Analysis of Redundancy(d) with Identical Replicas

Tim Hellemans and Benny Van Houdt
University of Antwerp
Middelheimlaan 1, B2020 Antwerp, Belgium
tim.hellemans@uantwerpen.be
benny.vanhoudt@uantwerpen.be

ABSTRACT

Queueing systems with redundancy have received considerable attention recently. The idea of redundancy is to reduce latency by replicating each incoming job a number of times and to assign these replicas to a set of randomly selected servers. As soon as one replica completes service the remaining replicas are cancelled. Most prior work on queueing systems with redundancy assumes that the job durations of the different replicas are i.i.d., which yields insights that can be misleading for computer system design.

In this paper we develop a differential equation, using the cavity method, to assess the workload and response time distribution in a large homogeneous system with redundancy without the need to rely on this independence assumption. More specifically, we assume that the duration of each replica of a single job is identical across the servers and follows a general service time distribution.

Simulation results suggest that the differential equation yields exact results as the system size tends to infinity and can be used to study the stability of the system.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability;

KEYWORDS

Redundancy(d), Large Scale Computer Network, Fixed Point Iteration

ACM Reference format:
Tim Hellemans and Benny Van Houdt. 2018. Analysis of Redundancy(d) with Identical Replicas. In *Proceedings of IFIP Performance 2018, Toulouse, France, December 2018,* 6 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Redundancy is regarded as an effective technique to reduce latency in a variety of systems including large scale computer clusters [9]. The idea of redundancy is to create a number of replicas of each incoming job and to assign these replicas to a set of random servers. When the first of these replicas is processed by a server, the remaining replicas get canceled. An attractive feature of this scheme is that the replicas can be assigned immediately without the need to consult the server states or the need to maintain such information. Queueing models to study the effect of redundancy on the job response time have been introduced recently (e.g., [1, 6]). One of the key assumptions to enable their analysis often exists in assuming that the processing times of the replicas are independent and identically distributed (i.i.d.) across servers. While this may be applicable in some contexts, this assumption may result in misleading insights in a computer systems setting. For instance this i.i.d. assumption suggests that mean response time reduces as a function of the number of replicas (for sufficiently variable job sizes), while without such an assumption the mean response time may increase sharply if too many replicas are used.

In this short paper we present a fixed point equation, based on the cavity process, to assess the workload and response time distribution of a queueing model with redundancy when the processing times of the replicas are assumed to be *identical* across servers as opposed to assuming they are i.i.d.. Next, we rewrite this fixed point equation as an Integro-Differential Equation (IDE) in case the job sizes are continuously distributed (i.e. has no atoms) and a Delayed Differential Equation (DDE) in case the job sizes are deterministic. We conjecture that this IDE/DDE has a unique solution (when the queueing system is stable) that corresponds to the limit of the workload distribution as the number of servers tends to infinity. We propose a numerical scheme to solve the IDE/DDE and illustrate that its accuracy improves with the system size for various job size distributions (i.e., for bounded Pareto, (hyper)exponential and deterministic job sizes) using simulation.

The model considered in the paper is introduced in Section 2. The cavity process associated to this queueing system is presented in Section 3, while the IDE/ODE are derived in Section 4. Numerical results are found in Section 5 and Section 6 discusses some future work.

2 MODEL DESCRIPTION

We consider a system with N identical servers (for large N), each having an infinite waiting room. Arrivals occur according to a Poisson process with rate λN. Each incoming job is replicated d times and each replica joins a random server (in total d, distinct, random servers receive an identical arrival). As soon as one replica finishes service, the remaining replicas are canceled (whether in service or not). Cancellation is assumed to be immediate, although this assumption can be relaxed. It is important to stress that the processing times of the d replicas of a job are identical in our setting and *not* assumed to be i.i.d. as in [6]. The service discipline at each server is assumed to be first-come-first-served (FCFS) and jobs are processed at a constant rate 1. The job sizes are distributed with cumulative distribution function (cdf) $G(\cdot)$, complementary cdf

IFIP Performance 2018, December 2018, Toulouse, France
© 2018 Copyright held by the owner/author(s).
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

(ccdf) $\bar{G}(\cdot)$, probability density function (pdf) $g(\cdot)$ (if it exists) and mean $\mathbb{E}[G]$. We assume $G(0) = 0$. In what follows we will generally employ the notation: a capital letter for cdf, a capital letter with an overline for ccdf, a lowercase letter for pdf and $\mathbb{E}$ for expectation.

The model described corresponds to the Redundacy(d) model with identical replicas, we use the notation Red(d) to denote this model. Just as in [7] the corresponding Markov process only needs to keep track of the workload at each of the N queues. The model is stable if $\lambda\mathbb{E}[G] < 1/d$ and unstable for $\lambda\mathbb{E}[G] \geq 1$, its stability is however unclear for $\lambda\mathbb{E}[G] \in (1/d, 1)$.

3 CAVITY PROCESS

We now apply the cavity process methodology introduced in [3] to Red(d). The cavity process intends to capture the evolution of the workload of one queue for the limiting system when the number of servers $N \to \infty$.

Definition 3.1 (Red(d) cavity process). Let $\mathcal{H}(t)$, $t \geq 0$, be a set of probability measures on $\mathbb{R}$ called the *environment process*. The *cavity process* $X^{\mathcal{H}(\cdot)}(t)$, $t \geq 0$, takes values in $\mathbb{R}$ and is defined as follows. Potential arrivals occur according to a Poisson process with rate λd. When a potential arrival of size x occurs at time t, we compare $x + X^{\mathcal{H}(\cdot)}(t-)$, where $X^{H(\cdot)}(t-)$ is the state just prior to time t, with the minimum of $d-1$ independent random variables with law $x + \mathcal{H}(t)$ (call this minimum Y). The potential incoming job is then of size $y = \min\left\{x + X^{\mathcal{H}(\cdot)}(t-), Y\right\} - X^{\mathcal{H}(\cdot)}(t-)$ provided that $Y > X^{\mathcal{H}(\cdot)}(t-)$ and of size $y = 0$ otherwise. Next, we immediately add the job to the queue, that is, $X^{\mathcal{H}(\cdot)}(t) = X^{\mathcal{H}(\cdot)}(t-) + y$. The cavity process decreases at rate one during periods without arrivals and is lower bounded by zero.

We now define the cavity process associated to the equilibrium environment process, which is such that the cavity process has distribution $\mathcal{H}(t)$ at time t:

Definition 3.2 (Equilibrium Environment). When a cavity process $X^{\mathcal{H}(\cdot)}(\cdot)$ has distribution $\mathcal{H}(t)$ for all $t \geq 0$, we say that $\mathcal{H}(\cdot)$ is an *equilibrium environment process*. Further, a probability measure $\mathcal{H}$ is called an *equilibrium environment* if $\mathcal{H}(t) = \mathcal{H}$ for all t and $X^{\mathcal{H}(\cdot)}(t)$ has distribution $\mathcal{H}$ for all t.

The modularized program for analyzing load balancing systems presented in [3] when applied to Red(d) involves the following steps (assuming stability for N large):

a. Asymptotic Independence. Demonstrate $\Pi^N \to \Pi$ as $N \to \infty$, where Π^N is the stationary distribution for the Red(d) system with N queues and Π is a stationary and ergodic distribution on $[0,\infty)^\infty$. Show that the limit Π is unique, depending only on the service time distribution. Show that, for every k:

$$\Pi^{(k)} = \bigotimes_{i=1}^{k} \Pi^{(1)},$$

where $\Pi^{(k)}$ is Π restricted to its first k coordinates.

b. The queue at the cavity. Let $\mathcal{B}_s^N$ denote the arrival size distribution (which may be zero with a non-zero probability) in case of a potential arrival when the queue at the cavity has workload s. Show that the arrival process of a queue in the system of size N converges to a Poisson process with rate λd and a job size distribution $\mathcal{B}_s$ that depends on the workload s at arrival time. Denote $\mathcal{B} = \{\mathcal{B}_s, s \geq 0\}$.

c. Calculations. Given $\mathcal{B}$, the arrival size distributions, analyze the queue at the cavity in the large N limit using queueing techniques to express $\Pi^{(1)}$ as a function of $\mathcal{B}$:

$$\Pi^{(1)} = T(\mathcal{B}).$$

The arrival size distribution is determined by the workload distribution $\Pi^{(1)}$ (as explained above) we thus have:

$$\mathcal{B} = H(\Pi^{(1)}).$$

We then must solve these two fixed point equations to obtain the equilibrium environment $\Pi^{(1)} = \mathcal{H}$.

In this work, we focus on **c**, the computational step of the program. We present a numerical method to compute the Equilibrium Environment $\mathcal{H}$ corresponding to Red(d) and validate it with simulation. Therefore we conjecture (numerical evidence to support this conjecture is presented in Section 5.1):

Conjecture 3.3. *Consider a load balancing system operating under the* Red(d) *policy on N servers, assume λ, d and G are such that this system is uniformly stable for sufficiently large N and the local service is FCFS. Then, in the large N limit, there is a unique equilibrium distribution. Under this distribution, any finite number of queues are independent. Moreover, this equilibrium can be found as the unique fixed point in step* **c**.

We now characterize the evolution of the cavity process associated with the equilibrium environment process. Let $f(t,s), t \in [0,\infty), s \in (0,\infty)$ describe the density at which a random server, at time t, has workload $s > 0$. Note that $f(t,\cdot)$ is not a real pdf as the probability that the server is empty is non-zero. Let $F(t,s) = F(t,0) + \int_0^s f(t,u)du$ denote the cdf of the workload of a random server, here $F(t,0) = 1 - \int_0^\infty f(t,s)$ is the probability that a random server is idle.

We define $c_d(t,s,r)$ as the double density that, if a potential arrival occurs at time t, the queue at the cavity has workload $s > 0$ and its workload is increased to $r > s$ by the potential arrival. Lastly we let $C_d(t,r)$ denote the density at which, if a potential arrival occurs at time t, the queue at the cavity has workload 0 and its workload is increased to $r > 0$.

We now obtain a partial IDE (PIDE) which describes the transient evolution of the cavity queue in function of c_d, C_d in a similar fashion as in [7].

Theorem 3.4. *The evolution of the cavity process associated to the equilibrium environment process of the Red(d) model is captured by the following set of equations:*

$$\frac{\partial f(t,s)}{\partial t} - \frac{\partial f(t,s)}{\partial s} = \lambda d \cdot \left(-\int_s^\infty c_d(t,s,r)dr + C_d(t,s) + \int_0^s c_d(t,u,s)du \right) \quad (1)$$

$$\frac{\partial F(t,0)}{\partial t} = -\lambda d F(t,0) + f(t,0^+), \quad (2)$$

for $s > 0$, where $f(x,z^+) = \lim_{y \downarrow z} f(x,y)$.

Proof. We first let $t, s > 0$ and $0 < \Delta < s$ be arbitrary. We now describe the possible evolution of the workload of the queue at the cavity in the interval $[t, t+\Delta]$ s.t. it has exactly workload s at time $t+\Delta$. We write:

$$f(t+\Delta, s) = Q_1 + Q_2 + Q_3 + o(\Delta), \tag{3}$$

and now describe how to obtain these Q_i.

(Q_1) First, we consider the case where the queue at the cavity has $s+\Delta$ work at time t and no potential arrivals in $[t, t+\Delta]$ make its workload increase. For this case we find:

$$Q_1 = f(t, s+\Delta) - \lambda d \int_0^{\Delta} \int_{s+\Delta-v}^{\infty} c_d(t+v, s+\Delta-v, r) dr dv.$$

(Q_2) Second, we consider the case in which the queue at the cavity is empty at time $t+v, v \in [0, \Delta]$ and its workload is increased to $s+(\Delta-v)$ by a potential arrival. This happens with density:

$$Q_2 = \lambda d \int_0^{\Delta} C_d(t+v, s+(\Delta-v)) dv.$$

(Q_3) Lastly, the queue at the cavity may be non-empty at time $t+v, v \in [0, \Delta]$ and its workload increases to $s+(\Delta-v)$ by a potential arrival. This case has density:

$$Q_3 = \lambda d \int_0^{\Delta} \int_v^{s+\Delta} c_d(t+v, u-v, s+(\Delta-v)) du dv.$$

We find from subtracting $f(t, s+\Delta)$, dividing by Δ and taking the limit $\Delta \longrightarrow 0$ on both sides of (3) that (1) indeed holds.

We have not yet considered the case $s = 0$, for this we need to consider which events on $[t, t+\Delta]$ result in the workload of the queue at the cavity to be 0 at time $t+\Delta$. To this end one readily shows:

$$F(t+\Delta, 0) = F(t, 0)(1 - \lambda d \Delta) + \int_0^{\Delta} f(t+v, \Delta-v) dv + o(\Delta).$$

Subtracting $F(t, 0)$, dividing by Δ and taking the limit $\Delta \longrightarrow 0$ on both sides results in (2). □

Remark. *The PIDE found in Theorem 3.4 could alternatively have been derived using the generalized Master Equation given in [8], (7.25-7.26).*

We still require an exact expression for c_d and C_d. Moreover, we need an efficient method to compute the quantities $\int_s^{\infty} c_d(t, s, r) \, dr$ and $\int_0^s c_d(t, u, s) \, du$. Therefore, in the next proposition, we describe how to determine c_d, C_d, where (4-6) are valid for general job size distributions and the latter three equalities hold for continuous job size distributions only (i.e., assuming G has a pdf g).

Proposition 3.5. *We have $c_d(t, s, r) = c_{d,1}(t, s, r) + c_{d,2}(t, s, r) + c_{d,3}(t, s, r)$ such that:*

$$\int_s^{\infty} c_{d,1}(t, s, r) dr = \bar{G}(s) f(t, s)(1 - \bar{F}(t, 0)^{d-1}) \tag{4}$$

$$\int_s^{\infty} c_{d,2}(t, s, r) dr = f(t, s)\bar{F}(t, s)^{d-1} \tag{5}$$

$$\int_s^{\infty} c_{d,3}(t, s, r) dr = (d-1) f(t, s) \left(\bar{F}(t, \cdot)^{d-2} f(t, \cdot) * \bar{G}(\cdot)\right)(s) \tag{6}$$

$$\begin{aligned}
\int_0^s c_{d,1}(t, u, s) du &= g(s) \cdot (F(t, s) - F(t, 0))(1 - \bar{F}(t, 0)^{d-1}) \\
\int_0^s c_{d,2}(t, u, s) du &= \left(g(\cdot) * f(t, \cdot)\bar{F}(t, \cdot)^{d-1}\right)(s) \\
\int_0^s c_{d,3}(t, u, s) du &= (d-1) F(t, s) \cdot \left(g(\cdot) * f(t, \cdot) \cdot \bar{F}(t, \cdot)^{d-2}\right)(s) \\
&\quad - (d-1)\left(g(\cdot) * F(t, \cdot) f(t, \cdot)\bar{F}(t, \cdot)^{d-2}\right)(s),
\end{aligned} \tag{7}$$

where $(f_1 * f_2)(s) = \int_0^s f_1(u) f_2(s-u) du$ denotes the convolution product. These quantities can all be computed quickly which simplifies solving (1-2) significantly. Lastly, we have $C_d(t, s) = F(t, 0) \cdot g(s)$.

Proof. First we define $c_{d,1}, c_{d,2}$ and $c_{d,3}$ as follows:

- At least one of the $d-1$ independent random variables with law $\mathcal{H}(t)$ is zero and the incoming job has size r. We find (for $s < r$):
$$c_{d,1}(t, s, r) = g(r) f(t, s)(1 - \bar{F}(t, 0)^{d-1}).$$
- The queue at the cavity is the queue with the minimal workload (i.e. s) and the size of the arrival is exactly $r - s$:
$$c_{d,2}(t, s, r) = g(r-s) f(t, s)\bar{F}(t, s)^{d-1}.$$
- The queue with minimal workload has $0 < u < s$ workload, where s is the workload of the queue at the cavity, and the arrival size is $r - u$:
$$c_{d,3}(t, s, r) = (d-1) f(t, s) \int_0^s g(r-u)\bar{F}(t, u)^{d-2} f(t, u) du.$$

now the claimed equalities all follow from direct computation and applying Fubini (which is allowed as all integrands are positive functions). It is trivial to derive the expression for C_d. □

Remark. *One can readily employ the strategy used in the proof of Proposition 3.5 to get a similar result for deterministic job sizes.*

The PIDE (1-2) can now be solved using an (improved) Euler scheme. This result is also of interest to obtain a fixed point equation for the equilibrium environment, i.e., workload distribution. In the subsequent section, we provide an efficient method to compute the equilibrium workload (and thus also response time) distribution.

4 EQUILIBRIUM REGIME

For the equilibrium we use the same notations as in the transient case, but we leave out the time dependence (e.g., we write $f(s)$ instead of $f(t, s)$) and set $\frac{\partial f(s)}{\partial t} = 0$. We now find from (1-2) a fixed point equation for $\bar{F}$, this is our first main result and is applicable for a general job size distribution.

Theorem 4.1. *The stationary workload distribution associated to an equilibrium environment satisfies the following fixed point equation:*

$$\bar{F}(s) = \bar{F}(0) + \lambda d \cdot \Bigg[\int_0^s \bar{G}(u) \left(\bar{F}(u)(1 - \bar{F}(0)^{d-1}) - (1 - \bar{F}(0)^d) \right) + (d-1)\bar{F}(u)(\bar{G} * f\bar{F}^{d-2})(u) - d(\bar{G} * f\bar{F}^{d-1})(u)\, du \Bigg] \tag{8}$$

Proof. Integrating (1-2) once, we find:

$$\begin{aligned} f(s) &= \lambda d \bigg(F(0) - \int_0^s C_d(u) du + \int_0^s \int_u^\infty c_d(u,r)\, dr\, du \\ &\quad - \int_0^s \int_0^u c_d(v,u)\, dv\, du \bigg) \\ &= \lambda d \left(F(0)\bar{G}(s) + \int_0^s \int_s^\infty c_d(u,v) dv du \right). \end{aligned}$$

Note that as in [2] the left hand side of this equality corresponds to the down-crossing rate through s and the right hand side corresponds to the up-crossing rate through s. Using (4-6) and integrating once more, we find the claimed equality (8). □

We now show that:

- If the job size distribution is continuous, the fixed point equation (8) can be written as an IDE.
- For deterministic job sizes, the fixed point equation (8) simplifies significantly and we obtain $\bar{F}$ as the solution of a simple DDE.

First we consider the case of continuous job sizes, this is our second main result:

Theorem 4.2. *The stationary workload distribution associated to the equilibrium environment satisfies the following IDE:*

$$\begin{aligned} \bar{F}'(s) = -\lambda d \bigg(&\bar{G}(s)(1 - \bar{F}(s)) \\ &+ \int_0^s g(u)\bar{F}^{d-1}(s-u)(\bar{F}(s-u) - \bar{F}(s))\, du \bigg). \end{aligned} \tag{9}$$

Proof. The claimed equality follows from differentiating (8) once and using integration by parts on $(\bar{G} * f\bar{F}^{d-2})$ and $(\bar{G} * f\bar{F}^{d-1})$. □

We have the following result for deterministic job sizes, this is our third (and last) main result:

Theorem 4.3. *Assume job sizes are deterministic of size* 1*, the stationary workload distribution associated to the equilibrium environment satisfies the following DDE:*

$$\bar{F}'(s) = \lambda d \cdot (\bar{F}(s) - 1) \qquad s \leq 1 \tag{10}$$

$$\bar{F}'(s) = \lambda d \cdot (\bar{F}(s) - \bar{F}(s-1))\bar{F}(s-1)^{d-1} \qquad s > 1. \tag{11}$$

Proof. In this case we have $\bar{G}(s) = 1$ if $s \leq 1$ and 0 otherwise. Substituting this into (8) and differentiating once, we obtain (10-11) □

Remark. *There is a striking resemblance between this DDE and the DDE presented in [7] for the stationary workload distribution in case of a Least Loaded policy. There we had the DDE:*

$$\bar{F}'(s) = \lambda \cdot (\bar{F}(s) - 1) \qquad s \leq 1$$

$$\bar{F}'(s) = \lambda \cdot (\bar{F}(s)^d - \bar{F}(s-1)^d) \qquad s > 1.$$

Note that Theorems 4.2 and 4.3 do not specify the boundary condition for $\bar{F}(0)$. This is not surprising as $\bar{F}(0)$ corresponds to the unknown actual system load (and exceeds $\lambda\mathbb{E}[G]$ as multiple replicas can be executed simultaneously). We can however simply look for the value $\bar{F}_0$ such that if we take $\bar{F}(0) = \bar{F}_0$, the solution of the associated IDE/DDE satisfies $\lim_{s\to\infty} \bar{F}(s) = 0$. Our numerical experiments suggest that a simple bisection algorithm can be used in order to find the value of $\bar{F}_0$ which has the desired property.

5 NUMERICAL EXPERIMENTS

In this section we use Theorem 4.2 and Theorem 4.3 to find the limiting workload distribution. We repeatedly solve the IDE/DDE with different initial conditions $\bar{F}(0)$ until we find an $\bar{F}(0)$ that satisfies $\lim_{s\to\infty} \bar{F}(s) = 0$ (up to an accuracy of 10^{-6}) using a bisection algorithm on $(\lambda\mathbb{E}[G], 1)$.

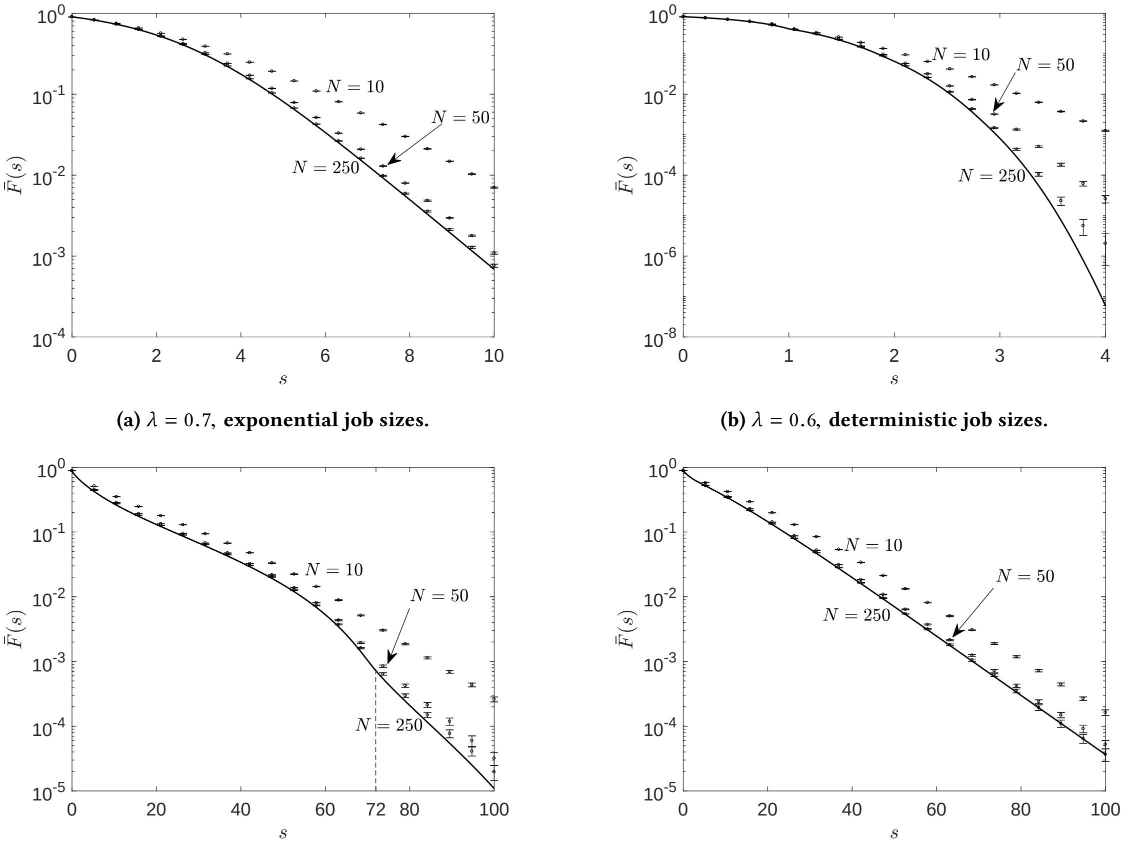
Throughout this section we will consider 4 job size distributions: exponential job sizes with mean one, deterministic job sizes equal to one, bounded Pareto job sizes with lower bound 0.2, upper bound 72 and $\alpha = 1.1$ (meaning, $\mathbb{E}[G] = 1$ and $\mathbb{E}[G^2] = 10$) and hyperexponential job sizes with two phases and balanced means, chosen such that $\mathbb{E}[G] = 1$ and $\mathbb{E}[G^2] = 10$.

5.1 Finite System Accuracy

We compare the equilibrium workload distribution with the simulated workload distribution for finite N. All simulation runs simulate the system up to time $t = 10^7/N$ and use a warm-up period of 30%. We simulate a system of $N = 10, 50, 250$ servers. In Figure 1 we see that as N increases the approximation provided by the IDE/DDE becomes more accurate (which supports Conjecture 3.3). Note that a similar figure can easily be made for the response time distribution by noting that the response time is given by $X + \min_{i=1}^d U_i$, where the cdf of X is G and U_i are i.i.d. with cdf F.

5.2 Performance of Redundancy(d)

This section is intended to illustrate the usefulness of our IDE/DDE, it is not intended as a detailed study of the performance of the Redundancy(d) policy. We show the actual workload $\bar{F}(0)$ and the mean response time $1 + \int_0^\infty \bar{F}(s)^d\, ds$ of the Red(d) policy in Figures 2 and 3 as a function of the arrival rate λ (recall $\mathbb{E}[G] = 1$). From Figure 2a, it is clear that the stability region not only depends on the mean and the variance of the job size distribution, but also on higher moments (as $\mathbb{E}[G^2] = 10$ for both the Bounded Pareto and hyperexponential). This makes the question of stability for Red(d) for general job size distributions a hard problem (which in turn makes proving Conjecture 3.3 hard). We can infer from the plot that the more variable the job size distribution, the lower the associated workload. From Figure 2b, it is obvious that $\lambda_{\max}$ (defined as the supremum of the arrival rates λ for which $\bar{F}(0) < 1$) decreases and the workload increases as a function of d (we have numerically



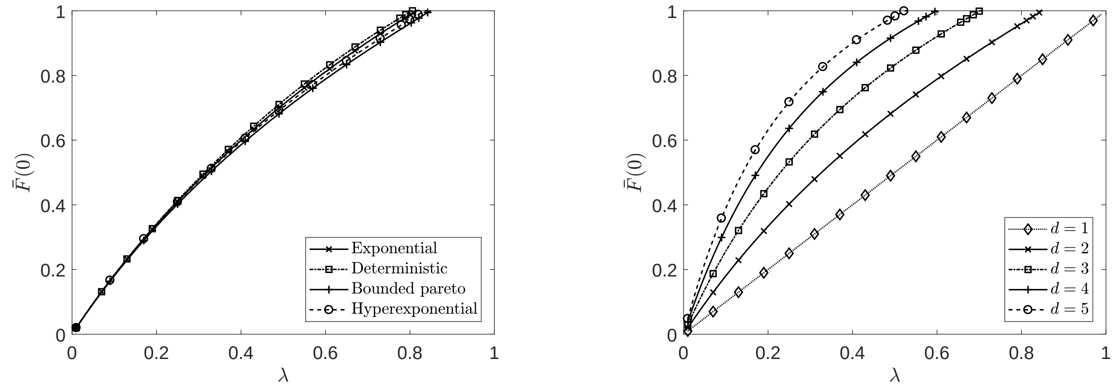
(a) $\lambda = 0.7$, **exponential job sizes.**

(b) $\lambda = 0.6$, **deterministic job sizes.**

(c) $\lambda = 0.7$, **bounded Pareto job sizes (max=72).**

(d) $\lambda = 0.7$, **hyperexponential job sizes.**

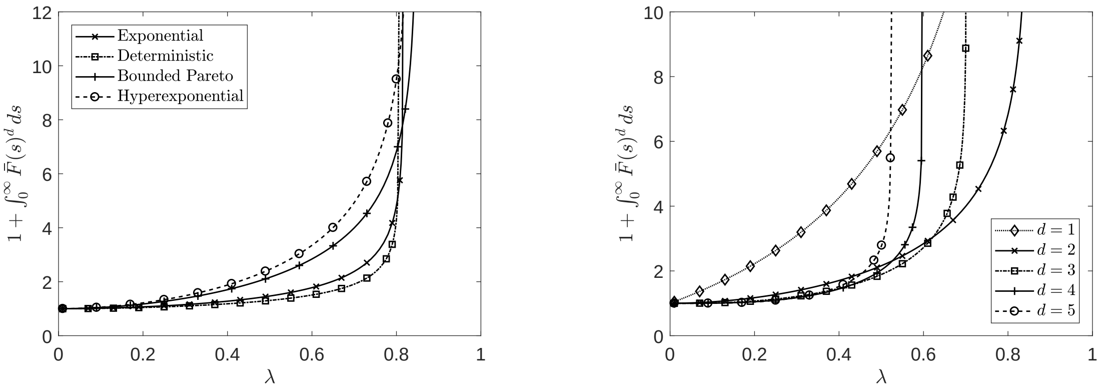
Figure 1: Limiting workload distribution vs. simulation for N servers with exponential, deterministic, bounded Pareto and hyperexponential job sizes. The full line represents the solution of the IDE/DDE, which is compared with the simulated 95% confidence intervals.



(a) $d = 2$ **and different job size distributions.**

(b) $d = 2, 3, 4, 5$ **and bounded Pareto job sizes.**

Figure 2: Workload $\bar{F}(0)$ in function of the arrival rate λ.



(a) $d = 2$ **and different job size distributions.**

(b) $d = 2, 3, 4, 5$ **and Bounded Pareto job sizes.**

Figure 3: Mean response time $\left(1 + \int_0^\infty \bar{F}(s)^d ds\right)$ **as a function of the arrival rate** λ**.**

verified that this also holds for the other job size distributions considered).

We show in Figure 3a that, despite the fact that the workload for the less variable jobs is consistently higher than that of the more variable ones, the same does not hold for the response times. We see that adding variability to the job size distribution also increases the mean response time (for λ sufficiently bounded away from instability). From Figure 3b it is clear that only for small values of λ there is a reduction in response time by increasing d: this reduction is due to the fact that for small arrival rates a job is more likely to find an idle server by increasing d, but as λ increases higher values of d cause too much extra load on the servers which causes an increased response time.

6 FUTURE WORK

An important generalization is to look at the *S&X* model of [5]. Our model corresponds to the *S&X* model with no slowdown (i.e., $S = 1$), which implies that the replica that starts execution first also finishes first. As such it is always better to cancel the other replicas as soon as one starts execution. However, with the *S&X* model different replicas may experience different slowdowns and cancellation-on-start may no longer be superior. It is not hard to obtain general expressions for $c_d(t, s, r)$ and $C_d(t, r)$ for the *S&X* model, which should lead to a similar differential equation with unknown boundary condition.

Proving Conjecture 3.3 would give a theoretical basis for the analysis provided here (as was done for other load balancing schemes in [4]). We note that this is also an open problem for the Redundancy(d) with i.d.d. replicas considered in [6].

It might be possible to explicitly solve the IDE (9) for certain job size distributions.

REFERENCES

[1] U. Ayesta, T. Bodas, and I. M. Verloop. 2018. On a unifying product form framework for redundancy models. (May 2018). https://hal.archives-ouvertes.fr/hal-01713937 Accepted at Performance 18, Toulouse, France.

[2] R. Bekker, S. Borst, O. Boxma, and O. Kella. 2004. Queues with workload-dependent arrival and service rates. *Queueing Systems* 46, 3-4 (2004), 537–556.

[3] M. Bramson, Y. Lu, and B. Prabhakar. 2010. Randomized load balancing with general service time distributions. In *ACM SIGMETRICS 2010*. 275–286. https://doi.org/10.1145/1811039.1811071

[4] M. Bramson, Y. Lu, and B. Prabhakar. 2012. Asymptotic independence of queues under randomized load balancing. *Queueing Systems* 71, 3 (2012), 247–292.

[5] K. Gardner, M. Harchol-Balter, A. Scheller-Wolf, and B. Van Houdt. 2017. A Better Model for Job Redundancy: Decoupling Server Slowdown and Job Size. *IEEE/ACM Trans. Netw.* 25, 6 (Dec. 2017), 3353–3367. https://doi.org/10.1109/TNET.2017.2744607

[6] K. Gardner, M. Harchol-Balter, A. Scheller-Wolf, M. Velednitsky, and S. Zbarsky. 2017. Redundancy-d: The Power of d Choices for Redundancy. *Operations Research* 65, 4 (2017), 1078–1094. https://doi.org/10.1287/opre.2016.1582

[7] T. Hellemans and B. Van Houdt. 2018. On the Power-of-d-choices with Least Loaded Server Selection. *Proc. ACM Meas. Anal. Comput. Syst.* 2 (2018), Article No. 27. Issue 2.

[8] Z. Schuss. 2009. *Theory and applications of stochastic processes: an analytical approach*. Vol. 170. Springer Science & Business Media.

[9] N. B. Shah, K. Lee, and K. Ramchandran. 2016. When Do Redundant Requests Reduce Latency? *IEEE Transactions on Communications* 64, 2 (Feb 2016), 715–722. https://doi.org/10.1109/TCOMM.2015.2506161