# The Price of Forgetting in Parallel and Non-Observable Queues

Jonatha Anselmi[1,2] and Bruno Gaujal[2]
[1] Basque Center for Applied Mathematics, Bizkaia Technology Park, Derio, 48170, Spain
[2] INRIA and Grenoble University, 51, Av. J. Kuntzmann, Montbonnot, 38330, France
anselmi@bcamath.org, bruno.gaujal@imag.fr

June 27, 2011

**Abstract**

We consider a broker-based network of non-observable parallel queues and analyze the minimum expected response time and the optimal routing policy when the broker has the memory of its previous routing decisions. We provide lower bounds on the minimum response time by means of convex programming that are tight, as follows by a numerical comparison with a proposed routing scheme. The "Price of Forgetting" (PoF), the ratio between the minimum response times achieved by a probabilistic broker and a broker with memory, is shown to be unbounded or arbitrarily close to one depending on the coefficient of variation of the service time distributions. In the case of exponential service times, the PoF is bounded from above by two, which is tight in heavy-traffic, and independent of the network size and heterogeneity. These properties yield a simple engineering product-form approximating tightly the minimum response time. Finally, we put our results in the context of game theory revisiting the "Price of Anarchy" (PoA) of parallel queues: It can be decomposed into the product of the PoA achieved by a probabilistic broker (already well understood) and the PoF.

## 1 Introduction

In the context of *-computing and manufacturing systems, users submit jobs without knowing which machine will handle their execution. A central broker is usually in charge of distributing incoming jobs to a set of resources to optimize some utility or cost function. The mean response time (simply response time in the following), the expected time it takes for a job to join the system and return to its issuer, is an important metric that is usually taken into account to achieve a better exploitation of resources and improve the average quality of service.

The problem of finding the routing strategy (or policy) that minimizes response time and the resulting response time itself are well-known problem in queueing theory and the literature is overwhelming. There are two main classes of routing strategies: Closed-loop policies, where the broker knows the state of each resource (number of jobs), and open-loop policies, where the broker does not know their state. Computing the best closed-loop policy is known to be a hard problem and a lot of work has been devoted to the computation of good approximations; see, e.g., [9, 27, 33] and the references therein.

The focus of this paper is on the open-loop case, and more precisely on off-line policies. Indeed, the state of the queues may not to be known on-line to the broker because of a number of reasons whose effects become worse and worse as the network size increases: i) Communicating the system state to the broker increases the network load, ii) the information received by the broker can be out of date, and iii) the synchronization that results from knowing the system state can degrade the performance [26]. A celebrated case is when the broker only knows the service time distributions of the network resources (or queues in the following) and dispatches jobs to queues according to an i.i.d. probabilistic law. The probabilistic nature of such broker yields tractable and accurate mathematical analyses of the problem; see, e.g., [8, 13, 16]. However, such a probabilistic broker ignores its past decisions that could be stored in a buffer (or memory). To design better routing strategies, real-world brokers do exploit local information about their previous routing decisions with a very limited cost. For instance, Round-Robin is optimal when the queues have identical service time distributions [32] and only requires the knowledge of the total number of queues and of the last queue where the last job has been sent.

Unfortunately, it is not known in general how the broker can exploit its buffer optimally: In this context, the assessment of the routing policy that minimizes response time as well as the analysis of such response time are current open problems; see, e.g., [18, 7, 13, 15]. In this framework, it is also open to assert whether the optimal policy is cyclic, even in the two queue case [15] and finding the optimal cyclic policy is NP-complete [7]. It is also unknown what the added-value of having a broker that optimally exploits its buffer with respect to its bufferless counterpart can be.

## 1.1 Our Contribution

In this paper, we analyze the minimum response time and the routing policy that minimizes response time of a broker-based system composed of $N$ parallel and non-observable queues. The broker can exploit the memory of its previous dispatching choices, and its routing policy can be deterministic, probabilistic or a mixture of them. The scheduling discipline of each queue is first-come-first-served (FCFS). This model is common in grid and volunteer computing [31, 1, 22], and supercomputing [30, 19]. To analyze the added-value of letting the router exploit its buffer, we introduce the Price of Forgetting (PoF), which we define as the ratio between the minimum response time achievable by a memoryless broker with respect to its counterpart with memory. In the remainder of the paper, the terms probabilistic, memoryless, and Bernoulli are synonyms.

First, we provide a lower bound on the minimum response time in terms of a convex optimization problem that is interpreted as the minimum response time of a parallel system of independent $\Gamma/GI/1$ queues. Then, we show that the PoF depends on the coefficients of variation of the service time distributions: It can be unbounded or be arbitrarily close to one. Explicit bounds for the PoF are given in heavy-traffic. The remainder of our contributions applies for the case of exponential service times, for which our lower bound is better captured by a simple convex program involving the Lambert $\mathcal{W}$ function. In heavy-traffic, this program allows for an explicit expression of the optimum that turns out to be half of the minimum response time achieved by a probabilistic broker. Here, we also prove that the PoF is bounded from above by two for any network load and size.

The efficiency of our bounds allows us to assess the quality of heuristics for the optimal routing. An exhaustive numerical analysis reveals that a broker deterministically forwarding jobs to queues according to a proposed *billiard* scheme [5, 21], a generalization of round-robin, yields a response time remarkably close to our lower bound. In other words, we empirically claim the proposed billiard scheme achieves the minimum response time that, in turn, is very-well captured by our bound and approximations. Our routing policy requires the solution of a convex optimization problem, and numerical results show that systems with thousands of queues can be solved in a few seconds. Our scheme has been successfully implemented in the context of a real-world volunteer-computing system [22].

We give numerical evidence of the fact that the PoF is an increasing function of the network load ($\rho$) only, meaning that it is insensitive to the network size ($N$) and heterogeneity. These structural properties entail that the minimum response time $R^{Opt}(\cdot)$ admits the product-form $R^{Opt}_{Bernoulli}(\cdot)/PoF(\rho)$ where i) $R^{Opt}_{Bernoulli}$ is the minimum response time achieved by a Bernoulli broker (which is well-understood [8]), and ii) $PoF(\rho)$, explicitly given, is increasing in $\rho$ and bounded from above by two.

We finally put our results in the context of game theory analyzing the "Price of Anarchy" (PoA) of our system [25], which measures the worst-case performance loss of a decentralized system with respect to its centralized counterpart in presence of non-cooperative users. In the analysis of the PoA, all existing works implicitly assume that the central broker behaves probabilistically. We revisit the PoA in the sense we let the broker exploit its memory. The main consequence is that the arrival process at each queue has in general different statistical properties than the corresponding one in the game-theoretic equilibrium. In both the centralized/decentralized situations, thus, the congestion functions are different. We show that our revisited PoA can be decomposed in the product of the PoA achieved by a probabilistic broker and the PoF.

This paper is organized as follows. Section 2 introduces the model under investigation and provides bounds on the minimum response time. The accuracy of our bounds is shown in Section 3, where we empirically show that the response time achieved by a particular billiard sequence is very close to our lower bound. Section 4 defines and analyzes the PoF, and Section 5 derives structural properties for the PoF

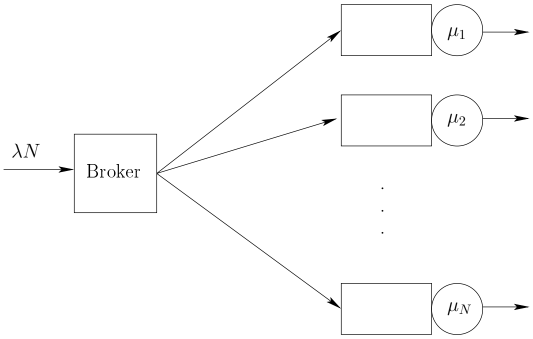

Figure 1: Queueing model under investigation. The broker dispatches incoming jobs to the queues according to some policy and with no delay. The mean arrival rate is proportional to the number of queues.

achieved with exponential service times. Section 6 introduces our revisited notion of PoA and relates it to the PoF. Finally, Section 7 draws the conclusions of this work.

Preliminary versions of this paper appeared in [4, 3], where we have limited the focus to queues having exponentially distributed service times.

## 2 Brokering in Parallel Non-Observable Queues

We consider a queueing system composed of $N$ infinite-room queues working in parallel as shown in Figure 2. Jobs arrive according to an external Poissonian source having intensity $\lambda N$ to a broker which instantaneously dispatches jobs to one of the $N$ queues according to a given *policy*, i.e., a routing rule.

In queue $i$, we assume that jobs require service for a random amount of time having mean $\mu_i^{-1}$ and standard deviation $s_i$ (both quantities are assumed to be independent of $N$, the number of queues). The service times of each queue are i.i.d. and independent of the arrival process. Initially, all queues are supposed to be empty (this assumption can be relaxed because the mean performance does not depend on initial states, but it is useful in our proofs for technical reasons). The scheduling discipline of each queue is assumed to be FCFS. In the remainder of the paper, index $i$ implicitly designates a queue and ranges from 1 to $N$, if not otherwise specified.

We denote by

$$\rho \overset{\text{def}}{=} \lambda N / \textstyle\sum_{i=1}^{N} \mu_i \tag{1}$$

the *network load*, an index measuring the network utilization. The considered queueing model is said to be *stable* if $\rho < 1$.

As for the broker policy, we focus on deterministic and/or randomized policies independent of the arrival process and of the service times. In other words, we restrict our study to policies that can be constructed off-line. Of course the policy may depend on the static parameters $\lambda$, $N$, $\mu_i$, and $s_i$, for all $i$.

The *routing policy* of jobs into queues is a random infinite sequence, denoted by $(A_n)_{n\in\mathbb{N}} \overset{\text{def}}{=} (A_n^1, \dots, A_n^N)_{n\in\mathbb{N}}$ and is such that the sequences $(A_n^i)_{n\in\mathbb{N}} \in \{0,1\}$, and $A_n^i = 1$ if the $n$-th arriving job is sent to queue $i$, and it is 0 otherwise. By definition, if $A_n^i = 1$ then $A_n^j = 0$ for all $j \neq i$, since a job is routed to a single queue. We assume that $A$ is independent of the arrival process. By definition of $A$, $\mathbb{P}(A_n^i = 1)$ is the probability that the $n$-th job is sent to queue $i$. Therefore, a routing policy is an infinite random process defined on the canonical probability space of infinite sequences, with values in a finite set of size $N$, namely $\{(1,0,\dots,0),(0,1,0,\dots,0),\dots,(0,\dots,0,1)\}$.

We also denote by $a$ a realisation of $A$, or any deterministic routing policy.

### 2.1 Bounding the Optimal Response Time

Let us consider $a$ to be a deterministic routing policy and let $Q_n^i(a_1^i, \dots, a_n^i)$ be the expected amount of work (measured in units of time) in queue $i$ after $n$ arrivals, where the expectation is taken over all arrival and service times.

We also denote by $R(a)$ the mean response time (also called sojourn time) of jobs in the system under policy $a$ (the dependence of $a$ will only be reported when necessary). Since all queues are FCFS, and using the limsup if the limit does not exist,

$$R(a) \stackrel{\text{def}}{=} \limsup_{K\to\infty} \frac{1}{K} \sum_{n=1}^{K} (a_n^1 Q_n^1(a_1^1, \ldots, a_n^1) + \cdots + a_n^N Q_n^N(a_1^N, \ldots, a_n^N)). \tag{2}$$

Once the function $R(\cdot) \in \mathbb{R}$ is defined over all deterministic policies, the mean response time under a randomized policy $A$ is $\mathbb{E}R(A)$.

The minimum mean response time is then defined as

$$R^{Opt} \stackrel{\text{def}}{=} \inf_{A} \mathbb{E}R(A). \tag{3}$$

It should be clear that for any randomized policy $A$, $\mathbb{E}R(A) \geq \inf_{a\in\mathcal{F}(A)} R(a)$ where $\mathcal{F}(A)$ is the set of all possible realisations of $A$. This implies that

$$R^{Opt} = \inf_{a} R(a), \tag{4}$$

where the infimum is taken over all deterministic routing policies.

Analogously, the minimum mean response time achievable by a *probabilistic* broker can be defined as

$$R^{Opt}_{Bernoulli} \stackrel{\text{def}}{=} \inf_{A\in\mathcal{B}} \mathbb{E}R(A), \tag{5}$$

where the infimum is taken over the set of all policies with Bernoulli distributions,
$\mathcal{B} \stackrel{\text{def}}{=} \{A : (A_n^1, \ldots, A_n^N) \text{ is i.i.d. } \forall n \in \mathbb{N}\}$, which ensures that each job is sent to some queue according to the same probability law independently of the others. Clearly, we have $R^{Opt}_{Bernoulli} \geq R^{Opt}$.

**Remark 1** *In the remainder of the paper, we say that a broker is probabilistic, Bernoulli or memoryless if its routing policy belongs to $\mathcal{B}$.*

**Remark 2** *Since our main goal is to study $R^{Opt}$, Equation* (4) *says that we can only focus on deterministic policies, i.e., the set of sequences $(a_n^1, \ldots a_n^N)_{n\in\mathbb{N}}$.*

**Remark 3** *Since our analysis only focuses on mean values of response times, in the remainder of the paper we omit the word "mean" each time that we refer to response time for simplicity.*

Now, let us consider queue $i$ in isolation. Let $R_i(a^i)$ be the Cesaro limit of $Q_n^i$, i.e.,

$$R_i(a^i) = \lim_{m\to\infty} \frac{1}{m} \sum_{n=1}^{m} Q_{\ell_n}^i(a_1^i, \ldots, a_{\ell_n}^i), \tag{6}$$

where $\ell_n$ is the index of the $n$-th job sent to queue $i$. This limit exists as soon as the system is ergodic.

For any $p$ and $\theta$ in $\mathbb{R}$, let us introduce the *Sturmian sequence with rate $p$ and phase $\theta$*, $\alpha(p,\theta) \stackrel{\text{def}}{=} \big(\alpha_n(p,\theta)\big)_{n\in\mathbb{N}}$ where, for all $n \geq 1$, $\alpha_n(p,\theta) \stackrel{\text{def}}{=} \lfloor np + \theta \rfloor - \lfloor (n-1)p + \theta \rfloor$ (see [2] and references therein for more details on Sturmian sequences). Note that $\alpha_n(p,\theta) \in \{0,1\}$ for all $n$ as long as $p \leq 1$ and it is periodic in $\theta$ with period 1.

**Theorem 1** *Under the foregoing notations, the response time of a job under any policy $a$ is bounded from below by a combination of response times in all queues using Sturmian sequences:*

$$R(a) \geq \inf_{\substack{p_1,\ldots,p_N\geq 0:\\ p_1+\cdots+p_N=1}} \big(p_1 R_1(\alpha(p_1,0)) + \cdots + p_N R_N(\alpha(p_N,0))\big).$$

This result is to be compared with [17], where Sturmian sequences with rate $r$ are proved to be optimal admission sequences in a single queue under the constraint that a proportion of at least $r$ jobs have to be admitted in the queue. The main difference comes from the fact that routing to several queues is more difficult than admitting to a single queue because one does not know whether the proportion of jobs sent to each queue by the optimal routing policy exists. This is still an open problem and Theorem 1 above does not answer to this question but just provides a lower bound on the response time of the optimal policy. On the other hand, the result stated in Theorem 1 is very close to Theorem 25 in [2]. The main difference is the fact that our cost is not additive, making the proof slightly more involved (see the appendix).

Let us consider a single queue $i$ and the arrival process induced by $\alpha(p_i, \theta)$ in queue $i$. Let $k \overset{\text{def}}{=} \lfloor 1/p_i \rfloor$. The inter-arrival times $\tau_1, \ldots, \tau_n, \ldots$ are made alternating sums of $k$ and $k+1$ i.i.d. exponential random variables with rate $N\lambda$. For example, if $p_i = 2/7$, then $k = 3$ and the arrival process in queue $i$ under policy $\alpha(2/7, 0) = 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, \ldots$ where the sequences $0, 0, 1$ and $0, 0, 0, 1$ alternate, is such that the distribution of the inter-arrival times alternates between the sum of three exponentials with rate $N\lambda$ and the sum of four exponentials with rate $N\lambda$. The analysis of response time when the arrival process at a queue has this alternating pattern is not easy. Therefore, we now develop a more tractable bound.

In general, the arrival rate in queue $i$ is $p_i N\lambda$. Now, considering a stationary i.i.d. arrival process $T_1, \ldots, T_n, \ldots$, with a Gamma distribution for inter-arrival times, with parameters $p_i$ and $N\lambda$. It should be clear that for any $n$, these two processes of the inter-arrival times compare for the convex ordering of random sequences, i.e., $(\tau_1, \ldots, \tau_n) \geq_{cx} (T_1, \ldots, T_n)$.

Using the fact that the response time of jobs is a convex increasing function of the input process, this implies that the response time in queue $i$ under a Sturmian arrival process with rate $p_i$ is larger than the response time in queue $i$ with Gamma-distributed inter-arrivals with rate $p_i N\lambda$. This argument and Theorem 1 yield the following result.

**Corollary 1** *Let $R_i^{\Gamma(a,b)/GI/1}$ be the response time of a job in queue $i$ having general i.i.d. service times and $\Gamma(a, b)$ i.i.d. inter-arrival times. Then,*

$$R^{Opt} \geq \inf_{\substack{\pi_1,\ldots,\pi_N \geq 0: \\ \pi_1+\cdots+\pi_N=1}} \sum_{i=1}^{N} \pi_i R_i^{\Gamma(1/\pi_i, N\lambda)/GI/1}. \tag{7}$$

In the following, we will use a coefficient that scales with $N$ for the proportion of jobs sent to queue $i$: We define $\beta_i \overset{\text{def}}{=} N\pi_i$, where $\beta_i$ is a positive constant. A lower bound on $R^{Opt}$ is finally obtained by solving the following optimization problem

$$\begin{array}{rll} GB(N) \overset{\text{def}}{=} & \min & \displaystyle\sum_{i=1}^{N} \frac{\beta_i}{N} R_i^{\Gamma(N/\beta_i, N\lambda)/GI/1} \\ & \text{s.t.} & \\ & & \sum_{i=1}^{N} \beta_i = N \\ & & U_i(\beta_i) \leq 1, \ \forall i \\ & & \beta_i \geq 0, \ \forall i, \end{array} \tag{8}$$

where

$$U_i = U_i(\beta_i) = \lambda\beta_i/\mu_i \tag{9}$$

and $GB(N)$ stands for Gamma-Bound with $N$ queues. By means of Little's law, the quantity $U_i$ is interpreted as the *utilization* of station $i$, and it represents the "proportion of time" in which station $i$ is busy (in the long term). To obtain a computable bound, one can use heavy-traffic bounds. Since the $\Gamma(N/\beta_i, N\lambda)$ distribution is increasing-failure-rate (because $N/\beta_i > 1$), we have (see [24, Formula 2.51] for details)

$$\begin{aligned} R_i^{\Gamma(N/\beta_i, N\lambda)/GI/1} & \geq \lambda\beta_i \frac{1/(\lambda^2 N\beta_i) + s_i^2}{2(1 - U_i(\beta_i))} - \frac{1}{2\lambda\beta_i}(\beta_i/N + U_i(\beta_i)) + \frac{1}{\mu_i} \\ & = \frac{1/(\lambda N) + \lambda\beta_i s_i^2}{2(1 - U_i(\beta_i))} - \frac{1}{2\lambda N} + \frac{1}{2\mu_i}. \end{aligned} \tag{10}$$

Substituting (10) in (8), we finally obtain

$$
\begin{array}{rll}
R^{Opt} \geq GB(N) \geq & \min & \displaystyle\frac{1}{2}\sum_{i=1}^{N}\frac{\beta_i}{N}\left[\frac{1/(\lambda N)+\lambda\beta_i s_i^2}{1-U_i(\beta_i)}+\frac{1}{\mu_i}\right]-\frac{1}{2\lambda N} \\
& \text{s.t.} & \\
& & \sum_{i=1}^{N}\beta_i = N \\
& & U_i(\beta_i) \leq 1,\ \forall i \\
& & \beta_i \geq 0,\ \forall i.
\end{array}
\tag{11}
$$

The structure of the optimization problem in (11) is identical to problem OR2-IID in [16], which is convex. Therefore, efficient polynomial-time algorithms can be applied for its solution [11].

## 2.2 Other Cost Functions: Convex Functions of Response Times

So far, we have considered the response time as our main cost function. However, it is possible to increase the expressive power of our approach by considering combinations of convex functions of the response times in each queue. For instance, these include the case where the queues have *holding costs*. More precisely, let $\phi_i : \mathbb{R} \to \mathbb{R}$ be arbitrary increasing convex functions for all $1 \leq i \leq N$.

We introduce an immediate cost for queue $i$ at step $n$, denoted by $\overline{Q_i^n}(a_1^i, \dots, a_n^i)$ and equal to the expectation of $\phi_i$ applied to the queue size under arrivals $(a_1^i, \dots, a_n^i)$. Now, let us consider the average cost in queue $i$ under policy $a$ to be $\overline{R_i}(a^i) \overset{\text{def}}{=} \lim_{m\to\infty} \frac{1}{m}\sum_{n=1}^{m} \overline{Q^i_{\ell_n}}(a_1^i, \dots, a_{\ell_n}^i)$, and the global cost of policy $a$ is

$$
\overline{R}(a) \overset{\text{def}}{=} \limsup_{K\to\infty} \frac{1}{K}\sum_{n=1}^{K}(a_n^1\overline{Q}_n^1(a_1^1,\dots,a_n^1)+\cdots+a_n^N\overline{Q}_n^N(a_1^N,\dots,a_n^N)).
\tag{12}
$$

Corollary 1 extends to that case.

**Corollary 2** *Using the foregoing notations,*

$$
\overline{R}^{Opt} \geq \inf_{\substack{\pi_1,\dots,\pi_N\geq 0:\\ \pi_1+\cdots+\pi_N=1}} \sum_{i=1}^{N}\pi_i\overline{R}_i^{\Gamma(1/\pi_i,N\lambda)/GI/1}.
\tag{13}
$$

However, note that Corollary 2 may not hold for cost functions that are arbitrary multi-dimensional convex functions of the buffer sizes.

Substituting Formula (10) in (13), we obtain a convex program that generalizes (11). In the following we will remove the functions $\phi_i$, however most of the results that follow can be extended to that convex case, in particular, when the functions $\phi_i$ are linear.

## 2.3 Bound and Approximation for Exponential Service Times

Assuming exponential service times, we can improve the accuracy of bound (11). The integration of the exact $\Gamma/M/1$-queue analysis, see [10], in the constraints of (8) renders a non-linear problem which seems to be difficult to analyze, e.g., in terms of convexity, and also yields numerical instabilities related to $O(N^N)$ terms. Therefore, we now address the development of simple approximations for the response time of $\Gamma/M/1$ queues better than (10). These hold in the regime where the job arrival rate to the broker proportionally grows with the number of queues, for which the *ideal* job arrival processes of each queue considered by our bound $GB(N)$ become more and more deterministic.

The following theorem provides bounds on $R_i^{\Gamma(N/\beta_i,N\lambda)/M/1}$ for any network load and size.

**Theorem 2** *Let $\sigma_i, {\sigma_i}^{+} \in [0,1)$, respectively, be the (unique) solutions of the equations*

$$
z\,exp(\tfrac{1-z}{U_i}) = 1
\tag{14}
$$

*and*

$$z exp(\tfrac{1-z}{U_i})\left(1-\frac{1}{2}\frac{(1-z)^2}{NU_i^2}\right)=1. \tag{15}$$

*Then,*

$$\frac{1}{\mu_i(1-\sigma_i)} \leq R_i^{\Gamma(N/\beta_i,N\lambda)/M/1} \leq \frac{1}{\mu_i(1-\sigma_i^+)}. \tag{16}$$

Given that, as $N \to \infty$, $\sigma_i^+ \to \sigma_i$, $\forall i$, the following corollary is straightforward.

**Corollary 3** *As $N \to \infty$,*

$$R_i^{\Gamma(N/\beta_i,N\lambda)/M/1}(\beta_i) \to \frac{1}{\mu_i(1-\sigma_i)} \tag{17}$$

*from above.*

**Lemma 1** *$\sigma_i \leq U_i^2$.*

Rewriting (14) as

$$-\frac{z}{U_i}e^{-\frac{z}{U_i}} = -\frac{1}{U_i}e^{-\frac{1}{U_i}} \tag{18}$$

and observing that it admits exactly two positive roots when $0 \leq U_i \leq 1$, where the largest one is at $z=1$, we note that $\sigma_i$ can be expressed in terms of the Lambert $\mathcal{W}$ function [14] if and only if $-z/U_i \geq -1 = -\mathcal{W}(-1/e)$, which is true by Lemma 1. Hence,

$$\sigma_i = -U_i\mathcal{W}(-\frac{1}{U_i}e^{-\frac{1}{U_i}}). \tag{19}$$

where $\mathcal{W}$ is the principal Lambert function (with $\mathcal{W}(0)=0$).

We recall that the Lambert $\mathcal{W}$ function [14], defined as the inverse function of $f(\mathcal{W})=\mathcal{W}\exp(\mathcal{W})$, over $[-1,+\infty)$, satisfies

$$0 \leq -\mathcal{W}\left(-\frac{1}{U_i}e^{-\frac{1}{U_i}}\right) \leq 1 \tag{20}$$

for all $0 \leq U_i \leq 1$. In particular, $-\mathcal{W}(-1/e)=1$ and $\mathcal{W}(0)=0$.

The rate of convergence of (17) is strictly related to the convergence of $(1+a/N)^N$, for $a$ fixed, to its limiting value $\exp(a)$, which is known to be $\Theta(1/N)$ as it can be shown by a Taylor expansion (see the proof of Theorem 2). In the experimental results section, we numerically show that this suffices to obtain very accurate response time estimates even when $N$ is relatively small and that it provides improved accuracy with respect to heavy-traffic approximations.

The simplicity of Formula (17) allows for the development of a simple optimization procedure. In fact, problem (8) can be rewritten as follows

$$\begin{array}{rll} GB^-(N) \overset{\text{def}}{=} & \min & \displaystyle\frac{1}{\lambda N}\sum_{i=1}^{N}\frac{U_i}{1-\sigma_i(U_i)} \\ & \text{s.t.} & \\ & & \sum_{i=1}^{N}\frac{\mu_i}{\lambda}U_i = N \\ & & 0 \leq U_i \leq 1,\ \forall i, \end{array} \tag{21}$$

where $\sigma_i(U_i)$ is given by (19).

**Remark 4** *By Theorem 2, $GB^-(N) \leq GB(N) \leq R^{Opt}(N)$ for any $N$.*

The following result ensures that efficient algorithms can be immediately applied to solve (21) in polynomial time [11].

**Theorem 3** *The optimization problem* (21) *is convex.*

In heavy-traffic, the structure of optimization problem (21) also allows for an explicit expression of $GB^-(N)$.

**Theorem 4** *If $\sum_{i=1}^N \mu_i - \lambda N = \epsilon$ for $\epsilon > 0$ sufficiently small, then*

$$GB^-(N) \sim \frac{1}{2}\frac{1}{\lambda N}\left(\frac{(\sum_{i=1}^N \sqrt{\mu_i})^2}{\sum_{i=1}^N \mu_i - \lambda N} - 1\right) \tag{22}$$

*and the corresponding utilizations of all queues are*

$$U_j \sim 1 - \frac{1}{\sqrt{\mu_j}}\frac{\sum_{i=1}^N \mu_i - \lambda N}{\sum_{i=1}^N \sqrt{\mu_i}}. \tag{23}$$

The following corollary of Theorem 4 is a direct consequence of the known explicit expression of $R_{Bernoulli}^{Opt}$ [20] and the fact that all queues must be used in heavy-traffic.

**Corollary 4** *If $\sum_{i=1}^N \mu_i - \lambda N = \epsilon$ for $\epsilon > 0$ sufficiently small, then $GB^-(N) \sim R_{Bernoulli}^{Opt}/2$.*

Here, it is worth anticipating one important property of our lower bound that will be analyzed in Section 5: Assuming that the $GB^-(N)$ bound is tight (this will be verified later), Corollary 4 implies that the impact of having memory in the broker (the ratio $R_{Bernoulli}^{Opt}/GB^-(N)$) is independent of the network heterogeneity (service rates) and size in heavy-traffic.

Let also $GB^+(N)$ be the optimum of (21) where $\sigma_i(U_i)$ is given by (15). Even though $GB^+(N)$, in general, does not seem to provide upper bounds on the minimum response time, the following result ensures that it always provides improved accuracy with respect to $GB^-(N)$ when estimating $R^{Opt}$.

**Theorem 5** $R^{Opt} - GB^-(N) > |R^{Opt} - GB^+(N)|$.

Even though more accurate approximations than $GB^-(N)$ and $GB^+(N)$ for $R^{Opt}$ can be derived (by taking into account more expansions terms, see the proof of Theorem 5), we numerically show that they suffice to obtain very accurate results. A numerical evaluation of its tightness and convergence speed is postponed in the experimental results section.

# 3 Optimal Routing

The framework introduced in Section 2.1 allows us to numerically inspect the average response-time gap between our bounds and heuristic strategies for the optimal routing. In this section, we first perform a validation of Formula (17) on several models. Then, we measure the performance achieved with a broker assigning jobs to queues according to a proposed billiard sequence. We show that the resulting distance from our formulas is remarkably small. Concisely, our empirical conclusions are that i) *the proposed routing scheme minimizes response time*, and ii) *our bounds and approximations on the minimum response time are tight*.

## 3.1 Accuracy of Formula (17)

We now measure the accuracy of asymptotic formula (17) by means of the percentage relative error

$$\frac{|R_{exact}^{\Gamma/M/1} - R_{approx}^{\Gamma/M/1}|}{R_{exact}^{\Gamma/M/1}}100\%, \tag{24}$$

where $R_{exact}^{\Gamma/M/1}$ is obtained numerically through the (exact) standard analysis of the $G/M/1$ queue, and $R_{approx}^{\Gamma/M/1}$ is given by (17). Numerical computations have been performed using Maple 13. We initially evaluate (24) by varying $N \in \{50, 100, 200, 1000\}$ and $U \in \{0.1, 0.2, \ldots, 0.9, 0.95\}$. Since the mean arrival rate $\lambda$ affects the percentage relative error (24) only through the utilization, it is not considered in our experiments. Figure 2 illustrates the quality of (24) in the above cases. As $N$ grows, we first note that the accuracy of (17) increases, which is expected because it is asymptotically exact. For $N = 50$, (17) is remarkably accurate and yields a relative error always less than 2%.

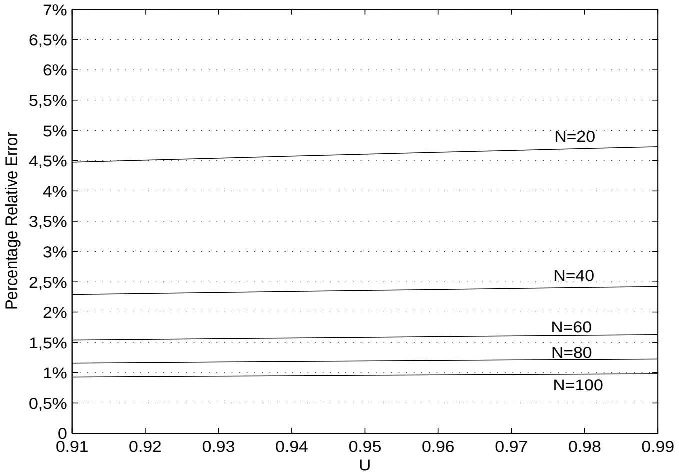

Figure 2: Accuracy evaluation of the asymptotic formula (17) through the error measure (24).

## 3.2 Quasi-Optimality of Billiard Sequences

We consider the case where the broker forwards jobs to queues according to billiard sequences, see [5, 21], that are constructed through the SG algorithm introduced in [21] (easily implementable in network brokers with a very limited cost). The SG algorithm takes as input the fraction of jobs to send to the queues (given by the solution of (21)) and an initial-position vector $x \in \mathbb{R}^N$ which we assume such that $x_i = 1$ if $\mu_i = \max_j \mu_j$ and 0 otherwise (we point the reader to [21] for further details on the SG algorithm and billiard sequences). Therefore, the buffer size of the broker must be no less than $O(\log N)$ bits, which is very small. Given that a numerical solution of the response time induced by billiard sequences is impractical to compute for a number of reasons, e.g., the aperiodicity of the resulting patterns, we use simulation. To measure the gap between the response time achieved with this routing scheme and our bounds/approximations, we assess the general quality of the percentage relative error

$$\mathrm{Err}_{\mathrm{App}} = \frac{|R_{\mathrm{App}} - R_{\mathrm{Sim}}|}{GB^{+}(N)} \cdot 100\% \tag{25}$$

where $R_{\mathrm{App}} \in \{GB^{-}(N), GB^{+}(N)\}$ (defined in Section 2.3) and $R_{\mathrm{Sim}}$ is the average response time computed by simulation. We measure percentage relative errors with respect to $GB^{+}(N)$ because it represents the closest approximation of $R^{Opt}$ (see Theorem 5). The measures of $R_{\mathrm{Sim}}$ refer to 99% confidence intervals having size no larger than 1% of $R_{\mathrm{Sim}}$ itself. For any pair $(N, \rho)$, $N \in \{20, 50, 100\}$ and $\rho \in \{0.10, 0.15, 0.20, \ldots, 0.95\}$, we generated 1,000 random models where the service rates $\mu_i$ have been drawn in the range $[0.01, 100]$ according to a uniform distribution. Larger values of $N$ have not been considered because of the strong computational requirements of simulation. In any case, the proposed analysis suffices to assess the accuracy of our approach.

The experimental results of this analysis are summarized in Figure 3.2, which refers to a total of nearly 50,000 experiments. In the figure, the dashed (continuous) lines refer to the error obtained with $GB^{-}(N)$ ($GB^{+}(N)$) for different network sizes. We clearly see that the response time achieved through a billiard routing is remarkably close to our approximation $GB^{+}(N)$ and also to our bound $GB^{-}(N)$. Given that the optimal response time achievable by the system must lie between our bound and the response time achieved by the billiard routing, we conclude, in an empirical sense, that billiard sequences are optimal for the response time that is very-well approximated by the solution of our optimization problems.

## 3.3 Computational Requirements

We illustrate the computational requirements for calculating our tight bound on $R^{Opt}$ through (21). These are important to know because the program (21) should be executed each time the network changes to

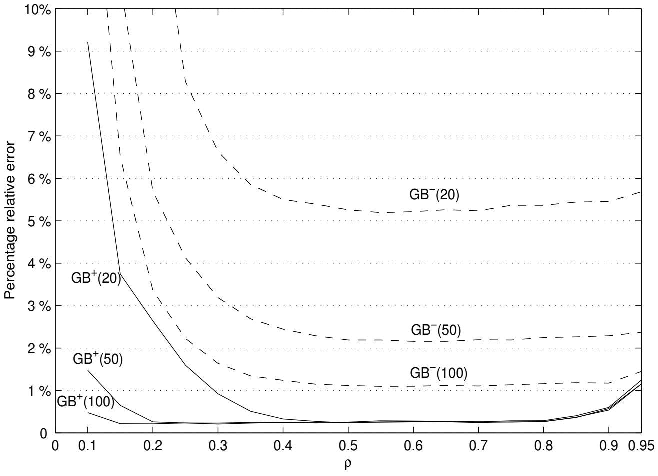

Figure 3: Plots of the error (25) averaged over a large number of experiments by varying the network load.

reinitialize the parameters of the optimal routing algorithm (e.g., addition or removal of one queue, variation of the arrival rate or service times, CPU frequency scaling). Experiments have been performed by running the IBM Ipopt optimization solver on a 2.80Ghz Intel Xeon processor with multi-threading technology. By varying $N$, we consider a wide test-bed of randomly generated models, where the service rates uniformly range in $[0.01, 100]$ and the arrival rate is such that the network load (1) uniformly ranges in $[0, 1]$. Table 1 illustrates the average and the standard deviation of the computation times (in seconds) required by the solution of (21), where each number refers to a sample of 1,000 models. From the results in the table, we conclude that the solution of (21) is almost online: Models of networks composed of a thousand of queues require less than three seconds in average.

| $N$ | 50 | 100 | 500 | 1,000 | 5,000 |
|---|---|---|---|---|---|
| Average time (sec) | 0.181 | 0.251 | 1.317 | 2.433 | 10.85 |
| Std. dev. | 0.038 | 0.059 | 0.187 | 0.312 | 2.901 |

Table 1: Seconds required by the computation of (21) by varying $N$.

# 4 Price of Forgetting

In this section, we obtain insights on the benefit of having memory in the broker. One of the major conclusions is that the added-value of having memory in the broker is negligible (significant) if the coefficients of variation of the service time distributions are large (small). We establish this property by introducing the Price of Forgetting (PoF) as the ratio between the optimal response time achieved with a Bernoulli broker and a broker with memory:

$$PoF \overset{\text{def}}{=} R^{Opt}_{Bernoulli}/R^{Opt}. \tag{26}$$

## 4.1 Heterogeneous Queues

In general, bounds on the PoF can be obtained numerically by computing our lower bound on $R^{Opt}$ through (11) and $R^{Opt}_{Bernoulli}$ as in, e.g., [13]. However, this approach prevents the understanding of its qualitative dependence with respect to general input parameters. In the heavy-traffic case, we are able to derive explicit bounds for the PoF. In contrast, in light-load conditions, i.e., $\rho \to 0$, we must have $PoF(N, \rho) = 1$, which is intuitive. In fact, if the queue lengths are almost empty, then the response time approaches the mean service time of the fastest queue.

We recall that the proposed bound $GB(N)$ is interpreted as the response time achieved when the arrival processes of all queues are independent and Gamma distributed. This means that the broker can be now thought as Bernoulli, provided that its job inter-arrival times are i.i.d. and Gamma distributed. Given that all queues become independent $\Gamma/GI/1$ queues, classic heavy-traffic analysis immediately applies to derive useful approximation and insights; see, e.g., [23, 16].

This corollary mainly follows by Theorem 1 and the heavy-traffic analysis of GI/GI/1 queues.

**Corollary 5** *For any $N$ and as $\rho \to 1$,*

$$PoF(N) \leq 1 + \frac{1}{\min_{i=1}^{N} \mu_i^2 s_i^2}. \tag{27}$$

Corollary 5 shows that the PoF can be arbitrarily close to one in heavy-traffic (the $\mu_i$'s are kept fixed and $\lambda$ is increased) if the smallest coefficient of variation of the service time distributions, i.e., $\mu_i s_i$, is large. In this case, the impact of memory is negligible.

**Remark 5** *If the coefficient of variation of all service time distributions is large, $R^{Opt} \approx R^{Opt}_{Bernoulli}$.*

This can be the case of web-server farms, where the squared coefficient of variation of the size of the incoming requests (and thus of their service times) is typically between 8 and 50.

On the other hand, the PoF can be unbounded if $\min_i \mu_i^2 s_i^2$ is small: If the service times are Erlang distributed with mean $\mu_i$ and $k$ phases, it follows that $PoF(N) \leq 1 + k$.

#### 4.1.1 Exponential service times

Under the assumption of exponential service times, Corollary 4 and 5 imply that $PoF(N) \leq 2$ in heavy-traffic. The following theorem extends this result to the non-heavy-traffic case.

**Theorem 6** *Assume that service times are exponentially distributed. For any $N$ and $\rho$, $PoF(N, \rho) \leq 2$.*

In words, the minimum response time achieved by a probabilistic broker can be at most twice larger than the minimum response time achieved by a broker with memory.

### 4.2 Homogeneous Queues

A scenario of practical interest is the case where the queues are *homogeneous* (or statistically equivalent), i.e., $\mu_1 = \cdots = \mu_N = \mu$ and $s_1 = \cdots = s_N = s$, for which we can draw additional results and easily compare with the case of Bernoulli brokers.

The following result is known in the literature (and also follows from Theorem 1 using a symmetry argument); see, e.g., [32] Prop. 8.3.4.

**Theorem 7 ([32])** *Under the foregoing assumptions, the round-robin policy minimizes response time for any $N$.*

Therefore, the PoF is the ratio between the response times of a $M/GI/1$ and of a $\Gamma/GI/1$ having the same mean arrival rate and service time distribution. This corollary follows algebraically by using i) the Pollaczek–Khintchine formula for the $M/GI/1$ queue, ii) Formula (10) for the upper bound, and iii) the inequality [24]

$$R^{\Gamma(N,N\lambda)/GI/1} \leq \lambda \frac{1/(\lambda^2 N) + s^2}{2(1-\rho)} + \frac{1}{\mu} \tag{28}$$

for the lower bound.

**Corollary 6** *If the queues are homogeneous, then*

$$\frac{\rho(\mu^2 s^2 - 1) + 2}{1/(\rho N) + \rho(\mu^2 s^2 - 2) + 2} \leq PoF(N, \lambda, \mu, s) \leq 1 + \frac{1 - 1/N}{1 + 1/N + (\mu^2 s^2 - 1)\rho} \tag{29}$$

*where $\rho = \lambda/\mu$.*

Thus, the PoF is bounded from below and above by two functions that, in heavy-traffic, converge to

$$1+\frac{1-1/N}{\mu^2 s^2+1/N}. \tag{30}$$

As stated above, the PoF depends on the squared coefficient of variation, i.e. $\mu^2 s^2$, of the service requirements: Since (28) is exact in heavy-traffic, (30) is the exact PoF of the system (as $\rho \to 1$). In the case of deterministic service times, i.e. $s=0$, it is easy to see that the heavy-traffic PoF grows linearly with $N$, meaning that

$$PoF(N,\lambda,\mu,0)=N. \tag{31}$$

### 4.2.1 Exponential service times

The results which follow in the remainder of this section are implicitly assumed to hold when i) service times are exponential, ii) queues are homogeneous and iii) $N \to \infty$, and provide upper bounds for the finite case. The main difference with the analysis above is that here we provide exact results even for the non-heavy-traffic case.

The following result is an immediate consequence of Theorems 2 and 7, and provides an asymptotically-exact formula for the PoF.

**Corollary 7**

$$PoF(\rho)=\frac{1+\rho\mathcal{W}(-exp(-1/\rho)/\rho)}{1-\rho} \tag{32}$$

*where* $\rho=\lambda/\mu$.

It can be shown that (32) converges to one in light-traffic and to two in heavy-traffic, which is in agreement with Corollary 4, Formula (30) and Theorem 6. Note that (30) becomes $1+\frac{1-1/N}{1+1/N}$ in the case of exponentially distributed service times. In contrast, the expression (32) depends on the network utilization and allows us to derive further properties.

**Corollary 8** *$PoF(\rho)$ is strictly increasing in $\rho$ and*

$$\lim_{\rho\to 0}\frac{dPoF(\rho)}{d\rho}=1,\quad \lim_{\rho\to 1}\frac{dPoF(\rho)}{d\rho}=0. \tag{33}$$

The limits in (33) and the monotonicity of the PoF show that i) the response-time benefits of a broker with memory are non-negligible even when the utilizations are small, and that ii) $PoF(\rho)$ is concave in heavy-load conditions (concavity does not hold for $PoF(\rho)$ in general), and, thus, large improvements can be obtained even in a non-negligible neighborhood of $\rho=1$.

# 5 Impact of Memory in the Broker: Structural Properties

We now measure the proposed upper bound on the PoF in order to numerically investigate its fundamental properties. Following the results of previous section, it is very tight. We infer an important structural property: *the PoF only depends on the network load $\rho$, meaning that it is independent of the network heterogeneity and size.*

## 5.1 Homogeneous Queues

In the case of homogeneous queues, the proposed bound boils down to the simple formula (32), which is asymptotically exact. By varying the utilization from 0.05 to 0.95 with step 0.05, Figure 4 illustrates i) the asymptotic PoF (32) (the dashed bold line), ii) the PoF obtained with a memoryless broker (the dashed-dotted line), and iii) for $N \in \{10,50,100,1000\}$, the exact PoF, which is obtained by applied standard analysis of the $E_N/M/1$ queue. In that figure, we first notice that the PoF is not concave and (slightly) increases as $N$ does converging to our asymptotic formula (32). The fact that the PoF increases with $N$ finds

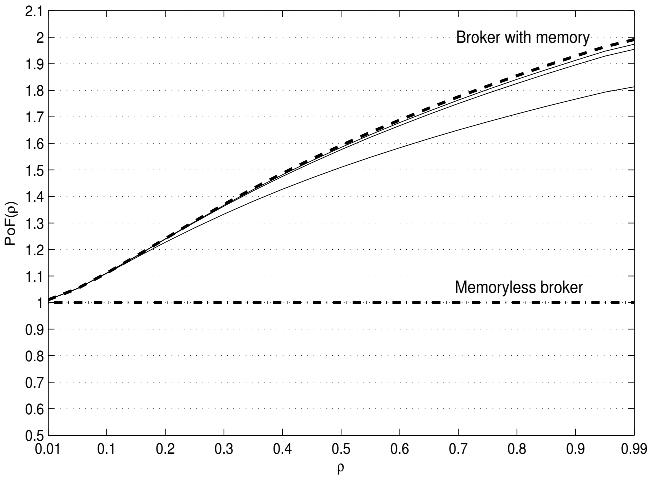

Figure 4: PoF by varying the queues utilization. The three continuous lines correspond to the exact PoF for increasing network sizes, where the lowest refers to $N = 10$ and the largest to $N = 1000$.

the simple intuition that adding new resources gives more and more freedom to the broker for optimizing the response time with respect to its Bernoulli counterpart. The exact PoF computed for $N = 100$ is very close to our asymptotic formula and, for $N = 10$, it has almost the same behavior. When $N = 100$ and $U = 0.85$, Figure 4 shows that a Bernoulli-based analysis underestimates the PoF of a factor 1.9. When the utilizations are 0.1, i.e., small, the Bernoulli PoF is 10% lower. These observations immediately quantify how large can be the worst-case impact of considering brokers with memory in the design of distributed or centralized systems, where utilizations usually range in $[0.6, 0.85]$.

## 5.2 Heterogeneous Queues: Independence of Network Heterogeneity and Size

We now measure the PoF in the heterogeneous case. We first consider an illustrative example which we use to inspect fundamental properties. Then, we carry out an extensive numerical analysis to give evidence of their correctness.

**An illustrative scenario** We consider a clustered network composed of $N$ queues where 1/10 of the queues have fast service rates $\mu_f = 100$, 2/10 of the queues have medium service rates $\mu_m = 50$, and the remaining ones have low service rates $\mu_l = 1$. By varying the network load ($\rho$) and size ($N$), we plot the resulting PoFs in Figure 5, which lets us draw two important hypotheses.

First, we observe that *our bound on the PoF is independent of the network size.* Second, if the ratios of Figure 5 are compared pointwisely to the corresponding ones of Figure 4 (where the concepts of network load and utilization are equivalent) we note that these points are very close each other. This suggests that *the PoF is not influenced by the heterogeneity of the considered scenario and depends on the network load only.* In Section 4.1.1, we showed that this property holds true in heavy-traffic and as $N$ grows.

**Exhaustive numerical investigation** We now carry out an extensive numerical analysis to give evidence of the independence of the PoF on the network size and heterogeneity. To do this, we focus on a very large test-bed of randomly generated models drawing the service rates $\mu_i$ in $[0.01, 100]$ uniformly. For any pair $(N, \rho)$, we generated 1,000 models computing average and standard deviation of the PoF. The results of this analysis are shown in Table 2, which refers to a total of 48,000 different models. The results presented in that table robustly confirm the two hypotheses arisen in previous section. When $N = 50$, we observe that the averages of the PoF are already settled to their asymptotic value. Furthermore, the standard deviations are very small and decreasing in both $N$ and $\rho$. This shows the independence with respect to the network heterogeneity. By varying $\rho$ and $U$ (for $\rho = U$), Figure 6 plots (32) and the average PoF shown in Table 2

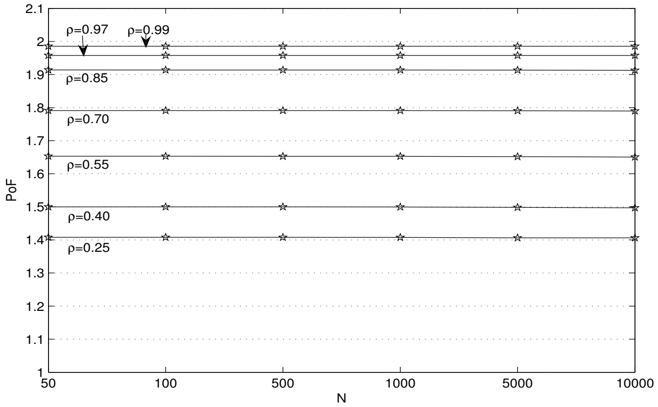

Figure 5: Insensitivity of the PoF with respect to network size.

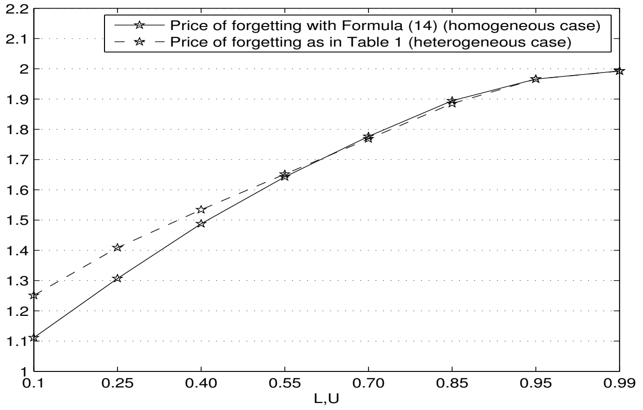

Figure 6: Comparison of Formula (32) with the averages of the PoFs in Table 2 by varying the network load.

to stress independence with respect to heterogeneity. Both curves are remarkably close each other, and they are almost equivalent when $\rho \geq 0.55$. In the figure, we observe that the slight gap achieved when $\rho$ is small must go to zero as $\rho \to 0$ because, in this regime, $R^{Opt} \to \max_i \mu_i$ and $R^{Opt}_{Bernoulli} \to \max_i \mu_i$.

Except for the heavy-traffic case, this is surprising because *the optimal fractions of jobs sent to each queue in the Bernoulli and non-Bernoulli settings are different* (see next section). These structural properties and the tightness of our GB bounds imply that the minimum response time $R^{Opt}$ can be seen as the product-form

$$R^{Opt} \approx R^{Opt}_{Bernoulli}/PoF(\rho), \tag{34}$$

where $PoF(\rho)$ is given by (32).

## 5.3 Optimal Routing Probabilities Comparison

We show the relation between the long-term fractions of jobs sent to each queue by the optimal Bernoulli broker ($p_i$) and of our bound (21) ($\pi_i$) by evaluating the distance $\sum_{i=1}^{N} |\pi_i - p_i|$ over the experiments performed in previous section. While in heavy-traffic the fractions of jobs in a memory/non-memory setting are equal (which is obvious) and the properties above could find some interpretation, this does not hold for the non-heavy-traffic case, for which a significant difference exists (see Table 3). Notwithstanding, the PoF is not affected by such difference as shown in previous section.

Averages

| $N$ | 50 | 100 | 500 | 1,000 | 5,000 | 10,000 |
|---|---|---|---|---|---|---|
| $\rho = 0.10$ | 1.252 | 1.254 | 1.254 | 1.254 | 1.253 | 1.253 |
| $\rho = 0.25$ | 1.408 | 1.409 | 1.409 | 1.409 | 1.409 | 1.409 |
| $\rho = 0.40$ | 1.534 | 1.534 | 1.535 | 1.534 | 1.535 | 1.534 |
| $\rho = 0.55$ | 1.652 | 1.652 | 1.652 | 1.652 | 1.652 | 1.652 |
| $\rho = 0.70$ | 1.768 | 1.768 | 1.768 | 1.768 | 1.768 | 1.768 |
| $\rho = 0.85$ | 1.885 | 1.885 | 1.885 | 1.885 | 1.885 | 1.885 |
| $\rho = 0.95$ | 1.966 | 1.966 | 1.966 | 1.966 | 1.966 | 1.966 |
| $\rho = 0.99$ | 1.992 | 1.992 | 1.992 | 1.992 | 1.992 | 1.992 |

Standard deviations

| $N$ | 50 | 100 | 500 | 1,000 | 5,000 | 10,000 |
|---|---|---|---|---|---|---|
| $\rho$ =0.10 | 3.0e-2 | 2.0e-2 | 8.3e-3 | 7.9e-3 | 6.9e-3 | 6.1e-3 |
| $\rho$ =0.25 | 1.4e-2 | 1.0e-2 | 5.3e-3 | 4.8e-3 | 2.8e-3 | 1.8e-3 |
| $\rho$ =0.40 | 9.5e-3 | 6.9e-3 | 3.1e-3 | 2.3e-3 | 2.1e-3 | 8.9e-4 |
| $\rho$ =0.55 | 5.3e-3 | 3.9e-3 | 1.7e-3 | 1.3e-3 | 7.1e-4 | 6.4e-4 |
| $\rho$ =0.70 | 2.9e-3 | 2.1e-3 | 1.0e-3 | 8.3e-4 | 5.5e-4 | 5.3e-4 |
| $\rho$ =0.85 | 2.1e-3 | 1.6e-3 | 7.7e-4 | 6.4e-4 | 4.8e-4 | 4.5e-4 |
| $\rho$ =0.95 | 7.3e-4 | 5.9e-4 | 4.7e-4 | 4.5e-4 | 4.4e-4 | 4.3e-4 |
| $\rho$ =0.99 | 4.8e-4 | 4.6e-4 | 4.3e-4 | 4.4e-4 | 4.3e-4 | 4.2e-4 |

Table 2: Averages and standard deviations of our bound on the PoF over the large number of tests (e-n reads $10^{-n}$).

| $\rho$ | 0.25 | 0.40 | 0.55 | 0.70 | 0.85 | 0.95 | 0.99 |
|---|---|---|---|---|---|---|---|
|  | 1.9e-1 | 7.7e-2 | 2.6e-2 | 6.2e-3 | 5.4e-4 | 1.3e-5 | $\leq$e-5 |

Table 3: $\sum_{i=1}^{N} |\pi_i - p_i|$ by varying the network load (e-n reads $10^{-n}$).

# 6 Price of Anarchy

The Price of Anarchy (PoA) [25] is an index measuring the inefficiency of a decentralized system with respect to its centralized counterpart in presence of selfish users. It is defined as the response-time ratio between the worst-case situation where users behave to maximize their own individual benefit, yielding some *game-theoretic equilibrium*, and the contrasting situation where users are controlled optimally by a central authority, e.g., a broker, yielding the *social optimum*. While the former identifies the equilibrium point for which any unilateral deviation of each job strategy does not lower its delay, the latter represents the optimal strategy for all users in a centralized setting.

In the context of queueing models, the interest for the PoA is currently growing because of its large spectrum of applications: *-computing, network routing, load balancing, peer-to-peer, wireless networks, server farms [28]. The great majority of existing works provide mathematical tools for characterizing and computing the response times in both the situations described above and try to relate the PoA to the network size in different settings. This lets designers estimate the loss of performance that occurs in shifting to decentralized solutions and subsequently perform a suitable dimensioning of the system. It is shown in [29] that the PoA is independent of the network topology as long as the mean job arrival rate is less than the mean service rate of the slowest queue, and, in this light-load regime, an upper bound is provided. When heterogeneous processor-sharing queues are considered, it is shown in [20, 34] that the PoA scales linearly with the network size, and it can only depend on the heterogeneity degree of the queues provided that these adopt the shortest-remaining-processing-time scheduling discipline [12]. In the case of multiple central authorities, which can be the case of large server farms, the PoA is shown to be lower bounded by the square root of the number of authorities [6].

We observe that a key point common to all the above works is that the broker is implicitly assumed to dispatch jobs to queues in a Bernoulli manner.

## 6.1 Wardrop Equilibrium and Social Optimum Revisited

Within our parallel model, we consider two different scenarios. In the first scenario, an infinite stream of jobs submitted by infinitely-many users that follows a Poisson process with intensity $\lambda$ joins the network selecting exactly one queue to minimize their individual response time. The state of the queues is not known to the users. Since each user carries an infinitesimally-small amount of traffic, the long-term average system dynamics is modeled by the *Wardrop equilibrium* [8, 20, 34], whose existence and uniqueness is ensured by the strict convexity of response times. The response time achievable in this scenario is thus denoted by $R^{We}$ and is obtained uniquely by Wardrop's principles [28]. In the second scenario, jobs are sent to queues by a central broker having the goal of minimizing the overall response time. We refer to this situation as *social optimization*, and the response time achieved in this scenario is denoted by $R^{Opt}$.

The two scenarios above reflect the conflicting situations where an infinite stream of non-cooperative jobs moves in an infrastructure with neither control nor shared information with respect to the case where a centralized object dictates the dynamics of the system to maximize the social welfare.

Given that no shared information is available in a fully-decentralized system before the arrivals of jobs, in a Wardrop equilibrium jobs make their decisions independently of the others according to some probabilistic law that is identical for each job (users have the same objective or utility function). As a consequence, existing works apply to our model in this case; see [8, 20, 34] for formulae and bounds on $R^{We}$.

On the other hand, the fact that in a Wardrop equilibrium jobs make their decisions in a i.i.d. manner does not imply that the optimal brokering strategy satisfies such condition. In fact, a real-world broker knows where previous jobs have been sent and it can clearly exploit this information to improve performance. Our notion of social optimum differs from the one considered in existing approaches in the sense that *we let the broker operate with the memory of its previous decisions*. The quantity $R^{Opt}$ is thus interpreted as in (4). As a consequence, the job arrival processes to the queues do not preserve the statistical properties of the job arrival process at the broker. For instance, with a broker implementing Round-Robin and assuming that the (external) stream of users follows a Poisson process, the inter-arrival times at each queue are Gamma distributed instead of exponentials as in the Wardrop equilibrium. This diversity impacts on the response times of all queues. In other words, the congestion functions of the queues in the centralized and decentralized scenario become different!

Within fixed inter-arrival and service time distributions, we measure the inefficiency of the Wardrop equilibrium with respect to the social optimum by means of the revisited PoA, which we define as

$$PoA(N) \overset{\text{def}}{=} \frac{R^{We}(N)}{R^{Opt}(N)} \geq 1. \tag{35}$$

Evidently, large values of $PoA$ indicate that the impact of a centralized control drastically improves the performance of the system, and vice versa. On the other hand, a centralized system is less scalable and reliable than a distributed one because it has a single point of failure.

The following connection between our revisited PoA and the PoA of a probabilistic broker immediately follows by multiplying the numerator and the denominator of (35) by $R^{Opt}_{Bernoulli}$.

**Proposition 1**

$$PoA(N) = PoA_{Bernoulli}(N) PoF(N). \tag{36}$$

Proposition 1 characterizes our revisited notion of the PoA and shows that the multiplicative factor $PoF$ should be taken into account by a Bernoulli analysis of the PoA to understand the impact of having memory in the broker.

Given that $PoA_{Bernoulli}(N)$ is well-understood, we can now apply the results on the PoF presented in Section 4 to obtain bounds on the revisited PoA.

In the case of exponentially distributed service times, using Theorem 6 and the fact that $PoA_{Bernoulli}(N) \leq N$ [20], we obtain

$$PoA(N) \leq 2N, \tag{37}$$

which is exact in heavy-traffic.

In the case of identical, exponentially distributed service times, fixing the load $\rho$ we obtain

$$PoA(N,\rho) = PoF(N,\rho) \leq \frac{1+\rho\mathcal{W}(-\exp(-1/\rho)/\rho)}{1-\rho}, \tag{38}$$

because $PoA_{Bernoulli}(N,\rho) = 1$ (see [20]), for any $\rho$ and $N$.

In the case of service times having a general distribution, the new PoA is arbitrarily large even when $N$ is fixed by means of (31).

# 7 Conclusions

In a network of parallel and non-observable queues, we have derived lower bounds on the minimum mean response time that allowed us to establish the quasi-optimality of a new and efficient routing strategy. We showed that the mean response time achieved by a broker dispatching jobs to queues according to our billiard sequence is very close to our lower bound (the error is less than 1%). Then, we have studied the maximum added-value of letting the broker operate with the memory of its previous decisions, i.e., the PoF. Such value is negligible if the coefficients of variation of the service time distributions are large and, contrariwise, it is unbounded. Assuming exponential service times, the PoF is always bounded from above by two, which is tight in heavy-traffic, and independent of the network heterogeneity and size. Finally, we have revisited the PoA of our system giving a more realistic definition of system inefficiency. The revisited PoA is bounded from above by the product of the PoA with a memoryless broker and the PoF. Our new vision of the PoA can be naturally applied to different games.

Most of the results presented in this paper can be generalized to networks with different topologies and more general arrival processes provided that they are stationary. An important extension of our analysis is when the broker takes into account also the size of the incoming jobs, provided that it is known, or when multiple classes of jobs enter the system.

# References

[1] I. Al-Azzoni and D. G. Down. Dynamic scheduling for heterogeneous desktop grids. In *GRID '08: Proceedings of the 2008 9th IEEE/ACM International Conference on Grid Computing*, pages 136–143, Washington, DC, USA, 2008. IEEE Computer Society.

[2] E. Altman, B. Gaujal, and A. Hordijk. *Discrete-Event Control of Stochastic Networks: Multimodularity and Regularity.* Number 1829 in LNM. Springer-Verlag, 2003.

[3] J. Anselmi and B. Gaujal. Optimal routing in parallel, non-observable queues and the price of anarchy revisited. In *22nd International Teletraffic Congress, ITC-22*, pages 1–8, 2010.

[4] J. Anselmi and B. Gaujal. The price of anarchy in parallel queues revisited. *SIGMETRICS Perform. Eval. Rev.*, 38(1):353–354, 2010.

[5] Y. Arian and Y. Levy. Algorithms for generalized round robin routing. *Oper. Res. Lett.*, 12:313–319, 1992.

[6] U. Ayesta, O. Brun, and B. J. Prabhu. Price of anarchy in non-cooperative load balancing. In *INFOCOM'10*, pages 436–440, Piscataway, NJ, USA, 2010. IEEE Press.

[7] A. Bar-Noy, R. Bhatia, J. S. Naor, and B. Schieber. Minimizing service and operation costs of periodic scheduling. *Math. Oper. Res.*, 27(3):518–544, 2002.

[8] C. H. Bell and S. Stidham. Individual versus social optimization in the allocation of customers to alternative servers. *Management Science*, 29(7):831–839, 1983.

[9] V. Berten and B. Gaujal. Brokering strategies in computational grids using stochastic prediction models. *Parallel Computing*, 33(4-5):238–249, 2007.

[10] U. N. Bhat. *An Introduction to Queueing Theory: Modeling and Analysis in Applications.* Birkhauser Verlag, 2008.

[11] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, March 2004.

[12] H.-L. Chen, J. R. Marden, and A. Wierman. The effect of local scheduling in load balancing designs. *SIGMETRICS Perf. Eval. Rev.*, 36(2):110–112, 2008.

[13] M. B. Combé and O. J. Boxma. Optimization of static traffic allocation policies. *Theor. Comput. Sci.*, 125(1):17–43, 1994.

[14] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jerey, and D. E. Knuth. On the lambert w function. *Adv. Comput. Math*, pages 329–359, 1996.

[15] B. Gaujal, E. Hyon, and A. Jean-Marie. Optimal routing in two parallel queues with exponential service times. *Discrete Event Dynamic Systems*, 16:71–107, January 2006.

[16] X. Guo, Y. Lu, and M. S. Squillante. Optimal probabilistic routing in distributed parallel queues. *SIGMETRICS Perf. Eval. Review*, 32(2):53–54, 2004.

[17] B. Hajek. The proof of a folk theorem on queueing delay with applications to routing in networks. *J. ACM*, 30:834–851, 1983.

[18] B. Hajek. Extremal splitting of point processes. *Math. Oper. Res.*, 10:543–556, 1986.

[19] M. Harchol-Balter. Task assignment with unknown duration. *J. ACM*, 49(2):260–288, 2002.

[20] M. Haviv and T. Roughgarden. The price of anarchy in an exponential multi-server. *Op. Res. Lett.*, 35(4):421–426, 2007.

[21] A. Hordijk and D. van der Laan. Periodic routing to parallel queues and billiard sequences. *Mathematical Methods of Operations Research*, 59(2):173–192, 2004.

[22] B. Javadi, D. Kondo, J.-M. Vincent, and D. P. Anderson. Mining for statistical models of availability in large-scale distributed systems: An empirical study of seti@home. *IEEE Transaction on Parallel and Distributed Systems, (to appear)*, 2010.

[23] J. F. C. Kingman. Some inequalities for the queue gi/g/1. *Biometrika*, 49(3/4):315–324, 1962.

[24] L. Kleinrock. *Queueing Systems, Volume 2: Computer Applications.* Wiley, 1976.

[25] E. Koutsoupias and C. Papadimitriou. Worst-case equilibria. In *STACS*, volume 1563 of *Proceedings of the 16th Annual Symposium on Theoretical Aspects of Computer Science. Springer Verlag, Berlin*, pages 404–413, January 1999.

[26] M. Mitzenmacher. How useful is old information? *IEEE Trans. Parallel Distrib. Syst.*, 11(1):6–20, 2000.

[27] J. Niño-Mora. Dynamic allocation indices for restless projects and queueing admission control: a polyhedral approach. *Math. Program.*, 93(3):361–413, 2002.

[28] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani. *Algorithmic Game Theory.* Cambridge University Press, New York, NY, USA, 2007.

[29] T. Roughgarden. The price of anarchy is independent of the network topology. *J. Comput. Syst. Sci.*, 67:341–364, September 2003.

[30] B. Schroeder and M. Harchol-Balter. Evaluation of task assignment policies for supercomputing servers: The case for load unbalancing and fairness. *Cluster Computing*, 7(2):151–161, 2004.

[31] T. Thanalapati and S. Dandamudi. An efficient adaptive scheduling scheme for distributed memory multicomputers. *IEEE Trans. Parallel Distrib. Syst.*, 12(7):758–768, 2001.

[32] J. Walrand. *An Introduction to Queueing Networks.* Prentice-Hall, 1988.

[33] R. Weber and G. Weiss. On an index policy for restless bandits. *Journal of Applied Probability*, 27(3):637–648, 1990.

[34] T. Wu and D. Starobinski. A comparative analysis of server selection in content replication networks. *IEEE/ACM Trans. Netw.*, 16(6):1461–1474, 2008.

# Appendix

## Proof of Theorem 1

First, note that $Q_n^i$ is only defined on integer points in $\{0,1\}^n$ and can be extended to $[0,1]^n$ by linear interpolation over simplexes, defined by the multimodular base $v_k = (0,\dots,0,-1,+1,0,\dots,0)$ (see [18]). Once this is done, it can been shown that $Q_n^i$ has the following properties [2]:

$(P_1)$ $Q_n^i$ is convex.

$(P_2)$ for all $m$, $Q_n^i(a_1^i,\dots,a_n^i) = Q_{n+m}^i(0,\dots,0,a_1^i,\dots,a_n^i)$.

$(P_3)$ for all $m < n$, $Q_n^i(a_1^i,\dots,a_n^i) \geq Q_m^i(a_{n-m+1}^i,\dots,a_n^i)$.

The last two properties are easy to prove because they are also true on each trajectory: Point $(P_2)$ is true because the system is initially empty and time-homogeneous, while the third property $(P_3)$ comes from the fact that adding a job in the past increases the load at time 0.

However, the first item $(P_1)$ is only true for the expected queue length.

For any $0 < \delta < 1$, let $p_\delta^i \overset{\text{def}}{=} (1-\delta)\sum_{k=1}^{\infty}\delta^{k-1}a_k^i$, that exists since all $a_n^i$ are bounded. By definition of $p_\delta^i$, $\sum_{i=1}^{N} p_\delta^i = 1$. Therefore, the set of limit points of $(p_\delta^1,\dots,p_\delta^N)$, when $\delta \to 1$, only contains vectors that sum to one.

Using these definitions and the properties (P1)-(P3) above, one has for any $M \in \mathbb{N}$,

$$\begin{aligned}
&\sum_{n=1}^{\infty}(1-\delta)\delta^{n-1}a_n^i Q_n^i(a_1^i,\dots,a_n^i)\\
&\geq \sum_{n=1}^{M}(1-\delta)\delta^{n-1}a_n^i Q_M^i(0,\dots,0,a_1^i,\dots,a_n^i) + \sum_{n=M+1}^{\infty}(1-\delta)\delta^{n-1}a_n^i Q_M^i(a_{n-M+1}^i,\dots,a_n^i)\\
&\geq \left(\sum_{n=1}^{\infty}(1-\delta)\delta^{n-1}a_n^i\right)Q_M^i\left(\sum_{n=1}^{M}(1-\delta)\delta^{n-1}(0,\dots,0,a_1^i,\dots,a_n^i) + \sum_{n=M+1}^{\infty}(1-\delta)\delta^{n-1}(a_{n-M+1}^i,\dots,a_n^i)\right)\\
&= p_\delta^i Q_M^i(\delta^M p_\delta^i, \delta^{M-1}p_\delta^i,\dots,p_\delta^i).
\end{aligned} \tag{39}$$

By definition of the mean response time,

$$R(a) = \limsup_{K\to\infty}\frac{1}{K}\sum_{n=1}^{K}(a_n^1 Q_n^1(a_1^1,\dots,a_n^1) + \cdots + a_n^N Q_n^N(a_1^N,\dots,a_n^N)). \tag{40}$$

Using the well-known fact that the Cesaro limit is always larger than the discounted limit with a discount going to one,

$$\begin{aligned}
R(a) &\geq \limsup_{\delta\to 1}(1-\delta)\sum_{n=1}^{\infty}\delta^{n-1}\big(a_n^1 Q_n^1(a_1^1\dots a_n^1) + \cdots + a_n^N Q_n^N(a_1^N\dots a_n^N)\big)\\
&\geq \limsup_{\delta\to 1}\sum_{i=1}^{N} p_\delta^i Q_M^i(\delta^M p_\delta^i, \delta^{M-1}p_\delta^i,\dots,p_\delta^i) \quad \text{(from (39))}\\
&\geq \inf_{p_1+\cdots+p_N=1}\sum_{i=1}^{N} p_i Q_M^i(p_i,p_i,\dots,p_i).
\end{aligned}$$

In the multimodular base introduced at the beginning of the proof, the point $(p_i, p_i, \ldots, p_i)$ belongs to the simplex of $\mathbb{R}^M$ whose extreme points are the $M+1$ prefixes of length $M$ of the Sturmian sequences $\{\alpha(p_i,\theta)\}_{0\leq\theta\leq 1}$. Since $Q^i_M$ is linear on each simplex, one gets $Q^i_M(p_i, p_i, \ldots, p_i) = \mathbb{E}_\theta Q^i_M(\alpha_1(p_i,\theta), \ldots, \alpha_M(p_i,\theta))$, where the expectation is taken over the uniform distribution on $0 \leq \theta \leq 1$.

Finally, by letting $M$ go to infinity, it can been shown (see Lemmas 7 and 8 in [2]) that one gets, for the Cesaro limit of any subsequence $\ell_n$,

$$\lim_{M\to\infty} Q^i_M(p_i, p_i, \ldots, p_i) = \lim_{m\to\infty} \frac{1}{m} \sum_{n=1}^{m} Q^i_{\ell_n}(\alpha_1(p_i,0)_1, \ldots, \alpha_{\ell_n}(p_i,0)). \tag{41}$$

This concludes the proof by noticing that this limit is the mean response time in queue $i$ under the admission sequence $\alpha(p_i, 0)$ (see Equation (6)).

## Proof of Corollary 2

Let us consider the functions $\overline{Q}^i_n \overset{\text{def}}{=} \phi_i \circ Q^i_n$ as defined in Section 2.2. It should be clear that these new functions satisfy the properties $(P_1), (P_2), (P_3)$ defined in the proof of Theorem 1.

Now, the remaining of the proof of Theorem 1 can be carried unchanged by replacing the functions $Q^i_n$ by $\overline{Q}^i_n$ everywhere. We get

$$\overline{R}(a) \geq \inf_{p_1+\cdots+p_N=1} \big(p_1\overline{R}_1(\alpha(p_1,0)) + \cdots + p_N\overline{R}_N(\alpha(p_N,0))\big).$$

Again, consider the inter-arrival times $\tau_1, \cdots, \tau_n, \cdots$ of a Sturmian routing sequence $\alpha(p_i, 0)$ and the inter-arrival times $T_1, \cdots T_n \cdots$ of a Gamma distribution with parameters $p_i$ and $N\lambda$. As mentioned in Section 2.1, these two inter-arrival processes compare for the convex ordering of random sequences: $(\tau_1, \cdots, \tau_n) \geq_{cx} (T_1, \cdots, T_n)$.

Now, since the mean response times $R_i$ are convex increasing functions of the input process, the same is true for its composition with $\phi_i$, so that by definition of the convex ordering,

$$\overline{R}_i(\alpha(p_i,0))\big) \geq \overline{R}_i^{\Gamma(1/\pi_i, N\lambda)/GI/1}.$$

Combining this with Equation (7) yields

$$\overline{R}^{Opt} \geq \inf_{\substack{\pi_1,\ldots,\pi_N\geq 0:\\ \pi_1+\cdots+\pi_N=1}} \sum_{i=1}^{N} \pi_i \overline{R}_i^{\Gamma(1/\pi_i, N\lambda)/GI/1}. \tag{42}$$

## Proof of Theorem 2

Applying the standard analysis of $G/M/1$ queues, e.g., [10], it follows that

$$R^{\Gamma(N/\beta_i, N\lambda)/M/1} = \frac{1}{\mu_i(1-x)} \tag{43}$$

where $x$ is the least positive solution of

$$z = \left(\frac{N\rho_i}{N\rho_i + 1 - z}\right)^{N/\beta_i}, \tag{44}$$

and $\rho_i = \lambda/\mu_i$, which we rewrite as

$$z\left(1 + \beta_i \frac{1-z}{NU_i}\right)^{N/\beta_i} = 1, \tag{45}$$

Assuming $a = (1-z)/U_i$, with a Maclaurin series expansion in $\beta_i/N$ we obtain

$$\begin{aligned}
\exp(-a)\left(1+\beta_i\frac{a}{N}\right)^{N/\beta_i} \;=\;& 1-\tfrac{1}{2}a^2\frac{\beta_i}{N} \\
&+\left(\tfrac{1}{3}a^3+\tfrac{1}{8}a^4\right)\frac{\beta_i^2}{N^2} \\
&-\left(\tfrac{1}{4}a^4+\tfrac{1}{6}a^5+\tfrac{1}{48}a^6\right)\frac{\beta_i^3}{N^3} \\
&+O(\beta_i^4/N^4)
\end{aligned} \tag{46}$$

where the coefficient of $(\beta_i/N)^{-i}$, $i \geq 0$, alternates because $a > 0$. Observing that $\sigma$ and $\sigma^+$ refer to truncations of the alternating series above, we must have $\sigma \leq x \leq \sigma^+$, which implies (16).

## Proof of Lemma 1

Let us refer to (14) through $f(z) = z\exp(-\frac{z}{U_i}) - \exp(-\frac{1}{U_i})$. One can easily check that:
i) $f(U_i) > 0$,
ii) $f'(z) = \exp(-\frac{z}{U_i})(1 - z/U_i) = 0$ if and only if $z = U_i$,
iii) $f''(z) = -\frac{1}{U_i}\exp(-\frac{z}{U_i})(2 + z/U_i) < 0$ $(z \geq 0)$, i.e., $f(z)$ is concave, and
iv) (14) has only two (positive) real roots (when $0 \leq U_i \leq 1$) where the largest one is $z = 1$.

Taking into account facts i)–iv), the statement can be proven by showing that $f(z)$ is non-negative when $z = U_i^2$. Substituting $U_i^2$ in $f$, we thus must have $U_i^2\exp(-\frac{U_i^2}{U_i}) \geq \exp(-\frac{1}{U_i})$, i.e., $\ln U_i^2 - U_i \geq -\frac{1}{U_i}$, i.e., (rearranging the terms)

$$h(U_i) \overset{\text{def}}{=} 2U_i \ln U_i - U_i^2 + 1 \geq 0. \tag{47}$$

Since $h(0) = 1$, $h(1) = 0$ and $h$ is decreasing (note that $\frac{1}{2}\frac{\mathrm{d}h(U_i)}{\mathrm{d}U_i} = \ln U_i + 1 - U_i < 0, \forall U_i \in [0,1)$, which easily follows by the change of variable $U_i = 1 - x_i$ and expanding the logarithm in Taylor series), we conclude that $h(U_i)$ must be strictly positive when $U_i \in [0,1)$.

## Proof of Theorem 4

Using Formula (17), the mathematical program (21) can be written as follows

$$\begin{aligned}
GB^-(N) = \min \quad & \frac{1}{\lambda N}\sum_{i=1}^{N}\frac{U_i}{1-\sigma_i} \\
\text{s.t.} \quad & \exp(\tfrac{\sigma_i - 1}{U_i}) = \sigma_i, \ \forall i \\
& \textstyle\sum_{i=1}^{N}\frac{\mu_i}{\lambda}U_i = N \\
& 0 \leq \sigma_i \leq U_i, \ \forall i.
\end{aligned} \tag{48}$$

Taking the logarithm, one obtains $\frac{U_i}{1-\sigma_i} = -\frac{1}{\ln \sigma_i}$, $\forall i$, that yields, together with the change of variable $\overline{\sigma}_i = 1 - \sigma_i$, the equivalent formulation

$$\begin{aligned}
GB^-(N) = \min \quad & \frac{1}{\lambda N}\sum_{i=1}^{N} -\frac{1}{\ln(1-\overline{\sigma}_i)} \\
\text{s.t.} \quad & \\
& \textstyle\sum_{i=1}^{N}\frac{\mu_i}{\lambda}\frac{-\overline{\sigma}_i}{\ln(1-\overline{\sigma}_i)} = N \\
& 0 < 1 - \overline{\sigma}_i \leq \frac{-\overline{\sigma}_i}{\ln(1-\overline{\sigma}_i)}, \ \forall i,
\end{aligned} \tag{49}$$

In heavy-traffic conditions $\overline{\sigma}_i$ must be small. This suggests to adopt, in (49), the following Laurent expansion $\frac{1}{\ln(1-x)} = -\frac{1}{x} + 0.5 + O(x)$, obtaining

$$\begin{aligned} GB^{-}(N) = \min \quad & \frac{1}{\lambda N} \sum_{i=1}^{N} \left( \frac{1}{\overline{\sigma}_i} - 0.5 \right) \\ \text{s.t.} \quad & \textstyle\sum_{i=1}^{N} \frac{\mu_i}{\lambda} \left(1 - 0.5\overline{\sigma}_i\right) = N \\ & 0 \leq \overline{\sigma}_i < 1, \ \forall i. \end{aligned} \tag{50}$$

The objective function in problem (50) is convex and differentiable over the feasible region. Therefore, a unique optimum exists and can be found by Lagrangian duality. Let us initially ignore the boundary constraints $0 \leq \overline{\sigma}_i < 1, \ \forall i$. The Lagrangian of (50) becomes

$$\mathcal{L}(\overline{\sigma}, x) = \frac{1}{\lambda N} \sum_{i=1}^{N} \left( \frac{1}{\overline{\sigma}_i} - 0.5 \right) + Nx - x \sum_{i=1}^{N} \frac{\mu_i}{\lambda} \left(1 - 0.5\overline{\sigma}_i\right) \tag{51}$$

and, differentiating with respect to $\overline{\sigma}_j$, we must have

$$-\frac{1}{N} \frac{1}{\overline{\sigma}_j^2} + x\frac{\mu_j}{2} = 0, \tag{52}$$

i.e.,

$$\overline{\sigma}_j = \sqrt{\frac{2}{\mu_j x N}} \tag{53}$$

because $\overline{\sigma}_j$ must be positive. Substituting in (51), we obtain the dual function

$$g(x) = \inf_{\overline{\sigma}} \mathcal{L}(\overline{\sigma}, x) = \frac{1}{\lambda N} \sum_{i=1}^{N} \left( x^{\frac{1}{2}} \sqrt{\frac{\mu_i N}{2}} - 0.5 \right) + x - \frac{\mu_i}{\lambda} \left( x - 0.5 \sqrt{\frac{2}{\mu_i N}} x^{\frac{1}{2}} \right) \tag{54}$$

and strong duality ensures that $GB^{-}(N) = \max_{x \in \mathbb{R}} g(x)$. Since

$$\begin{aligned} \frac{\mathrm{d}g(x)}{\mathrm{d}x} &= \sum_{i=1}^{N} \frac{1}{2\lambda} x^{-\frac{1}{2}} \sqrt{\frac{\mu_i}{2N}} + 1 - \frac{\mu_i}{\lambda} \left( 1 - \frac{1}{4} \sqrt{\frac{2}{\mu_i N}} x^{-\frac{1}{2}} \right) \\ &= x^{-\frac{1}{2}} \left[ \sum_{i=1}^{N} \frac{1}{2\lambda} \sqrt{\frac{\mu_i}{2N}} + \frac{\mu_i}{\lambda} \frac{1}{4} \sqrt{\frac{2}{\mu_i N}} \right] + N - \sum_{i=1}^{N} \frac{\mu_i}{\lambda}, \end{aligned} \tag{55}$$

after some algebra the maximizing $x$ becomes

$$x = \frac{1}{2N} \left[ \frac{\sum_{i=1}^{N} \sqrt{\mu_i}}{\sum_{i=1}^{N} \mu_i - \lambda N} \right]^2 \tag{56}$$

Substituting (56) in (53), the maximizing $\overline{\sigma}$ becomes

$$\overline{\sigma}_j = \frac{2}{\sqrt{\mu_j}} \frac{\sum_{i=1}^{N} \mu_i - \lambda N}{\sum_{i=1}^{N} \sqrt{\mu_i}}. \tag{57}$$

Now, in heavy-traffic we must have $\sum_{i=1}^{N} \mu_i - \lambda N \to 0$ (from above), meaning that $\overline{\sigma}_j \in (0, 1)$, i.e., the boundary constraints $0 \leq \overline{\sigma}_i < 1, \ \forall i$ are satisfied, and, thus, (57) optimizes the objective function of (50). The expression (22) and (23) follows by substitution of (57) in $g(x)$ and in $U_i = 1 - 0.5\overline{\sigma}_i$, respectively.

## Proof of Theorem 5

Within the $GB(N)$ bound (8), the mean response time of queue $i$ is $R^{\Gamma(N/\beta_i,N\lambda)/M/1} = 1/(\mu_i(1-s_i))$, where $s_i$ is the least positive root of

$$z\left(1+\frac{1-z}{N\rho_i}\right)^{N/\beta_i} = 1. \tag{58}$$

After a Maclaurin expansion in $\beta_i/N$ (see (46)) and taking the first two expansion terms (yielding (14) and (15)), we must have $GB(N)-GB^{-}(N) > GB^{+}(N)-GB(N)$. The statement follows by observing that $GB(N) \leq R^{Opt}$.

## Proof of Theorem 3

Given that the Hessian of the objective function is diagonal and the constraints are linear, to prove the convexity of (21), it suffices to show that $U_i/(1-s_i)$ is convex in $U_i$. Let $g = g(U_i) = -\exp(-1/U_i)/U_i$. From the expressions of the derivatives of the Lambert $\mathcal{W}$ function [14], we obtain

$$\frac{\mathrm{d}}{\mathrm{d}g}\mathcal{W}(g) = \frac{\mathcal{W}(g)}{g(1+\mathcal{W}(g))}, \tag{59}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}g^2}\mathcal{W}(g) = -\frac{\exp(-2\mathcal{W}(g))(g+2)}{(1+\mathcal{W}(g))^3} \tag{60}$$

and substituting $g = \mathcal{W}(g)\exp(\mathcal{W}(g))$ in the latter, which follows by the definition of the $\mathcal{W}$ function, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}U_i}\frac{U_i}{1+U_i\mathcal{W}(g)} = \frac{1}{(1+U\mathcal{W}(g))(1+\mathcal{W}(g))} \tag{61}$$

and

$$\frac{\mathrm{d}^2}{\mathrm{d}U_i^2}\frac{U_i}{1+U_i\mathcal{W}(g)} = \frac{-\mathcal{W}(g)}{(1+\mathcal{W}(g))^3U_i^2} \tag{62}$$

which is strictly positive because $0 < -\mathcal{W}(g) < 1$, for $0 < U < 1$.

## Proof of Corollary 5

Let $U_i = \lambda\beta_i/\mu_i$ and denote by $GB_H(N)$ the optimum of optimization problem (11). Using the heavy-traffic formula for $GI/GI/1$ queue (10), we obtain for $\sum_i \beta_i = N, \beta_i \geq 0$

$$GB_H(N) \quad \geq \min_{\beta_1,\ldots,\beta_N} \sum_{i=1}^{N} \frac{\beta_i}{N}\left[\lambda\beta_i\frac{s_i^2}{2(1-U_i)} + \frac{1}{2\mu_i}\right] - \frac{1}{2\lambda N}. \tag{63}$$

In the remainder of the proof and with a slight abuse of notation, we assume that the variables $\beta_1,\ldots,\beta_N$ yield the minimum in (63). With respect to such $\beta_i$'s , in the Bernoulli case we obtain

$$\begin{aligned} R^{Opt}_{Bernoulli} &\leq \sum_{i=1}^{N}\frac{\beta_i}{N}R_i^{M(\lambda\beta_i)/GI/1} \leq \sum_{i=1}^{N}\frac{\beta_i}{N}\left(\lambda\beta_i\frac{1/(\beta_i^2\lambda^2)+s_i^2}{2(1-U_i)} + \frac{1}{\mu_i}\right) \\ &= \sum_{i=1}^{N}\frac{\beta_i}{N}\left(\lambda\beta_i\frac{s_i^2}{2(1-U_i)} + \frac{1}{2\mu_i}\right) - \frac{1}{2\lambda N} + \sum_{i=1}^{N}\frac{\beta_i}{N}\left(\frac{1}{2\mu_i} + \lambda\beta_i\frac{1/(\beta_i^2\lambda^2)}{2(1-U_i)}\right) + \frac{1}{2\lambda N} \\ &= GB_H(N) + \sum_{i=1}^{N}\frac{\beta_i}{N}\left(\frac{1}{2\mu_i} + \frac{1/(\beta_i\lambda)}{2(1-U_i)}\right) + \frac{1}{2\lambda N}, \end{aligned} \tag{64}$$

where the first inequality follows from the fact that the $\beta_i$'s above are not necessarily optimal for the Bernoulli broker ($M(\lambda\beta_i)$ denotes a Poisson process with intensity $\lambda\beta_i$), and the second one by using a heavy-traffic upper bound of the $GI/GI/1$ queue [24]. Taking the ratio of the above expressions, we obtain

$$PoF \leq \frac{R^{Opt}_{Bernoulli}}{GB_H(N)} \leq 1 + \frac{\sum_{i=1}^N \frac{\beta_i}{N}\left(\frac{1/(\beta_i\lambda)}{2(1-U_i)} + \frac{1}{2\mu_i}\right) + \frac{1}{2\lambda N}}{\sum_{i=1}^N \frac{\beta_i}{N}\left[\lambda\beta_i \frac{s_i^2}{2(1-U_i)} + \frac{1}{2\mu_i}\right] - \frac{1}{2\lambda N}} \tag{65}$$

Now, keeping fixed the $\mu_i$ and increasing $\lambda$ such that $\rho \to 1$, we must have $\beta_i/N \to \mu_i/\sum_{j=1}^N \mu_j$, $\forall i$. Given that $(1-U_i)^{-1} \to \infty$, for each $i$, as $\rho \to 1$, we have

$$\begin{aligned} \lim_{\rho\to 1^-} PoF(\rho) \leq &\; 1 + \lim_{\rho\to 1^-} \frac{\sum_{i=1}^N [1-U_i]^{-1}}{\sum_{i=1}^N \lambda^2\beta_i^2 s_i^2 [1-U_i]^{-1}} \\ \leq &\; 1 + \lim_{\rho\to 1^-} (\lambda^2 \min_{j=1}^N \beta_j^2 s_j^2)^{-1} \\ = &\; 1 + (\min_{j=1}^N \mu_j^2 s_j^2)^{-1}. \end{aligned} \tag{66}$$

## Proof of Theorem 6

Let

$$f_1(\mathbf{U}) = \frac{1}{\lambda N}\sum_{i=1}^N \frac{1}{2}\frac{U_i}{1-U_i} \tag{67}$$

for $\mathbf{U} \in \mathbb{R}^N : \sum_{i=1}^N \frac{\mu_i}{\lambda} U_i = N,\ 0 \leq U_i < 1,\ \forall i$ and

$$f_2(\mathbf{U}) = \frac{1}{\lambda N}\sum_{i=1}^N \frac{U_i}{1+\mathcal{W}(g(U_i))U_i} \tag{68}$$

for $g(U_i) = -\exp(-1/U_i)/U_i$ and $\mathbf{U} \in \mathbb{R}^N : \sum_{i=1}^N \frac{\mu_i}{\lambda} U_i = N,\ 0 \leq U_i < 1,\ \forall i$. To prove the theorem, we show that $f_2(\mathbf{U}) \geq f_1(\mathbf{U})$, $\forall \mathbf{U}$, which is true if

$$1 + \mathcal{W}(g(U_i))U_i \leq 2(1-U_i),\ \forall i, \tag{69}$$

Since $-\mathcal{W}(g(U_i) \leq U_i$ by Lemma 1, (69) is satisfied because it holds with equality when $U_i = 1$ and strictly when $U_i = 0$, and

$$\frac{\mathrm{d}}{\mathrm{d}U_i}\left[\frac{1}{U_i} - 2 - \mathcal{W}(g(U_i))\right] = -\frac{1}{U_i^2} - \frac{(1-U_i)\mathcal{W}(g(U_i))}{U_i^2(1+\mathcal{W}(g(U_i)))} \tag{70}$$

is always negative.

## Proof of Corollary 8

**Monotonicity**. From the expressions (59) and (60), we have

$$\frac{\mathrm{d}PoA(\rho)}{\mathrm{d}\rho} = \frac{\rho\mathcal{W}(g)^2 + \mathcal{W}(g) + \mathcal{W}(g)\rho^2 + \rho}{\rho(1+\mathcal{W}(g))(1-\mathcal{W}(g))^2} > 0 \tag{71}$$

if and only if $\rho\mathcal{W}(g)(\mathcal{W}(g)+\rho) + \mathcal{W}(g) + \rho > 0$. Lemma 1 implies that $-\mathcal{W}(g) \leq \rho$ and proves that it is increasing in $\rho$.

**Limit as** $\rho \to 1$. Applying L'Hôpital's rule to (71), we obtain

$$\begin{aligned} &\lim_{\rho\to 1}\frac{\mathrm{d}PoA(\rho)}{\mathrm{d}\rho} \\ &= \lim_{\rho\to 1}\frac{\rho\mathcal{W}(g)^2 + \mathcal{W}(g) + \mathcal{W}(g)\rho^2 + \rho}{4(1+\mathcal{W}(g))} \\ &= \lim_{\rho\to 1}\frac{\mathcal{W}(g)^2 + 2\rho\mathcal{W}(g)\mathcal{W}'(g) + \mathcal{W}'(g) + 2\rho\mathcal{W}(g) + \mathcal{W}'(g)\rho^2 + 1}{4\mathcal{W}'(g)} \end{aligned} \tag{72}$$

Let $\mathcal{W}'(g) = \mathrm{d}\mathcal{W}(g(\rho))/\mathrm{d}\rho$. Taking into account (59) and applying again L'Hôpital's rule, we have

$$\begin{aligned}
\lim_{\rho\to 1} \mathcal{W}'(g) = & \lim_{\rho\to 1} \frac{\mathcal{W}(g)}{1+\mathcal{W}(g)} \frac{1-\rho}{\rho^2} \\
= & \lim_{\rho\to 1} \frac{\rho-1}{1+\mathcal{W}(g)} \\
= & \lim_{\rho\to 1} \frac{1}{\mathcal{W}'(g)},
\end{aligned} \tag{73}$$

which holds true if and only if $\lim_{\rho\to 1} \mathcal{W}'(g)$ is 1 or $-1$ (the existence of the limit (73) follows by continuity arguments and the fact that $\mathcal{W}'(g)$ is monotonically decreasing). Since $\mathcal{W}'(g)$ is monotonically decreasing, we conclude that $\lim_{\rho\to 1} \mathcal{W}'(g) = -1$. Substituting this in (72), we obtain zero.

**Limit as** $\rho \to 0$. From (71), we obtain

$$\begin{aligned}
\lim_{\rho\to 0} \frac{\mathrm{d}PoA(\rho)}{\mathrm{d}\rho} & = \lim_{\rho\to 0} \mathcal{W}(g)^2 + \frac{\mathcal{W}(g)}{\rho} + \mathcal{W}(g)\rho + 1 \\
& = 1 + \lim_{\rho\to 0} \mathcal{W}'(g) \\
& = 1 + \lim_{\rho\to 0} \frac{\mathcal{W}(g)}{1+\mathcal{W}(g)} \frac{1-\rho}{\rho^2} \\
& = 1 + \lim_{\rho\to 0} \frac{\mathcal{W}(g)}{\rho^2}
\end{aligned} \tag{74}$$

The Lambert $\mathcal{W}$ function admits the following Maclaurin expansion [14]

$$\mathcal{W}(g) = -\sum_{n\geq 1} \frac{n^{n-1}}{n!}(-g)^n \tag{75}$$

which is convergent $\forall g : |g| \leq 1/e$. The leading term of (75) is $\exp(-1/\rho)/\rho$, when $\rho \to 0$. Therefore, we have

$$\lim_{\rho\to 0} \frac{\mathcal{W}(g)}{\rho^2} = \lim_{\rho\to 0} \frac{\exp(-1/\rho)}{\rho^3} = 0. \tag{76}$$