[This is the accepted manuscript of the article that appeared in final form in PERFORMANCE EVALUATION (2017): 116, 26-52, which has been published in final form at http://dx.doi.org/10.1016/j.peva.2017.08.004.

Light traffic behavior under the power-of-two load balancing strategy: The case of heterogeneous servers.

A. Izagirre[a] and A.M. Makowski[b]

[a]Univ. of the Basque Country (UPV/EHU), Spain
[b] Department of Electrical and Computer Engineering, and Institute for Systems Research University of Maryland, College Park, MD 20742

Abstract

We consider a multi-server queueing system under the power-of-two policy with Poisson job arrivals, heterogeneous servers and a general job requirement distribution; each server operates under the first-come first-serve policy and there are no buffer constraints. We analyze the performance of this system in light traffic by evaluating the first two light traffic derivatives of the average job response time. These expressions point to several interesting structural features associated with server heterogeneity in light traffic: For unequal capacities, the average job response time is seen to decrease for small values of the arrival rate, and the more diverse the server capacities, the greater the gain in performance. These theoretical findings are assessed through limited simulations.

Keywords: Parallel servers, power-of-two scheduling, light traffic, heterogeneous servers

1. Introduction

Systems of parallel servers are commonly used to model resource sharing applications. These queueing models have been adopted in classical performance studies for supermarket cashiers, bank tellers and toll booths; they have also appeared in the context of computer systems and communication networks, and more recently in cloud computing infrastructures. A basic design issue for such systems is the *scheduling* (or *routing*) of incoming jobs, usually with an eye towards making the average job response time as small as can be. One possible choice is to randomly assign an incoming job to one of the available servers, a strategy which may lead to large delays but which has the advantage of requiring no state information. At the other extreme, the join-the-shortest-queue (JSQ) policy is known to possess certain optimality properties [13], but requires the queue length at each server to be available at the arrival epoch of every job.

JSQ and its variants have been extensively studied [1] (and references therein) with most of the work focusing on the *homogeneous* case when the servers have identical service capacities and use the same service discipline. In such cases job size variability is known to greatly affect average job performance under the first-come first-serve (FCFS) service discipline [1, Chapter 24]. However, the impact seems much reduced under the processor-sharing (PS) discipline, with near-insensitivity being reported in [2].

Much work has also been done to explore the *trade-off* between the information overhead to implement job scheduling and the resulting performance. An interesting alternative which interpolates between random assignment and JSQ is the following policy $SQ(d)$ (for some integer $d \geq 2$): Upon arrival, an

incoming job randomly selects a pool of d servers from amongst the set of available servers. The JSQ policy is then restricted to the d servers in this random pool (with a random tiebreaker) – Here, shortest queue refers to the queue in the random pool with the fewest jobs but other definitions (say in terms of workload) are possible; see Section 3 for additional comments.

This queueing system, sometimes known as the supermarket model, has been studied for some time now with special attention given to the case $d = 2$ (from which the terminology power-of-two derives); see the brief historical survey in [8, Section 1.1]. The analysis of this model is mathematically challenging due to the coupling created between queues by the use of *local* JSQ policies. This is so even when jobs arrive according to a Poisson process, servers are identical FCFS servers, and job requirements are exponentially distributed. In that setting, [8] and [12] (with $d = 2$), independently resorted instead to studying the limiting system when the number of servers becomes unboundedly large. Together their results point to a substantial improvement in performance over the case $d = 1$ (which corresponds to the random server assignment) *without* the full overhead of *global* JSQ.

In view of these encouraging results it is natural to inquire whether the policy $SQ(d)$ still provides a performance advantage when servers have *different* capacities. With $d = 2$, the paper [9] took a step in that direction: Following the same limiting strategy as in [8, 12] the authors discuss the average job response time for the $SQ(2)$ model under *heterogeneous* PS servers (but with a finite number of different server capacities), with Poisson arrivals and a general job requirement distribution.

In this paper we consider the policy $SQ(2)$ with FCFS servers with different capacities, Poisson job arrivals and a general job requirement distribution. Instead of looking at the many server asymptotics as in earlier papers, we focus instead on the *light traffic* regime under a *fixed* number of servers; this corresponds to the system operating with a very low traffic intensity. Using the framework developed by Reiman and Simon [11], we compute the first and second light-traffic derivatives of the average job response time; see Proposition 2.2 and Proposition 2.3 in Section 2. These derivatives already provide insights into the *structural* impact that server heterogeneity may have on job performance: For instance, at least in light traffic, the more diverse the server capacities, the greater the gain in performance. Moreover, job performance in $SQ(2)$ is *not* monotone in the traffic intensity (at least when this traffic intensity is small). See Section 3 for a short discussion of this issue.

The quadratic "approximation" (15) suggests itself naturally in Section 2.2 on the basis of the first two light-traffic derivatives. This *local* approximation will not be accurate in the moderate to heavy traffic regime, absent more information, e.g., the heavy traffic limit used by Reiman and Simon to generate light-traffic interpolations [10]. We nevertheless use (15) as a benchmark against simulations to illustrate the structural features revealed through the light traffic calculations; Section 4 contains a set of limited simulations exploring the theoretical findings. We stress that the light traffic methodology is used to provide a simplifying *lense* through which the performance of $SQ(2)$ can be explored in light traffic. Furthermore, the results presented here can be readily extended to a very large class of policies which can be obtained by tuning several design parameters, namely the *size* d of the random server pool, the selection *criterion* for selecting a server from the random pool of d servers and the *service discipline* at each of the servers. In many of these cases the light-traffic calculations are feasible, at least in principle, but are likely to be quite involved. This is especially so when computing the second derivative due to the combinatorial explosion inherent to the approach in [11]. We refer the reader to the end of Section 3 for details, possible conjectures and for future work.

The paper is organized as follows: The model and assumptions are introduced in Section 2.1, and the evaluation of the first two derivatives is presented in Section 2.2. Various comments on and implications of the results are given in Section 3, while in Section 4 we illustrate some of the theoretical findings with the help of limited simulations. In Section 5 we summarize the needed elements of the light traffic theory we use. In Section 6 we evaluate the light-traffic response time of a tagged job, the so-called $n = 0$ case in the Reiman-Simon theory. We start the technical discussion in Section 7 with an auxiliary result that greatly simplifies later computations of the first and second light-traffic derivatives. The first derivative is computed in Section 8. The calculations of the second derivative start in Section 9, and are developed through Sections 11–13. Additional computations are given in the Appendices A–C.

The results discussed here were announced without proofs in the conference paper [4]. Additional details concerning some of the calculations can be found in the Ph.D. thesis [3] and in the version [5].

2. Main results

The random variables (rvs) under consideration are all defined on the same probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. The construction of a sufficiently large probability triple carrying all needed rvs is standard and omitted in the interest of brevity. Probabilistic statements are made with respect to this probability measure $\mathbb{P}$, and we denote the corresponding operator by $\mathbb{E}$. Throughout let σ denote an $\mathbb{R}_+$-valued rv which is distributed according to some probability distribution function $F : \mathbb{R}_+ \to [0, 1]$, so that $F(x) = \mathbb{P}[\sigma \leq x]$ $(x \geq 0)$. We assume at minimum that $\mathbb{E}[\sigma] < \infty$. If E is a subset of Ω, then $\mathbf{1}[E]$ is the indicator of the set E with the usual understanding that $\mathbf{1}[E](\omega) = 1$ (resp. $\mathbf{1}[E](\omega) = 0$) if $\omega \in E$ (resp. $\omega \notin E$).

With any discrete set S which is non-empty and finite (so $0 < |S| < \infty$), we write $U \sim \mathcal{U}(S)$ to indicate that the rv U is uniformly distributed over S (under $\mathbb{P}$), namely

$$\mathbb{P}[U = u] = \frac{1}{|S|}, \quad u \in S.$$

2.1. Model and assumptions

The system comprises $K \geq 2$ parallel servers labelled $k = 1, \ldots, K$. Server k has capacity C_k (bytes/sec.), is attended by an infinite capacity buffer and operates in a FCFS manner. Jobs arrive according to a Poisson process $\{A(t),\ t \geq 0\}$ of rate $\lambda > 0$ with arrival epochs $\{T_n,\ n = 0, 1, \ldots\}$ – By convention we take $T_0 = 0$. For each $n = 0, 1, \ldots$, we refer to the job arriving at time T_n as the n^{th} job; this job brings a random amount of work σ_n (bytes). Upon arrival, the n^{th} job is assigned to one of the K servers according to the power-of-two load balancing scheme (so $SQ(d)$ with $d = 2$): Specifically, this incoming job randomly selects a pair Σ_n of distinct servers from the set of K servers. The JSQ policy is then used in isolation with these two servers; ties are broken randomly (but other choices are possible).

As usual, the Poisson arrival process $\{A(t),\ t \geq 0\}$, the sequence of job requirement rvs $\{\sigma_n,\ n = 0, 1, \ldots\}$ and the sequence of server selection rvs $\{\Sigma_n,\ n = 0, 1, \ldots\}$ are mutually independent collections of rvs. Throughout we make the following assumptions: (i) The rvs $\{\sigma_n,\ n = 0, 1, \ldots\}$ are i.i.d. rvs distributed according to the probability distribution F – The rv σ introduced earlier is therefore a generic element of this sequence of i.i.d. rvs; and (ii) The server selection rvs $\{\Sigma_n,\ n = 0, 1, \ldots\}$ are i.i.d. rvs, each of which is uniformly distributed over the collection of unordered pairs drawn from $\{1, \ldots, K\}$. Thus, with $\mathcal{P}_2(K)$ denoting the collection of unordered pairs drawn from $\{1, \ldots, K\}$, we have $\Sigma_n \sim \mathcal{U}(\mathcal{P}_2(K))$ with

$$\mathbb{P}[\Sigma_n = T] = \frac{1}{\binom{K}{2}}, \quad \begin{array}{l} T \in \mathcal{P}_2(K), \\ n = 0, 1, \ldots \end{array}$$

Assuming the system to be initially empty (for sake of convenience), for each $n = 0, 1, \ldots$, let $R_{n,\lambda}$ denote the response time of the n^{th} job when the arrival rate is λ. The stationary response time of a job when the arrival rate is λ is denoted by R_λ. The existence of R_λ (as an $[0, \infty]$-valued rv) can be established through classical semi-Markovian methods; details are omitted in the interest of brevity. We set

$$R(\lambda) = \mathbb{E}[R_\lambda].$$

We expect $\mathbb{E}[R_\lambda] < \infty$ over some non-degenerate interval $(0, \lambda^\star)$ for some finite $\lambda^\star > 0$, in which case $R_{n,\lambda} \Longrightarrow_n R_\lambda$ with $\Longrightarrow_n$ denoting weak convergence (also known as convergence in distribution) with n going to infinity; see also [8, Section 2.1, Lemma 1] and [9, Lemma 6]. In what follows we shall not be concerned with this issue any further since we are mainly interested in the situation where λ is very small (vanishingly so).

2.2. Evaluating the first two derivatives

Light-traffic analysis considers the performance of the system for small values of the arrival rate $\lambda > 0$. In that regime a so-called *light-traffic approximation* can often be constructed on the basis of the following Taylor series expansion argument; e.g. see the light-traffic interpolations discussed in [10].

Assume that for some positive integer L, the L first derivatives of the function $\lambda \to R(\lambda)$ exist in a neighborhood $(0, \lambda_\star)$ with $\lambda_\star > 0$. Whenever $0 < x < \lambda_\star$ with λ sufficiently small, Taylor's formula

$$R(x + \lambda) = R(x) + \lambda R^{(1)}(x) + \frac{\lambda^2}{2!} R^{(2)}(x) + \ldots + \frac{\lambda^L}{L!} R^{(L)}(x) + \text{Remainder}_L(x; \lambda)$$

holds where we use the notation

$$R^{(\ell)}(x) = \frac{d^\ell R}{d\lambda^\ell}(\lambda)\Big|_{\lambda=x}, \quad \ell = 1, \ldots, L.$$

For this discussion the exact form taken by the remainder term $\text{Remainder}_L(x;\lambda)$ is not important.

Assume further that all the limits

$$R(0+) = \lim_{\lambda\downarrow 0} R(\lambda) \quad \text{and} \quad R^{(\ell)}(0+) = \lim_{\lambda\downarrow 0} \frac{d^\ell R}{d\lambda^\ell}(\lambda), \quad \ell = 1, \ldots, L \tag{1}$$

exist (in $\mathbb{R}$) – We refer to these quantities as *light-traffic derivatives*. It is then natural to use the polynomial $R_{\text{App}}(\lambda) : (0,\infty) \to \mathbb{R}$ given by

$$R_{\text{App}}(\lambda) = R(0+) + \lambda R^{(1)}(0+) + \frac{\lambda^2}{2} R^{(2)}(0+) + \ldots + \frac{\lambda^L}{L!} R^{(L)}(0+), \quad \lambda > 0 \tag{2}$$

as a possible light-traffic approximation; this prompts us to write

$$R(\lambda) \simeq R_{\text{App}}(\lambda), \quad \lambda \simeq 0. \tag{3}$$

Here we shall use an approach proposed by Reiman and Simon [11] to compute successive light-traffic derivatives in the sense of (1). It requires that some *admissibility* condition be satisfied. Following the discussion in Appendix A in [11] we assume that the generic rv σ satisfies the condition

$$\mathbb{E}\left[e^{t\sigma}\right] < \infty \tag{4}$$

for some $t > 0$. This finite exponential moment condition on F entails admissibility; it is likely stronger than needed but its purpose here is to provide a convenient framework where calculations can be justified. In particular it ensures the requisite differentiability of $\lambda \to R(\lambda)$ as soon as $R(\lambda)$ is finite. We compute the first two derivatives in light traffic; these results were announced in the conference paper [4] without proofs.

Proposition 2.1. *Under the enforced assumptions, the limit* $\lim_{\lambda\downarrow 0} R(\lambda)$ *exists and is given by*

$$R(0+) \equiv \lim_{\lambda\downarrow 0} R(\lambda) = \frac{\Gamma}{K} \cdot \mathbb{E}\left[\sigma\right] \tag{5}$$

with

$$\Gamma = \sum_{k=1}^{K} \frac{1}{C_k} \tag{6}$$

We now turn to the first derivative.

Proposition 2.2. *Under the enforced assumptions, the function* $\lambda \to R(\lambda)$ *is differentiable in a small neighborhood of* $\lambda = 0$. *Furthermore,*

$$R'(0+) \equiv \lim_{\lambda\downarrow 0} \frac{dR}{d\lambda}(\lambda) = \frac{1}{K-1}\left(\left(\frac{\Gamma}{K}\right)^2 - \frac{1}{K}\sum_{k=1}^{K}\frac{1}{C_k^2}\right) \cdot \left(\mathbb{E}\left[\sigma\right]\right)^2. \tag{7}$$

The third result concerns the second derivative.

Proposition 2.3. *Under the enforced assumptions, the function* $\lambda \to R(\lambda)$ *is twice differentiable in a small neighborhood of* $\lambda = 0$. *Furthermore,*

$$R''(0+) \equiv \lim_{\lambda\downarrow 0} \frac{d^2 R}{d\lambda^2}(\lambda) = \frac{2}{K^2(K-1)^2}\left(\frac{\Gamma^3}{K} - 2\Gamma\sum_{k=1}^{K}\frac{1}{C_k^2} + K\sum_{k=1}^{K}\frac{1}{C_k^3}\right) \cdot \left(\mathbb{E}\left[\sigma\right]\right)^3. \tag{8}$$

3. Discussion

A probabilistic interpretation. The results of Propositions 2.1-2.3 can be expressed more compactly with the help of the following probabilistic interpretation: Let $X \equiv X(C_1, \ldots, C_K)$ denote a rv uniformly distributed over the set of values $\frac{1}{C_1}, \ldots, \frac{1}{C_K}$, i.e., $X \sim \mathcal{U}(\{\frac{1}{C_1}, \ldots, \frac{1}{C_K}\})$ with

$$\mathbb{P}\left[X = \frac{1}{C_1}\right] = \ldots = \mathbb{P}\left[X = \frac{1}{C_K}\right] = \frac{1}{K}.$$

With this notation it is easy to check that

$$\mathbb{E}\left[X^p\right] = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{C_k^p}, \quad p \geq 0.$$

The expressions (5), (7) and (8) can now be rewritten more compactly as

$$R(0+) = \mathbb{E}\left[X\right] \cdot \mathbb{E}\left[\sigma\right], \tag{9}$$

$$R'(0+) = -\frac{1}{K-1}\mathrm{Var}[X] \cdot \left(\mathbb{E}\left[\sigma\right]\right)^2 \tag{10}$$

and

$$\begin{aligned} R''(0+) &= \frac{2}{(K-1)^2}\left(\left(\mathbb{E}\left[X\right]\right)^3 - 2\mathbb{E}\left[X\right] \cdot \mathbb{E}\left[X^2\right] + \mathbb{E}\left[X^3\right]\right)\left(\mathbb{E}\left[\sigma\right]\right)^3 \\ &= \frac{2}{(K-1)^2}\left(\mathbb{E}\left[X^3\right] - \left(\mathbb{E}\left[X\right]\right)^3 - 2\mathbb{E}\left[X\right] \cdot \mathrm{Var[X]}\right)\left(\mathbb{E}\left[\sigma\right]\right)^3, \end{aligned} \tag{11}$$

respectively.

Equal capacities. From (10) it follows that $R'(0+) \leq 0$, with $R'(0+) = 0$ if and only if $\mathrm{Var}[X] = 0$, or equivalently, $C_1 = \ldots = C_K = C$. In that case the K servers all have the same capacity C, and we conclude $R''(0+) = 0$ upon substituting this fact into (8). It follows that

$$R(\lambda) = \frac{\mathbb{E}\left[\sigma\right]}{C} + o(\lambda^2)$$

if we assume the existence of a third derivative (via either the Lagrange or Cauchy form of the remainder).

Unequal capacities. When the capacities are different, then $R'(0+) < 0$ and $R(\lambda)$ is *decreasing* for small values of λ. This is a somewhat unexpected finding because most queueing systems are "monotone" in the sense that increasing the traffic intensity λ results in an increase in a performance metric such as the average job response time. This fact can be explained as follows: On the average, a job entering an empty system experiences a response time given by $R(0+)$ since the scheduling policy $SQ(2)$ assigns it to *any* of the K servers with probability $\frac{1}{K}$. However, when the servers have different capacities, the assigned server may not have been the fastest, therefore making it possible by the luck of the draw for subsequent jobs to be served by one of the faster servers. This will result in a decrease in the average job response time if the traffic intensity increases slightly but still allows for some faster server to be available with some non-negligible probability.

How much of a decrease? We see from (10) that the decrease in the average job response time will be more pronounced the larger the variance Var[X] of the rv X. It is therefore natural to wonder which set of capacity values $C_1, \ldots, C_K$ yield the largest value for this variance Var[X] under a given value for $\mathbb{E}\left[X\right]$, say $\mathbb{E}\left[X\right] = \frac{\Gamma}{K}$ for some $\Gamma > 0$. In Appendix A we assess the range of Var[X] under this constraint on $\mathbb{E}\left[X\right]$, and conclude that

$$0 \leq \mathrm{Var[X]} < \frac{\Gamma^2}{\mathrm{K}} - \left(\frac{\Gamma}{\mathrm{K}}\right)^2 = (\mathrm{K}-1)\left(\frac{\Gamma}{\mathrm{K}}\right)^2 \quad \text{with } \mathbb{E}\left[X\right] = \tfrac{\Gamma}{K}. \tag{12}$$

As mentioned earlier, the lower bound is achievable by the vector of capacities given by

$$\boldsymbol{C}_\star = \left(\frac{K}{\Gamma}, \ldots, \frac{K}{\Gamma}\right). \tag{13}$$

While the upper bound is *not* achievable by any vector of capacities satisfying the constraint, it is however tight in the following sense: For each $k = 1, \ldots, K$, let the vector $\boldsymbol{e}_k$ denote the K-dimensional vector with all zero entries except in the k^{th} position where it is one. The vectors of capacities given by

$$\boldsymbol{C}_{k,a} = \frac{1}{a\Gamma}\boldsymbol{e}_k + \frac{K-1}{(1-a)\Gamma}\sum_{\ell=1,\ \ell\neq k}^{K} \boldsymbol{e}_\ell, \qquad \begin{matrix} k = 1, \ldots, K \\ 0 < a < 1 \end{matrix} \tag{14}$$

can approach the upper bound value arbitrarily close by letting a go to 1; this is shown in Appendix A.

The lower bound is implemented by the most balanced capacity assignment (13) under the constraint $\mathbb{E}[X] = \frac{\Gamma}{K}$, whereas the upper bound is achieved, albeit asymptotically, by capacity assignments (14) that are as imbalanced as they can be under the constraint. In the limit these assignments correspond to $K-1$ servers that are infinitely fast with the remaining "slow" one of finite capacity.

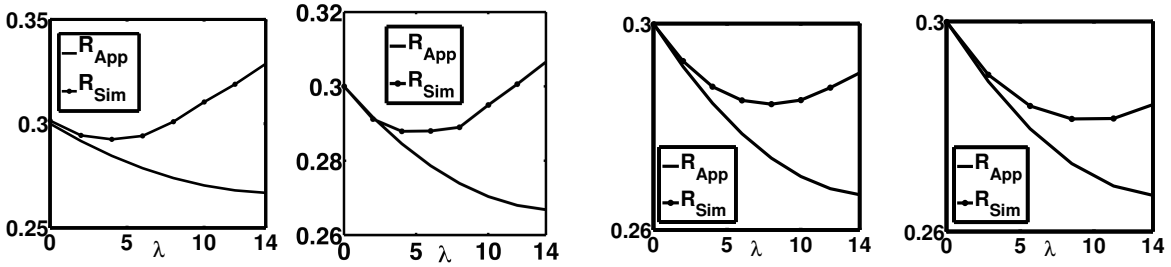
Only $\mathbb{E}[\sigma]$ matters. The two first derivatives at $\lambda = 0+$ depend only on the first moment of the rv σ, and could be read as a form of insensitivity in light traffic. This is in sharp contrast with other systems where the first light-traffic derivative depends on $\mathbb{E}\left[\sigma^2\right]$, e.g., $M|G|1$-like queues [10] and the discriminatory processor sharing model [6]. This is rather unexpected because the variance of the rv σ is known to be a key factor in shaping JSQ performance with homogeneous servers under FCFS scheduling [1, Chapter 24]. See the next item for a possible explanation.

FCFS vs. PS. Proposition 2.3 was established under the assumption that the servers operate under the FCFS discipline. It is easy to see that both (5) and (7) (but not (8)) are still valid if the servers all use the PS discipline: This is because in the cases $n = 0$ and $n = 1$ the tagged job will not share a server with another job under either discipline; see Section 6 and Section 8. However, this changes for the case $n = 2$ that involves three jobs. That the variability of the rv σ seems to play little role in light traffic is therefore consistent with the aforementioned fact that performance under the PS discipline is nearly insensitive to service variability [2].

Extensions. The results presented here concern the policy $SQ(2)$ where the server selection is implemented through a local JSQ policy applied to a pool of two randomly selected servers, and each of the K servers uses the FCFS discipline. Such policies have three distinct components: (i) The *size* d of the random pool of servers (here $d = 2$); (ii) The *criterion* used to select server from the random pool of d servers (here the incoming job is routed to the server with the shortest queue amongst the two randomly selected servers, namely a local JSQ policy); and (iii) The service discipline used by each of the servers (here FCFS). This structure naturally points to interesting extensions that are obtained by modifying one or more of the three components.

For instance, still under the FCFS policy and the shortest queue criterion, the Reiman-Simon approach can be applied to the policy $SQ(d)$ with arbitrary $d \geq 2$: With the subscript d indicating that a quantity is computed for $SQ(d)$, Proposition 2.1 remains unchanged with $R_d(0+) = \frac{\Gamma}{K} \cdot \mathbb{E}[\sigma] = R(0+)$. However, the evaluation of the light-traffic derivatives $R'_d(0+)$ and $R''_d(0+)$ become quite challenging, especially for the case $n = 2$, due to combinatorial considerations. As mentioned earlier in Section 1, our primary motivation for doing these calculations is not so much to construct light-traffic interpolations (although this is certainly a possibility), but rather to gain a better understanding of the impact of certain design choices on performance. For instance, the possible monotonic behavior of $R_d(\lambda)$ as d increases could be explored easily in light traffic – Comparing $R'_d(0+)$ against $R'_{d+1}(0+)$ would already shed some light on this issue, and lead to possible conjectures. In the same vein, it is also not immediately clear how the value of d affects the decrease in the average job response time when servers have different capacities. One could start answering this question through an analysis similar to the one carried earlier in this section on the light-traffic terms.

Staying with $d = 2$ and the shortest queue criterion, it is natural to wonder as to what happens to performance when the FCFS discipline is replaced by other service disciplines. The PS discipline is often used in this context; see the recent paper by Mukhopadhyay and Mazumdar [9]. As pointed out earlier, the light-traffic quantities $R(0+)$ and $R'(0+)$ under the PS discipline remain as calculated at (5) and (7)



(a) Hyper-exponential (b) Exponential (c) Weibull (d) Deterministic

Figure 1: Scenario 1

under the FCFS discipline. However, the evaluation of $R''(0+)$ is now much more involved due to the already large number of different cases to be considered, even with only three jobs present in the system. The complexity of the calculations further increases when considering $d > 2$.

Finally, we turn to the criterion for selecting the server in the random pool of d servers: For instance, if the servers have different capacities, the incoming job could be routed to the fastest server in the pool, e.g., the server with the largest capacity (with a suitable tie-breaker). Another possibility is to assign the incoming job to the "least loaded" server – There are multiple ways to define this idea, e.g., the server with the smallest workload (if the capacities were all identical) or the server with the smallest ratio of workload to its capacity.

The methodology used here provides a natural tool for comparing the performance of these various policies in light traffic, and for gaining a better understanding of the impact of various design choices. This rich research program will will be taken on elsewhere.

4. Limited simulations

As explained in Section 2.2 the second order polynomial

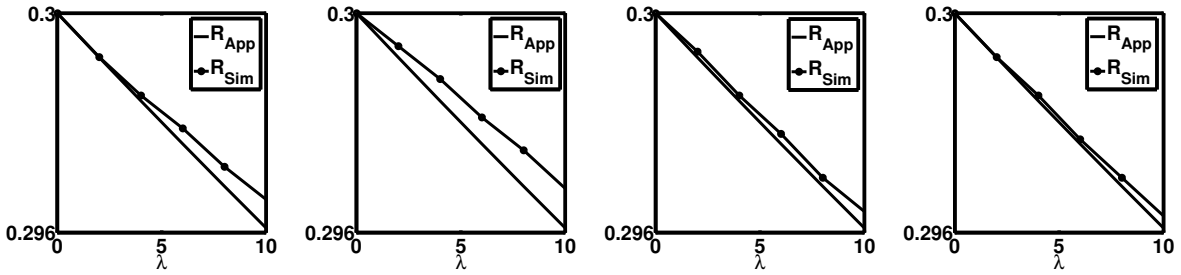
$$R_{\rm App}(\lambda) = R(0+) + \lambda R'(0+) + \frac{\lambda^2}{2} R''(0+), \quad \lambda \geq 0 \tag{15}$$

can be used as a local approximation to $R(\lambda)$ for small λ. As already pointed out by [10, 11], without additional information (e.g., heavy traffic information), we should not expect $R_{\rm App}(\lambda)$ to act as an accurate proxy for $R(\lambda)$ in medium to heavy traffic. This lack of accuracy is certainly apparent in the simulation results reported below.

We have carried out simulations for different distributions of σ, all with unit mean, namely hyperexponential (obtained by mixing the exponential rvs Exp(1/2) and Exp(2) with probability 1/3 and 2/3, respectively), exponential Exp(1) (of parameter 1), Weibull (with shape parameter 2 and scale parameter $\Gamma(3/2)^{-1}$) and deterministic. The simulation results are based on averaging 10 runs with each run comprising 10^5 busy periods. A busy period is defined as the interval of time between two consecutive time epochs when the system becomes empty, such points being regenerative points for the stochastic process of interest. We have verified that the simulation results obtained for a system with $K = 100$ homogeneous servers and exponential service requirements agree with those given by [8, Table 1].

We are interested in the behavior of the average job response time in lightly loaded situations, and stability is therefore not a concern here as mentioned earlier. Three different scenarios were explored. In Scenarios 1 and 2 there are two types of servers, namely slow servers with capacity $C_{\rm slow}$ bytes/sec and fast servers with capacity $C_{\rm fast}$ bytes/sec. In Scenario 3 all the servers have the same capacity:

- Scenario 1: $K = 10$ servers. 5 slow servers with capacity $C_{\rm slow} = 2$ bytes/sec and 5 fast servers with capacity $C_{\rm fast} = 10$ bytes/sec. See Figure 1.
- Scenario 2: $K = 100$ servers. 50 slow servers with capacity $C_{\rm slow} = 2$ bytes/sec and 50 fast servers with capacity $C_{\rm fast} = 10$ bytes/sec. See Figure 2.
- Scenario 3: $K = 10$ servers with $C_1 = \ldots = C_{10} = 10$ bytes/sec. See Figure 4.



(a) Hyper-exponential (b) Exponential (c) Weibull (d) Deterministic

Figure 2: Scenario 2

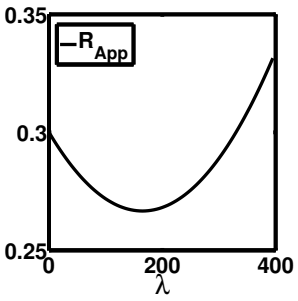


Figure 3: Scenario 2.

In the figures we use R_{Sim} to denote the average job response time obtained by simulation. Also the subscript $p\star$ in the quantities $R_{p\star}(0+)$ and $R'_{p\star}(0+)$ refers to Scenario p under distribution $\star$ where $\star$ corresponds to the hyper-exponential (H), exponential (E), Weibull (W) or deterministic (D) distribution, respectively.

Let $\text{CV}_{p\star}$ denote the coefficient of variation corresponding to Scenario p under distribution $\star = H, E, W, D$. Then $\text{CV}_{iH} = 1.4$, $\text{CV}_{iE} = 1$, $\text{CV}_{iW} = 0.52$ and $\text{CV}_{iD} = 0$ for $i = 1, 2, 3$. The simulations do confirm the structural insights gleaned from the light traffic derivatives for non-homogeneous servers; see Section 3: (i) For all distributions, the average job response time decreases as λ increases over a small neighborhood of $\lambda = 0$; (ii) Over that small interval, performance seems nearly insensitive to the variability of σ (as measured by its coefficient of variation).

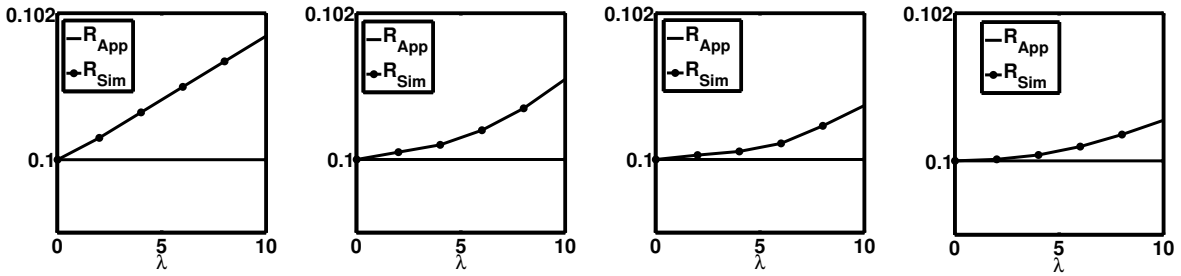
Although in Scenario 1 and Scenario 2 there is an equal proportion of slow and fast servers, with $R_{p\star}(0+) = 0.3000$ for $p = 1, 2$ and for $\star = H, E, W, D$, the impact of the variability in server capacitiess is seen to diminish with increasing K since $R'_{1\star}(0+) = -0.0044$ and $R'_{2\star}(0+) = -4.0404 \cdot 10^{-4}$ for $\star = H, E, W, D$.

Figure 2 is a zoom of Figure 3 that displays only this common approximation $R_{\text{App}}(\lambda)$. Although in Figure 3 the response time seems to be a straight line in a small interval of λ, after a while it also increases. Since $\mathbb{E}[\sigma] = 1$ for all four cases $\star = H, E, W, D$, and the approximation (15) that we use depends only on the first moment, Figure 3 is the same for all distributions considered here.

In Figure 4 we observe the aforementioned property for homogeneous servers; $R_{\text{App}}(\lambda)$ becomes a constant line while the simulation results show that the average response time of a job is increasing.

5. Review of the light traffic theory à la Reiman-Simon

The light traffic analysis presented here uses an ingenious approach proposed by [11] to compute successive light-traffic derivatives in the sense of (1): Imagine that the system starts at $t = -\infty$, so that its stationary regime will have been reached at time $t = 0$. Enters a *tagged* job at time $t = 0$ whose expected response time therefore coincides with the expected stationary response time. With this in mind, the n^{th} derivative of the expected stationary response time at $\lambda = 0+$ (namely (1)) is then shown to be computable in terms of the expected response time of the tagged job in a *scenario* where *exactly* n jobs (other than the tagged job) are allowed into the system. We refer the interested reader to the paper by Reiman and Simon [11] for details concerning this methodology.



(a) Hyper-exponential (b) Exponential (c) Weibull (d) Deterministic

Figure 4: Scenario 3

5.1. The framework

On the way to describing the Reiman-Simon approach to light traffic we find it convenient to introduce the following terminology and notation: With t in $\mathbb{R}$, a job arriving at time t, hereafter referred to as a t-job, has two rvs σ_t and Σ_t associated with it – The $\mathbb{R}_+$-valued rv σ_t stipulates the amount of work (in bytes) requested by the t-job from the system, while the rv Σ_t is an (unordered) pair of servers from amongst the K available servers. The t-job is assigned to a server ν_t selected in Σ_t according to the power-of-two policy (with a random tie-breaker). We shall refer to the rvs (σ_t, Σ_t) as the characteristic pair of the t-job.

As expected, we sometimes refer to the 0-job with characteristics (σ_0, Σ_0) as the *tagged* job. The Reiman-Simon approach to light traffic focuses on the performance of this tagged job under scenarios of increasing complexity. To define them, fix $n = 0, 1, \ldots$. Interpret every n-tuple $(t_1, \ldots, t_n)$ in $\mathbb{R}^n$ as the arrival epochs of n jobs into the system. For each $i = 1, \ldots, n$, we lighten the notation by denoting the characteristic pair $(\sigma_{t_i}, \Sigma_{t_i})$ of the t_i-job arriving at time t_i simply by (σ_i, Σ_i). Throughout the following conditions are assumed to be enforced:

1. The rvs $\{\sigma_0, \sigma_1, \ldots, \sigma_n\}$ are i.i.d. $\mathbb{R}_+$-valued rvs, each distributed according to the probability distribution F, namely
$$\mathbb{P}[\sigma_i \le x] = F(x), \qquad \begin{matrix} x \ge 0 \\ i = 0, 1, \ldots, n. \end{matrix}$$
2. The rvs $\{\Sigma_0, \Sigma_1, \ldots, \Sigma_n\}$ are i.i.d. $\mathcal{P}_2(K)$-valued rvs, each of which is uniformly distributed on $\mathcal{P}_2(K)$ with
$$\mathbb{P}[\Sigma_i = T] = \frac{1}{\binom{K}{2}}, \qquad \begin{matrix} T \in \mathcal{P}_2(K) \\ i = 0, 1, \ldots, n. \end{matrix}$$
3. The collections of rvs $\{\sigma_0, \sigma_1, \ldots, \sigma_n\}$ and $\{\Sigma_0, \Sigma_1, \ldots, \Sigma_n\}$ are mutually independent

We shall also have use for the rvs $\nu_0^\star, \nu_1^\star, \ldots, \nu_n^\star$ associated with the random pairs $\Sigma_0, \Sigma_1, \ldots, \Sigma_n$, and defined in the following manner: For each $i = 0, 1, \ldots, n$, *conditionally* on Σ_i, the rv $\nu_i^\star$ is an Σ_i-valued rv which is uniformly distributed on Σ_i – We shall write

$$[\nu_i^\star | \Sigma_i] \sim \mathcal{U}(\Sigma_i).$$

It is always understood that the rvs $\nu_0^\star, \nu_1^\star, \ldots, \nu_n^\star$ are conditionally mutually independent given the $2(n+1)$ rvs $\sigma_0, \sigma_1, \ldots, \sigma_n, \Sigma_0, \Sigma_1, \ldots, \Sigma_n$ with

$$[\nu_i^\star | \sigma_0, \ldots, \sigma_n, \Sigma_0, \ldots, \Sigma_n] \sim \mathcal{U}(\Sigma_i), \quad i = 0, 1, \ldots, n.$$

Under the enforced assumptions, we readily conclude that the rvs $\nu_0^\star, \nu_1^\star, \ldots, \nu_n^\star$ are i.i.d. rvs, each of which is uniformly distributed on $\{1, \ldots, K\}$ (as shown in Proposition 6.1).

5.2. Computing the derivatives

Fix $n = 1, 2, \ldots$. For each $(t_1, \ldots, t_n)$ in $\mathbb{R}^n$, let the rv $R_n(t_1, \ldots, t_n)$ denote the response time of the tagged job under the scenario that *in addition* to the tagged job, exactly n jobs are allowed to enter the system over $\mathbb{R}$, say at times $t_1, \ldots, t_n$, with characteristic pairs $(\sigma_1, \Sigma_1), \ldots, (\sigma_n, \Sigma_n)$ as defined earlier –

The times $t_1, \ldots, t_n$ being arbitrary in $\mathbb{R}$, these additional jobs can arrive either before or after the tagged job. Note that $R_n(t_1, \ldots, t_n)$ depends on the rvs $\{\sigma_0, \sigma_1, \ldots, \sigma_n\}$, $\{\Sigma_0, \Sigma_1, \ldots, \Sigma_n\}$ and $\{\nu_0^\star, \nu_1^\star, \ldots, \nu_n^\star\}$ in a complicated manner through the scheduling policy used. We shall write

$$\widehat{R}_n(t_1, \ldots, t_n) = \mathbb{E}\left[R_n(t_1, \ldots, t_n)\right]. \tag{16}$$

Under some appropriate integrability conditions, Reiman and Simon show that the light-traffic derivatives in the sense of (1) can be expressed in terms of the quantities (16) – Here we consider the cases $n = 0, 1, 2$: Using Theorems 1 and 2 in [11, pp. 29-30] for $n = 0, 1, 2$ we collect the expressions

$$R(0+) = \lim_{\lambda \downarrow 0} R(\lambda) = \widehat{R}_0 \tag{17}$$

with $\widehat{R}_0$ defined in Section 6,

$$R'(0+) = \lim_{\lambda \downarrow 0} \frac{dR}{d\lambda}(\lambda) = \int_{\mathbb{R}} \left(\widehat{R}_1(t) - \widehat{R}_0\right) dt \tag{18}$$

and

$$R''(0+) = \lim_{\lambda \downarrow 0} \frac{d^2R}{d\lambda^2}(\lambda) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} (\widehat{R}_2(s,t) - \widehat{R}_1(s) - \widehat{R}_1(t) + \widehat{R}_0) dt \right) ds. \tag{19}$$

6. The case $n = 0$

The case $n = 0$ is slightly different and corresponds to the scenario when besides the tagged job, no other job enters over the entire horizon $(-\infty, \infty)$. Let R_0 denote the response time of the tagged job under these circumstances. Obviously, under the power-of-two scheduling strategy, we have

$$R_0 = \frac{\sigma_0}{C_{\nu_0}} \quad \text{with } \nu_0 = \nu_0^\star \tag{20}$$

because in the absence of any other job in the system, the tagged job is necessarily assigned to server $\nu_0^\star$. Somewhat in analogy with earlier notation we write

$$\widehat{R}_0 = \mathbb{E}\left[R_0\right].$$

Proposition 6.1. *Under the enforced assumptions, the rv $\nu_0^\star$ is uniformly distributed over $\{1, \ldots, K\}$ with*

$$\mathbb{P}\left[\nu_0^\star = k\right] = \frac{1}{K}, \quad k = 1, \ldots, K \tag{21}$$

and the relation

$$\widehat{R}_0 = \left(\frac{1}{K} \sum_{k=1}^{K} \frac{1}{C_k} \right) \cdot \mathbb{E}\left[\sigma\right] \tag{22}$$

holds.

With Γ given at (6) it will often be convenient to write (22) more compactly as

$$\widehat{R}_0 = \frac{\Gamma}{K} \cdot \mathbb{E}\left[\sigma\right]. \tag{23}$$

Proof. For each $k = 1, \ldots, K$, the definition of $\nu_0^\star$ gives

$$\begin{aligned}
\mathbb{P}\left[\nu_0^\star = k\right] &= \sum_{\ell=1,\ \ell \neq k}^{K} \mathbb{P}\left[\Sigma_0 = \{k, \ell\}, \nu_0^\star = k\right] \\
&= \sum_{\ell=1,\ \ell \neq k}^{K} \mathbb{P}\left[\nu_0^\star = k | \Sigma_0 = \{k, \ell\}\right] \mathbb{P}\left[\Sigma_0 = \{k, \ell\}\right] \\
&= (K-1) \cdot \frac{1}{2} \cdot \frac{2}{K(K-1)} = \frac{1}{K}.
\end{aligned} \tag{24}$$

As pointed earlier, we necessarily have $\nu_0 = \nu_0^\star$. The rvs $\nu_0^\star$ and σ_0 being independent, we then obtain from (20) that

$$\widehat{R}_0 = \mathbb{E}\left[\frac{\sigma_0}{C_{\nu_0^\star}}\right] = \mathbb{E}\left[\sigma_0\right] \cdot \mathbb{E}\left[\frac{1}{C_{\nu_0^\star}}\right],$$

and the conclusion (22) readily follows from (21). ■

According to the Reiman-Simon theory, we have $R(0+) = \widehat{R}_0$ and Proposition 2.1 is established with the help of (22).

7. An auxiliary result

The cases $n = 1$ and $n = 2$ are computationally more involved. The technical result discussed next will simplify the presentation by isolating an evaluation which is repeatedly carried out during the analysis. This auxiliary result is given in a setting that mimics power-of-two scheduling with only two jobs present:

Fix $y < 0$. In addition to the tagged job arriving at time $t = 0$ with characteristic pair (σ_0, Σ_0), assume that another job arrives at time y with (random) service requirement τ. This y-job is then assigned to the server γ, with γ being some $\{1, \ldots, K\}$-valued rv, while the tagged job is assigned to the server γ_0 (in Σ_0) in accordance with the power-of-two scheduling policy. Thus, if $y + \frac{\tau}{C_\gamma} \leq 0$, then $\gamma_0 = \nu_0^\star$. On the other hand, if $y + \frac{\tau}{C_\gamma} > 0$, then the operational rules of the power-of-two scheduling policy will preclude the tagged job to be assigned to server γ: Indeed, if γ is not in Σ_0, then $\gamma_0 = \nu_0^\star$ again, while if γ is an element of Σ_0, then γ_0 is necessarily the other server in the pair Σ_0, i.e., the one different from γ. In this scenario γ is a rv given *a priori*, and should be thought as a place holder for a server assignment rv determined via power-of-two scheduling under various circumstances. On the other hand, γ_0 depends on y, τ, γ and Σ_0 (as well as $\nu_0^\star$). The explicit dependence on these quantities will be dropped from the notation.

Lemma 7.1 given next is used in the proofs of Proposition 8.1, Lemma 11.2 and Lemma 11.4. The formulation of Lemma 7.1 contains an event E (with indicator $\mathbf{1}[E]$) that will be specified in each of its applications.

Lemma 7.1. *Given are the rvs τ, γ, σ_0, Σ_0 and $\nu_0^\star$, together with the event E. We assume that (i) the rv $\nu_0^\star$ is uniformly distributed on Σ_0 conditionally on all the other rvs $\mathbf{1}[E]$, τ, γ, σ_0 and Σ_0; (ii) the collections of rvs $\{\mathbf{1}[E], \tau, \gamma\}$ and $\{\nu_0^\star, \Sigma_0, \sigma_0\}$ are independent; and (iii) the rvs Σ_0 and σ_0 are independent. Then, for each $y < 0$ and each $k = 1, \ldots, K$, we have*

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}[E]\,\mathbf{1}[\gamma = k]\,\frac{\sigma_0}{C_{\gamma_0}}\right] &= \mathbb{P}\left[E, \gamma = k, y + \frac{\tau}{C_k} \leq 0\right] \cdot \widehat{R}_0 \\ &\quad + \frac{1}{K-1}\mathbb{P}\left[E, \gamma = k, y + \frac{\tau}{C_k} > 0\right]\left(\Gamma - \frac{1}{C_k}\right) \cdot \mathbb{E}\left[\sigma_0\right] \qquad (25)\end{aligned}$$

with γ_0 as defined earlier.

Recall that under the enforced assumptions, the rv σ_0 is independent of the collection of rvs $\{\nu_0^\star, \Sigma_0\}$; see Section 5.1.

Proof. Fix $k = 1, \ldots, K$. We start with the natural decomposition

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}[E]\,\mathbf{1}[\gamma = k]\,\frac{\sigma_0}{C_{\gamma_0}}\right] \qquad (26)\\ &= \mathbb{E}\left[\mathbf{1}[E]\,\mathbf{1}[\gamma = k]\,\mathbf{1}\left[y + \frac{\tau}{C_\gamma} \leq 0\right]\frac{\sigma_0}{C_{\gamma_0}}\right] + \mathbb{E}\left[\mathbf{1}[E]\,\mathbf{1}[\gamma = k]\,\mathbf{1}\left[y + \frac{\tau}{C_\gamma} > 0\right]\frac{\sigma_0}{C_{\gamma_0}}\right].\end{aligned}$$

For the first term, the definition of γ_0 leads to

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}[E]\,\mathbf{1}[\gamma = k]\,\mathbf{1}\left[y + \frac{\tau}{C_\gamma} \leq 0\right]\frac{\sigma_0}{C_{\gamma_0}}\right] &= \mathbb{E}\left[\mathbf{1}[E]\,\mathbf{1}[\gamma = k]\,\mathbf{1}\left[y + \frac{\tau}{C_k} \leq 0\right]\frac{\sigma_0}{C_{\nu_0^\star}}\right] \\ &= \mathbb{E}\left[\mathbf{1}[E]\,\mathbf{1}[\gamma = k]\,\mathbf{1}\left[y + \frac{\tau}{C_k} \leq 0\right]\right]\mathbb{E}\left[\frac{\sigma_0}{C_{\nu_0^\star}}\right]\end{aligned}$$

$$= \quad \mathbb{P}\left[E, \gamma = k, y + \frac{\tau}{C_k} \leq 0\right] \cdot \widehat{R}_0 \tag{27}$$

since the collections $\{\mathbf{1}[E], \gamma, \tau\}$ and $\{\nu_0^\star, \sigma_0\}$ are independent under the enforced assumptions.

We further decompose the second term in (26) to obtain

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}[E]\,\mathbf{1}[\gamma = k]\,\mathbf{1}\left[y + \frac{\tau}{C_\gamma} > 0\right]\frac{\sigma_0}{C_{\gamma_0}}\right] \\
&= \quad \mathbb{E}\left[\mathbf{1}[E]\,\mathbf{1}[\gamma = k]\,\mathbf{1}\left[y + \frac{\tau}{C_k} > 0\right]\mathbf{1}[k \notin \Sigma_0]\frac{\sigma_0}{C_{\nu_0^\star}}\right] \\
&\qquad + \mathbb{E}\left[\mathbf{1}[E]\,\mathbf{1}[\gamma = k]\,\mathbf{1}\left[y + \frac{\tau}{C_k} > 0\right]\mathbf{1}[k \in \Sigma_0]\frac{\sigma_0}{C_{\gamma_0}}\right].
\end{aligned} \tag{28}$$

It is plain that

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}[E]\,\mathbf{1}[\gamma = k]\,\mathbf{1}\left[y + \frac{\tau}{C_k} > 0\right]\mathbf{1}[k \in \Sigma_0]\frac{\sigma_0}{C_{\gamma_0}}\right] \\
&= \quad \sum_{\ell=1,\ell\neq k}^{K} \mathbb{E}\left[\mathbf{1}[E]\,\mathbf{1}[\gamma = k]\,\mathbf{1}\left[y + \frac{\tau}{C_k} > 0\right]\mathbf{1}[\Sigma_0 = \{k,\ell\}]\frac{\sigma_0}{C_{\gamma_0}}\right] \\
&= \quad \sum_{\ell=1,\ell\neq k}^{K} \mathbb{E}\left[\mathbf{1}[E]\,\mathbf{1}[\gamma = k]\,\mathbf{1}\left[y + \frac{\tau}{C_k} > 0\right]\mathbf{1}[\Sigma_0 = \{k,\ell\}]\frac{\sigma_0}{C_\ell}\right] \\
&= \quad \frac{2}{K(K-1)}\sum_{\ell=1,\ell\neq k}^{K} \mathbb{P}\left[E, \gamma = k, y + \frac{\tau}{C_k} > 0\right]\frac{\mathbb{E}[\sigma_0]}{C_\ell} \\
&= \quad \frac{2}{K(K-1)}\left(\sum_{\ell=1,\ell\neq k}^{K}\frac{\mathbb{E}[\sigma_0]}{C_\ell}\right)\mathbb{P}\left[E, \gamma = k, y + \frac{\tau}{C_k} > 0\right] \\
&= \quad \frac{2}{K(K-1)}\left(\Gamma - \frac{1}{C_k}\right)\mathbb{P}\left[E, \gamma = k, y + \frac{\tau}{C_k} > 0\right]\cdot\mathbb{E}[\sigma_0]
\end{aligned} \tag{29}$$

since $\gamma_0 = \ell$ if $\Sigma_0 = \{k, \ell\}$ when $\gamma = k$ and $y + \frac{\tau}{C_k} > 0$.

On the other hand, the definition of $\nu_0^\star$ implies

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}[E]\,\mathbf{1}[\gamma = k]\,\mathbf{1}\left[y + \frac{\tau}{C_k} > 0\right]\mathbf{1}[k \notin \Sigma_0]\frac{\sigma_0}{C_{\nu_0^\star}}\right] \\
&= \quad \sum_{T\in\mathcal{P}_2(K)} \mathbb{E}\left[\mathbf{1}[E]\,\mathbf{1}[\gamma = k]\,\mathbf{1}\left[y + \frac{\tau}{C_k} > 0\right]\mathbf{1}[k \notin \Sigma_0]\,\mathbf{1}[\Sigma_0 = T]\frac{\sigma_0}{C_{\nu_0^\star}}\right] \\
&= \quad \sum_{a=1,a\neq k}^{K}\left(\sum_{b=1,b\neq k}^{a-1} \mathbb{E}\left[\mathbf{1}[E]\,\mathbf{1}[\gamma = k]\,\mathbf{1}\left[y + \frac{\tau}{C_k} > 0\right]\mathbf{1}[\Sigma_0 = \{a,b\}]\frac{\sigma_0}{C_{\nu_0^\star}}\right]\right) \\
&= \quad \sum_{a=1,a\neq k}^{K}\left(\sum_{b=1,b\neq k}^{a-1} \mathbb{P}\left[E, \gamma = k, y + \frac{\tau}{C_k} > 0\right]\mathbb{E}\left[\mathbf{1}[\Sigma_0 = \{a,b\}]\frac{\sigma_0}{C_{\nu_0^\star}}\right]\right) \\
&= \quad \sum_{a=1,a\neq k}^{K} \mathbb{P}\left[E, \gamma = k, y + \frac{\tau}{C_k} > 0\right]\left(\sum_{b=1,b\neq k}^{a-1}\mathbb{P}[\Sigma_0 = \{a,b\}]\cdot\frac{1}{2}\left(\frac{1}{C_a} + \frac{1}{C_b}\right)\right)\cdot\mathbb{E}[\sigma_0] \\
&= \quad \frac{1}{K(K-1)}\sum_{a=1,a\neq k}^{K} \mathbb{P}\left[E, \gamma = k, y + \frac{\tau}{C_k} > 0\right]\left(\sum_{b=1,b\neq k}^{a-1}\left(\frac{1}{C_a} + \frac{1}{C_b}\right)\right)\cdot\mathbb{E}[\sigma_0] \\
&= \quad \frac{1}{K(K-1)}\left(\sum_{a=1,a\neq k}^{K}\left(\sum_{b=1,b\neq k}^{a-1}\left(\frac{1}{C_a} + \frac{1}{C_b}\right)\right)\right)\mathbb{P}\left[E, \gamma = k, y + \frac{\tau}{C_k} > 0\right]\cdot\mathbb{E}[\sigma_0].
\end{aligned} \tag{30}$$

Uninteresting calculations (with details available in Appendix B of [5]) yield

$$\sum_{a=1,a\neq k}^{K}\left(\sum_{b=1,b\neq k}^{a-1}\left(\frac{1}{C_a}+\frac{1}{C_b}\right)\right)=(K-2)\left(\Gamma-\frac{1}{C_k}\right), \tag{31}$$

so that (30) can be written more compactly as

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}\left[E\right]\mathbf{1}\left[\gamma=k\right]\mathbf{1}\left[y+\frac{\tau}{C_k}>0\right]\mathbf{1}\left[k\notin\Sigma_0\right]\frac{\sigma_0}{C_{\nu_0^\star}}\right]\\ &=\quad\frac{K-2}{K(K-1)}\left(\Gamma-\frac{1}{C_k}\right)\mathbb{P}\left[E,\gamma=k,y+\frac{\tau}{C_k}>0\right]\cdot\mathbb{E}\left[\sigma_0\right].\end{aligned} \tag{32}$$

To conclude the proof, substitute (29) and (32) into (28). It yields

$$\mathbb{E}\left[\mathbf{1}\left[E\right]\mathbf{1}\left[\gamma=k\right]\mathbf{1}\left[y+\frac{\tau}{C_\gamma}>0\right]\frac{\sigma_0}{C_{\gamma_0}}\right]=\frac{1}{K-1}\left(\Gamma-\frac{1}{C_k}\right)\mathbb{P}\left[E,\gamma=k,y+\frac{\tau}{C_k}>0\right]\cdot\mathbb{E}\left[\sigma_0\right],$$

and combining this last expression with (27) we get the desired result (25) with the help of (26). ■

8. The case $n = 1$

The analysis of the first derivative is associated with the following scenario: The tagged job arrives at time $t = 0$ with characteristic pair (σ_0, Σ_0). With t in $\mathbb{R}$, in addition to the tagged job, a single job arrives during the entire horizon $(-\infty, \infty)$, say at time t with characteristic pair (σ_t, Σ_t). The tagged job and this t-job are assigned to the servers ν_0 (in Σ_0) and ν_t (in Σ_t), respectively, in accordance with the power-of-two scheduling policy.

8.1. Evaluating $\widehat{R}_1(t)$

For each t in $\mathbb{R}$, in accordance with (16) we have

$$\widehat{R}_1(t)=\mathbb{E}\left[R_1(t)\right]\quad\text{with } R_1(t)=\frac{\sigma_0}{C_{\nu_0}}. \tag{33}$$

However, with the presence of the t-job, ν_0 does not always coincide with $\nu_0^\star$, as the determination of ν_0 may be affected by whether the t-job completed service at the time the tagged job arrives.

First some notation: With t arbitrary in $\mathbb{R}$, set

$$H_k(t)=\mathbb{P}\left[C_kt+\sigma_t\leq 0\right]\cdot\widehat{R}_0+\frac{1}{K-1}\left(\Gamma-\frac{1}{C_k}\right)\mathbb{P}\left[C_kt+\sigma_t>0\right]\cdot\mathbb{E}\left[\sigma_0\right] \tag{34}$$

for each $k = 1, \ldots, K$. Note that

$$\begin{aligned}H_k(t)&=\widehat{R}_0\cdot\left(1-\mathbb{P}\left[C_kt+\sigma_t>0\right]\right)+\frac{1}{K-1}\left(\Gamma-\frac{1}{C_k}\right)\mathbb{E}\left[\sigma_0\right]\cdot\mathbb{P}\left[C_kt+\sigma_t>0\right]\\ &=\widehat{R}_0+\left(\frac{1}{K-1}\left(\Gamma-\frac{1}{C_k}\right)-\frac{\Gamma}{K}\right)\mathbb{E}\left[\sigma_0\right]\cdot\mathbb{P}\left[C_kt+\sigma_t>0\right]\\ &=\widehat{R}_0+\frac{1}{K-1}\left(\frac{\Gamma}{K}-\frac{1}{C_k}\right)\mathbb{P}\left[C_kt+\sigma_t>0\right]\cdot\mathbb{E}\left[\sigma_0\right]\end{aligned} \tag{35}$$

as we make use of the expression (23).

Proposition 8.1. *Under the enforced independence assumptions, we have $\widehat{R}_1(t) = \widehat{R}_0$ if $t > 0$, while for $t < 0$ it holds that*

$$\widehat{R}_1(t)=\frac{1}{K}\sum_{k=1}^{K}H_k(t). \tag{36}$$

Proof. Fix t in $\mathbb{R}$. As we seek to evaluate $\widehat{R}_1(t)$ as given by (33), two cases need to be examined: If $t > 0$, then $\nu_0 = \nu_0^\star$, whence $R_1(t) = R_0$, and the conclusion $\widehat{R}_1(t) = \widehat{R}_0$ follows.

If $t < 0$, then $\nu_t = \nu_t^\star$ and we are in the setting of Lemma 7.1 with $y = t$, $E = \Omega$, $\tau = \sigma_t$ and $\gamma = \nu_t^\star$ (so that $\gamma_0 = \nu_0$): For each $k = 1, \ldots, K$, the expression (25) becomes

$$\begin{aligned} & \mathbb{E}\left[\mathbf{1}\left[\nu_t^\star = k\right] \frac{\sigma_0}{C_{\nu_0}}\right] \\ = \; & \mathbb{P}\left[\nu_t^\star = k, t + \frac{\sigma_t}{C_k} \le 0\right] \cdot \widehat{R}_0 + \frac{1}{K-1}\mathbb{P}\left[\nu_t^\star = k, t + \frac{\sigma_t}{C_k} > 0\right]\left(\Gamma - \frac{1}{C_k}\right) \cdot \mathbb{E}\left[\sigma_0\right] \\ = \; & \frac{1}{K} \cdot H_k(t) \end{aligned} \tag{37}$$

since the rv $\nu_t^\star$ is independent of σ_t and uniformly distributed on $\{1, \ldots, K\}$ (as pointed out in Proposition 6.1). The desired result (36) now follows from (33) upon noting the decomposition

$$\widehat{R}_1(t) = \sum_{k=1}^{K} \mathbb{E}\left[\mathbf{1}\left[\nu_t^\star = k\right] \frac{\sigma_0}{C_{\nu_0}}\right].$$

■

8.2. A proof of Proposition 2.2

We can now complete the proof of Proposition 2.2: The expression (18) now takes the form

$$R'(0+) = \int_{\mathbb{R}} \left(\widehat{R}_1(t) - \widehat{R}_0\right) dt = \int_{-\infty}^{0} \left(\widehat{R}_1(t) - \widehat{R}_0\right) dt \tag{38}$$

as we recall that $\widehat{R}_1(t) = \widehat{R}_0$ for $t > 0$. Next, for $t < 0$, with the help of (35) and (36) we can rewrite the integrand as

$$\widehat{R}_1(t) - \widehat{R}_0 = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{K-1}\left(\frac{\Gamma}{K} - \frac{1}{C_k}\right) \mathbb{P}\left[C_k t + \sigma > 0\right] \cdot \mathbb{E}\left[\sigma\right] \tag{39}$$

as we recall that the rvs σ_t and σ_0 are both distributed like σ.

Inserting this expression back into (38) we get

$$\int_{-\infty}^{0} \left(\widehat{R}_1(t) - \widehat{R}_0\right) dt = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{K-1}\left(\frac{\Gamma}{K} - \frac{1}{C_k}\right) \frac{\mathbb{E}\left[\sigma\right]}{C_k} \cdot \mathbb{E}\left[\sigma\right] \tag{40}$$

upon noting that

$$\int_{-\infty}^{0} \mathbb{P}\left[C_k t + \sigma > 0\right] dt = \frac{1}{C_k} \int_{0}^{\infty} \mathbb{P}\left[\sigma > x\right] dx = \frac{\mathbb{E}\left[\sigma\right]}{C_k}, \quad k = 1, \ldots, K \tag{41}$$

by a simple change of variable. Uninteresting algebra on (40) readily yield (7) with the help of (23), and this completes the proof of Proposition 2.2 . ■

9. The case $n = 2$

The computation of the second derivative is given under the following scenario: The tagged job arrives at time $t = 0$ with characteristic pair (σ_0, Σ_0). With s and t in $\mathbb{R}$, in addition to the tagged job, exactly two jobs arrive over the entire horizon $(-\infty, \infty)$, say at times s and t with characteristic pairs (σ_s, Σ_s) and (σ_t, Σ_t), respectively. The tagged job, the s-job and the t-job are assigned to their respective servers ν_0 (in Σ_0), ν_s (in Σ_s) and ν_t (in Σ_t) in accordance with the power-of-two load balancing scheduling policy.

9.1. Evaluating $\widehat{R}_2(s,t)$

For each s and t in $\mathbb{R}$, we have

$$\widehat{R}_2(s,t) = \mathbb{E}\left[R_2(s,t)\right] \quad \text{with } R_2(s,t) = \frac{\sigma_0}{C_{\nu_0}}. \tag{42}$$

The server assignment rvs ν_0, ν_s and ν_t do not always coincide with $\nu_0^\star$, $\nu_s^\star$ and $\nu_t^\star$, respectively, because these rvs may be affected by whether earlier jobs have completed service by the time server selection needs to be determined.

Proposition 9.1. *Under the enforced independence assumptions, we have $\widehat{R}_2(s,t) = \widehat{R}_0$ for $0 < s < t$ and $\widehat{R}_2(s,t) = \widehat{R}_1(s)$ for $s < 0 < t$, while for $s < t < 0$, it holds that*

$$\begin{aligned}
\widehat{R}_2(s,t) \;=\; & \left(\frac{1}{K}\sum_{k=1}^{K}\mathbb{P}\left[s + \frac{\sigma_s}{C_k} \le t\right]\right)\cdot \widehat{R}_1(t) \\
& + \frac{1}{K(K-1)}\sum_{k=1}^{K}\left(\sum_{\ell=1,\ell\neq k}^{K}\mathbb{P}\left[t < s + \frac{\sigma_s}{C_\ell} \le 0\right]\right)\cdot H_k(t) \\
& + \frac{1}{K(K-1)^2}\sum_{k=1}^{K}\sum_{\ell=1,\ell\neq k}^{K}\left(\Gamma - \frac{1}{C_\ell}\right)\mathbb{P}\left[C_\ell s + \sigma_s > 0\right]\mathbb{P}\left[C_k t + \sigma_t \le 0\right]\cdot \mathbb{E}\left[\sigma_0\right] \\
& + \frac{1}{K^2(K-1)^2}\sum_{k=1}^{K}\sum_{\ell=1,\ell\neq k}^{K} J_{k\ell}\cdot \mathbb{P}\left[C_k s + \sigma_s > 0\right]\mathbb{P}\left[C_\ell t + \sigma_t > 0\right]\cdot \mathbb{E}\left[\sigma_0\right]
\end{aligned} \tag{43}$$

with

$$J_{k\ell} = (K+1)\Gamma - K\left(\frac{1}{C_k} + \frac{1}{C_\ell}\right), \quad k,\ell = 1,\ldots,K. \tag{44}$$

Before starting the proof of Proposition 9.1 in Section 11, we pause to give a more compact expression for (43).

9.2. Towards a more compact expression for (43)

As we focus on the last two terms in (43), interchange the dummy indices k and ℓ, and then change the order of summations in the resulting expression. We can readily check that

$$\begin{aligned}
& \frac{1}{K(K-1)^2}\sum_{k=1}^{K}\sum_{\ell=1,\ell\neq k}^{K}\left(\Gamma - \frac{1}{C_\ell}\right)\mathbb{P}\left[C_\ell s + \sigma_s > 0\right]\mathbb{P}\left[C_k t + \sigma_t \le 0\right]\cdot \mathbb{E}\left[\sigma_0\right] \\
& + \frac{1}{K^2(K-1)^2}\sum_{k=1}^{K}\sum_{\ell=1,\ell\neq k}^{K} J_{k\ell}\cdot \mathbb{P}\left[C_k s + \sigma_s > 0\right]\mathbb{P}\left[C_\ell t + \sigma_t > 0\right]\cdot \mathbb{E}\left[\sigma_0\right] \\
= \;& \frac{1}{K^2(K-1)^2}\sum_{k=1}^{K}\sum_{\ell=1,\ell\neq k}^{K} G_{k\ell}(s,t)\cdot \mathbb{E}\left[\sigma_0\right]
\end{aligned} \tag{45}$$

with

$$\begin{aligned}
G_{k\ell}(s,t) \;=\; & K\left(\Gamma - \frac{1}{C_k}\right)\mathbb{P}\left[C_k s + \sigma_s > 0\right]\mathbb{P}\left[C_\ell t + \sigma_t \le 0\right] \\
& + \left((K+1)\Gamma - K\left(\frac{1}{C_k} + \frac{1}{C_\ell}\right)\right)\cdot \mathbb{P}\left[C_k s + \sigma_s > 0\right]\mathbb{P}\left[C_\ell t + \sigma_t > 0\right] \\
= \;& K\left(\Gamma - \frac{1}{C_k}\right)\mathbb{P}\left[C_k s + \sigma_s > 0\right] \\
& + K\left(\frac{\Gamma}{K} - \frac{1}{C_\ell}\right)\mathbb{P}\left[C_k s + \sigma_s > 0\right]\mathbb{P}\left[C_\ell t + \sigma_t > 0\right]
\end{aligned} \tag{46}$$

for every $k, \ell = 1, \dots, K$. Upon substitution into (43), we then conclude that

$$\begin{aligned}
\widehat{R}_2(s,t) &= \left(\frac{1}{K}\sum_{k=1}^{K}\mathbb{P}\left[s+\frac{\sigma_s}{C_k}\le t\right]\right)\cdot\widehat{R}_1(t)\\
&\quad+\frac{1}{K(K-1)}\sum_{k=1}^{K}\left(\sum_{\ell=1,\ell\neq k}^{K}\mathbb{P}\left[t<s+\frac{\sigma_s}{C_\ell}\le 0\right]\right)\cdot H_k(t)\\
&\quad+\frac{1}{K(K-1)^2}\sum_{k=1}^{K}\sum_{\ell=1,\ell\neq k}^{K}\left(\frac{\Gamma}{K}-\frac{1}{C_\ell}\right)\mathbb{P}\left[C_k s+\sigma_s>0\right]\mathbb{P}\left[C_\ell t+\sigma_t>0\right]\cdot\mathbb{E}\left[\sigma_0\right]\\
&\quad+\frac{1}{K(K-1)}\sum_{k=1}^{K}\left(\Gamma-\frac{1}{C_k}\right)\mathbb{P}\left[C_k s+\sigma_s>0\right]\cdot\mathbb{E}\left[\sigma_0\right]. \qquad (47)
\end{aligned}$$

Next using (35) we get

$$\begin{aligned}
&\sum_{k=1}^{K}\left(\sum_{\ell=1,\ell\neq k}^{K}\mathbb{P}\left[t<s+\frac{\sigma_s}{C_\ell}\le 0\right]\right)\cdot H_k(t)\\
&= \sum_{k=1}^{K}\left(\sum_{\ell=1,\ell\neq k}^{K}\mathbb{P}\left[t<s+\frac{\sigma_s}{C_\ell}\le 0\right]\right)\left(\widehat{R}_0+\frac{1}{K-1}\left(\frac{\Gamma}{K}-\frac{1}{C_k}\right)\mathbb{P}\left[C_k t+\sigma_t>0\right]\cdot\mathbb{E}\left[\sigma_0\right]\right)
\end{aligned}$$

and the second term in (47) becomes

$$\begin{aligned}
&\frac{1}{K(K-1)}\sum_{k=1}^{K}\left(\sum_{\ell=1,\ell\neq k}^{K}\mathbb{P}\left[t<s+\frac{\sigma_s}{C_\ell}\le 0\right]\right)\cdot H_k(t)\\
&= \left(\frac{1}{K}\sum_{\ell=1}^{K}\mathbb{P}\left[t<s+\frac{\sigma_s}{C_\ell}\le 0\right]\right)\widehat{R}_0\\
&\quad+\frac{1}{K(K-1)^2}\sum_{k=1}^{K}\sum_{\ell=1,\ell\neq k}^{K}\left(\frac{\Gamma}{K}-\frac{1}{C_k}\right)\mathbb{P}\left[t<s+\frac{\sigma_s}{C_\ell}\le 0\right]\mathbb{P}\left[C_k t+\sigma_t>0\right]\cdot\mathbb{E}\left[\sigma_0\right].
\end{aligned}$$

Substituting this last expression into (47) we readily get the following more compact expression for (43).

Proposition 9.2. *Under the enforced independence assumptions, for $s<t<0$, it holds that*

$$\begin{aligned}
\widehat{R}_2(s,t) &= \left(\frac{1}{K}\sum_{\ell=1}^{K}\mathbb{P}\left[s+\frac{\sigma_s}{C_\ell}\le t\right]\right)\cdot\widehat{R}_1(t)+\left(\frac{1}{K}\sum_{\ell=1}^{K}\mathbb{P}\left[t<s+\frac{\sigma_s}{C_\ell}\le 0\right]\right)\cdot\widehat{R}_0\\
&\quad+\frac{1}{K(K-1)}\sum_{k=1}^{K}\left(\Gamma-\frac{1}{C_k}\right)\mathbb{P}\left[C_k s+\sigma_s>0\right]\cdot\mathbb{E}\left[\sigma_0\right]+\frac{1}{K(K-1)^2}\cdot H(s,t)
\end{aligned}$$

where we have set

$$H(s,t)=\sum_{\ell=1}^{K}\sum_{k=1,k\neq\ell}^{K}\left(\frac{\Gamma}{K}-\frac{1}{C_\ell}\right)\mathbb{P}\left[C_k t<C_k s+\sigma_s\right]\mathbb{P}\left[C_\ell t+\sigma_t>0\right]\cdot\mathbb{E}\left[\sigma_0\right].$$

10. A proof of Proposition 2.3

Our point of departure is the expression (19). For notational simplicity we shall write

$$R^\star(s,t)=\widehat{R}_2(s,t)-\widehat{R}_1(s)-\widehat{R}_1(t)+\widehat{R}_0,\quad s,t\in\mathbb{R}.$$

10.1. The integral to be evaluated

We start with

$$\begin{aligned} R''(0+) &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} R^\star(s,t)dt \right) ds \\ &= \int_{\mathbb{R}} \left(\int_{-\infty}^{s} R^\star(s,t)dt \right) ds + \int_{\mathbb{R}} \left(\int_{s}^{\infty} R^\star(s,t)dt \right) ds. \end{aligned} \tag{48}$$

The second term in this expression can be written as

$$\int_{\mathbb{R}} \left(\int_{s}^{\infty} R^\star(s,t)dt \right) ds = \int_{-\infty}^{0} \left(\int_{s}^{\infty} R^\star(s,t)dt \right) ds + \int_{0}^{\infty} \left(\int_{s}^{\infty} R^\star(s,t)dt \right) ds. \tag{49}$$

Now, by Propositions 8.1 and 9.1 we have $R^\star(s,t) = \widehat{R}_0 - \widehat{R}_0 - \widehat{R}_0 + \widehat{R}_0 = 0$ whenever $0 < s < t$, and the conclusion

$$\int_{0}^{\infty} \left(\int_{s}^{\infty} R^\star(s,t)dt \right) ds = 0 \tag{50}$$

follows.

Next, consider the decomposition

$$\int_{-\infty}^{0} \left(\int_{s}^{\infty} R^\star(s,t)dt \right) ds = \int_{-\infty}^{0} \left(\int_{s}^{0} R^\star(s,t)dt \right) ds + \int_{-\infty}^{0} \left(\int_{0}^{\infty} R^\star(s,t)dt \right) ds. \tag{51}$$

On the range $s < 0 < t$, Propositions 8.1 and 9.1 yield $\widehat{R}_2(s,t) = \widehat{R}_1(s)$ and $\widehat{R}_1(t) = \widehat{R}_0$, whence $R^\star(s,t) = \widehat{R}_1(s) - \widehat{R}_1(s) - \widehat{R}_0 + \widehat{R}_0 = 0$ again, so that

$$\int_{-\infty}^{0} \left(\int_{0}^{\infty} R^\star(s,t)dt \right) ds = 0.$$

Combining (49), (50) and (51), we conclude that the second term in (48) reduces to

$$\int_{\mathbb{R}} \left(\int_{s}^{\infty} R^\star(s,t)dt \right) ds = \int_{-\infty}^{0} \left(\int_{s}^{0} R^\star(s,t)dt \right) ds. \tag{52}$$

Finally, returning to the first term in the decomposition (48) we get

$$\begin{aligned} \int_{\mathbb{R}} \left(\int_{-\infty}^{s} R^\star(s,t)dt \right) ds &= \int_{\mathbb{R}} \left(\int_{t}^{\infty} R^\star(s,t)ds \right) dt \\ &= \int_{\mathbb{R}} \left(\int_{s}^{\infty} R^\star(t,s)dt \right) ds \\ &= \int_{\mathbb{R}} \left(\int_{s}^{\infty} R^\star(s,t)dt \right) ds \end{aligned} \tag{53}$$

as we note that $R^\star(t,s) = R^\star(s,t)$ for arbitrary s,t in $\mathbb{R}$ since the symmetry $\widehat{R}_2(t,s) = \widehat{R}_2(s,t)$ holds under the enforced statistical assumptions. It follows from (48) that

$$R''(0+) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} R^\star(s,t)dt \right) ds = 2 \int_{\mathbb{R}} \left(\int_{s}^{\infty} R^\star(s,t)dt \right) ds = 2 \int_{-\infty}^{0} \left(\int_{s}^{0} R^\star(s,t)dt \right) ds \tag{54}$$

as we combine (48), (52) and (53).

10.2. Computing the integrand in (54)

On the way to evaluating the integral (53) we consider $R^\star(s,t)$ for $s < t < 0$. On that range, applying (39) with t replaced by s yields

$$\widehat{R}_1(s) - \widehat{R}_0 = \frac{1}{K(K-1)} \sum_{k=1}^{K} \left(\frac{\Gamma}{K} - \frac{1}{C_k} \right) \mathbb{P}\left[C_k s + \sigma_s > 0 \right] \cdot \mathbb{E}\left[\sigma_0 \right]. \tag{55}$$

Using the expression for $\widehat{R}_2(s,t)$ in Proposition 9.2 and recalling the expression for $\widehat{R}_0$ we then readily get

$$\begin{aligned}
R^\star(s,t) &= \left(\widehat{R}_0 - \widehat{R}_1(s)\right) + \left(\widehat{R}_2(s,t) - \widehat{R}_1(t)\right) \\
&= -\frac{1}{K(K-1)} \sum_{k=1}^{K} \left(\frac{\Gamma}{K} - \frac{1}{C_k}\right) \mathbb{P}\left[C_k s + \sigma_s > 0\right] \cdot \mathbb{E}\left[\sigma_0\right] \\
&\quad + \left(\frac{1}{K} \sum_{\ell=1}^{K} \mathbb{P}\left[t < s + \frac{\sigma_s}{C_\ell} \leq 0\right]\right) \cdot \widehat{R}_0 + \left(\frac{1}{K} \sum_{\ell=1}^{K} \mathbb{P}\left[s + \frac{\sigma_s}{C_\ell} \leq t\right] - 1\right) \cdot \widehat{R}_1(t) \\
&\quad + \frac{1}{K(K-1)} \sum_{k=1}^{K} \left(\Gamma - \frac{1}{C_k}\right) \mathbb{P}\left[C_k s + \sigma_s > 0\right] \cdot \mathbb{E}\left[\sigma_0\right] + \frac{1}{K(K-1)^2} \cdot H(s,t) \\
&= \left(\frac{1}{K} \sum_{\ell=1}^{K} \mathbb{P}\left[t < s + \frac{\sigma_s}{C_\ell} \leq 0\right]\right) \cdot \widehat{R}_0 - \left(\frac{1}{K} \sum_{\ell=1}^{K} \mathbb{P}\left[s + \frac{\sigma_s}{C_\ell} > t\right]\right) \cdot \widehat{R}_1(t) \\
&\quad + \frac{1}{K} \sum_{k=1}^{K} \mathbb{P}\left[C_k s + \sigma_s > 0\right] \cdot \left(\frac{\Gamma}{K} \mathbb{E}\left[\sigma_0\right]\right) + \frac{1}{K(K-1)^2} \cdot H(s,t) \\
&= \left(\frac{1}{K} \sum_{k=1}^{K} \mathbb{P}\left[C_k t < C_k s + \sigma_s\right]\right) \cdot \widehat{R}_0 - \left(\frac{1}{K} \sum_{k=1}^{K} \mathbb{P}\left[s + \frac{\sigma_s}{C_k} > t\right]\right) \cdot \widehat{R}_1(t) \\
&\quad + \frac{1}{K(K-1)^2} \cdot H(s,t) \\
&= \left(\frac{1}{K} \sum_{k=1}^{K} \mathbb{P}\left[C_k t < C_k s + \sigma_s\right]\right) \cdot \left(\widehat{R}_0 - \widehat{R}_1(t)\right) + \frac{1}{K(K-1)^2} \cdot H(s,t). \qquad (56)
\end{aligned}$$

10.3. Evaluating (54)

Next, recall that in these expressions the rvs σ_s and σ_0 are distributed like σ. Thus, after a change of order of integration and a change of variable, we note that

$$\begin{aligned}
&\int_{-\infty}^{0} \left(\int_{s}^{0} \left(\frac{1}{K} \sum_{k=1}^{K} \mathbb{P}\left[C_k t < C_k s + \sigma\right]\right) \cdot \left(\widehat{R}_0 - \widehat{R}_1(t)\right) dt\right) ds \\
&= \frac{1}{K} \sum_{k=1}^{K} \int_{-\infty}^{0} \left(\int_{s}^{0} \mathbb{P}\left[C_k t < C_k s + \sigma\right] \cdot \left(\widehat{R}_0 - \widehat{R}_1(t)\right) dt\right) ds \\
&= \frac{1}{K} \sum_{k=1}^{K} \int_{-\infty}^{0} \left(\int_{-\infty}^{t} \mathbb{P}\left[C_k t < C_k s + \sigma\right] \cdot \left(\widehat{R}_0 - \widehat{R}_1(t)\right) ds\right) dt \\
&= \frac{1}{K} \sum_{k=1}^{K} \int_{-\infty}^{0} \left(\int_{-\infty}^{t} \mathbb{P}\left[C_k t < C_k s + \sigma\right] ds\right) \cdot \left(\widehat{R}_0 - \widehat{R}_1(t)\right) dt \\
&= \frac{1}{K} \sum_{k=1}^{K} \int_{-\infty}^{0} \left(\int_{0}^{\infty} \mathbb{P}\left[C_k x < \sigma\right] dx\right) \cdot \left(\widehat{R}_0 - \widehat{R}_1(t)\right) dt \\
&= \frac{1}{K} \sum_{k=1}^{K} \int_{-\infty}^{0} \frac{\mathbb{E}\left[\sigma\right]}{C_k} \cdot \left(\widehat{R}_0 - \widehat{R}_1(t)\right) dt \\
&= -\left(\int_{-\infty}^{0} \left(\widehat{R}_1(t) - \widehat{R}_0\right) dt\right) \cdot \frac{\Gamma}{K} \mathbb{E}\left[\sigma\right] \\
&= -\left(\frac{1}{K(K-1)} \sum_{k=1}^{K} \left(\frac{\Gamma}{K} - \frac{1}{C_k}\right) \frac{\mathbb{E}\left[\sigma\right]}{C_k} \cdot \mathbb{E}\left[\sigma\right]\right) \cdot \frac{\Gamma}{K} \mathbb{E}\left[\sigma\right] \\
&= \left(\frac{1}{K} \sum_{k=1}^{K} \frac{1}{C_k^2} - \left(\frac{\Gamma}{K}\right)^2\right) \cdot \frac{\Gamma}{K(K-1)} \left(\mathbb{E}\left[\sigma\right]\right)^3 \qquad (57)
\end{aligned}$$

where the step before last made used of the expression (40).

In a similar vein, we find that

$$\int_{-\infty}^{0}\left(\int_{s}^{0}H(s,t)dt\right)ds=\frac{1}{K(K-1)^{2}}\sum_{\ell=1}^{K}\sum_{k=1,k\neq\ell}^{K}\left(\frac{\Gamma}{K}-\frac{1}{C_{\ell}}\right)I_{k\ell}\cdot\mathbb{E}\left[\sigma\right] \tag{58}$$

with

$$\begin{aligned}
I_{k\ell} &= \int_{-\infty}^{0}\left(\int_{s}^{0}\mathbb{P}\left[C_{k}t<C_{k}s+\sigma\right]\mathbb{P}\left[C_{\ell}t+\sigma>0\right]dt\right)ds\\
&= \int_{-\infty}^{0}\left(\int_{-\infty}^{t}\mathbb{P}\left[C_{k}t<C_{k}s+\sigma\right]ds\right)\mathbb{P}\left[C_{\ell}t+\sigma>0\right]dt\\
&= \int_{-\infty}^{0}\left(\int_{0}^{\infty}\mathbb{P}\left[C_{k}x<\sigma\right]dx\right)\mathbb{P}\left[C_{\ell}t+\sigma>0\right]dt\\
&= \left(\int_{0}^{\infty}\mathbb{P}\left[C_{k}x<\sigma\right]dx\right)\left(\int_{-\infty}^{0}\mathbb{P}\left[C_{\ell}t+\sigma>0\right]dt\right)\\
&= \frac{\mathbb{E}\left[\sigma\right]}{C_{k}}\cdot\frac{\mathbb{E}\left[\sigma\right]}{C_{\ell}},\quad k,\ell=1,\ldots,K.
\end{aligned} \tag{59}$$

Therefore,

$$\int_{-\infty}^{0}\left(\int_{s}^{0}H(s,t)dt\right)ds=\sum_{\ell=1}^{K}\sum_{k=1,k\neq\ell}^{K}\left(\frac{\Gamma}{K}-\frac{1}{C_{\ell}}\right)\left(\frac{\mathbb{E}\left[\sigma\right]}{C_{k}}\cdot\frac{\mathbb{E}\left[\sigma\right]}{C_{\ell}}\right)\cdot\mathbb{E}\left[\sigma\right] \tag{60}$$

with

$$\begin{aligned}
\sum_{\ell=1}^{K}\sum_{k=1,k\neq\ell}^{K}\frac{\mathbb{E}\left[\sigma\right]}{C_{k}}\cdot\frac{\mathbb{E}\left[\sigma\right]}{C_{\ell}} &= \sum_{\ell=1}^{K}\frac{1}{C_{\ell}}\left(\Gamma-\frac{1}{C_{\ell}}\right)\cdot\left(\mathbb{E}\left[\sigma\right]\right)^{2}\\
&= \left(\Gamma^{2}-\sum_{\ell=1}^{K}\frac{1}{C_{\ell}^{2}}\right)\cdot\left(\mathbb{E}\left[\sigma\right]\right)^{2}
\end{aligned} \tag{61}$$

and

$$\begin{aligned}
\sum_{\ell=1}^{K}\sum_{k=1,k\neq\ell}^{K}\frac{1}{C_{\ell}}\left(\frac{\mathbb{E}\left[\sigma\right]}{C_{k}}\cdot\frac{\mathbb{E}\left[\sigma\right]}{C_{\ell}}\right) &= \sum_{\ell=1}^{K}\frac{1}{C_{\ell}^{2}}\left(\Gamma-\frac{1}{C_{\ell}}\right)\cdot\left(\mathbb{E}\left[\sigma\right]\right)^{2}\\
&= \left(\Gamma\sum_{\ell=1}^{K}\frac{1}{C_{\ell}^{2}}-\sum_{\ell=1}^{K}\frac{1}{C_{\ell}^{3}}\right)\cdot\left(\mathbb{E}\left[\sigma\right]\right)^{2}.
\end{aligned} \tag{62}$$

Substitute (61) and (62) into (60), and we find

$$\begin{aligned}
&\int_{-\infty}^{0}\left(\int_{s}^{0}H(s,t)dt\right)ds\\
&= \frac{1}{K(K-1)^{2}}\left(\frac{\Gamma}{K}\left(\Gamma^{2}-\sum_{\ell=1}^{K}\frac{1}{C_{\ell}^{2}}\right)-\left(\Gamma\sum_{\ell=1}^{K}\frac{1}{C_{\ell}^{2}}-\sum_{\ell=1}^{K}\frac{1}{C_{\ell}^{3}}\right)\right)\cdot\left(\mathbb{E}\left[\sigma\right]\right)^{3}\\
&= \frac{1}{K(K-1)^{2}}\left(\frac{\Gamma^{3}}{K}-\frac{K+1}{K}\cdot\Gamma\sum_{\ell=1}^{K}\frac{1}{C_{\ell}^{2}}+\sum_{\ell=1}^{K}\frac{1}{C_{\ell}^{3}}\right)\cdot\left(\mathbb{E}\left[\sigma\right]\right)^{3}.
\end{aligned} \tag{63}$$

Finally, return to (56) and collect (57) and (63): Uninteresting calculations show that

$$\int_{-\infty}^{0}\left(\int_{s}^{0}R_{2}^{\star}(s,t)dt\right)ds$$

$$
\begin{aligned}
&= \left(-\left(\frac{\Gamma}{K}\right)^2 + \frac{1}{K}\sum_{k=1}^{K}\frac{1}{C_k^2}\right)\cdot\frac{\Gamma}{K(K-1)}\left(\mathbb{E}\left[\sigma\right]\right)^3 \\
&\quad + \frac{1}{K(K-1)^2}\left(\frac{\Gamma^3}{K} - \frac{K+1}{K}\Gamma\sum_{\ell=1}^{K}\frac{1}{C_\ell^2} + \sum_{\ell=1}^{K}\frac{1}{C_\ell^3}\right)\left(\mathbb{E}\left[\sigma\right]\right)^3 \\
&= \left(-\frac{1}{K^3(K-1)} + \frac{1}{K^2(K-1)^2}\right)\Gamma^3\cdot\left(\mathbb{E}\left[\sigma\right]\right)^3 \\
&\quad + \left(\frac{1}{K^2(K-1)} - \frac{K+1}{K^2(K-1)^2}\right)\left(\sum_{k=1}^{K}\frac{1}{C_k^2}\right)\Gamma\cdot\left(\mathbb{E}\left[\sigma\right]\right)^3 \\
&\quad + \frac{1}{K(K-1)^2}\left(\sum_{\ell=1}^{K}\frac{1}{C_\ell^3}\right)\cdot\left(\mathbb{E}\left[\sigma\right]\right)^3 \\
&= \frac{1}{(K-1)^2}\left(\left(\frac{\Gamma}{K}\right)^3 - 2\left(\frac{1}{K}\sum_{k=1}^{K}\frac{1}{C_k^2}\right)\left(\frac{\Gamma}{K}\right) + \frac{1}{K}\sum_{\ell=1}^{K}\frac{1}{C_\ell^3}\right)\cdot\left(\mathbb{E}\left[\sigma\right]\right)^3, \qquad (64)
\end{aligned}
$$

and the expression (8) now follows from (54). ■

11. A proof of Proposition 9.1

The cases $0 < s < t$ and $s < 0 < t$ are straightforward by virtue of the operational assumptions of the power-of-two load balancing policy. Indeed, when $0 < s < t$, $\nu_0 = \nu_0^\star$, hence $R_2(s,t) = R_0$ and $\widehat{R}_2(s,t) = \widehat{R}_0$ holds. On the other hand, when $s < 0 < t$, the future t-job does not affect the selection of ν_0, hence has no impact on the performance of the tagged job. As only the s-job can possibily affect the choice of ν_0, we get $R_2(s,t) = R_1(s)$ and this shows that $\widehat{R}_2(s,t) = \widehat{R}_1(s)$.

From now on we assume $s < t < 0$, in which case we have $\nu_s = \nu_s^\star$. The selection of ν_t can in principle be affected by whether the s-job has completed its service by time t, while that of ν_0 will be determined by whether the s-job and t-job have completed service by the time the tagged job enters the system. Therefore, as the s-job completes at time $s + \frac{\sigma_s}{C_{\nu_s^\star}}$, several possibilities arise; they are captured in the decomposition

$$
\begin{aligned}
\mathbb{E}\left[R_2(s,t)\right] &= \mathbb{E}\left[\mathbf{1}\left[s + \frac{\sigma_s}{C_{\nu_s^\star}} \le t\right] R_2(s,t)\right] \\
&\quad + \mathbb{E}\left[\mathbf{1}\left[t < s + \frac{\sigma_s}{C_{\nu_s^\star}} \le 0\right] R_2(s,t)\right] \\
&\quad + \mathbb{E}\left[\mathbf{1}\left[s + \frac{\sigma_s}{C_{\nu_s^\star}} > 0\right]\mathbf{1}\left[t + \frac{\sigma_t}{C_{\nu_t}} \le 0\right] R_2(s,t)\right] \\
&\quad + \mathbb{E}\left[\mathbf{1}\left[s + \frac{\sigma_s}{C_{\nu_s^\star}} > 0\right]\mathbf{1}\left[t + \frac{\sigma_t}{C_{\nu_t}} > 0\right] R_2(s,t)\right]. \qquad (65)
\end{aligned}
$$

These four terms are evaluated separately in the next four lemmas.

Lemma 11.1. *With $s < t < 0$, we have*

$$
\mathbb{E}\left[\mathbf{1}\left[s + \frac{\sigma_s}{C_{\nu_s^\star}} \le t\right] R_2(s,t)\right] = \left(\frac{1}{K}\sum_{k=1}^{K}\mathbb{P}\left[s + \frac{\sigma_s}{C_k} \le t\right]\right)\cdot\widehat{R}_1(t) \qquad (66)
$$

Proof. When $s + \frac{\sigma_s}{C_{\nu_s^\star}} \le t$, the s-job will have completed service by the time the t-job arrives. Therefore, conditionally on $s + \frac{\sigma_s}{C_{\nu_s^\star}} \le t$, it holds that $R_2(s,t) =_{st} R_1(t)$, whence

$$
\mathbb{E}\left[\mathbf{1}\left[s + \frac{\sigma_s}{C_{\nu_s^\star}} \le t\right] R_2(s,t)\right] = \mathbb{E}\left[\mathbf{1}\left[s + \frac{\sigma_s}{C_{\nu_s^\star}} \le t\right] R_1(t)\right] = \mathbb{P}\left[s + \frac{\sigma_s}{C_{\nu_s^\star}} \le t\right]\cdot\widehat{R}_1(t) \qquad (67)
$$

with

$$\mathbb{P}\left[s+\frac{\sigma_s}{C_{\nu_s^\star}}\le t\right]=\frac{1}{K}\sum_{k=1}^{K}\mathbb{P}\left[s+\frac{\sigma}{C_k}\le t\right]$$

by the usual arguments. This completes the proof of (66). ■

Lemma 11.2. *With $s<t<0$, we have*

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}\left[t<s+\frac{\sigma_s}{C_{\nu_s^\star}}\le 0\right]R_2(s,t)\right]\\ &\quad=\quad\frac{1}{K(K-1)}\sum_{k=1}^{K}\left(\sum_{\ell=1,\ell\neq k}^{K}\mathbb{P}\left[t<s+\frac{\sigma_s}{C_\ell}\le 0\right]\right)\cdot H_k(t)\end{aligned}\tag{68}$$

with $H_k(t)$ given by (34) for all $k=1,\dots,K$.

Proof. When $t<s+\frac{\sigma_s}{C_{\nu_s^\star}}\le 0$, the s-job has not completed its service by time t, but will have completed it by the time the tagged job arrives. Thus, only the t-job can affect the definition of ν_0 (through σ_t and ν_t).

With this in mind, consider the decomposition

$$\mathbb{E}\left[\mathbf{1}\left[t<s+\frac{\sigma_s}{C_{\nu_s^\star}}\le 0\right]R_2(s,t)\right]=\sum_{k=1}^{K}\mathbb{E}\left[\mathbf{1}\left[t<s+\frac{\sigma_s}{C_{\nu_s^\star}}\le 0\right]\mathbf{1}\left[\nu_t=k\right]\frac{\sigma_0}{C_{\nu_0}}\right].\tag{69}$$

Fix $k=1,\dots,K$. We are in the setting of Lemma 7.1 with $y=t$, $E=[t<s+\frac{\sigma_s}{C_{\nu_s^\star}}\le 0]$, $\tau=\sigma_t$ and $\gamma=\nu_t$ so that $\gamma_0=\nu_0$: The expression (25) becomes

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}\left[t<s+\frac{\sigma_s}{C_{\nu_s^\star}}\le 0\right]\mathbf{1}\left[\nu_t=k\right]\frac{\sigma_0}{C_{\nu_0}}\right]\\ &\quad=\quad\mathbb{P}\left[t<s+\frac{\sigma_s}{C_{\nu_s^\star}}\le 0,\nu_t=k,t+\frac{\sigma_t}{C_k}\le 0\right]\cdot\widehat{R}_0\\ &\qquad+\frac{1}{K-1}\mathbb{P}\left[t<s+\frac{\sigma_s}{C_{\nu_s^\star}}\le 0,\nu_t=k,t+\frac{\sigma_t}{C_k}>0\right]\left(\Gamma-\frac{1}{C_k}\right)\cdot\mathbb{E}\left[\sigma_0\right]\\ &\quad=\quad\mathbb{P}\left[t<s+\frac{\sigma_s}{C_{\nu_s^\star}}\le 0,\nu_t=k\right]H_k(t)\end{aligned}\tag{70}$$

with $H_k(t)$ defined at (34). In the last step we used the fact that under the enforced independence assumptions, the rv σ_t is independent of the rvs $\{\sigma_s,\nu_s^\star,\nu_t\}$ when ν_t is generated by the power-of-two load balancing policy.

In Appendix B we show that

$$\mathbb{P}\left[t<s+\frac{\sigma_s}{C_{\nu_s^\star}}\le 0,\nu_t=k\right]=\frac{1}{K(K-1)}\sum_{\ell=1,\ell\neq k}^{K}\mathbb{P}\left[t<s+\frac{\sigma_s}{C_\ell}\le 0\right].\tag{71}$$

Inserting (71) back into (70) yields

$$\mathbb{E}\left[\mathbf{1}\left[t<s+\frac{\sigma_s}{C_{\nu_s^\star}}\le 0\right]\mathbf{1}\left[\nu_t=k\right]\frac{\sigma_0}{c_{\nu_0}}\right]=\left(\frac{1}{K(K-1)}\sum_{\ell=1,\ell\neq k}^{K}\mathbb{P}\left[t<s+\frac{\sigma_s}{C_\ell}\le 0\right]\right)\cdot H_k(t),$$

and the desired result is now obtained by making use of (69). ■

The last two terms in the decomposition (65) are more cumbersome to evaluate. Their expressions are given in the next two lemmas whose proofs can be found in Sections 12 and 13, respectively.

Lemma 11.3. *With $s < t < 0$, we have*

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}\left[s+\frac{\sigma_s}{C_{\nu_s^\star}}>0\right]\mathbf{1}\left[t+\frac{\sigma_t}{C_{\nu_t}}\le 0\right]R_2(s,t)\right]\\
&\quad= \frac{1}{K(K-1)^2}\sum_{k=1}^K\sum_{\ell=1,\ell\neq k}^K\left(\Gamma-\frac{1}{C_\ell}\right)\mathbb{P}\left[C_\ell s+\sigma_s>0\right]\mathbb{P}\left[C_k t+\sigma_t\le 0\right]\cdot\mathbb{E}\left[\sigma_0\right].
\end{aligned}\tag{72}$$

Lemma 11.4. *With $s < t < 0$, we have*

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}\left[s+\frac{\sigma_s}{C_{\nu_s^\star}}>0\right]\mathbf{1}\left[t+\frac{\sigma_t}{C_{\nu_t}}>0\right]R_2(s,t)\right]\\
&\quad= \frac{1}{K^2(K-1)^2}\sum_{k=1}^K\sum_{\ell=1,\ell\neq k}^K J_{k\ell}\cdot\mathbb{P}\left[C_k s+\sigma_s>0\right]\mathbb{P}\left[C_\ell t+\sigma_t>0\right]\cdot\mathbb{E}\left[\sigma_0\right]
\end{aligned}\tag{73}$$

with the constants $J_{k\ell},\ k,\ell=1,\ldots,$ given by (44).

12. A proof of Lemma 11.3

We are in the situation when $s < t < 0$. If $s+\frac{\sigma_s}{C_{\nu_s^\star}} > 0$ (hence $s+\frac{\sigma_s}{C_{\nu_s^\star}} > t$), then the s-job completes its service only after the tagged arrives, so that both the s-job and t-job can possibly affect the definition of ν_0. If in addition we have $t+\frac{\sigma_t}{C_{\nu_t}} \le 0$, then only the s-job can affect the selection ν_0.

In the usual manner we have the decomposition

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}\left[s+\frac{\sigma_s}{C_{\nu_s^\star}}>0\right]\mathbf{1}\left[t+\frac{\sigma_t}{C_{\nu_t}}\le 0\right]R_2(s,t)\right]\\
&\quad= \sum_{k=1}^K\mathbb{E}\left[\mathbf{1}\left[s+\frac{\sigma_s}{C_{\nu_s^\star}}>0\right]\mathbf{1}\left[t+\frac{\sigma_t}{C_k}\le 0\right]\mathbf{1}\left[\nu_t=k\right]\frac{\sigma_0}{C_{\nu_0}}\right]\\
&\quad= \sum_{k=1}^K\sum_{\ell=1,\ell\neq k}^K\mathbb{E}\left[\mathbf{1}\left[\nu_s^\star=\ell\right]\mathbf{1}\left[s+\frac{\sigma_s}{C_\ell}>0\right]\mathbf{1}\left[t+\frac{\sigma_t}{C_k}\le 0\right]\mathbf{1}\left[\nu_t=k\right]\frac{\sigma_0}{C_{\nu_0}}\right].
\end{aligned}\tag{74}$$

Pick distinct $k,\ell=1,\ldots,K$. This time we apply Lemma 7.1 with $y=s$, $E=[s+\frac{\sigma_s}{C_\ell}>0,\nu_t=k,t+\frac{\sigma_t}{C_k}\le 0]$, $\tau=\sigma_s$ and $\gamma=\nu_s^\star$ (so that $\gamma_0=\nu_0$). This leads to

$$\begin{aligned}
&\mathbb{E}\left[\left(\mathbf{1}\left[s+\frac{\sigma_s}{C_\ell}>0\right]\mathbf{1}\left[t+\frac{\sigma_t}{C_k}\le 0\right]\mathbf{1}\left[\nu_t=k\right]\right)\mathbf{1}\left[\nu_s^\star=\ell\right]\frac{\sigma_0}{C_{\nu_0}}\right]\\
&\quad= \mathbb{P}\left[s+\frac{\sigma_s}{C_\ell}>0,t+\frac{\sigma_t}{C_k}\le 0,\nu_t=k,\nu_s^\star=\ell,s+\frac{\sigma_s}{C_\ell}\le 0\right]\cdot\widehat{R}_0\\
&\qquad+\frac{1}{K-1}\mathbb{P}\left[s+\frac{\sigma_s}{C_\ell}>0,t+\frac{\sigma_t}{C_k}\le 0,\nu_t=k,\nu_s^\star=\ell,s+\frac{\sigma_s}{C_\ell}>0\right]\left(\Gamma-\frac{1}{C_\ell}\right)\cdot\mathbb{E}\left[\sigma_0\right]\\
&\quad= \frac{1}{K-1}\mathbb{P}\left[s+\frac{\sigma_s}{C_\ell}>0,t+\frac{\sigma_t}{C_k}\le 0,\nu_t=k,\nu_s^\star=\ell\right]\left(\Gamma-\frac{1}{C_\ell}\right)\cdot\mathbb{E}\left[\sigma_0\right]\\
&\quad= \frac{1}{K-1}\mathbb{P}\left[t+\frac{\sigma_t}{C_k}\le 0\right]\mathbb{P}\left[\nu_s^\star=\ell,s+\frac{\sigma_s}{C_\ell}>0,\nu_t=k\right]\left(\Gamma-\frac{1}{C_\ell}\right)\cdot\mathbb{E}\left[\sigma_0\right]
\end{aligned}\tag{75}$$

since the rvs σ_t is independent of the collection $\{\nu_s^\star,\sigma_s,\nu_t\}$ under the enforced independence assumptions. Next, we write

$$\begin{aligned}
&\mathbb{P}\left[\nu_s^\star=\ell,s+\frac{\sigma_s}{C_\ell}>0,\nu_t=k\right]\\
&\quad= \mathbb{P}\left[\nu_s^\star=\ell,s+\frac{\sigma_s}{C_\ell}>0,\ell\in\Sigma_t,\nu_t=k\right]+\mathbb{P}\left[\nu_s^\star=\ell,s+\frac{\sigma_s}{C_\ell}>0,\ell\notin\Sigma_t,\nu_t=k\right].
\end{aligned}\tag{76}$$

Taking terms in turn we first get

$$\begin{aligned}\mathbb{P}\left[\nu_s^\star=\ell, s+\frac{\sigma_s}{C_\ell}>0, \ell\in\Sigma_t, \nu_t=k\right] &= \mathbb{P}\left[\nu_s^\star=\ell, s+\frac{\sigma_s}{C_\ell}>0, \Sigma_t=\{\ell,k\}, \nu_t=k\right]\\ &= \mathbb{P}\left[\nu_s^\star=\ell, s+\frac{\sigma_s}{C_\ell}>0, \Sigma_t=\{\ell,k\}\right]\\ &= \frac{2}{K^2(K-1)}\mathbb{P}\left[s+\frac{\sigma_s}{C_\ell}>0\right] \qquad (77)\end{aligned}$$

since under the constraint $s+\frac{\sigma_s}{C_\ell}>0$, the fact that $\nu_s^\star$ is an element of Σ_t forces ν_t to be the other element in Σ_t. In a similar way, under the constraint $s+\frac{\sigma_s}{C_\ell}>0$, $\nu_s^\star$ not being in Σ_t implies $\nu_t=\nu_t^\star$, and this leads to

$$\begin{aligned}&\mathbb{P}\left[\nu_s^\star=\ell, s+\frac{\sigma_s}{C_\ell}>0, \ell\notin\Sigma_t, \nu_t=k\right]\\ &= \mathbb{P}\left[\nu_s^\star=\ell, s+\frac{\sigma_s}{C_\ell}>0, \ell\notin\Sigma_t, \nu_t^\star=k\right]\\ &= \sum_{a=1,a\neq k,a\neq\ell}^{K}\mathbb{P}\left[\nu_s^\star=\ell, s+\frac{\sigma_s}{C_\ell}>0, \Sigma_t=\{a,k\}, \nu_t^\star=k\right]\\ &= \sum_{a=1,a\neq k,a\neq\ell}^{K}\mathbb{P}\left[\nu_s^\star=\ell\right]\mathbb{P}\left[s+\frac{\sigma_s}{C_\ell}>0\right]\frac{1}{2}\cdot\frac{2}{K(K-1)}\\ &= \frac{K-2}{K^2(K-1)}\mathbb{P}\left[s+\frac{\sigma_s}{C_\ell}>0\right]. \qquad (78)\end{aligned}$$

Collecting (77) and (78) gives

$$\mathbb{P}\left[\nu_s^\star=\ell, s+\frac{\sigma_s}{C_\ell}>0, \nu_t=k\right]=\frac{1}{K(K-1)}\mathbb{P}\left[s+\frac{\sigma_s}{C_\ell}>0\right], \qquad (79)$$

and with the help of (75) we conclude that

$$\begin{aligned}&\mathbb{E}\left[\left(\mathbf{1}\left[s+\frac{\sigma_s}{C_\ell}>0\right]\mathbf{1}\left[t+\frac{\sigma_t}{C_k}\leq 0\right]\mathbf{1}\left[\nu_t=k\right]\right)\mathbf{1}\left[\nu_s^\star=\ell\right]\frac{\sigma_0}{C_{\nu_0}}\right]\\ &= \frac{1}{K(K-1)^2}\mathbb{P}\left[t+\frac{\sigma_t}{C_k}\leq 0\right]\mathbb{P}\left[s+\frac{\sigma_s}{C_\ell}>0\right]\left(\Gamma-\frac{1}{C_\ell}\right)\cdot\mathbb{E}\left[\sigma_0\right]. \qquad (80)\end{aligned}$$

Inserting this last expression into (74) we obtain (72) as desired. ■

13. A proof of Lemma 11.4

We are in the situation when $s<t<0$. If $s+\frac{\sigma_s}{C_{\nu_s^\star}}>0$ and $t+\frac{\sigma_t}{C_{\nu_t}}>0$, then ν_t is determined by the s-job and we must have $\nu_s^\star\neq\nu_t$. When the tagged job arrives, both $\nu_s(=\nu_s^\star)$ and ν_t would have already been selected, with both s-job and t-job still in service when ν_0 needs to be selected. In order to establish (73), we begin with the observation that

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}\left[s+\frac{\sigma_s}{C_{\nu_s^\star}}>0\right]\mathbf{1}\left[t+\frac{\sigma_t}{C_{\nu_t}}>0\right]R_2(s;t)\right]\\ &= \mathbb{E}\left[\mathbf{1}\left[s+\frac{\sigma_s}{C_{\nu_s^\star}}>0\right]\mathbf{1}\left[t+\frac{\sigma_t}{C_{\nu_t}}>0\right]\frac{\sigma_0}{C_{\nu_0}}\right]\\ &= \sum_{k=1}^{K}\sum_{\ell=1,\ell\neq k}^{K}\mathbb{E}\left[\mathbf{1}\left[\nu_s^\star=k\right]\mathbf{1}\left[s+\frac{\sigma_s}{C_k}>0\right]\mathbf{1}\left[\nu_t=\ell\right]\mathbf{1}\left[t+\frac{\sigma_t}{C_\ell}>0\right]\frac{\sigma_0}{C_{\nu_0}}\right]. \qquad (81)\end{aligned}$$

To take advantage of this decomposition, pick distinct $k, \ell = 1, \ldots, K$. As we keep in mind whether $\nu_s^\star$ and ν_t are in Σ_0, we shall have to consider four possible cases: First, if both $\nu_s^\star$ and ν_t are in Σ_0, then $\nu_0 = \nu_0^\star$ and we have

$$\begin{aligned}
& \mathbb{E}\left[\mathbf{1}\left[\nu_s^\star = k\right]\mathbf{1}\left[s + \frac{\sigma_s}{C_k} > 0\right]\mathbf{1}\left[\nu_t = \ell\right]\mathbf{1}\left[t + \frac{\sigma_t}{C_\ell} > 0\right]\mathbf{1}\left[k \in \Sigma_0, \ell \in \Sigma_0\right]\frac{\sigma_0}{C_{\nu_0}}\right] \\
= \; & \mathbb{E}\left[\mathbf{1}\left[\nu_s^\star = k\right]\mathbf{1}\left[s + \frac{\sigma_s}{C_k} > 0\right]\mathbf{1}\left[\nu_t = \ell\right]\mathbf{1}\left[t + \frac{\sigma_t}{C_\ell} > 0\right]\mathbf{1}\left[\Sigma_0 = \{k, \ell\}\right]\frac{\sigma_0}{C_{\nu_0^\star}}\right] \\
= \; & \mathbb{P}\left[\nu_s^\star = k, s + \frac{\sigma_s}{C_k} > 0, \nu_t = \ell, t + \frac{\sigma_t}{C_\ell} > 0\right]\mathbb{P}\left[\Sigma_0 = \{k, \ell\}\right]\frac{1}{2}\cdot\left(\frac{1}{C_k} + \frac{1}{C_\ell}\right)\cdot\mathbb{E}\left[\sigma_0\right] \\
= \; & \frac{1}{K(K-1)}\mathbb{P}\left[\nu_s^\star = k, s + \frac{\sigma_s}{C_k} > 0, \nu_t = \ell, t + \frac{\sigma_t}{C_\ell} > 0\right]\left(\frac{1}{C_k} + \frac{1}{C_\ell}\right)\cdot\mathbb{E}\left[\sigma_0\right]. \qquad (82)
\end{aligned}$$

Next, if $\nu_s^\star$ is not in Σ_0 but ν_t is in Σ_0, then ν_0 is the other element in Σ_0, and we get

$$\begin{aligned}
& \mathbb{E}\left[\mathbf{1}\left[\nu_s^\star = k\right]\mathbf{1}\left[s + \frac{\sigma_s}{C_k} > 0\right]\mathbf{1}\left[\nu_t = \ell\right]\mathbf{1}\left[t + \frac{\sigma_t}{C_\ell} > 0\right]\mathbf{1}\left[k \notin \Sigma_0, \ell \in \Sigma_0\right]\frac{\sigma_0}{C_{\nu_0}}\right] \\
= \; & \sum_{a=1, a\neq k, a\neq \ell}^{K}\mathbb{E}\left[\mathbf{1}\left[\nu_s^\star = k\right]\mathbf{1}\left[s + \frac{\sigma_s}{C_k} > 0\right]\mathbf{1}\left[\nu_t = \ell\right]\mathbf{1}\left[t + \frac{\sigma_t}{C_\ell} > 0\right]\mathbf{1}\left[\Sigma_0 = \{\ell, a\}\right]\frac{\sigma_0}{C_{\nu_0}}\right] \\
= \; & \sum_{a=1, a\neq k, a\neq \ell}^{K}\mathbb{E}\left[\mathbf{1}\left[\nu_s^\star = k\right]\mathbf{1}\left[s + \frac{\sigma_s}{C_k} > 0\right]\mathbf{1}\left[\nu_t = \ell\right]\mathbf{1}\left[t + \frac{\sigma_t}{C_\ell} > 0\right]\mathbf{1}\left[\Sigma_0 = \{\ell, a\}\right]\frac{\sigma_0}{C_a}\right] \\
= \; & \sum_{a=1, a\neq k, a\neq \ell}^{K}\mathbb{P}\left[\nu_s^\star = k, s + \frac{\sigma_s}{C_k} > 0, \nu_t = \ell, t + \frac{\sigma_t}{C_\ell} > 0\right]\mathbb{P}\left[\Sigma_0 = \{\ell, a\}\right]\cdot\frac{\mathbb{E}\left[\sigma_0\right]}{C_a} \\
= \; & \frac{2}{K(K-1)}\left(\sum_{a=1, a\neq k, a\neq \ell}^{K}\frac{1}{C_a}\right)\mathbb{P}\left[\nu_s^\star = k, s + \frac{\sigma_s}{C_k} > 0, \nu_t = \ell, t + \frac{\sigma_t}{C_\ell} > 0\right]\cdot\mathbb{E}\left[\sigma_0\right] \\
= \; & \frac{2}{K(K-1)}\left(\Gamma - \frac{1}{C_k} - \frac{1}{C_\ell}\right)\mathbb{P}\left[\nu_s^\star = k, s + \frac{\sigma_s}{C_k} > 0, \nu_t = \ell, t + \frac{\sigma_t}{C_\ell} > 0\right]\cdot\mathbb{E}\left[\sigma_0\right]. \qquad (83)
\end{aligned}$$

In a similar way, if $\nu_s^\star$ is in Σ_0 but ν_t is not in Σ_0, then ν_0 is necessarily the other element in Σ_0, and we get

$$\begin{aligned}
& \mathbb{E}\left[\mathbf{1}\left[\nu_s^\star = k\right]\mathbf{1}\left[s + \frac{\sigma_s}{C_k} > 0\right]\mathbf{1}\left[\nu_t = \ell\right]\mathbf{1}\left[t + \frac{\sigma_t}{C_\ell} > 0\right]\mathbf{1}\left[k \in \Sigma_0, \ell \notin \Sigma_0\right]\frac{\sigma_0}{C_{\nu_0}}\right] \\
= \; & \sum_{b=1, b\neq k, b\neq \ell}^{K}\mathbb{E}\left[\mathbf{1}\left[\nu_s^\star = k\right]\mathbf{1}\left[s + \frac{\sigma_s}{C_k} > 0\right]\mathbf{1}\left[\nu_t = \ell\right]\mathbf{1}\left[t + \frac{\sigma_t}{C_\ell} > 0\right]\mathbf{1}\left[\Sigma_0 = \{k, b\}\right]\frac{\sigma_0}{C_{\nu_0}}\right] \\
= \; & \frac{2}{K(K-1)}\left(\Gamma - \frac{1}{C_k} - \frac{1}{C_\ell}\right)\mathbb{P}\left[\nu_s^\star = k, s + \frac{\sigma_s}{C_k} > 0, \nu_t = \ell, t + \frac{\sigma_t}{C_\ell} > 0\right]\cdot\mathbb{E}\left[\sigma_0\right]. \qquad (84)
\end{aligned}$$

See details in [5].

Finally, when neither $\nu_s^\star$ nor ν_t are in Σ_0, then $\nu_0 = \nu_0^\star$, whence

$$\begin{aligned}
& \mathbb{E}\left[\mathbf{1}\left[\nu_s^\star = k\right]\mathbf{1}\left[s + \frac{\sigma_s}{C_k} > 0\right]\mathbf{1}\left[\nu_t = \ell\right]\mathbf{1}\left[t + \frac{\sigma_t}{C_\ell} > 0\right]\mathbf{1}\left[k \notin \Sigma_0, \ell \notin \Sigma_0\right]\frac{\sigma_0}{C_{\nu_0}}\right] \\
= \; & \mathbb{E}\left[\mathbf{1}\left[\nu_s^\star = k\right]\mathbf{1}\left[s + \frac{\sigma_s}{C_k} > 0\right]\mathbf{1}\left[\nu_t = \ell\right]\mathbf{1}\left[t + \frac{\sigma_t}{C_\ell} > 0\right]\mathbf{1}\left[k \notin \Sigma_0, \ell \notin \Sigma_0\right]\frac{\sigma_0}{C_{\nu_0^\star}}\right] \\
= \; & \sum_{a=1, a\neq k, a\neq \ell}^{K}\sum_{b=1, b\neq a, b\neq k, b\neq \ell}^{a-1}(\ldots)_{k\ell} \qquad (85)
\end{aligned}$$

with

$$(\ldots)_{k\ell} \;=\; \mathbb{E}\left[\mathbf{1}\left[\nu_s^\star = k\right]\mathbf{1}\left[s + \frac{\sigma_s}{C_k} > 0\right]\mathbf{1}\left[\nu_t = \ell\right]\mathbf{1}\left[t + \frac{\sigma_t}{C_\ell} > 0\right]\mathbf{1}\left[\Sigma_0 = \{a, b\}\right]\frac{\sigma_0}{C_{\nu_0^\star}}\right]$$

$$= \mathbb{P}\left[\nu_s^\star = k, s + \frac{\sigma_s}{C_k} > 0, \nu_t = \ell, t + \frac{\sigma_t}{C_\ell} > 0\right] \mathbb{P}\left[\Sigma_0 = \{a,b\}\right] \frac{1}{2} \cdot \left(\frac{1}{C_a} + \frac{1}{C_b}\right) \cdot \mathbb{E}\left[\sigma_0\right]$$

$$= \frac{1}{K(K-1)} \mathbb{P}\left[\nu_s^\star = k, s + \frac{\sigma_s}{C_k} > 0, \nu_t = \ell, t + \frac{\sigma_t}{C_\ell} > 0\right] \left(\frac{1}{C_a} + \frac{1}{C_b}\right) \cdot \mathbb{E}\left[\sigma_0\right]. \quad (86)$$

It then follows that

$$\mathbb{E}\left[\mathbf{1}\left[\nu_s^\star = k\right] \mathbf{1}\left[s + \frac{\sigma_s}{C_k} > 0\right] \mathbf{1}\left[\nu_t = \ell\right] \mathbf{1}\left[t + \frac{\sigma_t}{C_\ell} > 0\right] \mathbf{1}\left[k \notin \Sigma_0, \ell \notin \Sigma_0\right] \frac{\sigma_0}{C_{\nu_0}}\right]$$

$$= \frac{1}{K(K-1)} \cdot H_{k\ell} \cdot \mathbb{P}\left[\nu_s^\star = k, s + \frac{\sigma_s}{C_k} > 0, \nu_t = \ell, t + \frac{\sigma_t}{C_\ell} > 0\right] \cdot \mathbb{E}\left[\sigma_0\right] \quad (87)$$

where we have set

$$H_{k\ell} = \sum_{a=1, a\neq k, a\neq \ell}^{K} \left(\sum_{b=1, b\neq a, b\neq k, b\neq \ell}^{a-1} \left(\frac{1}{C_a} + \frac{1}{C_b}\right)\right). \quad (88)$$

Collecting terms (82)-(87), we conclude from (81) that

$$\mathbb{E}\left[\mathbf{1}\left[\nu_s^\star = k\right] \mathbf{1}\left[s + \frac{\sigma_s}{C_k} > 0\right] \mathbf{1}\left[\nu_t = \ell\right] \mathbf{1}\left[t + \frac{\sigma_t}{C_\ell} > 0\right] \frac{\sigma_0}{C_{\nu_0}}\right]$$

$$= \frac{1}{K(K-1)} \cdot H_{k\ell}^\star \cdot \mathbb{P}\left[\nu_s^\star = k, s + \frac{\sigma_s}{C_k} > 0, \nu_t = \ell, t + \frac{\sigma_t}{C_\ell} > 0\right] \cdot \mathbb{E}\left[\sigma_0\right] \quad (89)$$

with

$$H_{k\ell}^\star = H_{k\ell} + 4\Gamma - 3\left(\frac{1}{C_k} + \frac{1}{C_\ell}\right).$$

In Appendix C we show that

$$\mathbb{P}\left[\nu_s^\star = k, s + \frac{\sigma_s}{C_k} > 0, \nu_t = \ell, t + \frac{\sigma_t}{C_\ell} > 0\right] = \frac{1}{K(K-1)} \mathbb{P}\left[s + \frac{\sigma_s}{C_k} > 0\right] \mathbb{P}\left[t + \frac{\sigma_t}{C_\ell} > 0\right] \quad (90)$$

and the conclusion

$$\mathbb{E}\left[\mathbf{1}\left[\nu_s^\star = k\right] \mathbf{1}\left[s + \frac{\sigma_s}{C_k} > 0\right] \mathbf{1}\left[\nu_t = \ell\right] \mathbf{1}\left[t + \frac{\sigma_t}{C_\ell} > 0\right] \frac{\sigma_0}{C_{\nu_0}}\right]$$

$$= \frac{H_{k\ell}^\star}{K^2(K-1)^2} \cdot \mathbb{P}\left[s + \frac{\sigma_s}{C_k} > 0\right] \mathbb{P}\left[t + \frac{\sigma_t}{C_\ell} > 0\right] \cdot \mathbb{E}\left[\sigma_0\right] \quad (91)$$

follows. Uninteresting calculations (with details available in Appendix E of [5]) give

$$H_{k\ell} = (K-3)\left(\Gamma - \frac{1}{C_k} - \frac{1}{C_\ell}\right) \quad (92)$$

so that

$$\begin{aligned} H_{k\ell}^\star &= H_{k\ell} + 4\Gamma - 3\left(\frac{1}{C_k} + \frac{1}{C_\ell}\right) \\ &= (K-3)\left(\Gamma - \frac{1}{C_k} - \frac{1}{C_\ell}\right) + 4\Gamma - 3\left(\frac{1}{C_k} + \frac{1}{C_\ell}\right) \\ &= (K+1)\Gamma - K\left(\frac{1}{C_k} + \frac{1}{C_\ell}\right) = J_{k\ell} \end{aligned} \quad (93)$$

with $J_{k\ell}$ given by (44). Inserting this last expression into (91) yields the desired conclusion (73). ■

Acknowledgements

This research was partially carried out while the authors were in residence under the Saiotek Program on "Virtual Machines for the Traffic Analysis in High Capacity Networks" partially supported by Grant SA-2012/00331 (Department of Industry, Innovation, Trade and Tourism, Basque Government). The work of Ane Izagirre was also supported by grants from the École Doctorale Systèmes (EDSYS) and the Institut Nationale Polytechnique (INP) Toulouse.

References

[1] M. Harchol Balter. *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge University Press, Cambridge (UK), 2013.

[2] V. Gupta, M. Harchol Balter, K. Sigman, and W. Whitt. Analysis of join-the-shortest-queue routing for web server farms. *Performance Evaluation*, 64:1062–1081, 2007.

[3] A. Izagirre. Interpolation approximations for steady-state performance measures. *Ph.D Thesis, General Mathematics [math.GM]. INSA de Toulouse (2015). Available at* `https://tel.archives-ouvertes.fr/tel-01215869/document`, 2015.

[4] A. Izagirre and A.M. Makowski. Light traffic performance under the power of two load balancing strategy: The case of server heterogeneity. *ACM SIGMETRICS Performance Evaluation Review, Proceedings of PERFORMANCE 2014*, Torino (Italy), October 2014, 42:18–20, 2014.

[5] A. Izagirre and A.M. Makowski. Light traffic performance under the power of two load balancing strategy: The case of server heterogeneity. *arXiv:1701.06004v1 [cs.PF]*, 2017.

[6] A. Izagirre, U. Ayesta, and I.M. Verloop. Sojourn time approximations in a multi-class time-sharing server. *Proceedings of IEEE INFOCOM 2014*, Toronto (ON), April 2014,1:2786–2794, 2014.

[7] A.W. Marshall and I. Olkin. *Inequalities: Theory of Majorization and Its Applications*. Academic Press, San Diego (CA), 1979.

[8] M. Mitzenmacher. The power of two choices in randomized load balancing. *IEEE Transactions on Parallel and Distributed Systems*, TPDS-12:1094–1104, 2001.

[9] A. Mukhopadhyay and R.R. Mazumdar. Analysis of load balancing in large heterogeneous processor sharing systems. *IEEE Transactions on Control of Networked Systems*, TCNS-3:116 – 126, 2016.

[10] M.I. Reiman and B. Simon. An interpolation approximation for queueing systems with poisson input. *Operations Research*, 36:454–469, 1988.

[11] M.I. Reiman and B. Simon. Open queueing systems in light traffic. *Mathematics of Operations Research*, 14:26–59, 1989.

[12] N.D. Vvedenskaya, R.L. Dobrushin, and F.I. Karpelevich. Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problems of Information Transmission*, 32:20–34, 1996.

[13] W. Whitt. Deciding which queue to join: Some counterexamples. *Operations Research*, 34:55–62, 1986.

Appendix A: A proof of (12)

We are interested in assessing the range of values for Var[X] under the constraint $\mathbb{E}[X] = \frac{\Gamma}{K}$ for some $\Gamma > 0$. This amounts to considering the expression

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{C_k^2} - \left(\frac{1}{K}\sum_{k=1}^{K}\frac{1}{C_k}\right)^2, \quad \boldsymbol{C} = (C_1, \ldots, C_K) \in (0,\infty)^K$$

under the constraint

$$\sum_{k=1}^{K}\frac{1}{C_k} = \Gamma.$$

Defining the mapping $g : (0,\infty)^K \to \mathbb{R}_+$ as

$$g(\boldsymbol{C}) \equiv \sum_{k=1}^{K}\frac{1}{C_k^2}, \quad \boldsymbol{C} = (C_1, \ldots, C_K) \in (0,\infty)^K,$$

we need only focus on studying the range of $\{g(\boldsymbol{C}) : \boldsymbol{C} \in \mathcal{C}(\Gamma)\}$ where the constraint set $\mathcal{C}(\Gamma)$ is given by

$$\mathcal{C}(\Gamma) = \left\{ \boldsymbol{C} \in (0,\infty)^K : \sum_{k=1}^K \frac{1}{C_k} = \Gamma \right\}.$$

This issue is more easily understood with the help of the change of variables $T : (0,\infty)^K \to (0,\infty)^K$ given by

$$T(\boldsymbol{C}) = \left(\frac{1}{C_1}, \ldots, \frac{1}{C_K} \right), \quad \boldsymbol{C} = (C_1, \ldots, C_K) \in (0,\infty)^K.$$

The transformation T is a bijection from $(0,\infty)^K$ into itself (with inverse $T^{-1} = T$). Note that T puts the set $\mathcal{C}(\Gamma)$ into one-to-one correspondence with the set $\mathcal{X}(\Gamma)$ given by

$$\mathcal{X}(\Gamma) = \left\{ \boldsymbol{x} = (x_1, \ldots, x_K) \in (0,\infty)^K : \sum_{k=1}^K x_k = \Gamma \right\}.$$

If we define the mapping $h : \mathbb{R}_+^K \to \mathbb{R}_+$ by

$$h(\boldsymbol{x}) \equiv \sum_{k=1}^K x_k^2, \quad \boldsymbol{x} = (x_1, \ldots, x_K) \in \mathbb{R}_+^K,$$

then we obviously have

$$g(\boldsymbol{C}) = h(T(\boldsymbol{C})), \quad \boldsymbol{C} \in (0,\infty)^K. \tag{94}$$

Moreover it holds that $\{g(\boldsymbol{C}) : \boldsymbol{C} \in \mathcal{C}(\Gamma)\} = \{h(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}(\Gamma)\}$ since $T(\mathcal{C}(\Gamma)) = \mathcal{X}(\Gamma)$.

With $\prec$ denoting majorization [7, p. 7], whenever $\boldsymbol{x}_1 \prec \boldsymbol{x}_2$ in $\mathbb{R}_+^K$, we have

$$h(\boldsymbol{x}_1) \le h(\boldsymbol{x}_2) \tag{95}$$

by the Schur-convexity of the function $h : \mathbb{R}_+^K \to \mathbb{R}_+$ (inherited from the convexity of $t \to t^2$) [7, Prop. C.1, p. 64]; see also [7, p. 54] for the definition of Schur-convexity. The most "balanced" element of $\mathcal{X}(\Gamma)$ is the vector $\boldsymbol{x}_\star$ given by

$$\boldsymbol{x}_\star = \frac{\Gamma}{K} (1, \ldots, 1).$$

It represents the "smallest" element in the constraint set in the sense of majorization [7, p. 7] – We have $\boldsymbol{x}_\star \prec \boldsymbol{x}$ for any $\boldsymbol{x}$ in $\mathcal{X}(\Gamma)$, whence $h(\boldsymbol{x}_\star) \le h(\boldsymbol{x})$ by (95). With $\boldsymbol{C}_\star$ given by (13) we see from (94) that

$$g(\boldsymbol{C}_\star) \le g(\boldsymbol{C}), \quad \boldsymbol{C} \in \mathcal{C}(\Gamma)$$

since $\boldsymbol{C}_\star = T(\boldsymbol{x}_\star)$. This establishes the lower bound in (12) in agreement with the earlier discussion concerning the zero variance when all the capacities are identical.

We now turn to the upper bound: For each $k = 1, \ldots, K$, introduce the vector $\boldsymbol{x}_k^\star$ in $\mathbb{R}_+^K$ given by

$$\boldsymbol{x}_k^\star = \Gamma \boldsymbol{e}_k$$

with $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_K$ as defined in Section 3. It is well known [7, p. 7] that $\boldsymbol{x} \prec \boldsymbol{x}_k^\star$ for every $\boldsymbol{x}$ in $\mathcal{X}(\Gamma)$, so that $h(\boldsymbol{x}) \le h(\boldsymbol{x}_k^\star)$ for every $\boldsymbol{x}$ in $\mathcal{X}(\Gamma)$. Although $\boldsymbol{x}_k^\star$ is *not* an element of $\mathcal{X}(\Gamma)$, we nevertheless have

$$\sup \{h(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}(\Gamma)\} = \sup \left\{ h(\boldsymbol{x}) : \boldsymbol{x} \in \overline{\mathcal{X}(\Gamma)} \right\} = h(\boldsymbol{x}_k^\star)$$

by the continuity of h; the closure $\overline{\mathcal{X}(\Gamma)}$ of $\mathcal{X}(\Gamma)$ is given by $\overline{\mathcal{X}(\Gamma)} = \left\{ \boldsymbol{x} \in \mathbb{R}_+^K : \sum_{k=1}^K x_k = \Gamma \right\}$. In particular, we have

$$\sup \{g(\boldsymbol{C}) : \boldsymbol{C} \in \mathcal{C}(\Gamma)\} = h(\boldsymbol{x}_k^\star) = \Gamma^2. \tag{96}$$

Note that for each $k = 1, \ldots, K$ there is no vector $\boldsymbol{C}_k^\star$ in $\mathcal{C}(\Gamma)$ such that $\boldsymbol{x}_k^\star = T(\boldsymbol{C}_k^\star)$. However, there are vectors in $\mathcal{C}(\Gamma)$ whose value under g will come arbitrarily close to Γ^2. For instance, consider the vectors

$$\boldsymbol{x}_{k,a} = a\Gamma \boldsymbol{e}_k + \frac{(1-a)\Gamma}{K-1} \sum_{\ell=1, \ell \ne k}^K \boldsymbol{e}_\ell, \qquad \begin{matrix} 0 < a < 1 \\ k = 1, \ldots, K. \end{matrix}$$

These vectors are elements of $\mathcal{X}(\Gamma)$ with $\lim_{a\uparrow 1} \boldsymbol{x}_{k,a} = \boldsymbol{x}_k^\star$, hence $\lim_{a\uparrow 1} h(\boldsymbol{x}_{k,a}) = h(\boldsymbol{x}_k^\star)$ by continuity. As we recall the definition (14) we check that

$$\boldsymbol{C}_{k,a} = \frac{1}{a\Gamma}\boldsymbol{e}_k + \frac{K-1}{(1-a)\Gamma}\sum_{\ell=1,\ell\neq k}^{K} \boldsymbol{e}_\ell = T(\boldsymbol{x}_{k,a}), \qquad \begin{matrix} 0<a<1 \\ k=1,\dots,K \end{matrix}$$

so that $h(\boldsymbol{x}_{k,a}) = g(\boldsymbol{C}_{k,a})$ for all $0<a<1$. It follows that

$$\lim_{a\uparrow 1} g(\boldsymbol{C}_{k,a}) = \lim_{a\uparrow 1} h(\boldsymbol{x}_{k,a}) = h(\boldsymbol{x}_k^\star) = \Gamma^2$$

and this completes the discussion of the upper bound at (12). ■

Appendix B: A proof of (71)

We are in the situation $s<t<0$. Fix $k=1,\dots,K$. Our point of departure is the obvious decomposition

$$\mathbb{P}\left[t < s + \frac{\sigma_s}{C_{\nu_s^\star}} \leq 0, \nu_t = k\right] = \sum_{\ell=1,\ell\neq k}^{K} \mathbb{P}\left[\nu_s^\star = \ell, t < s + \frac{\sigma_s}{C_\ell} \leq 0, \nu_t = k\right]. \tag{97}$$

Pick $\ell = 1,\dots,K$ distinct from k, and note that

$$\begin{aligned}\mathbb{P}\left[\nu_s^\star = \ell, t < s + \frac{\sigma_s}{C_\ell} \leq 0, \nu_t = k\right] &= \mathbb{P}\left[\nu_s^\star = \ell, t < s + \frac{\sigma_s}{C_\ell} \leq 0, \ell \in \Sigma_t, \nu_t = k\right] \\ &\quad + \mathbb{P}\left[\nu_s^\star = \ell, t < s + \frac{\sigma_s}{C_\ell} \leq 0, \ell \notin \Sigma_t, \nu_t = k\right]. \end{aligned} \tag{98}$$

We examine each term in turn: First, when ℓ belongs to Σ_t with $\nu_s^\star = \ell$, then $\nu_t = k$ happens only if $\nu_s^\star = \ell$ and $\Sigma_t = \{k,\ell\}$, whence

$$\begin{aligned}\mathbb{P}\left[\nu_s^\star = \ell, t < s + \frac{\sigma_s}{C_\ell} \leq 0, \ell \in \Sigma_t, \nu_t = k\right] &= \mathbb{P}\left[\nu_s^\star = \ell, t < s + \frac{\sigma_s}{C_\ell} \leq 0, \Sigma_t = \{k,\ell\}, \nu_t = k\right] \\ &= \mathbb{P}\left[\nu_s^\star = \ell, t < s + \frac{\sigma_s}{C_\ell} \leq 0, \Sigma_t = \{k,\ell\}\right] \\ &= \frac{2}{K^2(K-1)}\mathbb{P}\left[t < s + \frac{\sigma_s}{C_\ell} \leq 0\right]. \end{aligned} \tag{99}$$

Next, we have $\nu_t = \nu_t^\star$ when $\nu_s^\star$ is not in Σ_t, so that

$$\begin{aligned}&\mathbb{P}\left[\nu_s^\star = \ell, t < s + \frac{\sigma_s}{C_\ell} \leq 0, \ell \notin \Sigma_t, \nu_t = k\right] \\ &= \sum_{a=1,a\neq k,a\neq \ell}^{K} \mathbb{P}\left[\nu_s^\star = \ell, t < s + \frac{\sigma_s}{C_\ell} \leq 0, \ell \notin \Sigma_t, \Sigma_t = \{k,a\}, \nu_t^\star = k\right] \\ &= \frac{1}{2}\sum_{a=1,a\neq k,a\neq \ell}^{K} \mathbb{P}\left[\nu_s^\star = \ell, t < s + \frac{\sigma_s}{C_\ell} \leq 0, \Sigma_t = \{k,a\}\right] \\ &= \frac{1}{2}\sum_{a=1,a\neq k,a\neq \ell}^{K} \frac{1}{K}\frac{2}{K(K-1)}\mathbb{P}\left[t < s + \frac{\sigma_s}{C_\ell} \leq 0\right] \\ &= \frac{1}{K^2(K-1)}\sum_{a=1,a\neq k,a\neq \ell}^{K} \mathbb{P}\left[t < s + \frac{\sigma_s}{C_\ell} \leq 0\right] \\ &= \frac{K-2}{K^2(K-1)}\mathbb{P}\left[t < s + \frac{\sigma_s}{C_\ell} \leq 0\right]. \end{aligned} \tag{100}$$

Inserting (99) and (100) back into (98) we get

$$\mathbb{P}\left[\nu_s^\star = \ell, t < s + \frac{\sigma_s}{C_\ell} \le 0, \nu_t = k\right] = \frac{1}{K(K-1)}\mathbb{P}\left[t < s + \frac{\sigma_s}{C_\ell} \le 0\right] \tag{101}$$

and the desired conclusion (71) follows with the help of (97). ■

Appendix C: A proof of (90)

Recall that we are in the situation $s < t < 0$. Fix distinct $k, \ell = 1, \ldots, K$. We need to show that

$$\mathbb{P}\left[\nu_s^\star = k, s + \frac{\sigma_s}{C_k} > 0, \nu_t = \ell, t + \frac{\sigma_t}{C_\ell} > 0\right] = \frac{1}{K(K-1)}\mathbb{P}\left[s + \frac{\sigma_s}{C_k} > 0\right]\mathbb{P}\left[t + \frac{\sigma_t}{C_\ell} > 0\right]. \tag{102}$$

By arguments used earlier we get

$$\begin{aligned}
& \mathbb{P}\left[\nu_s^\star = k, s + \frac{\sigma_s}{C_k} > 0, k \in \Sigma_t, \nu_t = \ell, t + \frac{\sigma_t}{C_\ell} > 0\right] \\
= \; & \mathbb{P}\left[\nu_s^\star = k, s + \frac{\sigma_s}{C_k} > 0, \Sigma_t = \{k, \ell\}, \nu_t = \ell, t + \frac{\sigma_t}{C_\ell} > 0\right] \\
= \; & \mathbb{P}\left[\nu_s^\star = k, s + \frac{\sigma_s}{C_k} > 0, \Sigma_t = \{k, \ell\}, t + \frac{\sigma_t}{C_\ell} > 0\right] \\
= \; & \frac{2}{K^2(K-1)}\mathbb{P}\left[s + \frac{\sigma_s}{C_k} > 0\right]\mathbb{P}\left[t + \frac{\sigma_t}{C_\ell} > 0\right]
\end{aligned} \tag{103}$$

under the enforced independence assumptions.

In a similar way, we find

$$\begin{aligned}
& \mathbb{P}\left[\nu_s^\star = k, s + \frac{\sigma_s}{C_k} > 0, k \notin \Sigma_t, \nu_t = \ell, t + \frac{\sigma_t}{C_\ell} > 0\right] \\
= \; & \sum_{a=1, a\neq k, a\neq \ell}^{K} \mathbb{P}\left[\nu_s^\star = k, s + \frac{\sigma_s}{C_k} > 0, \Sigma_t = \{a, \ell\}, \nu_t = \ell, t + \frac{\sigma_t}{C_\ell} > 0\right] \\
= \; & \frac{1}{2}\sum_{a=1, a\neq k, a\neq \ell}^{K} \mathbb{P}\left[\nu_s^\star = k, s + \frac{\sigma_s}{C_k} > 0, \Sigma_t = \{a, \ell\}, t + \frac{\sigma_t}{C_\ell} > 0\right] \\
= \; & \frac{1}{K^2(K-1)}\sum_{a=1, a\neq k, a\neq \ell}^{K} \mathbb{P}\left[s + \frac{\sigma_s}{C_k} > 0\right]\mathbb{P}\left[t + \frac{\sigma_t}{C_\ell} > 0\right] \\
= \; & \frac{K-2}{K^2(K-1)}\mathbb{P}\left[s + \frac{\sigma_s}{C_k} > 0\right]\mathbb{P}\left[t + \frac{\sigma_t}{C_\ell} > 0\right]
\end{aligned} \tag{104}$$

under the enforced independence assumptions. Collecting (103) and (104) we conclude to the validity of (102). ■