On the Power-of-d-choices with Least Loaded Server Selection

TIM HELLEMANS* and BENNY VAN HOUDT, University of Antwerp, Belgium

Motivated by distributed schedulers that combine the power-of-d-choices with late binding and systems that use replication with cancellation-on-start, we study the performance of the LL(d) policy which assigns a job to a server that currently has the least workload among d randomly selected servers in large-scale homogeneous clusters.

We consider general job size distributions and propose a partial integro-differential equation to describe the evolution of the system. This equation relies on the earlier proven ansatz for LL(d) which asserts that the workload distribution of any finite set of queues becomes independent of one another as the number of servers tends to infinity. Based on this equation we propose a fixed point iteration for the limiting workload distribution and study its convergence.

For exponential job sizes we present a simple closed form expression for the limiting workload distribution that is valid for any work-conserving service discipline as well as for the limiting response time distribution in case of first-come-first-served scheduling. We further show that for phase-type distributed job sizes the limiting workload and response time distribution can be expressed via the unique solution of a simple set of ordinary differential equations.

Numerical and analytical results that compare response time of the classic power-of-d-choices algorithm and the LL(d) policy are also presented and the accuracy of the limiting response time distribution for finite systems is illustrated using simulations.

CCS Concepts: • **Mathematics of computing → Probability and statistics**; • **Networks → Network performance modeling**;

Additional Key Words and Phrases: Scheduling; redundancy; late binding; large-scale systems

1 INTRODUCTION

Load balancing plays a crucial role in achieving low latency in large-scale clusters. A simple randomized approach, denoted as SQ(d), exists in assigning incoming jobs to a server that currently holds the fewest number of jobs among a set of d randomly selected servers, the so-called *power-of-d-choices* algorithm [2, 17, 18, 22]. While this approach yields short queues with high probability in case of first-come-first-served (FCFS) scheduling even for general job size distributions provided that d is chosen sufficiently large [6, 8], short queues do not guarantee low latency as the queue length is only a coarse indicator of the waiting time in the presence of high job size variability.

Authors' address: Tim Hellemans; Benny Van Houdt, University of Antwerp, Middelheimlaan 1, Antwerp, B-2020, Belgium, tim.hellemans@uantwerpen.be, benny.vanhoudt@uantwerpen.be.

The main issue is that under the FCFS discipline short jobs can get stuck behind a single long job which significantly increases the short job latency. In addition when multiple dispatchers are used to distribute the jobs, *race* conditions may occur where multiple schedulers concurrently place jobs on a server that appears lightly loaded [16].

To avoid these issues the notion of *late binding* was recently introduced in [19]. With late binding the dispatcher still probes d servers at random, but the servers do not immediately reply by sending their queue length information. Instead they place a reservation at the end of a local work queue and when the reservation reaches the front of the queue, the server requests the job associated to the reservation from the dispatcher. In this manner the job is assigned to the server that is able to launch the job the soonest among the d randomly selected servers. The downside of late binding is that the server always experiences some idle time in between the execution of two jobs, which implies some efficiency loss. However, whenever the network latencies are much smaller than the shortest job runtimes (and the system load is not extremely high), experiments on a 110-machine cluster show that a scheduler that relies on late binding performs close to an ideal scheduler [19].

Note that late binding as described above is equivalent to assigning the job to the server that has the least workload among d randomly selected servers, which is known as the LL(d) policy [7], provided that the network latencies are negligible[1].

The main objective of this paper is to study the large-scale limit of the server workload and response time distribution of the LL(d) policy when employed on a homogeneous cluster subject to Poisson job arrivals with general service times. For this purpose we introduce a partial integro-differential equation that captures the evolution of the so-called *cavity process* and study its equilibrium. The key observation, established in [7], is that under the LL(d) policy with general service time distributions, the workload distribution of any finite set of servers becomes asymptotically independent as the number of servers tends to infinity (provided that all the servers employ the same local non-idling service discipline, e.g., FCFS, PS, etc.). Moreover, the limit of the marginal workload distribution of a server corresponds to the unique equilibrium environment.

It is worth noting that the LL(d) policy is equivalent to the following system that uses replication with cancellation-on-start to reduce waiting times. Arriving jobs are replicated d times and are randomly assigned to d servers (that all operate in FCFS order). As soon as a single replica starts execution on a server, the remaining $d - 1$ replicas are killed (with the additional assumption that if multiple replicas start at exactly the same time, only one is executed). Prior work on replication was mainly done in the context of systems that experience server slowdown and therefore focused on replication with cancellation-on-job-completion [11, 12], which is considerably different from LL(d) as jobs are often (partially) executed on multiple servers in such case.

Another reason for studying the large-scale limit of the LL(d) policy exists in understanding how much benefit precise workload information gives in comparison to the coarser queue length information used by SQ(d).

The main contributions of the paper are as follows:

(1) A partial integro-differential equation to describe the transient evolution of the queue at the cavity under the LL(d) policy is derived.
(2) An integral equation for the limiting stationary workload distribution is presented together with a fixed-point iteration to compute its solution. Convergence of the fixed-point iteration is proven for system loads below $e^{-1/e} \approx 0.6922$.
(3) A simple explicit solution for the limiting workload and response time distribution is presented in case of exponential job sizes. For phase-type distributed job sizes we prove that the limiting

---

[1]When the network latencies are not negligible compared to the job runtimes, we can regard them as part of the workload of a job such that the job execution consists of two parts: fetching the job and executing it, see Section 7.3.

workload distribution can be computed easily by solving a simple set of ordinary differential equations.

(4) We present both analytical and numerical results that compare the response time of the LL(d) policy with the classic SQ(d) policy. These results illustrate that late binding offers a significant reduction in the response time under a very wide range of loads even when taking the idleness caused by late binding into account.

The paper is structured as follows. The model considered in this paper is described in Section 2. The partial integro-differential equation that captures the transient evolution of the workload is introduced in Section 3, while the integral equation for the limiting stationary workload and its associated fixed point equation are presented in Section 4. Sections 5 and 6 discuss the special cases of exponential and phase-type distributed job sizes, respectively. Section 7 compares the performance of the LL(d) and SQ(d) policies, while Section 8 briefly studies the accuracy of the limiting distributions for systems of finite size. Conclusions are drawn and future work is suggested in Section 9.

2 MODEL DESCRIPTION

We consider a system consisting of N single server queues each having an infinite waiting room. Arrivals occur into the system as a Poisson process with rate λN. For each incoming job d queues are selected uniformly at random (with replacement) and the job joins the queue that currently holds the least workload with ties being broken uniformly at random. The service discipline is such that the workload at any queue reduces at rate 1 when positive, that is, we do not put any restriction on the service discipline apart from the fact that it is non-idling and identical in each server (unless stated otherwise). The workload offered by a job has a general distribution with cdf $G(\cdot)$, pdf $g(\cdot)$, mean $\mathbb{E}[G]$ and is such that $G(0) = 0$. We define $\rho = \lambda \mathbb{E}[G]$ and assume that $\rho < 1$.

The above model corresponds to the so-called least-loaded supermarket model, denoted as LL(d) in [6, 7]. Note that the corresponding Markov process that keeps track of the workloads of the N queues is positive Harris recurrent and has a unique stationary probability measure $\mathcal{E}^{(N)}$ whenever the queueing system is subcritical, that is, when $\rho < 1$, as noted at the end of Section 5 in [5]. In fact, this result is a special case of [10, Theorem 2.5].

3 CAVITY PROCESS

We start by introducing the cavity process from [6, 7] for the LL(d) supermarket model. The process is intended to capture the evolution of the workload of a single queue for the limiting system where the number of servers N tends to infinity.

Definition 3.1 (LL(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 occurs at time t, we compare the state $X^{\mathcal{H}(\cdot)}(t-)$ just prior to time t with the states of $d-1$ independent random variables with law $\mathcal{H}(t)$. The potential incoming job is assigned to the state among these d states that has the lowest value, where ties are broken uniformly at random. If the job is assigned to state $X^{\mathcal{H}(\cdot)}(t-)$, we immediately add the job to the queue, that is, $X^{\mathcal{H}(\cdot)}(t) = X^{\mathcal{H}(\cdot)}(t-) + x$ where x is the size of the incoming job. Otherwise, the job immediately leaves the system, i.e., $X^{\mathcal{H}(\cdot)}(t) = X^{\mathcal{H}(\cdot)}(t-)$. Clearly, if $X^{\mathcal{H}(\cdot)}(t-)$ has law $\mathcal{H}(t)$ a potential arrival at time t joins the queue with probability $1/d$. Finally, the cavity process decreases at rate one during periods without arrivals and is lower bounded by zero.

Definition 3.2 (Equilibrium Environment Process). 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*.

Definition 3.3 (Equilibrium Environment). 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.

Theorem 3.4 (due to Theorem 2.2 of [7]). *Consider the LL(d) supermarket model with N queues, general service times (with mean $\mathbb{E}[G]$), Poisson arrivals with rate $\lambda N < N/\mathbb{E}[G]$ and an identical non-idling service discipline at each queue. Let $\mathcal{E}^{(N,N')}$ be the projection of the stationary measure $\mathcal{E}^{(N)}$ of the N workloads into the workloads of the first N' queues, then $\mathcal{E}^{(N,N')}$ converges in total variation to the N'-fold convolution of $\mathcal{E}^{(\infty,1)}$ (in an appropriate metric space) as N tends to infinity. Moreover, $\mathcal{E}^{(\infty,1)}$ is the unique equilibrium environment of the LL(d) supermarket model.*

In other words the above theorem indicates that the workload distributions of any finite set of N' queues becomes asymptotically independent as N tends to infinity and the marginal workload distribution of any queue is given by the *unique* equilibrium environment $\mathcal{H}$ of the LL(d) supermarket model. Thus, there is no need to interchange the limits when relying on Theorem 3.4 when studying the limit of the stationary distributions. We do note that the order of these limits can in fact be reversed (when starting from an empty system) as shown in [7, Section 8].

In this section we characterize the evolution of the cavity process associated with the equilibrium environment process $\mathcal{H}(\cdot)$ of the LL(d) supermarket model. This derivation is based on definitions 3.1 and 3.2 only and does not make use of Theorem 3.4, which is leveraged in the next section when studying the limiting stationary workload distribution.

Let $f(t, s)$ for $s \in (0, \infty)$ describe the density of servers which, at time t, have workload s. Note that $f(t, \cdot)$ is not a real probability density function (pdf) as some of the servers may be idle, denote $F(t, 0) = 1 - \int_0^\infty f(t, s)ds$ (where $f(t, 0)$ may be defined arbitrarily). In the following we will refer to $f(t, \cdot)$ as a density, and we define its cumulative distribution function (cdf) $F(t, \cdot)$ as $F(t, s) = F(t, 0) + \int_0^s f(t, u)du$.

For any $d \in \{2, 3, \dots\}$, we define the function $c_d(t, u)$ as the density at which a potential arrival at time t joins the cavity queue with workload $u > 0$. By definition of the cavity process associated to the equilibrium environment, this density is given by:

$$c_d(t, u) = f(t, u)(1 - F(t, u))^{d-1} = f(t, u)\bar{F}(t, u)^{d-1}, \tag{1}$$

where we use the notation $\bar{F}(t, u) = 1 - F(t, u)$ for the complementary cdf (ccdf). We further denote the probability that a potential arrival at time t joins the cavity queue with workload at most u by $C_d(t, u)$. In this case we have, as ties are broken uniformly at random:

$$C_d(t, u) = F(t, 0) \sum_{k=0}^{d-1} \binom{d-1}{k} \frac{F(t, 0)^k \bar{F}(t, 0)^{d-1-k}}{k+1} + \int_{v=0}^u c_d(t, v)dv$$

$$= \frac{1 - \bar{F}(t, 0)^d}{d} + \int_{v=0}^u c_d(t, v)dv = \frac{1 - \bar{F}(t, u)^d}{d}. \tag{2}$$

In particular, $C_d(t, 0)$ is the probability that a potential arrival joins an empty cavity queue.

Theorem 3.5. *The evolution of the cavity process associated to the equilibrium environment process of the LL(d) supermarket model is captured by the following set of equations:*

$$\frac{\partial f(t, s)}{\partial t} - \frac{\partial f(t, s)}{\partial s} = \lambda d \int_0^s c_d(t, u)g(s - u)du + \lambda d C_d(t, 0)g(s) - \lambda d c_d(t, s) \tag{3}$$

$$\frac{\partial F(t, 0)}{\partial t} = f(t, 0^+) - \lambda d C_d(t, 0), \tag{4}$$

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

Proof. Assume $s > 0$ and let $s > \Delta > 0$ be arbitrary. In order to have a workload of s at time $t + \Delta$ we need to consider three possible cases: no arrivals in $[t, t + \Delta]$, an arrival occurs in $[t, t + \Delta]$ when the workload is non-zero and an arrival occurs in an idle server in $[t, t + \Delta]$. Hence, we can write

$$f(t + \Delta, s) = Q_1 + Q_2 + Q_3. \tag{5}$$

The terms Q_i, for $i = 1, 2$ and 3 are discussed next.

1) No arrivals in the interval $[t, t + \Delta]$: if the cavity queue at time t has a workload exactly equal to $s + \Delta$ and has no arrivals in $[t, t + \Delta]$, it will have a workload equal to s at time $t + \Delta$. The density of having a workload $s + \Delta$ at time t is given by $f(t, s + \Delta)$ and the density at which an arrival occurs at the cavity queue at time $t + v, v \in [0, \Delta]$, when it has workload $s + \Delta - v$, is equal to $\lambda d c_d(t + v, s + \Delta - v)$. Therefore we find:

$$Q_1 = f(t, s + \Delta) - \lambda d \int_{v=0}^{\Delta} c_d(t + v, s + \Delta - v) dv + o(\Delta).$$

2) A single arrival occurs when the cavity queue is not idle: in this case at some time $t + v, v \in [0, \Delta]$ an arrival of size $s + \Delta - u$ at the cavity queue which has workload $u - v$ for some $u \in [v, s + \Delta]$ occurs. We find:

$$Q_2 = \lambda d \int_{v=0}^{\Delta} \int_{u=v}^{s+\Delta} c_d(t + v, u - v) g(s + \Delta - u) du dv + o(\Delta).$$

3) A single arrival occurs when the cavity queue is empty: in this case a job of size $s + \Delta - v$ arrives at time $t + v$ for some $v \in [0, \Delta]$. Hence,

$$Q_3 = \lambda d \int_{v=0}^{\Delta} C_d(t + v, 0) g(s + \Delta - v) dv + o(\Delta).$$

By subtracting $f(t, s + \Delta)$, dividing by Δ and letting Δ decrease to zero, we find (3) from (5).

We still require a differential equation for $F(t, 0)$, a server may be idle at time $t + \Delta$ by remaining idle in $[t, t + \Delta]$ or having a workload equal to $\Delta - v, v < \Delta$ at time $t + v$. We therefore find:

$$F(t + \Delta, 0) = F(t, 0) - \lambda d \int_{v=0}^{\Delta} C_d(t + v, 0) dv + \int_{v=0}^{\Delta} f(t + v, \Delta - v) du + o(\Delta),$$

subtracting $F(t, 0)$, dividing by Δ and letting Δ tend to zero yields (4). □

Remark. *The set of equations given by (3-4) can be solved numerically using the following scheme:*

$$f(t + \delta, 0^+) = \lambda d C_d(t, 0),$$

$$f(t + \delta, s) = f(t, s + \delta) + \lambda d \delta \int_0^s c_d(t, u) g(s - u) du + \lambda d \delta C_d(t, 0) g(s) - \lambda d \delta c_d(t, s),$$

for $s \geq \delta$. As a boundary condition, we may impose that we start with all servers being idle, i.e., for $s > 0$ we set $f(0, s) = 0$ and $F(0, 0) = 1$. The main objective of Theorem 3.5 is however to use it to characterize the unique equilibrium environment $\mathcal{H}$ of the LL(d) policy.

Remark. *A propagation of chaos result was established in [7, Section 7] that states that the workload distributions of any set of $N' \leq N$ queues at time t become asymptotically independent, provided that the workload distribution at time 0 in the N queues is i.i.d. and does not depend on N, e.g., if the system is empty at time zero. As such Theorem 3.5 characterizes the limiting transient behavior.*

4 LIMITING WORKLOAD DISTRIBUTION

As indicated in the previous section, the limiting stationary workload distribution is given by the unique equilibrium environment. Let $F(s)$ be the cdf of the limiting workload distribution, that is, $F(s)$ represents the probability that the workload is at most s and let $f(s)$ be its density for $s > 0$. Furthermore, similar to (1) and (2), define

$$c_d(u) = f(u)\bar{F}(u)^{d-1}, \tag{6}$$

and

$$C_d(u) = \frac{1 - (1 - F(u))^d}{d}. \tag{7}$$

THEOREM 4.1. *The stationary workload distribution is the unique distribution that obeys the following integral equation:*

$$F(s) = (1 - \rho) + \lambda \cdot \left(\int_0^s (1 - \bar{F}(u)^d)(1 - G(s - u))du \right) \tag{8}$$

PROOF. By demanding that the derivatives with respect to t are zero in (3-4), we find

$$\frac{\partial f(s)}{\partial s} = \lambda d \left(c_d(s) - \int_0^s c_d(u) g(s - u) du - C_d(0) g(s) \right), \tag{9}$$

and

$$f(0^+) = \lambda d C_d(0). \tag{10}$$

Integrating (9) once (and relying on the assumption that $G(0) = 0$) we find:

$$f(s) = K - \lambda d \cdot \left(\frac{1}{d} - C_d(s) + C_d(0)G(s) + \int_0^s c_d(u)G(s - u)du \right), \tag{11}$$

for an appropriate constant K. As we know from (10) that $f(0^+) = \lambda d C_d(0)$, we see that we should set K equal to λ. We may therefore conclude that

$$f(s) = \lambda d \cdot \left(C_d(s) - C_d(0)G(s) - \int_0^s c_d(u)G(s - u)du \right) \tag{12}$$

Integrating equation (12) once more and using the fact that $F(0) = 1 - \rho$, yields

$$F(s) = (1 - \rho) + \lambda d \cdot \left(\int_0^s C_d(u)(1 - G(s - u))du \right)$$

The uniqueness follows from the fact that there exists a unique equilibrium environment for the LL(d) supermarket model as stated earlier. □

REMARK. *For d large, we have $\bar{F}(u)^d \approx 1$ and therefore (8) yields*

$$F(s) \approx (1 - \rho) + \lambda \int_0^s \bar{G}(u)du,$$

meaning $f(s) \approx \lambda \bar{G}(s)$. Note that $\lambda \int_0^s \bar{G}(u)du$ can be recognized as the equilibrium distribution of the forward recurrence time of a renewal process with inter-arrival time distribution G.

REMARK. *The cavity process evolves as the workload of an M/G/1 queue with a workload dependent arrival rate, we can therefore also apply Theorem 2.1 in [4] to the LL(d) cavity process. In this manner we obtain that*

$$f(s) = \lambda d \left(C_d(0)(1 - G(s)) + \int_0^s c_d(u)(1 - G(s - u))du \right),$$

which can easily be shown to be equivalent to (12) by using the fact that $c_d(u) = \frac{d}{du} C_d(u)$. The interpretation of this equation is as follows. The left-hand side of the equation corresponds to the downcrossing rate through level s, while the right-hand side denotes the upcrossing rate through s.

4.1 Fixed point iteration

We propose to use the following simple fixed point iteration to solve the integral equation (8):

$$F_{n+1}(s) = (1 - \rho) + \lambda \cdot \left(\int_0^s (1 - \bar{F}_n(u)^d)(1 - G(s - u)) du \right),$$

which we prove converges to the unique fixed point provided that $\rho < d^{-1/d}$. In Section 6 we further show that if the job sizes follow a phase-type distribution, we can directly compute the limiting workload distribution $F(s)$ by solving a simple set of differential equations (for any $\rho < 1$), meaning there is no need to make use of the above fixed point iteration.

Define the space $\mathrm{CDF}_{1-\rho} \subseteq [1 - \rho, 1]^{[0,\infty)}$ to be the space of cumulative distribution functions starting in $1 - \rho$, i.e., the space of functions which satisfy:

- $F(0) = 1 - \rho$,
- $\lim_{s \to \infty} F(s) = 1$,
- for $s, h > 0 : F(s + h) \geq F(s)$,
- $\lim_{h \to 0^+} F(s + h) = F(s)$.

On this space we can define an operator $T_d : \mathrm{CDF}_{1-\rho} \longrightarrow \mathbb{R}^{[0,\infty)}$ defined by:

$$T_d F : [0, \infty) \to \mathbb{R} : s \mapsto (1 - \rho) + \lambda d \cdot \left(\int_0^s C_d(u)(1 - G(s - u)) du \right).$$

LEMMA 4.2. *For $F \in \mathrm{CDF}_{1-\rho}$, we have $T_d F \in \mathrm{CDF}_{1-\rho}$.*

PROOF. The only non-trivial part is to show that $\lim_{s \to \infty} T_d F(s) = 1$. We find:

$$\lim_{s \to \infty} \left| \int_0^s dC_d(u) \cdot (1 - G(s - u)) du \right| \leq \lim_{s \to \infty} \int_0^s (1 - G(s - u)) du = \mathbb{E}[G],$$

which shows that $\lim_{s \to \infty} T_d F(s) \leq 1$. To obtain the other inequality observe that for any $\varepsilon > 0$, we can find a $U > 0$ for which:

$$\lim_{s \to \infty} \int_U^s (1 - G(s - u)) du > \sqrt{1 - \varepsilon} \mathbb{E}[G], \quad C_d(u) \geq \sqrt{1 - \varepsilon},$$

for $u > U$. We thus find:

$$\lim_{s \to \infty} \int_0^s dC_d(u)(1 - G(s - u)) du \geq \lim_{s \to \infty} \int_U^s dC_d(u)(1 - G(s - u)) du = \geq (1 - \varepsilon) \mathbb{E}[G]$$

this shows that $\lim_{s \to \infty} T_d F(s) \geq 1$ □

REMARK. *Due to the above lemma we may write $T_d : \mathrm{CDF}_{1-\rho} \to \mathrm{CDF}_{1-\rho}$.*

REMARK. *We can define an order on $CDF_{1-\rho}$ by stating that $F_1 \preceq F_2 \Leftrightarrow \forall s \in [0, \infty) : F_1(s) \leq F_2(s)$, then a simple application of the Knaster-Tarski theorem [21] also guarantees the existence of a fixed point of T_d. Indeed note that we have $F_1 \preceq F_2 \Rightarrow T_d F_1 \preceq T_d F_2$.*

THEOREM 4.3. *For any $F_1, F_2 \in \mathrm{CDF}_{1-\rho}$ we have:*

$$d_K(T_d F_1, T_d F_2) \leq d\rho^d \cdot d_K(F_1, F_2),$$

where d_K denotes the uniform (or Kolmogorov) metric, i.e., $d_K(F_1, F_2) = \sup_s |F_1(s) - F_2(s)|$.

Proof. Let $\varepsilon > 0$ be arbitrary and let s^* be such that:

$$\sup_s \int_0^s |(1 - F_1(u))^d - (1 - F_2(u))^d|(1 - G(s - u))du$$

$$< \int_0^{s^*} |(1 - F_1(u))^d - (1 - F_2(u))^d|(1 - G(s^* - u))du + \varepsilon.$$

We therefore have that $d_K(T_d F_1, T_d F_2)$ is bounded above by:

$$\lambda \int_0^{s^*} |(1 - F_2(u))^d - (1 - F_1(u))^d|(1 - G(s^* - u))du + \varepsilon.$$

We now use the fact (which can be shown by applying the mean value theorem) that for any $x, y \in [0, \rho)$ we have $|x^d - y^d| \le d\rho^{d-1} \cdot |x - y|$. This shows by applying the above that we have:

$$d_K(T_d F_1, T_d F_2) < \lambda \int_0^{s^*} d\rho^{d-1}|F_1(u) - F_2(u)|(1 - G(s^* - u))du + \varepsilon$$

$$\le \lambda d\rho^{d-1} d_K(F_1, F_2) \int_0^{s^*} (1 - G(s^* - u))du + \varepsilon$$

$$\le d\rho^d d_K(F_1, F_2) + \varepsilon,$$

which completes the proof.   □

Remark. *In particular for $\rho < e^{-1/e} \approx 0.6922$ the above theorem shows by the Banach fixed-point theorem that T_d admits a unique fixed point which can be found by our proposed fixed point iteration with speed of convergence $d_K(F^*, F_n) \le \frac{d^n \rho^{nd}}{1 - d\rho^d} d_K(F_1, F_0)$. This follows from the fact that $d^{-1/d}$ attains a minimum in e. For higher values of ρ, d must be such that $d\rho^d < 1$ to guarantee convergence via Theorem 4.3. Numerical experiments using both light-tailed and heavy-tailed distributions suggest that the fixed point iteration converges quickly for any $\rho < 1$.*

5 EXPONENTIAL JOB SIZES

In the previous section we established an integral equation for the limiting stationary workload distribution (for any non-idling service discipline). In this section we derive an explicit expression for this distribution in case of exponential job sizes with mean 1, that is, when $G(s) = 1 - e^{-s}$ and $\rho = \lambda$. In addition we also derive an explicit expression for the limiting response time distribution in case the service discipline is first-come-first-served.

5.1 Limiting workload distribution

Theorem 5.1. *The ccdf of the limiting stationary workload distribution for the LL(d) policy for any non-idling service discipline with exponential job sizes with mean 1 is given by:*

$$\bar{F}(s) = (\lambda + (\lambda^{1-d} - \lambda)e^{(d-1)s})^{\frac{1}{1-d}}. \tag{13}$$

Proof. Using (8) with $G(s) = 1 - e^{-s}$ and $\rho = \lambda$, we have

$$F(s) = (1 - \lambda) + \lambda d \int_0^s C_d(u)e^{u-s}du, \tag{14}$$

Taking the derivative on both sides and using Leibniz integral rule, we find the following simple ODE for $F(s)$:

$$
\begin{aligned}
F'(s) &= \lambda(1 - \bar{F}(s)^d) - \lambda \int_0^s (1 - \bar{F}(u)^d)e^{u-s}du \\
&= \lambda(1 - \bar{F}(s)^d) - (F(s) - (1 - \lambda)) \\
&= \bar{F}(s) - \lambda\bar{F}(s)^d,
\end{aligned}
\tag{15}
$$

with boundary condition $F(0) = 1 - \lambda$, equivalently:

$$
\bar{F}'(s) = \lambda\bar{F}(s)^d - \bar{F}(s),
$$

with $\bar{F}(0) = \lambda$. This ODE can be solved explicitly and one easily verifies that the solution $\bar{F}(s)$ is given by:

$$
\bar{F}(s) = (\lambda + (\lambda^{1-d} - \lambda)e^{(d-1)s})^{\frac{1}{1-d}}.
$$

□

REMARK. *There is a striking and unexpected similarity between the limiting workload distribution of the LL(d) policy and the response time distribution of the replication with cancellation-on-completion [11, Section 5] in case of exponential job sizes in the sense that the response time distribution of the latter system solves exactly the same ODE as in* (15), *except that it is subject to the boundary condition* $\bar{F}(0) = 1$.

REMARK. *For d large, $\bar{F}(s)$ can be approximated by λe^{-s} as $(\lambda^{1-d} - \lambda)^{1/(1-d)}$ is close to λ for large d. This result is expected as for large d we expect that a fraction λ of the servers contains exactly one job and the remaining workload of any such job is exponentially distributed due to the memoryless nature of the exponential distribution.*

In order to obtain an expression for the expected workload of a server, we first recall the following integral representation for the analytic continuation of the hypergeometric function $_2F_1(a, b; c; z)$ [1, Chapter 15]

$$
_2F_1(a, b; c; z) = \frac{1}{B(b, c - b)} \int_0^1 x^{b-1}(1 - x)^{c-b-1}(1 - zx)^{-a}dx,
\tag{16}
$$

where $B(x, y) = \int_0^1 t^{x-1}(1 - t)^{y-1}dt$ is the Beta function. This integral expression is valid for any $c > b > 0$ and $z < 1$. When $|z| < 1$ this function can be represented as an infinite sum using the Pochhammer symbol (or falling factorial) $(q)_n = \prod_{k=0}^{n-1}(q + k)$ when $n > 0$ and $(q)_0 = 1$:

$$
_2F_1(a, b; c; z) = \sum_{n=0}^{\infty} \frac{(a)_n(b)_n}{(c)_n} \frac{z^n}{n!}.
\tag{17}
$$

THEOREM 5.2. *The mean $W_d(\lambda)$ of the limiting workload distribution of a server under the LL(d) policy with exponential job sizes with mean 1 is given by:*

$$
W_d(\lambda) = \sum_{n=0}^{\infty} \frac{\lambda^{dn+1}}{1 + n(d - 1)},
\tag{18}
$$

in particular we find:

$$W_2(\lambda) = -\frac{\log\left(1 - \lambda^2\right)}{\lambda},$$

$$W_3(\lambda) = -\frac{1}{\sqrt{\lambda}} \cdot \log\left(\frac{\sqrt{1 - \lambda^3}}{\lambda^{3/2} + 1}\right).$$

PROOF. We employ the notation $b = \lambda^{1-d} - \lambda$. We begin by computing (using $y = e^{-s}$ and $x = y^{d-1}$):

$$W_d(\lambda) = \int_0^\infty \bar{F}(s)ds$$

$$= \int_0^1 \frac{1}{(\lambda y^{d-1} + b)^{1/(d-1)}}dy$$

$$= \frac{1}{b^{1/(d-1)}} \frac{1}{(d-1)} \int_0^1 \frac{x^{-(d-2)/(d-1)}}{(1 + \frac{\lambda}{b}x)^{1/(d-1)}}dx$$

Hence, by (16) this last integral can be expressed via the hypergeometric function $_2F_1$ as

$$W_d(\lambda) = \frac{1}{b^{1/(d-1)}} \cdot {}_2F_1\left(\frac{1}{d-1}, \frac{1}{d-1}; 1 + \frac{1}{d-1}; -\frac{\lambda}{b}\right).$$

Note that we cannot directly use the sum representation of $_2F_1$ as λ/b may become greater than 1 (which happens when λ gets close to one). Therefore we now employ the well-known linear transformation formulas:

$$_2F_1(a, b; c; z) = (1 - z)^{c-a-b} \cdot {}_2F_1(c - a, c - b; c; z)$$

$$_2F_1(a, b; c; z) = (1 - z)^{-a} \cdot {}_2F_1\left(a, c - b; c; \frac{z}{z - 1}\right). \tag{19}$$

Using these indicates that

$$W_d(\lambda) = \frac{1}{b^{1/(d-1)}}\left(1 + \frac{\lambda}{b}\right)^{-\frac{1}{d-1}} \cdot {}_2F_1\left(1, \frac{1}{d-1}; 1 + \frac{1}{d-1}; \lambda^d\right) = \lambda \cdot {}_2F_1\left(1, \frac{1}{d-1}; 1 + \frac{1}{d-1}; \lambda^d\right)$$

As $\lambda^d \in (0, 1)$, we can use the sum representation given by (17) to find that

$$W_d(\lambda) = \sum_{n=0}^\infty \frac{\lambda^{nd+1}}{1 + n(d-1)},$$

as $(1)_n = n!$ and $(1/(d-1))_n/(1/(d-1)+1)_n = 1/(1 + n(d-1))$. The expressions for $d = 2, 3$ can be either found directly by looking at the Taylor expansion of the logarithm or by solving the integral representation of $W_d(\lambda)$. □

5.2 Limiting response time distribution

We now focus on the limiting response time distribution R in case the service discipline is first-come-first-served and denote its cdf as $F_R(s)$.

THEOREM 5.3. *The ccdf of the limiting response time distribution of the LL(d) policy with FCFS service and exponential job sizes with mean 1 is given by:*

$$\bar{F}_R(s) = \left(\lambda^d + (1 - \lambda^d)e^{(d-1)s}\right)^{\frac{1}{1-d}}. \tag{20}$$

PROOF. Let E be an exponential random variable with mean 1 and let T_i, $i = 1, \ldots, d$ denote the d independent workloads of the d randomly selected servers. We find:

$$\bar{F}_R(s) = \mathbb{P}\left\{E + \min_{i=1}^{d} T_i > s\right\} = e^{-s} + \int_0^s \bar{F}(s-t)^d e^{-t} dt.$$

Due to (13) and using standard integration techniques, this integral can be simplified to:

$$\bar{F}_R(s) = e^{-s} \cdot \left(1 + \frac{1}{\lambda b^{1/(d-1)}} \cdot \int_{\left(\frac{b}{\lambda}\right)^{1/(d-1)}}^{e^s \left(\frac{b}{\lambda}\right)^{1/(d-1)}} (1 + x^{d-1})^{d/(1-d)} dx\right),$$

where $b = \lambda^{1-d} - \lambda$ as before. This is an integral that can be solved exactly to prove the statement. □

REMARK. *It is easy to verify that the workload and response time distributions $F(s)$ and $F_R(s)$ have the same increasing failure rate $r(s) = f(s)/\bar{F}(s) = f_R(s)/\bar{F}_R(s)$.*

REMARK. *For d large, $F_R(s) \approx e^{-s}$, as expected.*

THEOREM 5.4. *The mean of the limiting response time distribution for the LL(d) policy with FCFS service and exponential job sizes with mean 1 is given by:*

$$T_d(\lambda) = \sum_{n=0}^{\infty} \frac{\lambda^{dn}}{1 + n \cdot (d-1)}. \tag{21}$$

PROOF. Let $E \sim \text{Exp}(1)$, we find:

$$T_d(\lambda) = \mathbb{E}[E + \min\{T_1, \ldots, T_d\}] = 1 + \int_0^{\infty} \bar{F}(s)^d ds.$$

Using (13) and standard integration techniques (mainly substitution), we can reduce this expression to:

$$T_d(\lambda) = 1 + \frac{1}{\lambda^{d/(d-1)} \cdot (d-1)} \cdot \int_0^{\lambda/b} \frac{v^{1/(d-1)}}{(1+v)^{d/(d-1)}} dv.$$

Using the substitution $y = \frac{v}{1+v}$, one can show that the above integral reduces to

$$1 + \frac{\lambda^d}{d} \cdot {}_2F_1\left(\frac{d}{d-1}, 1; 1 + \frac{d}{d-1}; \lambda^d\right).$$

As $\lambda^d \in (0, 1)$, one can use (17) and the claimed equality follows as $(1)_n = n!$ and $(d/(d-1))_n/(1 + d/(d-1))_n = d/((n+1)(d-1)+1)$. □

REMARK. *In the proof of Theorem 5.4 it is also possible to directly use (20) instead of relying on (13).*

REMARK. *Note that $W_d(\lambda) = \lambda T_d(\lambda)$, which is expected due to Little's law and the fact that the mean workload of a server under the LL(d) policy for exponential job sizes with mean 1 is equal to the mean number of jobs in such a server. The relation $W_d(\lambda) = \lambda T_d(\lambda)$ also yields simple formulas for $T_2(\lambda)$ and $T_3(\lambda)$ due to Theorem 5.2. It is possible do derive similar expressions for larger d values, but these become more and more complex as d increases.*

REMARK. *In [11] the mean of the limiting response time distribution in case of exponential job sizes of the replication with cancellation-on-completion policy (under the assumption of the independence ansatz) was argued to be equal to*

$$\mathbb{E}[T^{RR(d)}] = \frac{{}_2F_1(1, 1; 1 + \frac{d}{d-1}; \frac{-\rho}{1-\rho})}{\mu d(1-\rho)}.$$

This expression can be reduced to a simple sum formula as follows (using (19) *and* (17) *as* $\rho \in (0, 1)$*)*

$$_2F_1(1, 1; 1 + \frac{d}{d-1}; \frac{-\rho}{1-\rho}) = (1-\rho)\,_2F_1(1, \frac{d}{d-1}; 1 + \frac{d}{d-1}; \rho) = (1-\rho) \sum_{n=0}^{\infty} \frac{(1)_n \left(\frac{d}{d-1}\right)_n}{\left(1 + \frac{d}{d-1}\right)_n} \frac{\rho^n}{n!},$$

which allows us to conclude that

$$\mathbb{E}[T^{RR(d)}] = \frac{1}{\mu} \sum_{n=0}^{\infty} \frac{\rho^n}{n(d-1)+d}.$$

Note that $E[T^{RR(d)}]$ *converges to* $1/(d\mu)$ *as* ρ *tends to zero due to the independent execution times of the replicas in* [11]*.*

6 PHASE-TYPE AND DETERMINISTIC JOB SIZES

In Section 4.1 we proposed a fixed point iteration to compute the limiting workload distribution $F(s)$ under LL(d) for any job size distribution G, that was proven to converge if $d\rho^d < 1$. We now show that $F(s)$ can also be directly obtained as the solution of a set of coupled ordinary differential equations (ODEs) for any $\rho < 1$, provided that the job lengths follow a phase-type (PH) distribution. PH distributions are distributions with a modulating finite state background Markov chain [15] and any general positive-valued distribution can be approximated arbitrary closely with a PH distributions. Further, various fitting tools are available online for phase-type distributions (e.g., [14, 20]). A PH distribution with $G(0) = 0$ is fully characterized by a stochastic vector $\alpha = (\alpha_i)_{i=1}^n$ and a subgenerator matrix $A = (a_{i,j})_{i,j=1}^n$ such that $\bar{G}(s) = \alpha e^{As} \mathbf{1}$, where $\mathbf{1}$ is a column vector of ones.

THEOREM 6.1. *Suppose the job lengths have a PH distribution characterized by* (α, A)*, then the ccdf of the limiting workload distribution under the LL(d) policy satisfies:*

$$\bar{F}'(s) = -\lambda((1 - \bar{F}(s)^d) + \alpha Ah(s)),$$

$$h'(s) = (1 - \bar{F}(s)^d)\mathbf{1} + Ah(s),$$

with $\bar{F}(0) = \rho$, $h(0) = 0$ *and* $h(s) : \mathbb{R} \to \mathbb{R}^{n \times 1}$*.*

PROOF. For $i \in \{1, \ldots, n\}$ we define:

$$h_i(s) = \int_0^s (1 - \bar{F}(u)^d) e_i^T e^{(s-u)A} \mathbf{1} du,$$

where e_i^T is the i-th row of the identity matrix I_n. First note that $h_i(0) = 0$. We now derive a differential equation for $h_i(s)$. Using the equality $I_n = \sum_{k=1}^n e_k e_k^T$ we find :

$$h_i'(s) = (1 - \bar{F}(s)^d) + \int_0^s (1 - \bar{F}(u)^d) e_i^T A I_n e^{(s-u)A} \mathbf{1} du$$

$$= (1 - \bar{F}(s)^d) + \sum_{k=1}^n \int_0^s (1 - \bar{F}(u)^d) e_i^T A e_k e_k^T e^{(s-u)A} \mathbf{1} du$$

$$= (1 - \bar{F}(s)^d) + \sum_{k=1}^n a_{i,k} h_k(s).$$

In matrix notation this yields:

$$h'(s) = (1 - \bar{F}(s)^d)\mathbf{1} + Ah(s).$$

Due to (8) and $\bar{G}(s - u) = \alpha e^{(s-u)A}\mathbf{1}$, we have $\bar{F}'(s) = -\lambda\alpha h'(s)$, which yields the equation for $\bar{F}'(s)$. □

We now generalize this result to the case where the service times are the sum of a deterministic random variable and a PH distribution.

THEOREM 6.2. *Assume the service times are the sum of a deterministic random variable with mean τ and a phase-type distribution characterized by (α, A), i.e., $\bar{G}(s) = I_{\{s \le \tau\}} + I_{\{s > \tau\}}\alpha e^{(s-\tau)A}\mathbf{1}$, then the ccdf of the limiting workload distribution under the LL(d) policy satisfies:*

$$\bar{F}'(s) = \lambda(\bar{F}(s)^d - 1), \qquad\qquad s \le \tau,$$
$$\bar{F}'(s) = -\lambda((1 - \bar{F}(s)^d) + \alpha A h(s - \tau)), \qquad\qquad s > \tau,$$
$$h'(s) = (1 - \bar{F}(s)^d)\mathbf{1} + A h(s),$$

with $h(0) = 0$ and $\bar{F}(0) = \rho = \lambda(\tau + \alpha(-A)^{-1}\mathbf{1})$.

PROOF. We distinguish two cases: first let $s \in [0, \tau]$, we find that $\bar{F}(s) = \rho - \lambda\int_0^s 1 - \bar{F}(u)^d du$, deriving this equation once yields the first equation.

For the second note that we have (using the notation from the proof of Theorem 6.1):

$$\bar{F}(s) = \rho - \lambda\alpha h(s - \tau) - \lambda\int_{s-\tau}^s (1 - \bar{F}(u)^d)du.$$

Taking the derivative and using the expression for $h'(s)$ found in Theorem 6.1 completes the proof. □

THEOREM 6.3. *If the job sizes are deterministic and equal to one, the ccdf $\bar{F}(s)$ is determined by $\bar{F}(0) = \lambda$, and*

$$\bar{F}'(s) = \lambda(\bar{F}(s)^d - 1) \qquad\qquad s \in [0, 1),$$
$$\bar{F}'(s) = \lambda(\bar{F}(s)^d - \bar{F}(s - 1)^d) \qquad\qquad s \ge 1.$$

PROOF. The proof is similar to the proof of Theorem 6.2. □

REMARK. *We note that the ODEs and DDEs presented in this section have a unique solution: the existence follows from the fact that (8) solves the ODE/DDE, while the uniqueness follows from [9, Section 23, theorem A].*

REMARK. *It is easy to compute the ccdf of the response time distribution $\bar{F}_R(s)$ given $\bar{F}(s)$ as the probability that a new arrival joins a queue with a workload exceeding s is given by $\bar{F}(s)^d$ under the LL(d) policy.*

7 LL(D) VERSUS SQ(D)

The aim of this section is to study the margin of improvement that can be achieved by using exact workload information as opposed to the coarser queue length information used by SQ(d). This margin of improvement is of interest to understand the possible response time improvements offered by schedulers that implement late binding (as discussed in the introduction). Furthermore, we also compare the SQ(d) policy with the LL(d) policy where the job sizes of the latter take the late binding overhead into account. We start by focusing on exponential job sizes, for which we can also establish some closed form results.

7.1 Exponential job sizes

In this subsection we compare the limiting response time of the LL(d) and SQ(d) policies for exponential job sizes with mean 1 and FCFS service. This comparison provides an answer on the reduction in the response times that can be obtained if the workloads at the different servers are known instead of the coarser queue length information. To distinguish between the response times of both policies we make use of the superscripts $^{(LL)}$ and $^{(SQ)}$. For the SQ(d) policy the mean of the limiting response time distribution is given by [17]

$$T_d^{(SQ)}(\lambda) = \frac{1}{\lambda} \sum_{k=1}^{\infty} \lambda^{\frac{d^k-1}{d-1}}.$$

THEOREM 7.1. *The mean of the limiting response time distribution for the LL(d) policy is smaller than the mean for the SQ(d) policy for exponential job sizes with mean* 1, *moreover*

$$T_d^{(SQ)}(\lambda) - T_d^{(LL)}(\lambda) = \frac{1}{\lambda} \sum_{k=1}^{\infty} A_k,$$

where for $\lambda \in (0, 1)$:

$$A_k = \lambda^{\frac{d^{k+1}-1}{d-1}} - \sum_{n=1}^{d^k} \frac{\lambda^{nd+1+\frac{d^{k+1}-d^2}{d-1}}}{1 + n(d-1) + (d^k - d)} > 0.$$

PROOF. Due to (21), we need to show:

$$\sum_{n=1}^{\infty} \frac{\lambda^{dn+1}}{1 + n(d-1)} \leq \sum_{k=2}^{\infty} \lambda^{\frac{d^k-1}{d-1}}.$$

To see this, we group the terms on the left hand side with $n \in \{\sum_{s=0}^{k-1} d^{s-1}, \ldots, \sum_{s=1}^{k} d^s\}$ together and compare their sum with the term $\lambda^{\frac{d^{k+1}-1}{d-1}}$ on the right hand side for $k \geq 1$. We have

$$\sum_{n=1+\cdots+d^{k-1}}^{d+\cdots+d^k} \frac{\lambda^{nd+1}}{1 + n(d-1)} < \sum_{n=1+\cdots+d^{k-1}}^{d+\cdots+d^k} \frac{\lambda^{d(1+d+\cdots+d^{k-1})+1}}{(1 + \cdots + d^{k-1})(d-1)} = \lambda^{1+d+\cdots+d^k} = \lambda^{\frac{d^{k+1}-1}{d-1}}.$$

Hence, the result follows.

$\square$

REMARK. *For any fixed* λ *the above theorem shows that for any* $\epsilon > 0$, *there exists a* $d(\lambda, \epsilon)$ *such that for* $d > d(\lambda, \epsilon)$ *we have* $T_d^{(SQ)}(\lambda) - T_d^{(LL)}(\lambda) < \epsilon$.

THEOREM 7.2. *For the ratio of the mean of the limiting response time distribution of SQ(d) and LL(d) for exponential job sizes with mean* 1 *we have*

$$\lim_{\lambda \to 1} T_d^{(SQ)}(\lambda)/T_d^{(LL)}(\lambda) = \frac{d-1}{\log(d)}.$$

PROOF. Let $K \in \mathbb{N}$ be arbitrary and define:

$$U_K(\lambda) = \frac{1 + \sum_{k=1}^{K} \sum_{n=1+\cdots+d^{k-1}}^{d+\cdots+d^k} \frac{\lambda^{nd+1}}{1+n(d-1)}}{1 + \sum_{k=1}^{K} \lambda^{\frac{d^{k+1}-1}{d-1}}}.$$

We note that we have:

$$\lim_{\lambda \to 1} \lim_{K \to \infty} U_K(\lambda) = \lim_{\lambda \to 1} \frac{T_d^{(LL)}(\lambda)}{T_d^{(SQ)}(\lambda)}.$$

On the other hand (with ψ the Digamma function [1, Chapter 6]) we have:

$$\lim_{K \to \infty} \lim_{\lambda \to 1} U_K(\lambda) = \lim_{K \to \infty} \frac{\sum_{k=0}^{K} \sum_{n=\frac{d^{k}-1}{d-1}}^{d \frac{d^{k}-1}{d-1}} \frac{1}{1+n(d-1)}}{\sum_{k=0}^{K} 1} = \frac{1}{d-1} \lim_{K \to \infty} \frac{\sum_{k=0}^{K} \psi\left(\frac{d^{k+1}}{d-1}\right) - \psi\left(\frac{d^{k}}{d-1}\right)}{\sum_{k=0}^{K} 1}.$$

Since $\lim_{k \to \infty} \psi\left(\frac{d^{k+1}}{d-1}\right) - \psi\left(\frac{d^{k}}{d-1}\right) = \log(d)$, we may apply the Stolz-Cesaro theorem to assert that

$$\lim_{K \to \infty} \lim_{\lambda \to 1} U_K(\lambda) = \frac{\log(d)}{d-1}.$$

If we may interchange the limits this would incur:

$$\lim_{\lambda \to 1} \frac{T_d^{(LL)}(\lambda)}{T_d^{(SQ)}(\lambda)} = \frac{\log(d)}{d-1}.$$

An application of the Moore-Osgood theorem [13, p100] implies that we may indeed interchange limits: as U_K and $U = \lim_{K \to \infty} U_K$ are continuous functions defined on the compact set $[0, 1]$ and U_K converges pointwise to U, it follows that this convergence is also uniform. Moreover, we trivially have pointwise convergence of $\lim_{\lambda \to 1} U_K(\lambda)$.   □

REMARK. *As $(d-1)/\log(d)$ tends to infinity as d becomes large, we note that for any $c > 0$ there exists a λ and d such that the ratio $T_d^{(SQ)}(\lambda)/T_d^{(LL)}(\lambda) > c$. In other words, for arbitrary λ and d, there is no bound on how much worse the SQ(d) policy performs than the LL(d) policy.*

In Figure 1 we plot the ratio $T_d^{(SQ)}(\lambda)/T_d^{(LL)}(\lambda)$ as a function of λ. We note that this ratio increases with λ and approaches a constant as λ approaches one. Looking at this figure, the limit values for the ratio $T_d^{(SQ)}(\lambda)/T_d^{(LL)}(\lambda)$ as λ tends to one may appear to be less than $(d-1)/\log(d)$ (as shown in Theorem 7.2), but this is simply due to the fact that this ratio still increases significantly between 0.999 and 1. From this figure we may conclude that the increase in the mean of the limiting response time distribution by using the coarser queue length information instead of the exact workload is below 50% when $d = 2$ for exponential job sizes. For larger d we see a more significant increase under high load.

We further note that the curves for different d values cross one another. Intuitively this can be understood by noting that for λ small many jobs select an idle server and when an idle server is selected knowing the queue length is equally good as knowing the workload. When d increases it becomes more likely that an idle server is selected and thus we expect the mean response time ratio to decrease with increasing d when λ is small. For large λ it becomes unlikely that one of the selected queues is idle and SQ(d) has to rely on the coarser queue length information. When λ is large, we therefore see a larger loss of more information as d increases and thus the mean response time ratio now increases with increasing d.

Apart from comparing the mean response times, we can also easily compare the response time distribution of the LL(d) and SQ(d) policy. For the SQ(d) policy it is not hard to establish that the ccdf of the limiting response time distribution can be written as

$$\bar{F}_R^{(SQ)}(s) = \sum_{k=1}^{\infty} \left(\lambda^{(d^{k-1}-1)d/(d-1)} - \lambda^{(d^{k}-1)d/(d-1)} \right) \sum_{n=0}^{k-1} \frac{s^n}{n!} e^{-s} = \sum_{n=0}^{\infty} \frac{s^n}{n!} e^{-s} \lambda^{(d^n-1)d/(d-1)}, \tag{22}$$

by noting that a job that joins a queue of length $k - 1$ has an Erlang-k distributed response time for exponential job sizes. Figure 2 depicts the response time distributions for $\lambda = 0.95$ and $d = 2, 3$ and
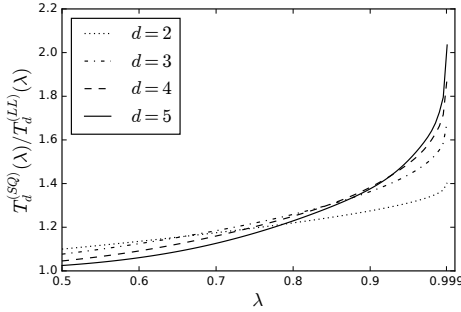
Fig. 1. Ratio of the mean of the limiting response time distribution of SQ(d) and LL(d) for exponential job sizes with mean 1, FCFS service as a function of λ.
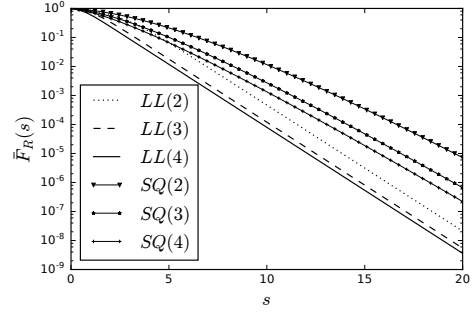
Fig. 2. Limiting response time distribution of SQ(d) and LL(d) for exponential job sizes with mean 1, FCFS service and $\lambda = 0.95$.

4. We note that $\bar{F}_R(s)$ decreases as a function of d and $\bar{F}_R^{(SQ)}(s)$ dominates $\bar{F}_R^{(LL)}(s)$ for all $s > 0$. The next theorem proves an even stronger result.

THEOREM 7.3. *The function* $f(s) = \bar{F}_R^{(SQ)}(s)/\bar{F}_R^{(LL)}(s)$ *is non-decreasing on* $[0, \infty)$*, thus* $\bar{F}_R^{(SQ)}(s) \geq \bar{F}_R^{(LL)}(s)$ *for all s.*

PROOF. It suffices to show that $f'(s) \geq 0$ for $s > 0$ (as $\bar{F}_R^{(SQ)}(0) = \bar{F}_R^{(LL)}(0) = 1$). Denote $\mu = \lambda^d$. Using (22) and (20), the condition $f'(s) \geq 0$ can be restated as

$$\frac{\sum_{k=0}^{\infty} \mu^{\frac{d^k-1}{d-1}} (\mu^{d^k} - 1) \frac{s^k}{k!}}{\sum_{k=0}^{\infty} \mu^{\frac{d^k-1}{d-1}} \frac{s^k}{k!}} + \frac{(1-\mu)e^{(d-1)s}}{\mu + (1-\mu)e^{(d-1)s}} \geq 0.$$

By rearranging terms this is equivalent to showing:

$$e^{(d-1)s} \left(\sum_{k=0}^{\infty} \mu^{d^k} \mu^{\frac{d^k-1}{d-1}} \frac{s^k}{k!} \right) \geq \frac{\mu}{1-\mu} \sum_{k=0}^{\infty} \mu^{\frac{d^k-1}{d-1}} (1 - \mu^{d^k}) \frac{s^k}{k!}.$$

For the left hand side we find, by using the Taylor expansion of $e^{(d-1)s}$ and applying Merten's theorem (which states that if $\sum_n a_n$ converges to A and $\sum_n b_n$ converges to B, then the Cauchy product converges to AB if at least one of the two sequences converges absolutely):

$$e^{(d-1)s} \left(\sum_{k=0}^{\infty} \mu^{d^k} \mu^{\frac{d^k-1}{d-1}} \frac{s^k}{k!} \right) = \sum_{n=0}^{\infty} \frac{s^n}{n!} \sum_{k=0}^{n} \binom{n}{k} (d-1)^{n-k} \mu^{d^k} \mu^{\frac{d^k-1}{d-1}}.$$

It therefore suffices to show that the inequality holds for all coefficients of $\frac{s^n}{n!}$, i.e. it remains to show that:

$$\frac{\mu}{1-\mu} \mu^{\frac{d^n-1}{d-1}} (1 - \mu^{d^n}) \leq \sum_{k=0}^{n} \binom{n}{k} (d-1)^{n-k} \mu^{d^k} \mu^{\frac{d^k-1}{d-1}}.$$

By noting that $\frac{1-\mu^{d^n}}{1-\mu} \leq d^n$, the result follows if the following holds

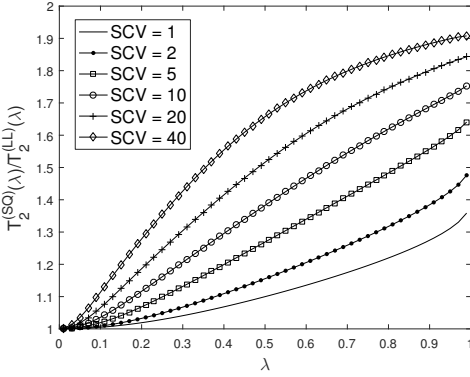$$d^n \leq \sum_{k=0}^{n} \binom{n}{k} (d-1)^{n-k} \mu^{\frac{d^{k+1}-1}{d-1} - \frac{d^n-1}{d-1} - 1},$$

Fig. 3. Ratio of the mean of the limiting response time distribution of SQ(2) and LL(2) for hyperexponential job sizes with mean 1, shape parameter $f = 1/2$ and FCFS service as a function of λ.
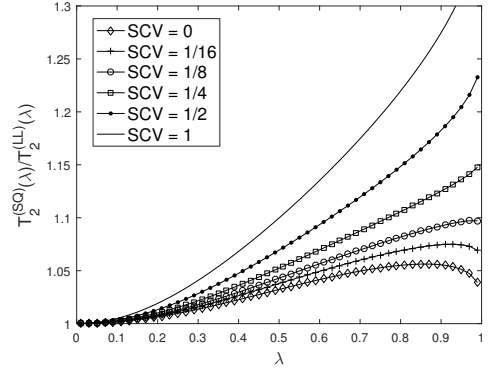
Fig. 4. Ratio of the mean of the limiting response time distribution of SQ(2) and LL(2) for Erlang job sizes with mean 1 and FCFS service as a function of λ.

We clearly have an equality in $\mu = 1$ (and for $n = 0$). It therefore suffices to show that the right hand side decreases for $\mu \in [0, 1]$ for $n > 0$. The first n terms are all convex decreasing, while the last term is convex increasing. The derivative of the sum of the first and last term in $\mu = 1$ is $(d^n - 1)(1 - (d - 1)^{n-1}) \leq 0$. Since the derivative of a convex function on $[0, 1]$ is maximized in 1, the sum of the first and last term is decreasing and we may conclude that $f'(s) \geq 0$. □

7.2 Impact of job variability

In this subsection we study the impact of the job size variability on the ratio of the mean of the limiting response time distribution of SQ(d) and LL(d). In real systems a significant part of the total workload is often offered by a small fraction of long jobs, while the remaining workload consists mostly of (very) short jobs [19]. For simplicity we represent these workloads as a hyperexponential (HEXP) distribution (with 2 phases) such that we can vary the job size variability in a systematic manner. More precisely, with probability p a job is a type-1 job and has an exponential length with parameter $\mu_1 > 1$ and with the remaining probability $1 - p$ a job is a type-2 job and has exponential length with parameter $\mu_2 < 1$. Hence, the type-2 jobs are longer on average and we therefore sometimes refer to the type-2 jobs as the *long* jobs. The parameters p, μ_1 and μ_2 are set such that the following three values are matched: (i) mean job length (set to one), (ii) the squared coefficient of variation (SCV) and (iii) a shape parameter f, using the following equations:

$$\mu_1 = \frac{SCV + (4f - 1) + \sqrt{(SCV - 1)(SCV - 1 + 8f\bar{f})}}{2f(SCV + 1)},$$

$$\mu_2 = \frac{SCV + (4\bar{f} - 1) - \sqrt{(SCV - 1)(SCV - 1 + 8f\bar{f})}}{2\bar{f}(SCV + 1)},$$

with $\bar{f} = 1 - f$ and $p = \mu_1 f$. The shape parameter $f \in (0, 1)$ represents the fraction of the workload that is offered by the type-1 jobs.

The mean of the limiting response time distribution for the LL(d) policy can be computed in a fraction of a second for any $\rho < 1$ by making use of Theorem 6.1. For the SQ(d) policy we use a fixed point iteration to determine the stationary queue length distribution of the cavity process associated to the equilibrium environment [6]. More specifically, we determine the queue length

distribution of a sequence of M/G/1 FCFS queues with a queue length dependent arrival rate λ, where the queue length distribution determined during the n-th iteration determines the arrival rates of the $n + 1$-th iteration, until the queue length distribution converges (starting from the empty distribution). While the queue length distribution of such a queue can be computed in a very fast manner when the job sizes follow a phase-type distribution (or are deterministic), the number of iterations needed increases sharply as ρ approaches 1. This prevents us from studying what happens in the limit as ρ tends to one.

Figure 3 depicts the ratio of the mean of the limiting response time distribution of the SQ(d) and LL(d) policies when $d = 2$ and $f = 1/2$ (meaning half of the workload is offered by the *long* jobs). This ratio increases when the jobs sizes become more variable, which is expected as having precise workload information should be more valuable when jobs vary significantly in size. The results indicate that a mechanism like late binding can offer substantial gains even at fairly low loads if the job sizes vary significantly (and the round-trip time to fetch the job can be neglected). The results for $f = 1/10$, which implies that 90% of the workload is offered by the *long* jobs, are very similar (and therefore not depicted). For $d > 2$ these ratios tend to increase under sufficiently high loads as in the exponential case.

For completeness we also present some results for job sizes with an SCV below 1 in Figure 4. In this case we cannot make use of a hyperexponential distribution and therefore consider Erlang-k distributed and deterministic job sizes instead. This figure shows that as λ approaches 1 the ratio of the means of the limiting response time distribution starts to decrease for sufficiently small SCVs. In fact, studying this ratio for λ values closer to 1 as depicted in Figure 4 suggests that this ratio decreases to 1 for deterministic job sizes. This seems to make sense intuitively as for λ approaching one, the queue lengths become long and knowing the coarser queue length information is almost as good as knowing the exact workload.

7.3 Late binding overhead

In the previous subsection we shed light on the margin of improvement that late binding can provide compared to the classic SQ(d) policy assuming that the jobs can be fetched from the dispatchers in negligible time. In this section we take the idleness caused by late binding into account. We do this by comparing the mean of the limiting response time distribution of the SQ(d) policy with the mean of the LL(d) policy, where the size of each job under the LL(d) policy is incremented by a deterministic quantity τ that represents the overhead, that is, the time that the server remains idle under late binding while fetching the job. We denote the mean of the limiting response time in the latter case as $T_{d,\tau}^{(LL)}$ and rely on Theorem 6.2 for its computation. We consider the same job size distributions (with average job size equal to one) as in the previous section.

In Figure 5 the ratio $T_d^{(SQ)}/T_{d,\tau}^{(LL)}$ is shown as a function of λ for the case where $\tau = 0.05$, meaning each job induces an idle server period with a length equal to 5% of the mean job size. It indicates that for a very wide range of arrival rates λ, late binding offers substantial gains over the SQ(d) policy even with an overhead of 5%. For systems with high job size variability, this range even includes arrival rates above 0.9. Note that the overhead of the scheduler implementation in [19] was estimated to be below 2%. We further note that late binding requires storing the jobs at the dispatcher(s) until a notification from one of the servers arrives, which may be regarded as a drawback compared to SQ(d) which allows immediate dispatching as soon as the queue length information is obtained.

In fact for medium loads much higher amounts of overhead can be tolerated by the LL(d) policy before it becomes inferior to SQ(d). This is illustrated in Figure 6, where we plot the largest τ value
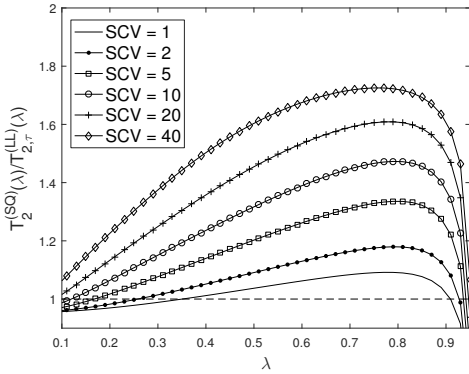
Fig. 5. Ratio of the mean of the limiting response time distribution of SQ(2) and LL(2) with 5% overhead (i.e., $\tau = 0.05$) for hyperexponential job sizes with mean 1, shape parameter $f = 1/2$ and FCFS service as a function of λ.
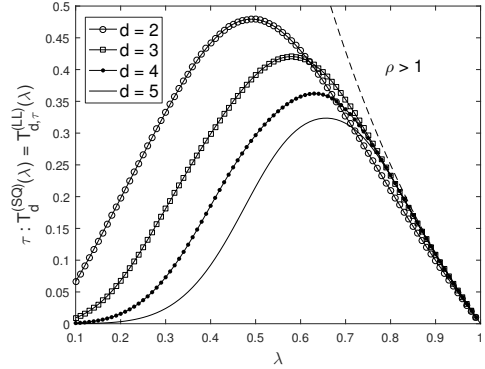
Fig. 6. Degree of delay that the LL(d) policy can tolerate without being outperformed by SQ(d) as a function of λ for hyperexponential job sizes with mean 1, SCV = 20 and $f = 1/2$, i.e., the largest τ such that $T_{d,\tau}^{(LL)} \leq T_d^{(SQ)}$.

for which $T_{d,\tau}^{(LL)} \leq T_d^{(SQ)}$ when the SCV was set to 20. We observe that overheads of 25% and more can be tolerated for system workloads around 50%.

8 FINITE SYSTEM ACCURACY

In this section we briefly compare the limiting response time distribution with simulation experiments where the number of servers N is finite. All simulation runs simulate the system up to time $t = 10^7/N$ and use a warm-up period of 30%.

Figure 7a compares the expression for the limiting response time distribution given by (20) for exponential job sizes with simulation experiments. In the simulation the number of servers equals $N = 100$ servers, the 95% confidence intervals are computed based on 10 runs that each start from an empty system. The agreement with simulation is very good (except for high loads combined with a small d) considering that we are simulating a system with only 100 servers.

In Figure 7b we look at the impact of the number of simulated servers N under high loads when $d = 2$. We note that the limiting distribution is not necessarily a good match for the tail probabilities of the response time when N is small, e.g., $N = 20$, but the accuracy quickly improves as the number of servers increases.

In Figure 7c and 7d we look at a similar setting as in Figure 7a and 7b, but the job sizes now follow a hyperexponential distribution with $f = 1/2$ (see Section 7.2 for details). In this case the 95% confidence intervals are computed based on 25 runs. We note that even though the job sizes are now substantially more variable, the accuracy seems quite similar to the exponential case. Thus, more variable job size distributions do not necessarily imply worse accuracy for a fixed N.

Figure 8 illustrates the accuracy of the limiting response time distribution in case of power law and deterministic job sizes (computed via the fixed point iteration in Section 4.1). More specifically, for the power law distribution we used $\bar{G}(s) = s^{-\beta}$ with $\beta = 2$. This implies that the mean job size is finite and equal to 2, while the variance of the job size distribution is infinite. In the deterministic case the job size equals 1. The figure indicates that somewhat larger N values are needed to closely match the limiting response time distribution compared to the (hyper)exponential case.
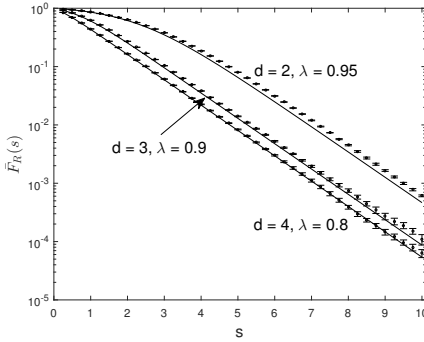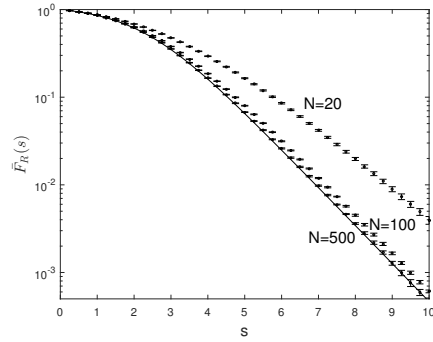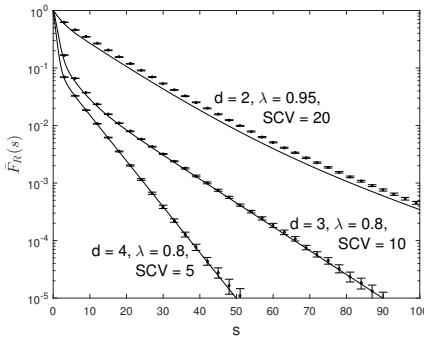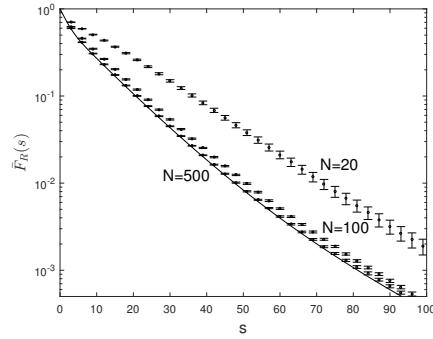
(a) Exponential job sizes, $N = 100$

(b) Exponential job sizes, $d = 2, \lambda = 0.95$

(c) Hyperexponential job sizes, $N = 100$

(d) Hyperexponential job sizes, $d = 2, \lambda = 0.95$, $SCV = 20$

Fig. 7. Limiting response time distribution vs. simulation for N servers with (hyper)exponential job sizes with mean 1. The full line represents the limiting response time distribution.

9 CONCLUSIONS AND FUTURE WORK

In this paper we studied the limiting workload and response time distribution of the LL(d) policy which assigns an incoming job to a server with the least work left among d randomly selected servers. We introduced a fixed point iteration to determine the limiting workload distribution for general job size distributions and any non-idling service discipline and studied its convergence. We derived a closed form expression for both the workload and response time distribution (for FCFS service) in case of exponential job sizes and indicated that these distributions can be computed easily by solving a set of ordinary differential equations for phase-type distributed job sizes.

We provided insight into the gains that can be expected when exact workload information is used instead of the coarser queue length information by comparing the performance of the LL(d) policy with the classic SQ(d) policy. Such a comparison is relevant to understanding the performance gains offered by schedulers implementing *late binding*. In this regard we demonstrated that late binding offers significant gains over SQ(d) for a wide range of arrival rates, even when taking the late binding overhead into account.

We intend to generalize the results presented in this paper to a setting where each incoming job consists of k tasks and these k tasks can be executed simultaneously on different servers. For each incoming job d servers are still selected at random, but now the k tasks are assigned to the k

(a) Power law job sizes with $\bar{G}(s) = x^{-2}, d = 2, \rho = 0.8$

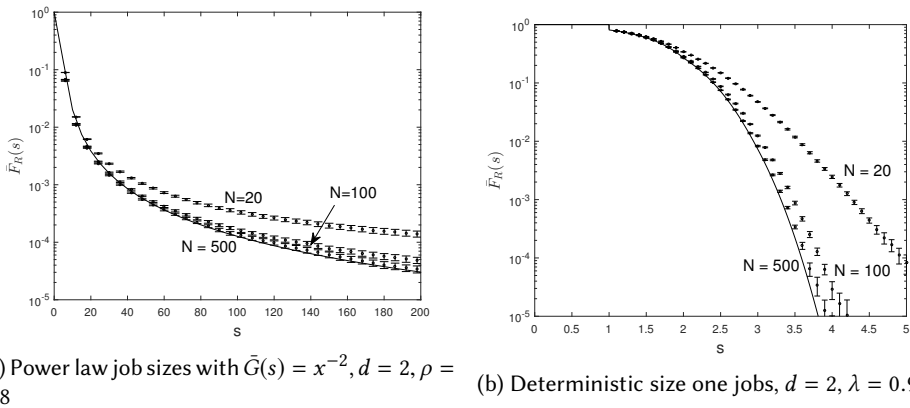(b) Deterministic size one jobs, $d = 2, \lambda = 0.9$

Fig. 8. Limiting response time distribution vs. simulation for N servers with power law and deterministic job sizes with mean 1. The full line represents the limiting response time distribution.

servers (among the d selected servers) with the least workload left. Some of the challenges in this setting concern proving the stability condition and asymptotic independence of the workloads.

Furthermore, some of the results presented in this paper should be extendable to systems where the incoming jobs belong to different job classes and not all servers can necessarily serve all job classes. Another possible, but challenging, direction for future work on the LL(d) policy is to perform a worst-case analysis as was done for the SQ(d) policy in [3].

REFERENCES

[1] M. Abramowitz and I. A. Stegun. 1964. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (ninth dover printing, tenth gpo printing ed.). Dover, New York.

[2] R. Aghajani, X. Li, and K. Ramanan. 2017. The PDE Method for the Analysis of Randomized Load Balancing Networks. *Proc. ACM Meas. Anal. Comput. Syst.* 1, 2, Article 38 (Dec. 2017), 28 pages. https://doi.org/10.1145/3154497

[3] Y. Azar, A.Z. Broder, A.R. Karlin, and E. Upfal. 1999. Balanced Allocations. *SIAM J. Comput.* 29, 1 (1999), 180–200.

[4] R. Bekker, S.C. Borst, O.J. Boxma, and O. Kella. 2004. Queues with Workload-Dependent Arrival and Service Rates. *Queueing Systems* 46, 3 (01 Mar 2004), 537–556. https://doi.org/10.1023/B:QUES.0000027998.95375.ee

[5] M. Bramson. 2011. Stability of join the shortest queue networks. *Ann. Appl. Probab.* 21, 4 (2011), 1568–1625. https://doi.org/10.1214/10-AAP726

[6] 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

[7] M. Bramson, Y. Lu, and B. Prabhakar. 2012. Asymptotic independence of queues under randomized load balancing. *Queueing Syst.* 71, 3 (2012), 247–292. https://doi.org/10.1007/s11134-012-9311-0

[8] M. Bramson, Y. Lu, and B. Prabhakar. 2013. Decay of tails at equilibrium for FIFO join the shortest queue networks. *Ann. Appl. Probab.* 23, 5 (10 2013), 1841–1878. https://doi.org/10.1214/12-AAP888

[9] R. D. Driver. 1977. *Ordinary and Delay Differential Equations*. Springer-Verlag, Berlin-Heidelberg-New York.

[10] S. Foss and N. Chernova. 1998. On the Stability of a Partially Accessible Multi-station Queue with State-dependent Routing. *Queueing Syst. Theory Appl.* 29, 1 (May 1998), 55–73. https://doi.org/10.1023/A:1019175812444

[11] 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

[12] K. Gardner, S. Zbarsky, M. Harchol-Balter, and A. Scheller-Wolf. 2016. The Power of d Choices for Redundancy. In *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science, Antibes Juan-Les-Pins, France, June 14-18, 2016*. 409–410. https://doi.org/10.1145/2896377.2901497

[13] L. M. Graves. 1946. *The theory of functions of real variables*. McGraw-Hill book company, inc.

[14] J. Kriege and P. Buchholz. 2014. *PH and MAP Fitting with Aggregated Traffic Traces*. Springer International Publishing, Cham, 1–15. https://doi.org/10.1007/978-3-319-05359-2_1

[15] G. Latouche and V. Ramaswami. 1999. *Introduction to Matrix Analytic Methods and stochastic modeling.* SIAM, Philadelphia.

[16] M. Mitzenmacher. 2000. How Useful Is Old Information? *IEEE Trans. Parallel Distrib. Syst.* 11, 1 (Jan. 2000), 6–20. https://doi.org/10.1109/71.824633

[17] M. Mitzenmacher. 2001. The Power of Two Choices in Randomized Load Balancing. *IEEE Trans. Parallel Distrib. Syst.* 12 (October 2001), 1094–1104. Issue 10.

[18] D. Mukherjee, S. Borst, J. van Leeuwaarden, and P. Whiting. 2016. Universality of Power-of-d Load Balancing Schemes. *SIGMETRICS Perform. Eval. Rev.* 44, 2 (Sept. 2016), 36–38. https://doi.org/10.1145/3003977.3003990

[19] K. Ousterhout, P. Wendell, M. Zaharia, and I. Stoica. 2013. Sparrow: Distributed, Low Latency Scheduling. In *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles (SOSP '13).* ACM, New York, NY, USA, 69–84. https://doi.org/10.1145/2517349.2522716

[20] A. Panchenko and A. Thümmler. 2007. Efficient Phase-type Fitting with Aggregated Traffic Traces. *Perform. Eval.* 64, 7-8 (Aug. 2007), 629–645. https://doi.org/10.1016/j.peva.2006.09.002

[21] A. Tarski. 1955. A lattice-theoretical fixpoint theorem and its applications. *Pacific J. Math.* 5, 2 (1955), 285–309. https://projecteuclid.org:443/euclid.pjm/1103044538

[22] N.D. Vvedenskaya, R.L. Dobrushin, and F.I. Karpelevich. 1996. Queueing System with Selection of the Shortest of Two Queues: an Asymptotic Approach. *Problemy Peredachi Informatsii* 32 (1996), 15–27.