0 \leq t < t_0$. However, $\varphi_i(\overline{\boldsymbol{Q}}^a(t)) \geq \varphi_i(\overline{\boldsymbol{Q}}^b(t))$, for all $i = 1, \ldots, M$ and $0 \leq t < t_0$, from which it follows that $\overline{Q}_i^a(t) \leq \overline{Q}_i^b(t)$ for all $i = 1, \ldots, M$ and $0 \leq t < t_0$. □

The key property of these fluid limits is that all the components of the fluid limit corresponding to all the virtual queues remain equal at all times when starting with the same initial conditions. More importantly, even the fluid limits of the lower- and upperbound systems coincide when starting with the same initial conditions, since Equations (1) and (2) coincide.

Property 1. *If $\overline{Q}_i(0) = q(0)$ for all $i = 1, \ldots, M$, then $\overline{Q}_i(t) = q(t)$ for all $i = 1, \ldots, M$ and $t \geq 0$, where*

$$q(t) = q(0)(1 - G(\overline{\eta}(0, t))) + \frac{\lambda}{M}\int_{s=0}^{t}(1 - F_{X_{min}}(\overline{\eta}(s, t)))ds,$$

with

$$\overline{\eta}(s, t) = \frac{1}{\binom{N-1}{d-1}}\int_{u=s}^{t}\varphi(q(u))du,$$

and

$$\varphi(q(u)) = \frac{1}{q(u)},$$

for $q(u) > 0$ and $\varphi(0) = 0$. The expression for the attained service process follows from $\varphi_i(\boldsymbol{Q}(s)) = \frac{1}{\binom{N-1}{d-1}q(s)}$ for all $i = 1, \ldots, M$. Note that this is the fluid limit of an $M/X_{min}/1/PS$ queue with arrival rate λ/M and server speed $1/\binom{N-1}{d-1}$.

The proof of Property 1 follows by contradiction. Assume that there exists an $i \in \{1, \ldots, M\}$ such that t_0 is the first time for which $\overline{Q}_i(t_0) \neq q(t_0)$. Then there must exist an $i^* \in \{1, \ldots, M\}$ such that $\overline{\eta}_{i^*}(s, t_1) \neq \overline{\eta}(s, t_1)$ with $0 \leq s \leq t_1 < t_0$. However, $\overline{Q}_i(t) = q(t)$ for all $i = 1, \ldots, M$ and $t \leq t_1$ from which it follows $\overline{\eta}_i(s, t) = \overline{\eta}(s, t)$ for all $i = 1, \ldots, M$ and $0 \leq s \leq t \leq t_1$.

Lemma 2. *A necessary stability condition for the fluid-limit model of both the lower- and upperbound system is given by $\tilde{\rho} \leq 1$.*

Proof. For the minimum of the fluid limit components we have that

$$\min_{i\in\{1,\ldots,M\}} \overline{Q}_i(t) \geq \min_{i\in\{1,\ldots,M\}} \overline{Q}_i^0(t) = \frac{\lambda}{M}\Big(t - \int_{s=0}^{t} F_{X_{\min}}(\overline{\eta}^*(s, t))\mathrm{d}s\Big),$$

where $\overline{Q}_i^0(t)$ denotes the fluid limit and $\overline{\eta}^*$ the attained service at virtual queue i when starting with initial condition 0 at all virtual queues. The inequality follows from Lemma 1 and the equality follows by Property 1. Note that the above expression holds both for the lower- and upperbound system, with the service rate per job given by Equation (1) and (2), respectively.

Thus, the minimum of the fluid limit components is lower bounded by the fluid limit of an $M/X_{\min}/1/PS$ queue with arrival rate λ/M and server speed $1/\binom{N-1}{d-1}$. This fluid limit goes to infinity for $\tilde{\rho} = \frac{d\lambda \mathbb{E}[\min\{X_1,\ldots,X_d\}]}{N} > 1$, since $\binom{N-1}{d-1}/M = d/N$. □

Theorem 2. *A necessary stability condition for redundancy-d systems with the processor-sharing discipline and general job size distributions is given by* $\tilde{\rho} \leq 1$.

Proof. As stated in the proof of Lemma 2 it follows that for $\tilde{\rho} = \frac{d\lambda \mathbb{E}[\min\{X_1,\ldots,X_d\}]}{N} > 1$ the fluid limit of the lowerbound system goes to infinity, which implies that the stochastic lowerbound system is unstable as well, see [12]. □

Lemma 3. *A sufficient stability condition for the fluid-limit model of both the lower- and upperbound system is given by* $\tilde{\rho} < 1$.

Proof. For the maximum of the fluid limit components we have that

$$\begin{aligned}\max_{i\in\{1,\ldots,M\}} \overline{Q}_i(t) &\leq \max_{i\in\{1,\ldots,M\}} \overline{Q}_i^q(t) \\ &= \left(\max_{j\in\{1,\ldots,M\}} \overline{Q}_j(0)\right)\cdot(1 - G(\overline{\eta}^*(0,t))) + \frac{\lambda}{M}\Big(t - \int_{s=0}^{t} F_{X_{\min}}(\overline{\eta}^*(s,t))\mathrm{d}s\Big),\end{aligned}$$

where $\overline{Q}_i^q(t)$ denotes the fluid limit and $\overline{\eta}^*$ the attained service at virtual queue i when starting with initial condition $q = \max_{j\in\{1,\ldots,M\}} \overline{Q}_j(0)$ at all virtual queues. The inequality follows from Lemma 1 and the equality follows by Property 1. Once again, note that the above expression holds both for the lower- and upperbound system, with the service rate per job given by Equation (1) and (2), respectively.

Thus, the maximum of the fluid limit components is upper bounded by the fluid limit of an $M/X_{\min}/1/PS$ queue with arrival rate λ/M and server speed $1/\binom{N-1}{d-1}$. This fluid limit goes to zero in finite time for $\tilde{\rho} = \frac{d\lambda \mathbb{E}[\min\{X_1,\ldots,X_d\}]}{N} < 1$. □

In [13] it is proved that the stability of the fluid limit implies the stability of the stochastic process in case of general job sizes. However, this result assumes *head-of-the-line* service disciplines and does not cover the processor-sharing discipline that we consider. For exponential job sizes the *head-of-the-line* and processor-sharing service disciplines are equivalent and for the special cases of identical and i.i.d. replicas it follows that $X_{\min}$ is exponentially distributed with mean $\mathbb{E}[X]$ and $\mathbb{E}[X]/d$, respectively. Therefore, in this special case, we can use the results in [13] to obtain the (sufficient) stability condition for our stochastic upperbound system and recover the results in [3].

While it seems plausible that the stability of the fluid model for general job sizes implies positive Harris recurrence, there is, to the best of our knowledge, no proof. See also [14, 15] for a more extensive discussion on this stability issue. This leads to the next conjecture.

Conjecture 1. *A sufficient stability condition for redundancy-d with the processor-sharing discipline and general job size distributions is given by* $\tilde{\rho} < 1$.

3.4. Alternative upperbound

An alternative analytically tractable upperbound system that would (partly) give the stability condition of our stochastic model is the system where all the d replicas are fully served, as opposed to at least one server (out of the d sampled servers) fully completing the replica, see also [3, Section 6.2.2]. Each individual server in this alternative system can be viewed as an $M/X/1/PS$ queue with arrival rate $\frac{d\lambda}{N}$ and expected job size $\mathbb{E}[X]$.

Remark 3. *For $d = 1$ and $d = N$ this upperbound system is exactly the same as the original system.*

Theorem 3. *A sufficient stability condition for redundancy-d systems with the processor-sharing discipline and general job size distributions is given by $\rho < \frac{1}{d}$, with $\rho = \frac{\lambda\mathbb{E}[X]}{N}$.*

Proof. Follows by comparison of the original system with the $M/X/1/PS$ queue. □

In general this sufficient stability condition does not coincide with the necessary stability condition from Theorem 2. However, for identical replicas it does as in this case $d\rho = \tilde{\rho}$. Here, the PS discipline causes the worst possible reduction among all work-conserving service disciplines for the stability condition, since the stability condition is equal to that in a system where all replicas are fully served as is the case in the alternative upperbound system described in this subsection.

Now, we discuss the conjectured stability condition for the PS discipline in the special case of i.i.d. replicas in more detail.

Definition 3. *Consider a non-negative random variable X with support denoted by R_X and cumulative distribution function (cdf) $F_X(x)$. Let $\bar{F}_X(x) = 1 - F_X(x)$ denote the complementary cumulative distribution function (ccdf). Then, X is New-Better-than-Used (NBU) if for all $t_1, t_2 \in R_X$,*

$$\bar{F}_X(t_1 + t_2) \leq \bar{F}_X(t_1)\bar{F}_X(t_2). \tag{5}$$

On the other hand, X is New-Worse-than-Used (NWU) if for all $t_1, t_2 \in R_X$,

$$\bar{F}_X(t_1 + t_2) \geq \bar{F}_X(t_1)\bar{F}_X(t_2). \tag{6}$$

Note that

$$\mathbb{E}[\min\{X_1, \dots, X_d\}] = \int_{[0,1]^d} \min\{F_X^{-1}(u_1), \dots, F_X^{-1}(u_d)\} \mathrm{d}C(u_1, \dots, u_d),$$

where C denotes the copula model describing the dependency structure, see also [10]. In the special case of i.i.d. replica sizes and a job size that is NBU (NWU) we have

$$\mathbb{E}[\min\{X_1, \dots, X_d\}] = \int_{t=0}^{\infty} (1 - F_X(t))^d \mathrm{d}t \geq (\leq) \int_{t=0}^{\infty} (1 - F_X(dt))\mathrm{d}t = \frac{1}{d}\mathbb{E}[X].$$

This implies that for i.i.d. replica sizes the stability threshold, compared to the stability threshold for the exponential distribution, is larger for NWU distributions and smaller for NBU distributions.

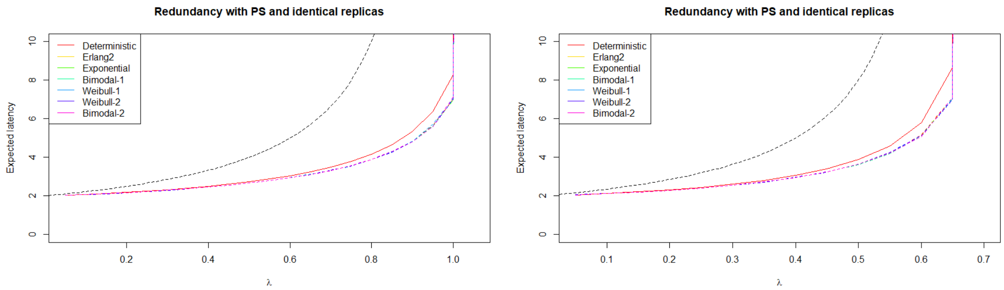

Figure 3: The expected latency in the original system and in the lower- and upperbound systems (dashed lines) for processor sharing with identical replicas, $N = 4$, $\mathbb{E}[X] = 2$, various job size distributions (see Appendix A.2) and $d = 2$ (left) and $d = 3$ (right).

Remark 4. *For independent exponentially distributed job sizes the stability conditions for FCFS and PS are the same. However, for the scaled Bernoulli job size distribution, which is a NWU distribution, this is not the case. If*

$$X = \begin{cases} K, & w.p.\ 1-p, \\ 0, & w.p.\ p, \end{cases}$$

where $p = 1 - 1/K$, then the asymptotic stability condition for FCFS is given by $\lambda/K^{d-1} < 1$, see [9]. The stability condition for PS is $(\lambda d)/(NK^{d-1}) < 1$. Thus for this job size distribution, the stability threshold for PS is larger than for FCFS.

4. Numerical results

In Section 4.1 the accuracy of the bounds for both identical and i.i.d. replicas is studied. In Section 4.2 we show a near-insensitivity result for the expected latency in the case of identical replicas. In Section 4.3 we discuss the load of the system in more detail.

4.1. Accuracy bounds

In Figure 3 the expected latency for the lowerbound system, upperbound system and original system with $N = 4$ (homogeneous) servers and various job size distributions is depicted for $d = 2$ and $d = 3$ identical replicas. Only the alternative upperbound system from Section 3.4 is analytically tractable and the expected latency in the other systems is obtained via simulation. For identical replicas it can be seen that the lowerbound is quite accurate. Moreover, note that the expected latency in the upperbound system is only dependent on the mean of the job size distribution and not its higher moments. At first instance, it seems that the expected latency is insensitive to the job size distribution, but in Section 4.2 we show that the expected latency in fact slightly differs for the various job size distributions, which is called near-insensitivity.

In Figure 4 the expected latency in the lowerbound system, upperbound system and original system with $N = 4$ (homogeneous) servers and various job size distributions is depicted for $d = 2$ and $d = 3$ i.i.d. replicas. For i.i.d. replicas it can be seen that especially the upperbound is quite accurate for various job size distributions. Moreover, note that the lowerbound for deterministic job sizes is tight, see also Remark 3.

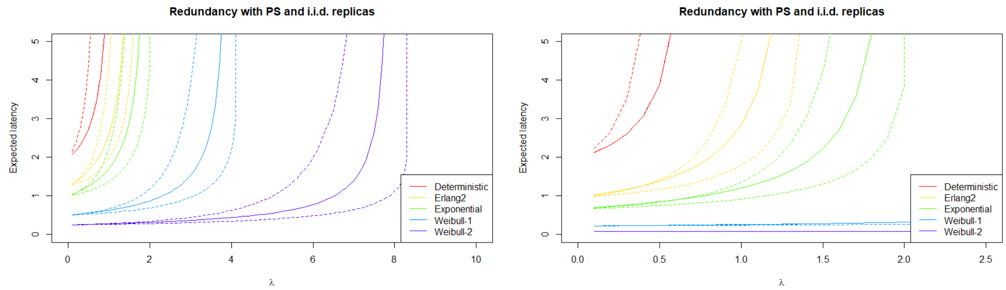

Figure 4: The expected latency in the original system and in the lower- and upperbound systems (dashed lines) for processor sharing with i.i.d. replicas, $N = 4$, $\mathbb{E}[X] = 2$, various job size distributions (see Appendix A.2) and $d = 2$ (left) and $d = 3$ (right).

4.2. *Near-insensitivity identical replicas*

It is well known that the expected latency in an ordinary $M/G/1/PS$ system is insensitive to the job size distribution given its mean. In this subsection our aim is to show that the expected latency is nearly insensitive to the job size distribution in redundancy-d systems with processor sharing and identical replicas. To demonstrate this claim, we consider the system with various job size distributions. We run the simulation 50 times, where each run consists of 10^7 arrivals. See also [16] in which the authors observe a similar near-insensitivity for an $M/G/N/JSQ/PS$ system.

In Figure 5 it can be seen that the 95% confidence intervals for the expected latency differ and are not overlapping for the various job size distributions. However, looking at the y-axis of the figure, we conclude that the expected latency only differs by 0.05, which is approximately 1.5%, between the job size with the lowest and highest variance.

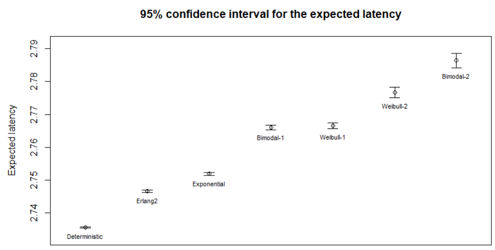

Figure 5: Insensitivity results for various job size distributions (see Appendix A.2) for the setting $d = 2$, $N = 4$, $\tilde{\rho} = 0.5$.

4.3. *Load*

In Section 3 it is proved that the necessary stability condition for the PS discipline is given by $\tilde{\rho} \leq 1$ and conjectured that $\tilde{\rho} < 1$ is the sufficient stability condition. However, when fixing the load $\tilde{\rho}$, the expected latency still differs when varying d and N. In Figure 6 it can be seen that, especially for identical replicas, even with a relatively high load of $\tilde{\rho} = 0.75$ for $d = N/2$

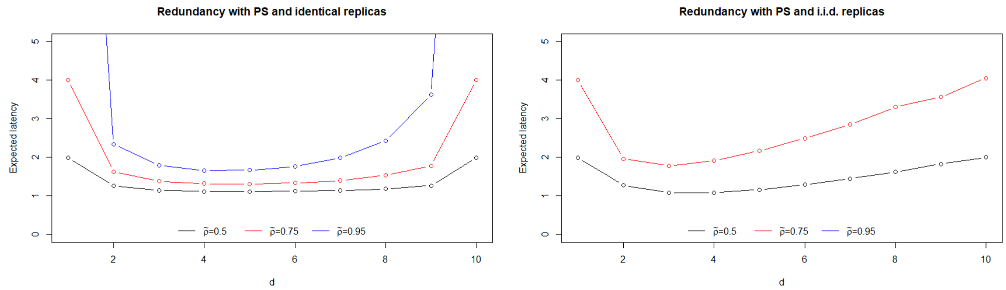

Figure 6: Simulated expected latency for processor sharing with $N = 10$, exponential job sizes X with $\mathbb{E}[X] = 1$, various loads and identical replicas (left) and i.i.d. replicas (right).

the job does (most of the times) not experience delay from other jobs, since $\mathbb{E}[T] \approx \mathbb{E}[X] = 1$. Even for $\tilde{\rho} = 0.95$, which we would call heavy traffic in the classical setting, the expected latency $\mathbb{E}[T]$ for the number of replicas d between 3 and 7 is approximately equal to 2, while for $d = 1$ and $d = N = 10$, i.e., equivalent to the classical setting, the expected latency $\mathbb{E}[T] = 20$. Note that for i.i.d. replicas and $\tilde{\rho} = 0.95$, the expected latency is between 5 and 10 and the blue line is therefore not visible in Figure 6. Also observe that for identical replicas, the expected latency varies significantly more than for i.i.d. replicas.

5. Conclusion

In this paper we established the stability condition for redundancy-d systems where all servers follow a processor-sharing discipline. We allow for generally distributed job sizes with possible dependence among the replica sizes of a job being governed by some joint distribution. The stability condition is characterized by the expectation of the minimum of d replica sizes being less than the mean interarrival time per server. In the special case of identical replicas the stability condition is insensitive to the job size distribution given its mean. Moreover, the stability threshold is inversely proportional to the number of replicas. Thus, in this case a higher degree of redundancy reduces stability. In the special case of i.i.d. replicas the stability threshold decreases (increases) in the number of replicas for job size distributions that are NBU (NWU).

In further research we could allow for heterogeneous server speeds. Note that in this scenario the lower- and upperbound systems from Sections 3.1 and 3.2 continue to apply. Again, we could consider the fluid limits of these systems. Observe that the heterogeneity of the servers is reflected in the service rate per job and consequently in the attained service process. As a consequence, Property 1 that played a key role in establishing the stability condition of the fluid models is not valid anymore. Moreover, it is no longer true that the fluid limits of the lower- and upperbound systems coincide when starting with the same initial conditions. Our simulations of the lower- and upperbound systems with heterogeneous server speeds reveal that the stability conditions do not coincide, which means that (most probably) the stability conditions of the fluid models also do not coincide. In addition, we observed that for heterogeneous server speeds the stability condition of the original stochastic system is sensitive, i.e., not only dependent on the expectation of the minimum of d replica sizes. For the PS discipline, the model of heterogeneous server speeds could be even further generalized to also include various job types, each with their own server speed realizations. One example in such a direction is the S&X model discussed in [4].

The stability condition for FCFS and ROS in the case of general job sizes is still an open problem. By simulation we observed that the stability condition for PS gives a reasonable approximation for the stability condition for FCFS. Moreover, in all the simulations that we performed, we observed that for i.i.d. replicas the stability threshold for PS is smaller (larger) than the stability threshold when the job size distribution is NBU (NWU). Note that this observation is in agreement with Remark 4. Intuitively, this could be explained since for the PS discipline every replica starts at the same time and the queue lengths are approximately equal. Thus for a job size distribution that is NBU (NWU) the PS discipline maximizes the wastage (improvement) of the server capacity. For the FCFS discipline this is not necessarily the case since not all replicas start at the same time or start service at all. However, formal proofs of these statements remain as a challenge for further research.

Acknowledgments

The work in this paper is supported by the Netherlands Organisation for Scientific Research (NWO) through Gravitation grant NETWORKS 024.002.003.

References

[1] G. Ananthanarayanan, A. Ghodsi, S. Shenker, I. Stoica, Effective straggler mitigations: Attack of the clones., Proc. 10th USENIX conference on Networked Systems Design and Implementation (2013) 185–198.

[2] A. Vulimiri, P. Godfrey, R. Mittal, J. Sherry, S. Ratnasamy, S. Shenker, Low latency via redundancy., Proc. 9th ACM conference on Emerging Networking Experiments and Technologies (2013) 283–294.

[3] E. Anton, U. Ayesta, M. Jonckheere, I. Verloop, On the stability of redundancy models., ArXiv 1903.04414 (2019).

[4] K. Gardner, M. Harchol-Balter, A. Scheller-Wolf, B. V. Houdt, A better model for job redundancy: Decoupling server slowdown and job size., IEEE ACM Transactions on Networking 25 (6) (2017) 3353–3367.

[5] K. Gardner, M. Harchol-Balter, A. Scheller-Wolf, M. Velednitsky, S. Zbarsky, Redundancy-d: The power of d choices for redundancy., Operations Research 65 (4) (2017) 1078–1094.

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

[7] T. Hellemans, B. V. Houdt, Analysis of redundancy(d) with identical replicas., ACM SIGMETRICS Performance Evaluation Review 46 (3) (2018) 74–79.

[8] Y. Raaijmakers, S. Borst, O. Boxma, Delta probing policies for redundancy., Performance Evaluation 127-128 (2018) 21–35.

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

[10] R. Nelsen, An Introduction to Copulas., Springer, New York, 1999.

[11] T. Bonald, A. Proutière, On stochastic bounds for monotonic processor sharing networks., Queueing Systems 47 (2004) 81–106.

[12] J. Dai, A fluid limit model criterion for instability of multiclass queueing networks., The Annals of Applied Probability 6 (3) (1996) 751–757.

[13] J. Dai, On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models., The Annals of Applied Probability 5 (1) (1995) 49–77.

[14] J. Dai, J. Harrison, Processing Networks: Fluid Models and Stability, Cambridge University Press, 2020.

[15] N. Walton, M. Mandjes, A stability conjecture on bandwidth sharing networks., Queueing Systems 68 (2011) 237–250.

[16] V. Gupta, M. Harchol-Balter, K. Sigman, W. Whitt, Analysis of join-the-shortest-queue routing for web server farms., Performance Evaluation 64 (9-12) (2007) 1062–1081.

[17] R. Egorova, Sojourn time tails in processor-sharing systems, Ph.D. thesis, Eindhoven University of Technology (2009).

[18] H. Gromoll, P. Robert, A. Zwart, Fluid limits for processor sharing queues with impatience., Mathematics of Operations Research 33 (2008) 375–402.

Appendix A. Appendix

Appendix A.1. Proof fluid limits

The proof of Theorem 1 is very similar to the proof of Theorem 5.2.1 in [17]. The latter proof only relies on the property that $\overline{\eta}_i^n(s,t)$ is decreasing in s and $\overline{\eta}(\cdot,t)$ is continuous on $[\xi_i(t)+\epsilon, t]$, where $\xi_i(t) = \sup(u \in [0,t] : \overline{Q}_i(u) = 0)$. In order to keep the paper self-contained, we give here the proof.

Applying the fluid scaling to each term in Equation (3) gives

$$\overline{Q}_i^n(t) = \frac{1}{n}\sum_{l=1}^{n\overline{Q}_i^n(0)} \mathbb{1}_{\{v'_{il}>\overline{\eta}_i^n(0,t)\}} + \frac{1}{n}\sum_{k=1}^{A_i^n(t)} \mathbb{1}_{\{v_{ik}>\overline{\eta}_i^n(U_{ik}^n,t)\}} := I_i^n + J_i^n.$$

We now proceed to derive $\lim_{n\to\infty}\overline{Q}_i^n(t)$, and distinguish two cases depending on whether $\overline{Q}_i(u) > 0$ for all $u \in [0,t]$ or not. Let $\xi_i(t) = \sup(u \in [0,t] : \overline{Q}_i(u) = 0)$. It is useful to distinguish two further cases, depending on whether $\xi_i(t) < t$ or $\xi_i(t) = t$. We start with the former case, and fix $\epsilon > 0$ such that $\xi_i(t) + \epsilon < t$. Now

$$\begin{aligned} J_i^n &= \frac{1}{n}\sum_{k=1}^{A_i^n(\xi_i(t)+\epsilon)} \mathbb{1}_{\{v_{ik}>\overline{\eta}_i^n(U_{ik}^n,t)\}} \\ &\quad + \frac{1}{n}\sum_{k=A_i^n(\xi_i(t)+\epsilon)+1}^{A_i^n(t)} \mathbb{1}_{\{v_{ik}>\overline{\eta}_i^n(U_{ik}^n,t)\}} := J_{1,i}^n + J_{2,i}^n. \end{aligned}$$

We first determine $\lim_{n\to\infty} J_{2,i}^n$. By definition $\overline{Q}_i(u) > 0$ for all $[\xi_i(t)+\epsilon, t]$. Hence, the bounded convergence theorem yields

$$\lim_{n\to\infty} \overline{\eta}_i^n(u,v) = \overline{\eta}_i(u,v),$$

for all $t \ge v \ge u \ge \xi_i(t)+\epsilon$. Since $\overline{\eta}_i^n(s,t)$ is decreasing in s and $\overline{\eta}_i(\cdot,t)$ is continuous on $[\xi_i(t)+\epsilon, t]$, the convergence is uniform on $[\xi_i(t)+\epsilon, t]$, i.e., for any $\delta > 0$ there exists an n_δ such that

$$\sup_{s\in[\xi_i(t)+\epsilon,t]} |\overline{\eta}_i^n(s,t) - \overline{\eta}_i(s,t)| \le \delta, \quad \text{for all } n \ge n_\delta. \tag{A.1}$$

We partition the interval $[\xi_i(t)+\epsilon, t]$ into N_1 subintervals $[t_{j-1}^{N_1}, t_j^{N_1}]$, $j = 1, \ldots, N_1$, for some integer $N_1 \ge 1$, in such a way that
$\max_{j=0,\ldots,N_1}(t_j^{N_1} - t_{j-1}^{N_1}) \to 0$ as $N_1 \to \infty$. Then,

$$J_{2,i}^n = \frac{1}{n}\sum_{j=1}^{N_1}\sum_{k=A_i^n(t_j^{N_1})+1}^{A_i^n(t_{j+1}^{N_1})} \mathbb{1}_{\{v_{ik}>\overline{\eta}_i^n(U_{ik}^n,t)\}}.$$

Suppose that $t_{j-1}^{N_1} \le U_{ik}^n \le t_j^{N_1}$ for some $j \in \{1,\ldots,N_1\}$, some $k \in \{A_i^n(\xi_i(t)+\epsilon)+1, \ldots, A_i^n(t)\}$, and some $n > n_\delta$. It then follows from (A.1) that for $n > n_\delta$

$$\overline{\eta}_i(t_j^{N_1}, t) - \delta \le \overline{\eta}_i^n(U_{ik}^n, t) \le \overline{\eta}_i(t_{j-1}^{N_1}, t) + \delta,$$

which yields

$$\frac{1}{n}\sum_{j=1}^{N_1}\sum_{k=A_i^n(t_{j-1}^{N_1})+1}^{A_i^n(t_j^{N_1})}\mathbb{1}_{\{v_{ik}>\overline{\eta}_i(t_j^{N_1},t)-\delta\}}\le J_{2,i}^n$$

$$\le\frac{1}{n}\sum_{j=1}^{N_1}\sum_{k=A_i^n(t_{j-1}^{N_1})+1}^{A_i^n(t_1^{N_1})}\mathbb{1}_{\{v_{ik}>\overline{\eta}_i(t_{j-1}^{N_1},t)+\delta\}}.$$

Using Lemma 5.1 in [18], we obtain

$$\limsup_{n\to\infty}J_{2,i}^n\le\frac{\lambda}{\binom{N}{d}}\sum_{j=1}^{N_1}(t_j^{N_1}-t_{j-1}^{N_1})\mathbb{P}(v_{ik}>\overline{\eta}_i(t_{j-1}^{N_1},t)+\delta),$$

$$\liminf_{n\to\infty}J_{2,i}^n\ge\frac{\lambda}{\binom{N}{d}}\sum_{j=1}^{N_1}(t_j^{N_1}-t_{j-1}^{N_1})\mathbb{P}(v_{ik}>\overline{\eta}_i(t_j^{N_1},t)-\delta).$$

For $s\in[\xi_i(t)+\epsilon,t]$ the bounded convergence theorem implies that

$$\lim_{N_1\to\infty}\sum_{j=1}^{N_1}\mathbb{1}_{[t_j^{N_1}-t_{j-1}^{N_1})}(s)\mathbb{P}(v_{ik}>\overline{\eta}_i(t_{j-1}^{N_1},t)+\delta)=\mathbb{P}(v_{ik}>\overline{\eta}_i(s,t)+\delta),$$

$$\lim_{N_1\to\infty}\sum_{j=1}^{N_1}\mathbb{1}_{[t_j^{N_1}-t_{j-1}^{N_1})}(s)\mathbb{P}(v_{ik}>\overline{\eta}_i^n(t_j^{N_1},t)-\delta)=\mathbb{P}(v_{ik}>\overline{\eta}_i(s,t)-\delta).$$

Letting $N_1\to\infty$, we deduce

$$\limsup_{n\to\infty}J_{2,i}^n\le\frac{\lambda}{\binom{N}{d}}\int_{\xi_i(t)+\epsilon}^{t}\mathbb{P}(v_{ik}>\overline{\eta}_i(s,t)+\delta)\mathrm{d}s,$$

$$\liminf_{n\to\infty}J_{2,i}^n\ge\frac{\lambda}{\binom{N}{d}}\int_{\xi_i(t)+\epsilon}^{t}\mathbb{P}(v_{ik}>\overline{\eta}_i(s,t)-\delta)\mathrm{d}s.$$

Passing $\delta\downarrow 0$ and $\epsilon\downarrow 0$, we obtain because of continuity,

$$\lim_{n\to\infty}J_{2,i}^n=\frac{\lambda}{\binom{N}{d}}\int_{\xi_i(t)}^{t}\mathbb{P}(v_{ik}>\overline{\eta}_i(s,t))\mathrm{d}s. \tag{A.2}$$

We now determine $\lim_{n\to\infty}J_{1,i}^n$ and $\lim_{n\to\infty}I_i^n$. Fatou's lemma and the fact that $\xi_i(t)<t$ imply

$$\liminf_{n\to\infty}\overline{\eta}_i^n(0,t)\ge\int_{u=0}^{t}\liminf_{n\to\infty}\varphi(\overline{Q}_i^n(u))\mathrm{d}u=\overline{\eta}_i(0,t)=\infty. \tag{A.3}$$

We partition the interval $[0, \xi_i(t)]$ into N_2 subintervals $[s_{j-1}^{N_2}, s_j^{N_2}]$, $j = 1, \ldots, N_2$, for some integer $N_2 \geq 1$, in such a way that $\max_{j=1,\ldots,N_2}(s_j^{N_2} - s_{j-1}^{N_2}) \to 0$ as $N_2 \to \infty$. Suppose $s_{j-1}^{N_2} \leq U_{ik}^n \leq s_j^{N_2}$ for some $j \in \{1, \ldots, N_2\}$ and $k \in \{1, \ldots, A_i^n(\xi_i(t))\}$. It then follows from (A.3) that

$$\liminf_{n\to\infty} \overline{\eta}_i^n(U_{ik}^n, t) \geq \liminf_{n\to\infty} \overline{\eta}_i^n(s_j^{N_2}, t) \geq \overline{\eta}_i(s_j^{N_2}, t) = \infty. \tag{A.4}$$

For $J_{1,i}^n$ we have

$$0 \leq J_{1,i}^n \leq \frac{1}{n} \sum_{k=1}^{A_i^n(\xi_i(t))} \mathbb{1}_{\{v_{ik} > \overline{\eta}_i^n(U_{ik}^n, t)\}} + \frac{1}{n}(A_i^n(\xi_i(t) + \epsilon) - A_i^n(\xi_i(t))).$$

The first term on the right-hand side tends to 0 by (A.4), while the second term converges to $\frac{\lambda}{\binom{N}{d}}\epsilon$ according to Lemma 5.1 in [18]. Passing $\epsilon \downarrow 0$, we obtain

$$\lim_{n\to\infty} J_{1,i}^n = 0 = \frac{\lambda}{\binom{N}{d}} \int_{s=0}^{\xi_i(t)} \mathbb{P}(v_{ik} > \overline{\eta}_i(0, t))\mathrm{d}s. \tag{A.5}$$

The term I_i^n follows from (A.3):

$$\lim_{n\to\infty} \frac{1}{n} \sum_{l=1}^{n\overline{Q}_i^n(0)} \mathbb{1}_{\{v'_{il} > \overline{\eta}_i^n(0,t)\}} = 0 = \overline{Q}_i(0)\mathbb{P}(v'_{il} > \overline{\eta}_i(0, t)). \tag{A.6}$$

Taking the sum of (A.5), (A.2) and (A.6), yields the right-hand side of (4). The limit on the left-hand side is $\lim_{n\to\infty} \overline{Q}_{i,n}(t) = \overline{Q}_i(t)$. This proves (4) in case $\xi_i(t) < t$.

In case $\xi_i(t) = t$, Equation (4) immediately follows from the fact that $\overline{\eta}_i^n(s, t) \to \infty$ for any $s \in [0, t]$.

It remains to treat the case when $\overline{Q}_i(u) > 0$ for all $u \in [0, t]$. Then $\overline{\eta}_i^n(u, v)$ converges uniformly to $\overline{\eta}_i(u, v)$ for any $u, v \in [0, t]$, while the expression J_i^n follows by the same argument as used in $J_{2,i}^n$, on the interval $[\xi_i(t) + \epsilon, t]$. For any $\epsilon > 0$, there exists an n_ϵ such that $\overline{\eta}_i^n(0) \in (\overline{\eta}_i(0, t) - \epsilon, \overline{\eta}_i(0, t) + \epsilon)$ for all $n > n_\epsilon$. Multiplying and dividing I_i^n by $\overline{Q}_{i,n}(0)$, we deduce

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{l=1}^{n\overline{Q}_i^n(0)} \mathbb{1}_{\{v'_{il} > \overline{\eta}_i^n(0)\}} \geq \overline{Q}_i(0)\mathbb{P}(v'_{il} > \overline{\eta}_i(0, t) - \epsilon),$$

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{l=1}^{n\overline{Q}_i^n(0)} \mathbb{1}_{\{v'_{il} > \overline{\eta}_i^n(0)\}} \leq \overline{Q}_i(0)\mathbb{P}(v'_{il} > \overline{\eta}_i(0, t) + \epsilon).$$

Letting $\epsilon \downarrow 0$, we find that

$$\lim_{n\to\infty} I_i^n = \overline{Q}_i(0)\mathbb{P}(v'_{il} > \overline{\eta}_i(0, t)).$$

This completes the proof.

Appendix A.2. Distributions

In this paper we use the following job size distributions.

1. Deterministic: point mass at 2 (variance= 0)

2. Erlang2: sum of two exponential random variables with mean 1 (variance= 2)

3. Exponential: exponential distribution with mean 2 (variance= 4)

4. Bimodal-1: (mean= 2, variance= 9)

$$X = \begin{cases} 1 & w.p. \quad 0.9 \\ 11 & w.p. \quad 0.1 \end{cases}$$

5. Weibull-1: Weibull with shape parameter= 0.5 and scale parameter= 1 (heavy-tailed, mean= 2, variance=20)

6. Weibull-2: Weibull with shape parameter= $\frac{1}{3}$ and scale parameter= $\frac{1}{3}$ (heavy-tailed, mean= 2, variance=76)

7. Bimodal-2: (mean= 2, variance= 90)

$$X = \begin{cases} 1 & w.p. \quad 0.99 \\ 101 & w.p. \quad 0.01 \end{cases}$$

Appendix A.3. Simulation study near-insensitivity identical replicas

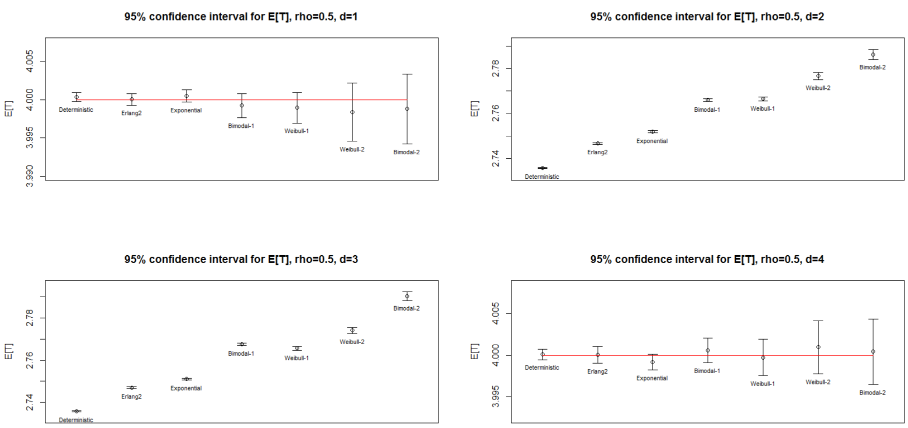

Figure A.7: Insensitivity results for various job size distributions for the setting $N = 4$, $\tilde{\rho} = 0.5$.

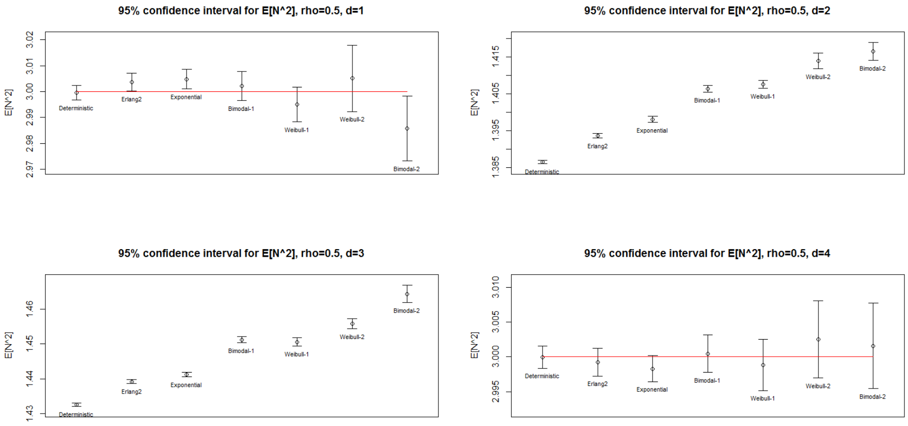

Figure A.8: Insensitivity results for various job size distributions for the setting $N = 4$, $\tilde{\rho} = 0.5$.

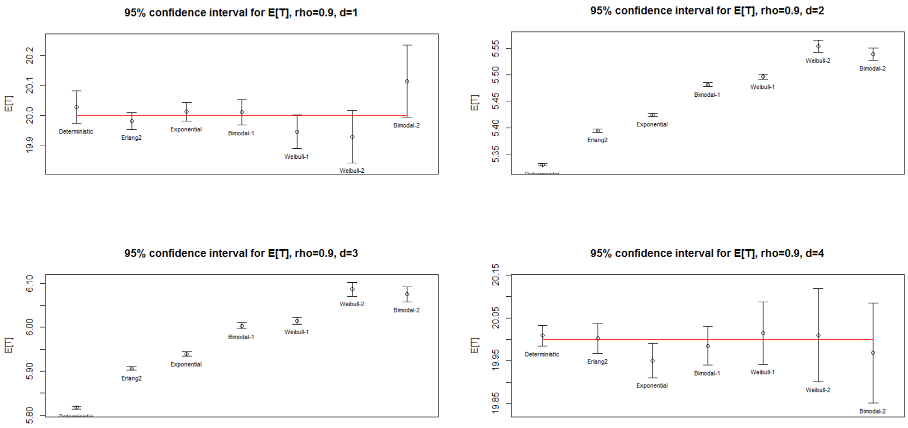

Figure A.9: Insensitivity results for various job size distributions for the setting $N = 4$, $\tilde{\rho} = 0.9$.

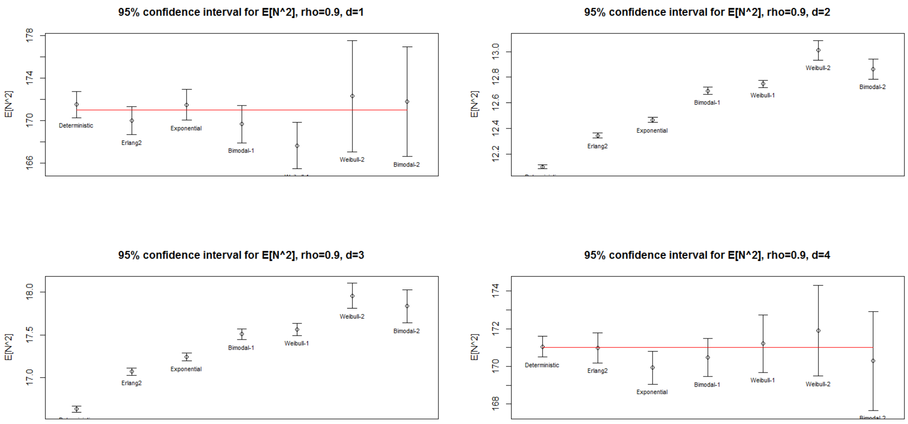

Figure A.10: Insensitivity results for various job size distributions for the setting $N = 4$, $\tilde{\rho} = 0.9$.