Optimal Routing in two parallel Queues with exponential service times

Bruno Gaujal, Emmanuel Hyon, Alain Jean-Marie

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Optimal Routing in two parallel Queues with exponential service times

Bruno Gaujal — Emmanuel Hyon — Alain Jean-Marie

N° 5109

Février 2004

THÈME 1

ISSN 0249-6399 ISRN INRIA/RR--5109--FR+ENG

Optimal Routing in two parallel Queues with exponential service times

Bruno Gaujal* , Emmanuel Hyon† , Alain Jean-Marie ‡

Thème 1 —Réseaux et systèmes
Projet TRIO

Rapport de recherche n° 5109 —Février 2004 —28 pages

Abstract: In this paper we investigate the problem of the effective computation of the optimal routing sequence in a queuing system made of two queues with exponential service time in parallel. We first show that the optimal policy (minimizing the expected waiting time) is a Sturmian sequence and we establish several qualitative properties of this policy (monotonicity, continuity, convexity) Then, we propose an algorithm to compute the optimal routing sequence. We address the issues of time complexity as well as numerical stability of this algorithm. We then run an extensive set of experiments which show several interesting features of the optimal policy with apparent discontinuities and a fractal behavior.

Key-words: Open-Loop routing, parallel queues, Sturmian words.

* INRIA LIP, ENS Lyon, 46 allée d'Italie, F-69364 Lyon, France. E-mail Bruno.Gaujal@ens-lyon.fr
† LORIA, 615 route du jardin botanique, BP 101, F-54606 Villers-les-Nancy, France. E-mail: Emmanuel.Hyon@loria.fr.
‡ LIRMM, Université Sciences et Techniques du Languedoc, 161 Rue Ada, F-34392 Montpellier, France. E-mail ajm@lirmm.fr

Unité de recherche INRIA Lorraine
LORIA, Technopôle de Nancy-Brabois, Campus scientifique,
615, rue du Jardin Botanique, BP 101, 54602 Villers-Lès-Nancy (France)
Téléphone : +33 3 83 59 30 00 —Télécopie : +33 3 83 27 83 19

Routage optimal en boucle ouverte dans 2 files en parallèle avec services exponentiels

Résumé : On s'intéresse dans ce papier au calcul effectif de la suite optimale des routages dans un système de deux files en parallèle dont les temps de service sont distribués suivant une loi exponentielle. On montre tout d'abord que la politique optimale (minimisant le temps de séjour moyen) est un mot de Sturm et nous établissons plusieurs propriétés qualitatives (monotonie, continuité, convexité) de cette politique. Nous proposons alors un algorithme qui calcule cette politique. Nous présentons les problèmes de complexité ou de stabilité liés à cet algorithme. Nous donnons ensuite un certain nombre d'expériences numériques qui illustrent certaines caractéristiques intéressantes du comportement de la politique optimale notamment un comportement fractal et des discontinuité apparentes.

Mots-clés : Routage, Files d'attente en parallèle, mots de Sturm.

1 Introduction

The problem of finding the optimal routing policies in networks of queues is of uttermost importance as fas as the new requirements of quality of service are considered. Indeed, the increasing the bandwitdh forever will not be enough to face all the problems (such as long delays). This is why, load balancing policies between streams of customer in a network and admission control policies should be introduced to outperform the *Best effort* policy currently applied. It turns out, that this problem can be treated by the mean of the routing in parallel queues which models efficiency many situations arising in the communication networks. However, the problem of constructing the optimal routing policy is most often unsolved.

We assume here that the input process in the router is Poisson and the services are exponential. Most of the time, the problem is considered with closed-loop (also called dynamic) routing point of view. This means that the decision depends on information in the system. This problem can be expressed as a Markov decision processes and the optimal policies are threshold policies (see [9] for the full information case in two queues and [1] for a recent overview). However, the full information case is often unrealistic in practice and furthermore taking into account the delays or the information losses leads to intractable problems for routers.

An alternative approach is to consider an open-loop control in which the only informations available to the decision maker are those available at time 0. For an infinite number of customers, an algorithm to compute the optimal policy when both services and queues are deterministic can be found in [6] for parallel queues when the number of distinct services is two, in [20] for more than two distinct services and in [7] for two parallel networks of tandem queues. When the number of customers is finite and both services and inter arrivals are exponential an algorithm using dynamical programming methods is provided in [17]. But in the general case with infinite number of customers only approximations of the optimal policy are given [5]. It must be added that the complexity of the problem is not only due to difficulties related to the control, but also to performance evaluation issues. This is the reason why the computational part of the problem is the object of special treatments here.

In this work, we consider a system made of two parallel queues denoted by Q_1 and Q_2 with exponential service times. The problem addressed is to assign, the infinite arriving customers in the system, in one of the two queue. Our aim is to minimize the average waiting time or the average sojourn time of the customers.

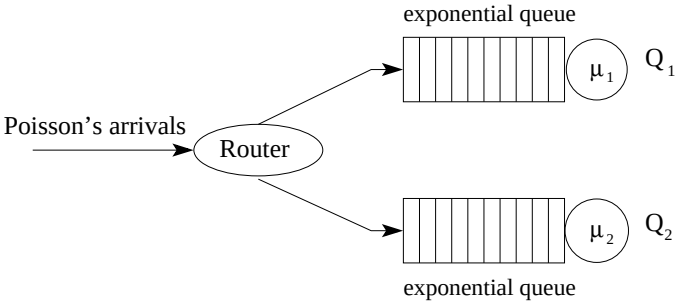

Figure 1: Model

The paper is structured as follows, in Section 2 we describe the input process induced by the routing and we explain how to compute the performances associated with the routing protocol. Section 3 is dedicated to computational issues of the expected performances and presents some improvements of the computations. We also discuss the comparisons with standard methods to compute the stationary distributions. In Section 4 the effective computation of the optimal is presented. Numerous numerical experiments are made and some heuristics which approximates the optimal policy as well as Bernoulli routing are detailed and compared.

2 Input process and routing sequences

2.1 Construction of the Input Process

In this section we will introduce more precisely the input process of the system as well as the routing protocol.

We consider an arrival process $\mathcal{A}$ which is stationary and ergodic of intensity λ. The two queues have a FIFO discipline and the service times in Q_1 and Q_2 are stationary ergodic sequences with rates μ_1 and μ_2 respectively.

The routing policy is determined by an infinite binary sequence which will be called a *routing word* is the following. The routing works as follows. Suppose that m is a word over the alphabet $\{0,1\}$. If the ith letter of m : $m(i)$ is a one then the customer is sent in Q_1 otherwise the customer is sent to Q_2. Henceforth, the input processes in each queue are sampled processes of the global input process.

Starting from the routing word m we can build that the *sampling words* in each queue called m_1 and m_2 respectively. The letters in each of these words are a one when the customer is admitted and zero otherwise. It should be clear that $m_1 = m$ and that m_2 is the complementary word of m.

We call $(\mathcal{A}|m_j, \sigma_j|m_j)$ the marked point process which is the input process in Q_j when the sampling is made according to m_j.

Definition 1 (Notations related to words). *Let m be a word (finite or not), we denote by $m_{[n]}$ the prefix of length n of m. If the word if finite $|m|$ is its size while $|m|_1$ and $|m|_0$ are the number of ones and the number of zeros of m respectively. The slope of a finite word is the quantity $|m|_1/|m|$ and if m is an infinite word the slope is the limit when n goes to infinity of*

$$\sup_n \frac{|m_{[n]}|_1}{n}.$$

In the rest of the paper the slope of a word m is denoted by $\mathfrak{s}(m)$.

Let m be an infinite word, we call shift (of size k) of m the word $\mathcal{S}_k(m)$ such that the nth letter of $\mathcal{S}_k(m)$ is $\big(\mathcal{S}_k(m)\big)(n) = m(n+k)$, where k is any non-negative integer number.

Definition 2. *We denote by $T(n)$ the epoch of the nth arrival in the whole system (i.e. before any routing), and $\tau(n) \stackrel{def}{=} T(n+1) - T(n)$ the nth inter-arrival between two customers. We recall that the sequence $\tau(n)_{n\in\mathbb{N}}$ is a stationary and ergodic sequence.*

We denote by $\sigma_j(n)$ the nth service in Q_j. We define by $T_j(n)$ the epoch of the nth arrival in the queue Q_j, as well we define by $\tau_j(n) \stackrel{def}{=} T_j(n+1) - T_j(n)$ the nth inter-arrival. By assumption, the processes σ_j and τ_j are independent.

The point process $\mathcal{A}|m_j$ may not be stationary. This depends on the sampling word m_j. However a stationary ergodic point process with the same properties as $\mathcal{A}|m_j$ can always be built.

Lemma 3.
-i) Let m_j be periodic of period ℓ and slope α. Let U be a random variable uniformly distributed on $\{1,\cdots,\ell\}$, the process $\big(\mathcal{A}|\mathcal{S}_U(m_j), \sigma_j|\mathcal{S}_U(m_j)\big)$ is a stationary and ergodic marked point process.

Proof. (sketch) It should be clear that the infinite sequence $\mathcal{S}_U(m_j)$ is stationary and ergodic (the probability that the ith letter is one is equal to α).

From this point on, the result follows from mutual independence, ergodicity and stationarity of the sequences τ_j, m_j and σ_j. □

Corollary 4. *When the routing is stationary, the queue Q_j is stable if $\mathfrak{s}(m_j)\lambda/\mu_j < 1$, where λ is the intensity of $\mathcal{A}$ and μ_j is the speed of server j. If $\mathfrak{s}(m_j)\lambda/\mu_j > 1$ the queue is unstable.*

Proof. Note that the input process in queue Q_j has intensity $\mathfrak{s}(m_j)\lambda$, where λ is the intensity of the overall arrival process. Then, this is a classical result for stationary and ergodic $G/G/1$ queues (see for example [4]). □

2.2 Stationary distribution

In order to find one optimal policy, we will compute the average performances of any periodic routing policy. This implies the knowledge of the average waiting time for each queue for the sampled process of both queue. This is why, this section focuses on a single queue, say Q_1. To simplify the notations, the index of the queue (say 1) is omitted in this section. Moreover in this section, we assume that all the words are periodic, and, with a slight abuse of notation, we all call m the smallest period of an infinite periodic word m. This smallest period has length $\ell \stackrel{def}{=} |m|$ and the number of ones is $a \stackrel{def}{=} |m|_1$. The input process before the routing $\mathcal{A}$ is a Poisson process of intensity λ and the services are all independent and identically distributed with an exponential distribution of parameter μ.

We now show that under the sampling by a periodic word, the number of customers in a queue can be modeled by a Markov Process.

The behavior of the number of customers in the system of this $G/M/1/\infty$ queue is given by a continuous time Markov chain X_t (sometimes called Markov process) which state space is equal to $\mathbb{N} \times 1, \ldots, \ell$. The first entry represents the number of customers in the system at time t while the second entry represents the current letter of the sampling m (modulo ℓ).

Proposition 5. *The continuous time Markov chain X_t is a quasi birth and death process whose generator Q is given by*

$$Q = \begin{bmatrix} C & A_0 & 0 & 0 & \ldots \\ A_2 & A_1 & A_0 & 0 & \ldots \\ 0 & A_2 & A_1 & A_0 & \ldots \\ 0 & 0 & A_2 & A_1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

with $A_0 = \lambda J^+$, $A_1 = \lambda J^0 - (\lambda + \mu)I$, $A_2 = \mu I$ and $C = \lambda J^0 - \lambda I$, where I is the identity matrix of size ℓ and the matrices J^0 and J^+ are matrices which entries are respectively $(J^0)_{i,k} = (1 - m_j(i))\delta_{i,k+1}$ and $(J^+)_{i,k} = m_j(i)\delta_{i,k+1}$. The index k is taken modulo ℓ and the term $\delta_{i,k}$ is the Kronecker's symbol (1 if $i = k$ and 0 otherwise).

Proof. (sketch) Let us assume that we are in state (i, k) with $i > 0$ and $n \in \{1, \ldots, \ell\}$. Let us detail the states which could be reached from (i, k). The service rate being μ, the process jumps from (i, k) to $(i - 1, k)$ with rate μ as long as $i - 1$ remains non-negative.
The global arrival rate is λ and two cases can occur:
either $m(k) = 1$ (the current arrival is routed in the queue) and the process jumps in state $(i+1, k+1 \mod \ell)$, or $m(k) = 0$ (the current arrival is not routed in the queue) and the process jumps in state $(i, k+1 \mod \ell)$. □

Example 6 (Generator Q for $m = 110$). *Here is an example of the infinitesimal generator Q of X_t when the sample is $m = (110)^\infty$.*

$$Q = \begin{bmatrix} -\lambda & 0 & 0 & 0 & \lambda & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \\ 0 & -\lambda & 0 & 0 & 0 & \lambda & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \\ \lambda & 0 & -\lambda & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \\ \mu & 0 & 0 & -b & 0 & 0 & 0 & \lambda & 0 & 0 & 0 & 0 & \ldots \\ 0 & \mu & 0 & 0 & -b & 0 & 0 & 0 & \lambda & 0 & 0 & 0 & \ldots \\ 0 & 0 & \mu & \lambda & 0 & -b & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \\ 0 & 0 & 0 & \mu & 0 & 0 & -b & 0 & 0 & 0 & \lambda & 0 & \ldots \\ 0 & 0 & 0 & 0 & \mu & 0 & 0 & -b & 0 & 0 & 0 & \lambda & \ldots \\ 0 & 0 & 0 & 0 & 0 & \mu & \lambda & 0 & -b & 0 & 0 & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

with the notation $b = \lambda + \mu$.

Let π be the invariant measure of the process X_t (when it exists). This probability satisfies

$$\pi Q = 0. \tag{1}$$

We now refine the notation by introducing block vectors π_n of dimension ℓ whose ith entry $(\pi_n(i))$ represents the stationary probability to have n customers in the system when we stand in the routing decision $m(i)$. Hence $\pi_k(1) + \cdots + \pi_k(\ell)$ is the stationary probability to have k customer in the system.

We will not try to compute π directly which can be quite hard, but we will rather determine its generating function. Some discussions on this choice are postponed to the following sections.

Definition 7 (Generating Function). *Let* $\overline{D}(0,1)$ *be the closed unit disk. The generating function of* π *is the function* $\Pi(z)$ *from* $\overline{D}(0,1)$ *to* $\mathbb{C}^\ell$ *defined by*

$$\Pi(z) = \sum_{n=0}^{\infty} z^n \pi_n.$$

The following theorem will be used to make sure that the stationary distribution (as well as the function $\Pi(z)$) exist.

Lemma 8 (Positive recurrence of the QBD). *The Quasi Birth and Death process* X_t *is positive recurrent if and only if*

$$\frac{a\lambda}{\ell\mu} < 1. \tag{2}$$

Proof. This proof is based on Theorem 1.3.2 of [19] which states that any QBD is positive recurrent if and only if

$$\boldsymbol{p} A_2 \mathbf{1} > \boldsymbol{p} A_0 \mathbf{1},$$

where $\mathbf{1}$ is the column vector with all entries equal to one, and $\boldsymbol{p}$ is the stationary distribution vector of the finite generator $A = A_0 + A_1 + A_2$, *i.e* $\boldsymbol{p}A = 0$ and $\boldsymbol{p}\mathbf{1} = 1$.

Let us compute $\boldsymbol{p}$. Using formulas of A_0, A_1 and A_2, we get $A = -\lambda I + \lambda J$, where J is the matrix which entries are given by $(J)_{i,k} = \delta_{i,k+1}$. It should be clear that $\boldsymbol{p}_1 = \boldsymbol{p}_2 = \ldots = \boldsymbol{p}_i = \ldots = \boldsymbol{p}_\ell = 1/\ell$.

Hence, the stability condition becomes

$$\boldsymbol{\pi} A_2 \mathbf{1} = \mu \boldsymbol{\pi} I \mathbf{1} = \mu, \quad \text{and} \quad \boldsymbol{\pi} A_0 \mathbf{1} = \lambda \sum_{k=1}^{k=l} m(k) \boldsymbol{\pi}_k = \lambda \frac{a}{l}.$$

□

Remark 9. *Note that Corollary 4 is not sufficient to show the existence of a unique stationary distribution since it only proves recurrence and not positive recurrence.*

The next lemma shows that the performance of the system is invariant when the Markov chain is constructed using any period of the routing word instead of the smallest period.

Lemma 10. *Let* m *and* m' *be two words such that* $m' = m^k$ *for some* $k \geq 2$. *If chain is built using the sampling sequences* m *or* m', *then the behaviors are equal.*

Proof. The proof is made for $k = 2$, but remains similar for higher values of k. Let π (resp. π') the stationary distribution when the sampling is m (resp. m'). The probability π' satisfies (by mere verification of (1))

$$\pi'_i(n) = \pi'_i(l+n) = \frac{\pi_i(n)}{2} \qquad \forall n,\ 1 \leq n \leq l, \qquad \forall i \geq 0. \tag{3}$$

Now, all performance measures in both cases are the same. For example, the expected number of customers is equal to

$$\sum_{n=1}^{2l}\sum_{i=0}^{\infty} i\pi'_i(n) = 2\sum_{n=1}^{l}\sum_{i=0}^{\infty} i\frac{\pi_i(n)}{2} = \sum_{n=1}^{l}\sum_{i=0}^{\infty} i\pi_i(n),$$

by the use of (3) providing the equality in both systems. □

Hence, we can restrict our investigations to one period of m (say the smallest period). Moreover it must be noticed that the performances of a shift $\mathcal{S}_k(m)$ (with $k \leq l$) of a finite word m repeated infinitely are equal to the performances of m, when the queue is stable. Indeed, since the word $\mathcal{S}_k(m)$ is repeated infinitely, we can consider that the sampling given by $\mathcal{S}_k(m)$ is identical to the sampling given by m with an additional finite prefix composed of the $l-k$ last letters of m. If the queue is stable, then the perturbation induced by the adding of the prefix to the sampling word does not have any effects on the stationary behavior since this prefix can be assimilated to an initial load.

2.3 Kernel Method

We use now the kernel method to compute the generating function. This method can be decomposed in two major steps. First we determine a functional equation, depending on a parameter, which is satisfied by the generating function. The second step uses the singularities of the kernel of the equation to compute some parameters of the equation which will induce a complete computation of the generating function.

Lemma 11. *Let us define $\rho = \lambda/\mu$. Let $K(z)$ be the $\ell \times \ell$ matrix $K(z)$ defined by*

$$K(z) = \begin{bmatrix} 1-(1+\rho)z & \rho z^{1+m_j(1)} & & \\ \ddots & \ddots & & \\ & 1-(1+\rho)z & \rho z^{1+m_j(\ell-1)} \\ \rho z^{1+m_j(\ell)} & & 1-(1+\rho)z \end{bmatrix}.$$

Then the generating function satisfies the functional equation with kernel $K(z)$,

$$\Pi(z)K(z) = \pi_0(1-z). \tag{4}$$

Proof. Using the global balance equation (1) we get the induction

$$\pi_0 C + \pi_1 A_2 = 0, \tag{5}$$

$$\pi_{n-1}A_0 + \pi_n A_1 + \pi_{n+1}A_2 = 0, \quad \forall n \geq 1. \tag{6}$$

By multiplying the second equation by z^{n+1} and by summing it follows

$$\Sigma_{n=1}^{\infty}\pi_{n-1}z^{n-1}z^2A_0 + \pi_n z^n z A_1 + z^{n+1}\pi_{n+1}A_2 = 0,$$

that is

$$\Pi(z)(z^2A_0 + zA_1 + A_2) - \pi_1 A_2 z - \pi_0 A_2 - \pi_0 A_1 z = 0,$$

which gives

$$\Pi(z)(z^2A_0 + zA_1 + A_2) = \pi_0\mu(1-z).$$

using the definitions of the matrices A_0, A_1, A_2 and C. Dividing by μ yields (4). □

Let us now study the zeros of $K(z)$. More precisely we will focus on the zeros inside the unit disk since $\Pi(z)$ is a power series with radius of convergence one. Let us call $\Delta(z)$ the determinant of the matrix $K(z)$.

By using the definition of the matrices A_0, A_1 and A_2, one gets after direct computations

$$\Delta(z) = (-1)^{\ell+1}\rho^\ell z^{\ell+a} + (1-(1+\rho)z)^\ell.$$

Lemma 12. *If $\frac{a\rho}{\ell} < 1$ then $\Delta(z)$ has ℓ roots inside the unit disk.*

Proof. Let ε be a positive real number. We define by C the circle of center 0 and radius $1+\varepsilon$. Let us consider two functions $f(z)$ and $g(z)$ from $\mathbb{C}$ to $\mathbb{C}$ defined by :

$$f(z) = \Big(\frac{\rho+1}{\rho}\Big)^{\ell}(z - \frac{1}{1+\rho})^{\ell}, \qquad g(z) = f(z) - z^{\ell+a}.$$

Let us compute now the values of ℓ, a and ρ such that the following inequality is satisfied

$$|f(z) - g(z)| < |f(z)|, \qquad \forall z \text{ s.t. } |z| = 1+\varepsilon. \tag{7}$$

Using the fact that

$$|f(z) - g(z)| = (1+\varepsilon)^{\ell+a},$$

and

$$|f(z)| = \Big(\frac{\rho+1}{\rho}\Big)^{l}|z - \frac{1}{1+\rho}|^{\ell} \geq (1+\varepsilon - \frac{1}{1+\rho})^{\ell}(\frac{\rho+1}{\rho})^{\ell},$$

and developing both expressions in ε, one gets

$$|f(z) - g(z)| = 1 + (\ell+a)\varepsilon + o(\varepsilon^2)$$

and

$$|f(z)| \geq 1 + \frac{\ell(\rho+1)\varepsilon}{\rho} + o(\varepsilon^2),$$

after simplifications, we get for small ε,

$$\frac{a\rho}{\ell} < 1 \Rightarrow |f(z) - g(z)| \leq |f(z)|.$$

Since f has all its ℓ roots inside the unit disk since $1/(\rho+1) \leq 1$, using Rouché's theorem shows that the polynomial $g(z)$ as well as $\Delta(z)$ have only ℓ roots inside the closed unit disk $\overline{D}(0,1)$. □

Theorem 13. *If z_i is the ith root of $\Delta(z)$ in the unit disk and v_i is the right eigenvector of the eigenvalue 0 of $K(z_i)$, then π_0 is a solution of the system :*

$$\left\{ \begin{array}{ll} (1-z_i)\pi_0 v_i = 0, & \forall i \in \{1,\dots,l\} \text{ s.t. } z_i \neq 1 \\ \pi_0 \mathbf{1} = 1 - \frac{a\rho}{l}, & \text{when } z_i = 1 \end{array} \right\}. \tag{8}$$

where $\mathbf{1}$ *is the column vector with all its components equal to* 1.

Proof. If $|z_i| < 1$, then it comes

$$(1-z_i)\pi_0 v_i = 0.$$

Note that the rank of the matrix $K(z_i)$ is $\ell - 1$ so that the vector v_i is unique according one multiplicative constant.

The case $z_i = 1$ has to be handled differently since $(1-z_i) = 0$ and $K(1)\mathbf{1} = \mathbf{0}$, with $\mathbf{0}$ the vector of dimension ℓ with all its entries equal to 0. Let us begin to notice that

$$\frac{K(z)\mathbf{1} - K(1)\mathbf{1}}{1-z} = \big(1 - \rho z \frac{1 - z^{m_j(i)}}{1-z}\big)_{i=1,\dots,l},$$

and that, with (4), we have

$$\Pi(z)\frac{K(z)\mathbf{1}}{1-z} = \pi_0\mathbf{1}.$$

Therefore it comes

$$\pi_0\mathbf{1} = \lim_{z\to 1}\frac{\Pi(z)K(z)\mathbf{1}}{1-z} = \lim_{z\to 1}\Pi(z)\frac{K(z)\mathbf{1}-K(1)\mathbf{1}}{1-z} = \lim_{z\to 1}\Pi(z)\left(1-\rho z\frac{1-z^{m_j(i)}}{1-z}\right)_{i=1,\ldots,l}$$
$$= \Pi(1)\left(1-\rho m_j(i)\right)_{i=1,\ldots,l} = \frac{1}{l}\sum_{k=1}^{k=l}(1-\rho m_j(k)).$$

Since $\Pi(1)(i)$ is the probability to be in $m(i)$ and since this probability equals $1/\ell$, it follows $\pi_0\mathbf{1} = 1-\frac{a\rho}{l}$. □

Corollary 14. *If the set of equations (8) has a unique solution, this allows us to get the function $\Pi(z)$ as a vectorial analytical function in the closed unit disk, given by*

$$\Pi(z) = (1-z)\pi_0 K(z)^{-1}.$$

However, this system may not have full rank, in which case our approach fails. Among the several thousand cases we considered in our experimental runs, they all had full rank.

2.3.1 Computation of the average waiting time

Let $\mathbb{E}(N_j(m_j))$ and $\mathbb{E}(W_j(m_j))$ be respectively the expected number of customers and the expected waiting time in the queue Q_j when the sample is made according m_j.

Since $\mathbb{E}(N_j(m_j)) = \frac{d\Pi(1)\mathbf{1}}{dz}|_{z=1}$ introducing the vector $\hat{K}(z)$ which verifies $K(z)\hat{K}(z) = \mathbf{1}$ it follows

$$\mathbb{E}(N_j(m_j)) = \frac{d}{dz}\big(\Pi(z)\mathbf{1}\big)|_{z=1} = \pi_0\frac{d}{dz}\big((1-z)\hat{K}(z)\big)|_{z=1}. \tag{9}$$

This gives the expected waiting time :

$$\mathbb{E}(W_j(m_j)) = \frac{l}{a\lambda}\mathbb{E}(N_j(m_j)) - \frac{1}{\mu}. \tag{10}$$

3 Computational Issues

One important point one have to keep in mind in the computation of stationary distribution is the efficiency and the numerical robustness of the algorithms. Indeed, one could easily reach the limits (both in memory and precision) of the current computers.

In this work, one of the critical point is the length of the periods of the words which induce large matrices A_0, A_1 and A_2. The experiments reported in following sections often lead to matrices which sizes often exceed 4000). This is why it is critical to address numerical computation issues.

In Section 3.2, we will compare the efficiency of the kernel method used in this paper, with one of the classical method, namely the matrix geometric method.

3.1 Numerical computations of the roots

We focus on the computations of the roots of $\Delta(z)$ since the rest is classical linear algebra. Solving numerically $\Delta(z) = 0$ in the complex plane could be very delicate when the degree of $\Delta(z)$ is large. We provide an alternative way where all computations are done using appropriate polar coordinates and finding roots of real functions.

As we saw before the determinant is equal to

$$\Delta(z) = (-1)^{\ell+1}\rho^{\ell}z^{\ell+a} + (1-(1+\rho)z)^{\ell}\ .$$

Therefore the equation $\Delta(z) = 0$ could be rewritten as follow

$$z^{\ell+a} = \big(\frac{\rho+1}{\rho}\big)^{\ell}\big(z - \frac{1}{\rho+1}\big)^{\ell}. \tag{11}$$

Let $x_0 \stackrel{def}{=} 1/(\rho+1)$. We can now give a description of the location of the roots of $\Delta(z)$.

Lemma 15. *The roots of $\Delta(z)$ inside the unit disk are located on the curve Γ, given in shifted polar coordinates (i.e. $z = x_0 + re^{i\theta}$) by*

$$x_0^2 + 2rx_0\cos(\theta) + r^2 = \Big(1+\frac{1}{\rho}\Big)^{2\ell/(\ell+a)} r^{2\ell/(\ell+a)}. \tag{12}$$

This curve is an oval, symmetric with respect to the x axis which intersects this real axis in two points : $(1,0)$ and in a second point denoted by $(r_0,0)$, r_0 being the solution in $\mathbb{R}$ of

$$z^{2(\ell+a)} = \left(\frac{(\rho+1)}{\rho}\right)^{2\ell}(z-x_0)^{2\ell}. \tag{13}$$

When ℓ is even, r_0 is the second real roots of $\Delta(z) = 0$. Furthermore the curve Γ is located inside the unit disk of center $(x_0,0)$ and radius $\rho/(\rho+1)$ and contains the circle with center $(x_0,0)$ and with radius $x_0 - r_0$.

Proof. Introducing x_0 in Equation (11) yields

$$\left|z^{\ell+a}\right| = \Big(\frac{\rho+1}{\rho}\Big)\,|z-x_0|^{\ell}$$

which can be rewritten in polar coordinates by

$$\big(x_0^2 + 2rx_0\cos(\theta) + r^2\big)^{(\ell+a)/2} = \Big(1+\frac{1}{\rho}\Big)^{\ell} r^{\ell},$$

which can be simplified into

$$x_0^2 + 2rx_0\cos(\theta) + r^2 = \Big(1+\frac{1}{\rho}\Big)^{2\ell/(\ell+a)} r^{2\ell/(\ell+a)}.$$

The cosine function being even, the curve is symmetric with respect to the x axis.

On the other hand Equation (11) also implies

$$|z|^{\ell+a} = \Big(\frac{\rho+1}{\rho}\Big)^{\ell}|z-x_0|^{\ell}. \tag{14}$$

and therefore $|z| \le 1$ implies $|z-x_0| \le \frac{\rho}{\rho+1}$.

Letting $z = x+iy$, it comes

$$(x^2+y^2)^b = \Big(\frac{\rho+1}{\rho}\Big)^{2\ell}\big((x-x_0)^2+y^2\big)^{\ell}.$$

Thus, the points for which y is equal to zero satisfy Equation (13). Since

$$\left|z^{2(\ell+a)} - \left(\frac{(\rho+1)}{\rho}\right)^{2\ell}(z-x_0)^{2\ell}\right|_{z=1} = 0,$$

there exist at less two real roots. Next, we want to show that $0 < r_0 < x_0$. Consider Equation (13) when $z = 0$ and when $z = x_0$. We get

$$\left| z^{2(\ell+a)} - \left(\frac{(\rho+1)}{\rho}\right)^{2\ell} (z-x_0)^{2\ell} \right|_{z=0} = -(\frac{1}{2\ell})^{2\ell} < 0,$$
$$\left| z^{2(\ell+a)} - \left(\frac{(\rho+1)}{\rho}\right)^{2\ell} (z-x_0)^{2\ell} \right|_{z=x_0} = \left(\frac{1}{\rho+1}\right)^{2(\ell+a)} > 0,$$

We claim, with the help of the following lemma, that r_0 is unique and furthermore that the smallest module of the complex numbers belonging to Γ is $x_0 - r_0$. □

Example 16 (Curve of the roots located in $\overline{D}(0,1)$ for $\rho = 5/6$ and $l = 120$.). *In this example the set of the roots of $\Delta(z)$ for $\rho = 5/6$, $l = 120$ and $a = 26$ is displayed. On Figure 2, the "oval" as well as the circles of center $(x_0, 0)$ and radius $\rho/(1+\rho)$ and radius $x_0 - r_0$ are displayed.*

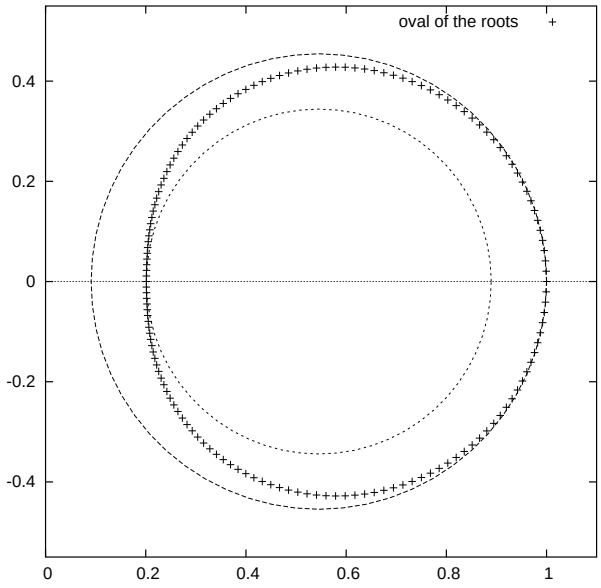

Figure 2: Location of the roots for the example 16.

Lemma 17. *The function $\theta \mapsto r$ where $z = x_0 + re^{i\theta}$ and $z \in \Gamma$ is bijective.*

Proof. The first part of the proof consists in finding a lower bound for the interval in which r varies (the upper bound being $\rho/(\rho+1)$. We introduce the function h defined on $\mathbb{R}^2 \to \mathbb{R}$ such that

$$h(r,\theta) = r^2 - (\frac{\rho+1}{\rho})^{2\ell/(\ell+a)} r^{2\ell/(\ell+a)} + 2rx_0\cos(\theta) + x_0^2.$$

This function is continuous and infinitely differentiable on $]0, \rho/(\rho+1)] \times [0,\pi]$.

We have $\forall r \in]0, \rho/(\rho+1)]$,

$$\frac{\partial h(r,\theta)}{\partial r} = 2r - 2\frac{\ell}{\ell+a}\left(\frac{\rho+1}{\rho}\right)^{2\ell/(\ell+a)} r^{(\ell-a)/(\ell+a)} + 2x_0\cos(\theta),$$
$$\frac{\partial^2 h(r,\theta)}{\partial r^2} = \frac{1}{(\ell+a)^2}\left(2(\ell+a)^2 - 2\big(\ell+a\big)\ell\Big(\frac{\rho+1}{\rho}\Big)^{2\ell/(\ell+a)} r^{-2a/(\ell+a)}\right),$$
$$\frac{\partial h(r,\theta)}{\partial r^3} = \frac{4a\ell(\ell-a)}{(\ell+a)^3}\Big(\frac{\rho+1}{\rho}\Big)^{2\ell/(\ell+a)} r^{(-\ell-3a)/(\ell+a)}.$$

Therefore, for any $r > 0$, $\frac{\partial h(r,\theta)}{\partial r^3}$ is strictly positive. Let us introduce

$$\tilde{r} \stackrel{def}{=} \left(\frac{(\ell+a)^2}{\ell(\ell+a)}\right)^{-(\ell+a)/2a} \left(\frac{\rho}{\rho+1}\right)^{\ell/a}.$$

Since

$$\left|\frac{\partial^2 h(r,\theta)}{\partial r^2}\right|_{r=\tilde{r}} = 0,$$

then $\frac{\partial h(r,\theta)}{\partial r}$ is decreasing for r in the interval from 0 to $\tilde{r}$ and increasing next. We will now show that this derivative is negative. This first derivative is decreasing in θ, and takes the values $2x_0 cos(\theta)$ in $r = 0$. Hence, for any θ greater than $\pi/2$, the derivative $\partial h(r,\theta)/\partial r$ is negative in $r = 0$ and in particular, when $(r,\theta) = (\rho/(\rho+1), 0)$ the first derivative takes the value

$$2\left(\frac{\rho}{\rho+1} - \frac{\ell}{\ell+a}\frac{\rho+1}{\rho} + x_0\cos(\theta)\right) \leq 2\frac{\rho}{\rho+1} - 2\frac{\ell}{\ell+a}\frac{\rho+1}{\rho} < 2\frac{\rho}{\rho+1} - 2 < 0,$$

because of the stability condition. Thus, for any $\theta \geq \pi/2$, the derivative $\partial h(r,\theta)/\partial r$ is strictly negative. Henceforth, $h(r,\theta)$ is strictly decreasing in r and there is an unique solution of the equation $h(r,\theta) = 0$ for a fixed θ. In particular $x_0 - r_0$ is the unique solution of $h(r,\pi) = 0$.

Since

$$\frac{\partial h(r,\theta)}{\partial \theta} = -2r\sin(\theta) \leq 0,\ \forall r > 0,\ \forall \theta \in [0,\pi],$$

then the two partial derivatives of h are non positive, and h is decreasing. This implies that $x_0 - r_0$ is the smallest module of the points belonging to Γ for $\theta \geq \pi/2$.

We will now show that for any $\theta \leq \pi/2$, r can not be smaller than $x_0 - r_0$. We assume that $r < x_0 - r_0$, thanks to Equation (13), we get

$$\left(1+\frac{1}{\rho}\right)^{2\ell} r^{2\ell} < \left(1+\frac{1}{\rho}\right)^{2\ell} \left(r_0 - x_0\right)^{2\ell} = r_0^{2(\ell+a)},$$

which yields $x_0^2 + 2rx_0\cos(\theta) + r^2 < r_0^2$. But $x_0 > r_0$ for all $\theta \leq \pi/2$, and for all r non negative $x_0^2 + 2rx_0\cos(\theta) + r^2$ is greater than x_0^2. Thus we are led to an absurdity and therefore $r > x_0 - r_0$. Thus the smallest module of the points belonging to Γ is $x_0 - r_0$.

We are now able to compute the sign of the derivatives. We resume the study of the function. It follows

$$\frac{\partial h(x_0 - r_0, \theta)}{\partial r} \leq \frac{\partial h(x_0 - r_0, 0)}{\partial r} < 0$$

and

$$\frac{\partial h(\rho/(\rho+1), \theta)}{\partial r} \leq \frac{\partial h(\rho/(\rho+1), 0)}{\partial r} < 0,$$

then h is strictly decreasing in r. This also proves the uniqueness of the module r for any $z \in \Gamma$ for θ given. On the other hand, h being also monotone decreasing in θ the module r of any point on Γ is decreasing in θ.

Finally, the function h being strictly monotonic for both r and θ. The function $\theta \mapsto r$ such that $h(r,\theta) = 0$ is a bijection, by the theorem of implicit functions. □

Lemma 18. *We define the function $\theta(r)$ by*

$$\theta(r) = \arccos\Big[\frac{1}{2rx_0}\Big(\big(\frac{\rho+1}{\rho}\big)^{2\ell/(\ell+a)} r^{2\ell/(\ell+a)} - r^2 - x_0^2\Big)\Big]. \tag{15}$$

The function $\theta(r)$ is continuous and decreasing from π to 0 on the interval $[x_0 - r_0, \rho/(\rho+1)]$.

Proof. From the preceding lemma we know that r varies from $x_0 - r_0$ to $\rho/(\rho+1)$.

Let us define $\eta(r)$ by

$$\eta(r) = \big(\frac{\rho+1}{\rho}\big)^{2\ell/\ell+a} \frac{r^{(\ell-a)/(\ell+a)}}{2x_0} - \frac{r}{2x_0} - \frac{x_0}{2r},$$

for all $x_0 - r_0 < r \leq \rho/(\rho+1)$. The partial derivative of $\eta(r)$ according the variable r is called $\eta'(r)$ and is described by

$$\eta'(r) = \frac{\ell-a}{\ell+a}\big(\frac{\rho+1}{\rho}\big)^{2\ell/\ell+a} \frac{1}{2x_0 r^{2a/\ell+a}} + \frac{x_0}{2r^2} - \frac{1}{2x_0}.$$

This function is decreasing in r. Indeed, when $r = \rho/(\rho+1)$ we get

$$\eta'(1+\frac{1}{\rho}) = \frac{\ell-a}{\ell+a}(\rho+1) + (1-\rho).$$

The stability condition $a\rho/\ell < 1$ implies that $(\ell-a)(\rho+1) + (1-\rho)(\ell+a) > 0$, thus, $\eta'(1+\frac{1}{\rho}) > 0$. This leads to the fact that $\eta'(r)$ is strictly positive $\forall r \in [x_0 - r_0, \rho/(\rho+1)]$.

Therefore, $\eta(r)$ is continuous and strictly increasing from $\eta(x_0 - r_0) = -1$ to $\eta(1+1/\rho) = 1$. Since the function $x \to \arccos(x)$ is continuous and decreasing, then $\theta(r)$ is continuous and decreasing from π to 0. □

Corollary 19. *The roots of $\Delta(z)$ are the points $(r, \theta(r))$ such that there exists $k \in \mathbb{Z}$ verifying*

$$\theta(r) = \frac{\ell+a}{\ell} \arctan\Big(\frac{r\sin\theta(r)}{x_0 + r\cos\theta(r)}\Big) + \frac{2k\pi}{\ell}. \tag{16}$$

Proof. From the preceding lemmas, we know that the roots of the determinant verify Equation (11) and are located on the curve Γ. The angle, $\theta(r)$, of the points on Γ is given by Equation (15).

From Equation (11) we get

$$\begin{aligned} \arg z^{\ell+a} &= \arg\big(\frac{\rho+1}{\rho}\big)^{\ell}(x-x_0)^{\ell} + 2k\pi, \\ (\ell+a)\arg z &= \ell\theta(r) + 2k\pi, \\ (\ell+a)\arctan\Big(\frac{r\sin\theta(r)}{x_0 + r\cos\theta(r)}\Big) &= \ell\theta(r) + 2k\pi, \end{aligned}$$

hence the result. □

Since the roots are conjugated, we can restrict our attention to the value of θ within the interval $[0, \pi]$. This also allows us to dismiss all the problems related to the discontinuity of $\arctan$.

Let us summarize all this section by giving the algorithm to compute of π_0.

Algorithm 20.

1. Solve for x in $\mathbb{R}$ the following equation to get r_0 :

$$x^{2(\ell+a)} = \left(\frac{\rho+1}{\rho}\right)^{2\ell} (x-x_0)^{2\ell}.$$

2. Solve (16) in r for k which varies from 0 to $\lfloor \ell/2 \rfloor$.
3. Compute the conjugated roots.
4. Solve the linear system (8).

3.2 Comparisons with the matrix geometric method

This section is dedicated to the comparison with the *matrix geometric* method. The *matrix geometric* method was largely studied by M.F. Neuts in [18, 19] and is classically used in the numerical computations of the stationary probability of quasi birth and death processes and $M/G/1$ queues. Let us give the general principle of the method. To compute the stationary distribution we assume the existence of a matrix R such that the distribution of the states follows a geometrical distribution and is given by $\pi_{n+1} = \pi_n R$. This matrix R is called the *rate matrix* and is the minimal non negative solution of the quadratic equation $A_0 + RA_1 + R^2 A_2 = 0$. However, this quadratic equation is very difficult to solve. This is why, iterative methods are applied to approximate R. In order to compare with the kernel method,, we have implemented one of the most efficient (or considered so) iterative method, presented in [13]. Let us briefly detail this method. The aim is to compute the dual matrix G which is the minimal solution of the quadratic equation $A_2 + A_1 G + A_0 G^2$. and which is related to R by $RA_2 = A_0 G$. This matrix G can be computed by $G = \lim_{N\to\infty} G(N)$ with

$$G(N) = \sum_{i=0}^{N} \left(\prod_{k=0}^{i-1} B_0(k) \right) B_2(i),$$

where $B_i(0) = (I - A_1)^{-1} A_i$ and $B_i(k+1) = (I - B_0(k)B_2(k) - B_2(k)B_0(k))^{-1} B_i(k)^2$ for $i = 0, 2$ and $k \geq 0$. Once G is obtained, we compute $R = A_0(I - A_1 + A_0 G)^{-1}$, the relation $RA_2 = A_0 G$ being intractable when A_2 is singular.

Since this matrix G is stochastic as soon as the QBD is recurrent and since the sequence of matrices $\{G(N)\}_{N\in\mathbb{N}}$ form a non decreasing sequence converging to G ([13]), a stopping criteria follows: $\|\mathbf{1} - G(N)\mathbf{1}\|_\infty < \varepsilon$ where ε is the wished precision and where $\|M\|_\infty$, is the greatest absolute value of all the entries of the matrix M. The number of iterations to reach the wished precision denoted by I_{ex} is shown, in [13], to be related with the value of a special probability to come back in the level 0. This value must exceed $1 - \varepsilon$.

3.2.1 Comparisons of theoritical complexities

The kernel method computations can be decomposed in three main steps : Step 1 : computations of the $\lceil \frac{\ell}{2} \rceil$ roots, which is made using a dichotomy. Step 2 : computations of π_0 which is equivalent to a resolution of a linear system. Step 3 : computation of $d/dz\big((1-z)\hat{K}(z)\big)|_{z=1}$, which is equivalent after straightforward transformations to two resolutions of triangular linear systems, each with two terms on each line. Thus the theoretical complexity of the kernel method is

$$O\big(\lceil \ell/2 \rceil) \log_2(\varepsilon^{-1})\big) + O(\ell^3)$$

where ε is the precision used in the dichotomy. For example, when ε is chosen to be 10^{-9} and the size of the matrix is $\ell = 100$, which are somehow typical in the following experiments, the value of the first term $\Big(\log_2(10^{-9}) \times (\lceil \ell/2 \rceil)\Big)$ is 1500, while the second ℓ^3 is 10^6. Therefore, the bottleneck of the kernel method is the size of the matrix more than the precision of the dichotomy.

Now, let us consider the complexity of the iterative matrix geometric method of [13]. The complexity of this method is $O(I_{ex}\ell^3)$ to get G from which one can retrieve R. The computation of R from G requires one matrix inversion, two sums and two products of matrices. Once R is obtained, we have to solve a linear system of dimension 2ℓ and to make two products and one inversion to get the expected waiting time. In addition, note that all the iterative methods suffer from a major problem which is the speed of convergence since there does not exist any bounds of I_{ex} : the number of iterations required. Since this is related with a time probability to come back in 0 then, when the queue is heavily loaded it may happen that the time of convergence is very high.

Henceforth, even if the orders of complexity of the two methods are equivalent, the constant term is much greater for the iterative method even though the number of iterations is reduced to one.

3.2.2 Numerical Stability

In addition to time complexity, the numerical stability of the algorithms alters their overall performance in a very sensible way as shown below.

For the kernel method there are three main points to take into account. First, the precision with which r_0 is computed. Indeed if r_0 is computed with a bad precision, the value given by $f(x_0 - r_0)$ is in absolute value greater than one, where $f(r) = \frac{1}{2rx_0}\Big(\big(\frac{1+\rho}{\rho}\big)^{2l/(l+a)} - r^2 - x_0^2 \Big)$. But we are induced to treat $f(x_0 - r_0)$ along the computations of the roots namely since $\theta(r) = \arccos\big(f(r)\big)$.

The absolute value of the pivot is a critical feature for numerical stability of a Gauss resolution. However we consider that it is beyond the scope of this paper to study the numerical values in the matrices involved. We have noticed that with an appropriate byte coding number (the type *long double* in C) which provides a maximum precision of 10^{-20}, the computations of π_0 are made with a relatively good precision. This is essentially due to the fact that most of the time all the terms in the matrix defined by Equations (8) never became all smaller than 10^{-6} even for matrices of big sizes.

The third point concerns the computations of $d/dz\big((1-z)\hat{K}(z)\big)|_{z=1}$ and appear to be the most crucial since it imposes us to keep formal computations to get an acceptable precision. Indeed, even for medium values of ρ and small values of l if we do not use formal computations we suffer from big numerical errors which are not compensated and may end up giving negatives values of the expected waiting time.

As for the matrix geometric method, the main problems related to the Latouche and Ramaswami's method lies in the fact that the matrix involved in the computation of G sometimes becomes equal to zero, up to machine precision, before we reach an acceptable precision for G. The second problems given by [21] is related to the fact that roundoff errors of inversion matrix are carried during all the I_{ex} iterations. If the method of Latouche *et al* is said numerically stable, it however induces to treat very small values which affect, as we will see below, the numerical precision.

3.2.3 Numerical comparisons

We now give some results summarized in Table 1 and in Table 2 of computations made in order to evaluate the numerical precision of the two methods. These computations are made on Maple in order to keep exact values as much as possible in order to minimize the roundoff errors. During all the computations, the precision of the machine is fixed to be at least 10^{-20}.

For the iterative method owing to the fact that the matrix G is stochastic we evaluate $\|\mathbf{1} - G(N)\mathbf{1}\|_\infty$. We also give the number of iterations before reaching the limits of the machine precision as well as the infinite norm (given by Maple) at the last iteration possible of the matrix denoted by PI in [13] which represents at step i the product $\prod_{k=0}^{i-1} B_0(k)$.

Parameters	$\|\mathbf{1} - G\mathbf{1}\|_\infty$	$\|PI\|_\infty$	Iterations
$\lambda = 1, \mu = 20, a = 2, l = 9$	.115	$2.1\ 10^{-218}$	3
$\lambda = 1, \mu = 20, a = 3, l = 4$	.26	$5.6\ 10^{-121}$	4
$\lambda = 4, \mu = 5, a = 2, l = 9$	.12	$2.6\ 10^{-61}$	6
$\lambda = 4, \mu = 5, a = 3, l = 4$	.32	$5.3\ 10^{-106}$	7

Table 1: Numerical stability of the iterative method

For this set of parameters, the iterative method never provided a precision below 10^{-2} for the expected waiting time (comparatively to the most precise that we have, presented in the following).

As for the kernel method we computed the difference in absolute values between the expected waiting time computed formally when the word m is fixed and short, by keeping the parameters λ and μ under an algebraic form as much as possible (hence the expected waiting time does not suffer from an important numerical

error) and the expected time computed for m^{200}. A star shows the cases where r_0 can not be computed correctly. However, computing r_0 with a precision chosen to be 10^{-25} did not yield any problem during all the computations of section 4.3.2.

	$\lambda = 1, \mu = 20$		$\lambda = 4, \mu = 5$	
	$a = 2, l = 9$	$a = 3, l = 4$	$a = 2, l = 9$	$a = 3, l = 4$
$\|m\|$	1800	800	1800	800
Precision	$8.1\ 10^{-14}$	$*$	$8.6\ 10^{-15}$	$1.7\ 10^{-14}$

Table 2: Numerical stability of the kernel method

4 Optimal Routing

This section is dedicated to the search of one optimal routing policy when we consider an open-loop control in a system made of two parallel queues Q_1 and Q_2. We assume that the inputs in the whole system follow a Poisson process while the service times have exponential distributions with parameters μ_1 in Q_1 and μ_2 in Q_2. Our aim is to find one optimal routing policy, optimal in the sense that the average waiting time (or more generally any increasing convex function of the average waiting time) is minimized. In an open-loop routing since the routing does not depend on the state of the system , the routing sequence m is a binary word given before the system starts. This sequence will route the customers as described earlier in section 2.

The average waiting time when the routing sequence is m is denoted by $\mathbf{W}(m)$ given by

$$\mathbf{W}(m) = \lim \sup_{N\to\infty} \frac{1}{N} \sum_{k=1}^{k=N} W(m)_k,$$

where $W(m)_k$ denotes the waiting time of the kth customer.

The problem can be decomposed in two steps. One qualitative step which consists in the characterization of optimal policies. The other is a quantitative step which consists in the determination of one optimal policy among all the policies characterized in the previous step.

4.1 Optimal allocations sequences

This section is devoted to the presentation of the class of the optimal policy. We recall that the binary alphabet is $\{0, 1\}$. Let us determine now one class of policies which contains an optimal policy among all the binary policy. For this, we introduce the mechanical words.

Definition 21 (Mechanical words). *For a real number x, we denote by $\lfloor x \rfloor$ (resp. $\lceil x \rceil$) the largest (resp. smallest) integer smaller (resp. larger) than x.*

The upper mechanical word with slope α is the infinite word $\overline{m}_\alpha$ where the n^{th} letter, with $n \geq 0$, is :

$$\overline{m}_\alpha(n) = \lceil (n+1) \times \alpha \rceil - \lceil n \times \alpha \rceil. \tag{17}$$

The lower mechanical word with slope α is the infinite word $\underline{m}_\alpha$ where the n^{th} letter, with $n \geq 0$, is :

$$\underline{m}_\alpha(n) = \lfloor (n+1) \times \alpha \rfloor - \lfloor n \times \alpha \rfloor. \tag{18}$$

When the slope is a rational number then the mechanical words are periodic words while when the slope is an irrational number then the mechanical words are aperiodic infinite words.

Definition 22 (Dominant subset for admissible policies). *A dominant subset of admissible (or feasible) policies is a subset of all admissible policies which contains one optimal policy.*

Theorem 23 ([2]). *The lower mechanical words are dominant.*

Proof. The proof of this theorem can be found in [2]. □

This means that among all the binary policies the average waiting time is minimal for a given lower mechanical word with a given slope α_{opt}. Also note that the conditions under which this theorem applies are much more general than the case at hand here. However, this theorem does not give any clue on how to actually compute one optimal policy. This will be done in the next section with the determination of the optimal slope α_{opt}.

4.2 Optimal Ratio

In this part we focus to the effective computations of the ratio of the customers sent in each queue. We begin by giving which ratios of customers are possible (*i.e* that let the system stable).

For convenience and without loss of generality from now on we assume than Q_2 is the fastest queue this means that we assume that $\mu_2 > \mu_1$.

The system is stable if the traffic intensity in the whole system ρ_s is strictly smaller than one *i.e.*

$$\rho_s = \frac{\lambda}{\mu_1 + \mu_2} < 1$$

Once the parameters are given, the slope α of the routing word can vary in an interval that keeps the queue stable. This interval is called the interval of stability denoted by I_s. Owing to Equation (2),

$$I_s = \left]1 - \frac{\mu_2}{\lambda}, \frac{\mu_1}{\lambda}\right[\cap [0,1].$$

Using Theorem 23, we restrict our investigation to routing sequences which are mechanical and which have slopes. Therefore, when the routing sequence is $\underline{m}_\alpha$ the cost function becomes by conditioning over the route taken by each customer, and calling N_1 the number sent in Q_1 among the first N customers,

$$\mathbf{W}(\overline{m}_\alpha) = \overline{\lim}_{N\to\infty} \frac{1}{N} \sum_{N_1=1}^{N} \left(\mathbf{1}\left(\sum_{k=1}^{k=N} (\overline{m}_\alpha(k) = N_1) \right) \left(\sum_{i=1}^{N_1} W_1(\overline{m}_\alpha)_i + \sum_{i=1}^{N-N_1} W_2(\underline{m}_{1-\alpha})_i \right) \right),$$

where $W_j(\overline{m}_\alpha)_k$ represents the waiting time of the k-th customer introduced in the queue Q_j, for $j = 1, 2$ and where $\mathbf{1}(A)(x)$ is the function equal to 1 when $x \in A$ and 0 otherwise. Since the upper mechanical word is a shift of the lower mechanical word [16], they have the same asymptotic performances (see Lemma 3). The fact that $\overline{m}_\alpha$ can be seen as a stationary and ergodic sequence by randomizing the initial shift and $N_1 = \alpha N$ imply that

$$\mathbf{W}(\overline{m}_\alpha) = \alpha \overline{\lim}_{N\to\infty} \frac{1}{N} \left(\sum_{i=1}^{i=N} W_1((\overline{m}_\alpha)_i \right) + (1-\alpha) \overline{\lim}_{N\to\infty} \frac{1}{N} \left(\sum_{k=1}^{k=N} W_2(\overline{m}_{1-\alpha})_k \right).$$

The use of the ergodic theorem yields

$$\lim\sup_{N\to\infty} \frac{1}{N} \sum_{i=1}^{i=N} W_j(\overline{m}_\alpha)_i = \mathbb{E}\left(W_j(\overline{m}_\alpha)\right).$$

This implies that

$$\mathbf{W}(\overline{m}_\alpha) = \alpha \mathbb{E}(W_1(\overline{m}_\alpha)) + (1-\alpha)\mathbb{E}(W_2(\overline{m}_{1-\alpha})). \tag{19}$$

When α is a rational number ($\alpha \stackrel{def}{=} a/\ell$) then m_j is periodic and $\mathbb{E}\left(W_j(m_j)\right)$ is given by (10).

Our aim is now to compute the optimal slope α_{opt} such that

$$\alpha_{opt} = \arg\min_{\alpha \in I_s} \mathbf{W}(\overline{m}_\alpha).$$

4.2.1 Properties of average waiting times with mechanical routing sequences

Lemma 24. *The function $\alpha \mapsto \mathbb{E}(W_j(\overline{m}_\alpha))$:*
- i) is increasing.
- ii) goes to infinity when $\alpha\rho$ goes to 1.

Proof. The proof of *i)* uses coupling arguments. Let $\alpha = p/q$ and $\alpha' = p'/q'$ be two rational numbers, such that $\alpha < \alpha'$. We call g the least common multiple of q and q' thus $g = \beta q = \beta' q'$. The words $(\overline{m}_\alpha)^\beta$ and $(\overline{m}_{\alpha'})^{\beta'}$ have the same length g but the number of "ones" differs. Indeed, $\mathfrak{s}\left((\overline{m}_\alpha)^\beta\right) = p/q$, $\mathfrak{s}\left((\overline{m}_{\alpha'})^{\beta'}\right) = p'/q'$, and $\mathfrak{s}\left((\overline{m}_\alpha)^\beta\right) = (\beta p)/g$, $\mathfrak{s}\left((\overline{m}_{\alpha'})^{\beta'}\right) = (\beta' p')/g$. Since $\alpha < \alpha'$ this shows that $|(\overline{m}_\alpha)^\beta|_1 < |(\overline{m}_{\alpha'})^{\beta'}|_1$.

Let us introduce now the binary word m obtained from $(\overline{m}_{\alpha'})^{\beta'}$ by replacing by "zero" the $(\beta' p' - \beta p)$ "ones" which follow the first "one" of $(\overline{m}_{\alpha'})^{\beta'}$. The slope of the word m is now α. Since the service times are all independent and identically distributed we can make a coupling between the service times under the routing words m and $(\overline{m}_{\alpha'})^{\beta'}$ over one period. This means that the value of the nth service under m is equal to the nth service $(\overline{m}_{\alpha'})^{\beta'}$. By construction for all $n \geq 1$, the nth arrival of $\mathcal{A}(m)$ will take place later than the nth arrival of $\mathcal{A}\left((\overline{m}_{\alpha'})^{\beta'}\right)$. Therefore it comes

$$W_j(m) \leq W_j\left((\overline{m}_{\alpha'})^{\beta'}\right), a.s.$$

But by [3], among all the words of slope α the mechanical words have the smallest average performances this yields $W_j\left((\overline{m}_\alpha)^\beta\right) \leq W_j(m)$, a.s.. We then conclude that

$$\alpha < \alpha' \Rightarrow \mathbb{E}W_j(\overline{m}_\alpha) = \mathbb{E}W_j(\overline{m}_\alpha^\beta) \leq \mathbb{E}W_j(\overline{m}_{\alpha'}^{\beta'}) = \mathbb{E}W_j(\overline{m}_{\alpha'}).$$

-ii). This follows from [10], where it is shown that the Markov Arrival Process determined by the mechanical routing word over a Poisson process has worse expected performances than an input process with same intensity given by i.i.d inter-arrivals with a Gamma distribution. By the use of the Pollacek-Khinchine Formula for the Gamma distributed , we get that when $\alpha\rho$ goes to 1 then the expected waiting time in the queue goes to infinity. This concludes the proof. □

Proposition 25 ([20]). *The function $\alpha \mapsto \alpha\mathbb{E}\Big(W_j(\overline{m}_\alpha)\Big)$ is convex in α.*

Remark 26 (Performances for irrational slopes). *As a consequence of the convexity of the function $\alpha \mapsto \alpha\mathbb{E}\Big(W_j(\overline{m}_\alpha)\Big)$, the function is continuous in α over I_s. This allows us to approximate the performances of non-periodic routing words by using the kernel method on a sequence of periodic routing words whose slopes go to the slope of the irrational one.*

Corollary 27. *The function $\alpha \mapsto \mathbf{W}(\overline{m}_\alpha)$ admits an unique minimum in I_s.*

Proof. The convexity implies that the minimum is unique. The point *ii*) of 24 implies that the function $\alpha \mapsto \mathbf{W}(\overline{m}_\alpha)$ goes to infinity when either α goes to $\frac{\mu_1}{\lambda}$ or $1 - \frac{\mu - 1 - \alpha}{\lambda}$. This shows that the minimum is reachable in the interior of I_s. □

Example 28 (Curve of $\mathbf{W}(\overline{m}_\alpha)$ for $\lambda = 1$, $\mu_1 = 1/3$ and $\mu_2 = 4/5$). *In this example, the parameters take the following values $\lambda = 1$, $\mu_1 = 1/3$ and $\mu_2 = 4/5$. One can check that the system is stable since $\lambda/(\mu_1 + \mu_2) = 15/17 < 1$. The interval of stability is $]1/5, 1/3[$. We represent on figure 3 the cost function $W(\overline{m}_\alpha)$ when α varies in I_s. The limits of the interval of stability are represented by the two vertical straight lines of equation $\alpha = 1/5$ and $\alpha = 1/3$.*

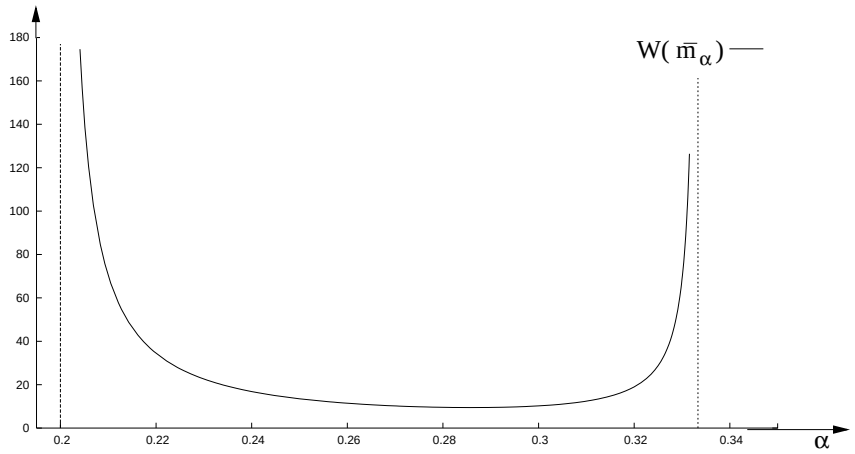

Figure 3: Curve $\alpha \mapsto W(\overline{m}_\alpha)$

4.2.2 Continuity issues

One of the major open problem related to this work is the dominance of periodic policies. This mainly depends on the continuity of the function $(\lambda, \mu_1, \mu_2) \mapsto \arg\min W(\overline{m}_\alpha)$. If this function is continuous over some (possibly very small) domain, then, this implies that for some values of (λ, μ_1, μ_2), the slope α_{opt} is an irrational number which implies in turn that no periodic policy can be optimal.

The fact that no optimal policy is periodic has an impact on, the validity of the assumptions in [11, 17] and more generally in all contributions based on dynamic programming that assume that optimal policies are periodic.

Lemma 29. *Let us consider $\alpha_{opt}(\lambda, \mu_1, \mu_2)$ as a function of the parameters. If the cost function $\alpha \mapsto \alpha W(\overline{m}_\alpha)$ is strictly convex, then the function $\alpha_{opt}(\lambda, \mu_1, \mu_2)$ is continuous.*

Proof. This proof is inspired from similar results in [15]. We denote by $g(\alpha, \lambda, \mu_1, \mu_2)$ the cost function defined by $g(\alpha, \lambda, \mu_1, \mu_2) = \alpha \mathbb{E}W_1(\overline{m}_\alpha) + (1-\alpha)\mathbb{E}W_2(\overline{m}_{1-\alpha})$. We assume that this function is strictly convex in α.

Let us assume that the function $\alpha_{opt}(\lambda, \mu_1, \mu_2)$ is not continuous. There exists a sequence $(\lambda_n)_{n\in\mathbb{N}}$ such that $\lambda_n \to \lambda$ and $\delta > 0$ such that

$$\alpha_n \stackrel{def}{=} \alpha_{opt}\big(\lambda_n, \mu_1, \mu_2\big) \to \alpha_\infty \quad \text{and} \quad |\alpha_{opt} - \alpha_\infty| > \delta.$$

First, $g(\alpha_n, \lambda_n, \mu_1, \mu_2) \le g(\alpha_{opt}, \lambda_n, \mu_1, \mu_2)$, by definition of α_n. The function g is continuous in λ, μ_1 and μ_2 since $\mathbb{E}W_1(\overline{m}_\alpha)$ is continuous in λ/μ_1 and $\mathbb{E}W_2(\overline{m}_{1-\alpha})$ is continuous in λ/μ_2 (this is a direct consequence of the computation of $\mathbb{E}W_j(\overline{m}_\alpha)$, using the kernel method, for example). By taking the limit $n \to \infty$ in the previous inequality,

$$g\big(\alpha_\infty, \lambda, \mu_1, \mu_2\big) \le g\big(\alpha_{opt}, \lambda, \mu_1, \mu_2\big).$$

But by definition of α_{opt} we know that $g\big(\alpha_\infty, \lambda, \mu_1, \mu_2\big) \ge g\big(\alpha_{opt}, \lambda, \mu_1, \mu_2\big)$, and therefore

$$g(\alpha_\infty, \lambda, \mu_1, \mu_2\big) = g\big(\alpha_{opt}, \lambda, \mu_1, \mu_2\big).$$

Thus g is constant on an interval of size greater than δ. This implies that g is not strictly convex. This is a contradiction. □

4.3 Numerical Experiments

This section is devoted to the computations of the optimal ratios. The algorithm given here finds numerically α_{opt} with an arbitrary precision.

4.3.1 Algorithm

The framework of the algorithm is the following:

1. In a first step, we compute the two rational numbers with the smallest denominators in I_s denoted by α and α', with $\alpha < \alpha'$. We also compute two rational numbers β_l and β_u which are respectively a rational lower bound and a rational upper bound of α_{opt}. The lower bound β_l is the rational with smallest denominator in $[0, \max(0, 1 - \mu_2/\lambda)]$ and the upper bound β_u is the rational with smallest denominator in $[\min(1, \mu_1/\lambda), 1]$.

2. We compute $\mathbf{W}(\overline{m}_\alpha)$ and $\mathbf{W}(\overline{m}_{\alpha'})$ using the algorithm presented in 20. Since the cost function is convex, if $\mathbf{W}(\overline{m}_\alpha) \leq \mathbf{W}(\overline{m}_{\alpha'})$ then $\alpha_{opt} \leq \alpha'$ and we have to find a new point in $]\beta_l, \alpha'[\cap I_s$. Otherwise α becomes the lower bound and we have to find a new point in $]\alpha, \beta_u[\cap I_s$.

3. Once the new point is selected we iterate the same process until the size of the search interval is smaller than the precision (given before hand).

The remaining question is the choice of the new point in the search interval.

One possibility is to choose the new point in the middle of the interval. This guarantees a fast convergence (the size of the interval decreases exponentially fast) but a very large cost per iteration. the size of the matrices used in 20 increase super-exponentially fast (the log of the size of the matrices increases exponentially fast).

Another possibility is to choose the rational number with the smallest denominator in the remaining interval. This means that the time per iteration increases more slowly (the size of the matrices only grows exponentially) but convergence can be very slow because the number with the smallest denominator could be very close to one end of the search interval, as shown in Example 30.

It is possible to find a compromise between this two extreme possibilities by keeping the best of both. Namely, using an ad-hoc dichotomy to keep the exponential decrease of the search interval while making sure that the size of the matrices involved in the new steps only grows exponentially.

Example 30. *Here is a simple example that shows that choosing at each step the rational number with the smallest denominator can be bad for the speed of convergence. Assume that the initial interval is* $[0, 1]$ *and that the optimal fraction* α_{opt} *is very close to* 0*. The sequence of search intervals constructed by the algorithm would be* $[0, 1/2], [0, 1/3], [0, 1/4], \cdots [0, 1/n]$ *until* $1/n$ *is smaller than the precision. The convergence is very slow (linear in the number of steps) and can be improved by introducing a little bit of dichotomy in the choice of the new point.*

We chose to select the points as close as possible to a classical dichotomy by selecting the rational number with the smallest denominator in a zone "around" the middle of the current search interval. The size of that zone is computed as hs, where s is the size of current interval and h is a parameter smaller than 1. This method is very efficient as the two following lemmas show.

Lemma 31. *Let* n *be the number of computations of* $\mathbf{W}(\overline{m}_\alpha)$*, let* k *be a non negative finite integer number that does not exceed* 3*. The size of the search interval after* n *steps is in the worst case equal to* $C\frac{1}{2}^{n/2-k}$ *and in the best case in* $C\frac{1}{3}^{n/2-k}$*, where* C *depends on the parameter* h*.*

Proof. We first study the slightly modified algorithm which differs by the first step where the two points are real numbers such that I_s is shared in three equal intervals. We also modify the research of the new point in each step by taking the exact real number in the middle of the interval inside which we look for the new point. We claim that a such modified algorithm has in the worst case a speed of decrease equal to $\frac{1}{2}^{(n-1)/2}$ and in the best case a speed of decrease of $\frac{1}{3}^{(n-1)/2}$. Indeed during the odd steps the interval is reduced by $2/3$ and this reduction requires only one computation of the objective function. On the other hand, during the even steps the interval is reduced either by a factor $3/4$ in the worst case or by a factor $1/2$ in the best case and this reduction

also requires only one computation. Hence, in two steps the interval is reduced either by $1/3$ or by $1/2$, the observation that the first step use two computations concludes the claim.

Let us study now the effect of the choice of the two rational numbers with smallest denominator, at the beginning, on the efficiency of the algorithm. Indeed, these two points could be very far of the required position to start the algorithm. But it can be checked that after at most three iterations of the process the points are in a situation close to a normal (without rational number constraints) work of the algorithm. □

Lemma 32. *The denominators of the extreme points of the search interval only grow exponentially fast.*

Proof. Let the current search interval be $[p_1/q_1, p_2/q_2]$. Let m be the middle point: $m = \frac{p_1q_2+p_2q_1}{2q_1q_2}$. Let $s = \frac{-p_1q_2+p_2q_1}{q_1q_2}$ be the size of the interval. Let M be the interval centered in m of size hs: $M = [m-hs/2, m+hs/2]$. Let r be the rational number in M with the smallest denominator. By definition, $r = c_n = u_n/v_n$, where c_n is the first convergent in M of $m = [0, a_1 \cdots a_t]$. Using the relations $v_n = a_nv_{n-1} + v_{n-2}$ and $c_n - c_{n-1} = (-1)^{n+1}/v_nv_{n-1}$ shows that

$$1/v_n^2 \geq hs/2a_n.$$

This implies that $v_n \leq \sqrt{2a_n/h}\sqrt{q_1q_2} \leq \sqrt{2a_n/h}(q_1 + q_2)$ so that v_n grows exponentially fast in the worse case.

□

By choosing the precision between 10^{-4} and 10^{-6} and assuming that the worse case occurs at each step, one can compute the optimal value of $h \approx 0.08 \approx 1/12$. During our computations the parameter h has been chosen to be $h = 1/5$ for it provided the best convergence times on test beds.

4.3.2 Effective computations of α_{opt} for the minimization of the average waiting time

In this part we focus on the study of several series of numerical experiments, in order to see how the optimal ratio for the average waiting time behaves when the parameters of the system vary.

Behavior of α_{opt} when ρ_s varies The first series of computations is displayed in Figure 4. This figure presents the values of the optimal proportion α_{opt} (on the y-axis) as a function of the variation of the total load ρ_s (on the x-axis). During these computations, we let λ vary such that the total load of the system ρ_s ranges from 0 to 1 while the service times μ_1 and μ_2 are fixed. The value of μ_1 is taken to be $7/16$ and the value of $\mu_2 = 3\mu_1 = 21/16$. The $\arg\min$ is computed with a precision of 10^{-4}. The different values taken by α_{opt} are displayed in Figure 4.

A very interesting feature of the behavior of α_{opt} is that it remains constant over long intervals. Moreover the optimal values are most often rational number with small denominators such as $1/2$, $1/3$, $2/7$, $1/4$ and $2/9$ (this is not an artefact, since this curve is computed with a high precision).

The dotted line on Figure 4, corresponds to the proportion $\alpha = 1/4$. This proportion corresponds to a simple load balancing policy which sends 3 times more customers in Q_2 than in Q_1. This seems natural since Q_2 is three times faster than Q_1. However, as seen in the figure this policy is not always optimal.

It can also be noticed that for very small loads the Round Robin policy is optimal. The optimality of Round Robin in such cases is rather unexpected.

Continuity of the $\arg\min$ We will now refine the computations presented Figure 4, in order to get some clues on the open problem given in Section 4.2.2. For this, we present two zooms of Figure 4 given in Figure 5. The uppest zoom of Figure 5 represents the values α_{opt} when the load varies between 0.35 and 0.37 while the second zoom shows the values of α_{opt} when ρ_s varies between 0.65 and 1.

At the sight of these two zooms it seems difficult to conclude of the continuity issue, since in one case the function does not look continuous and in the other case the function does look continuous.

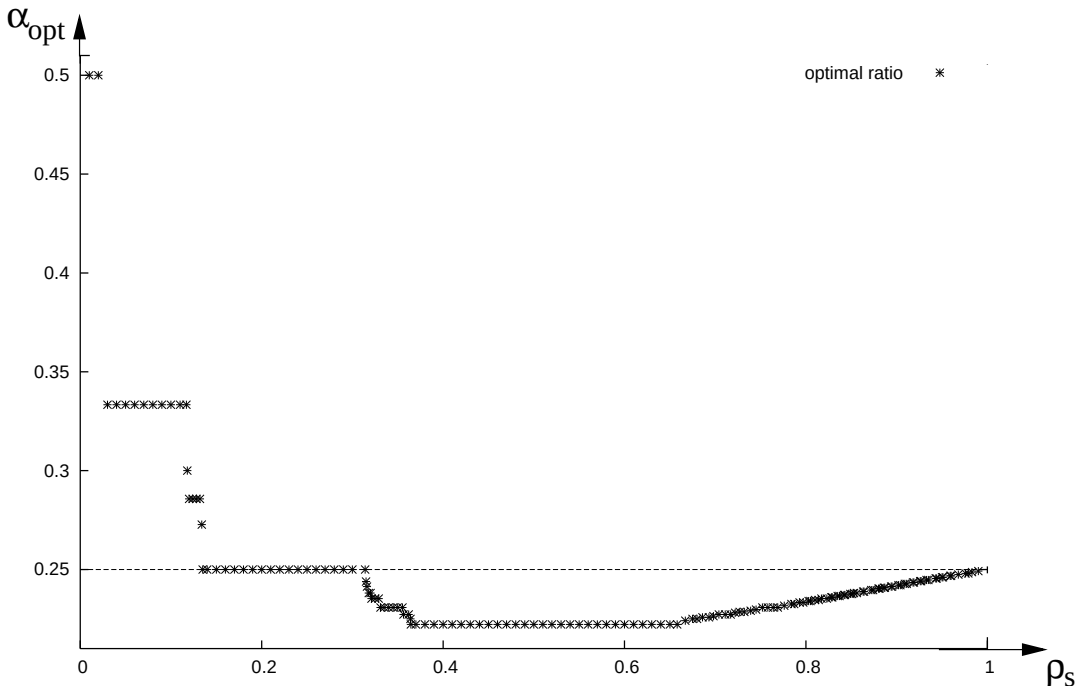

Figure 4: Curve of $\rho \mapsto \alpha_{opt}$

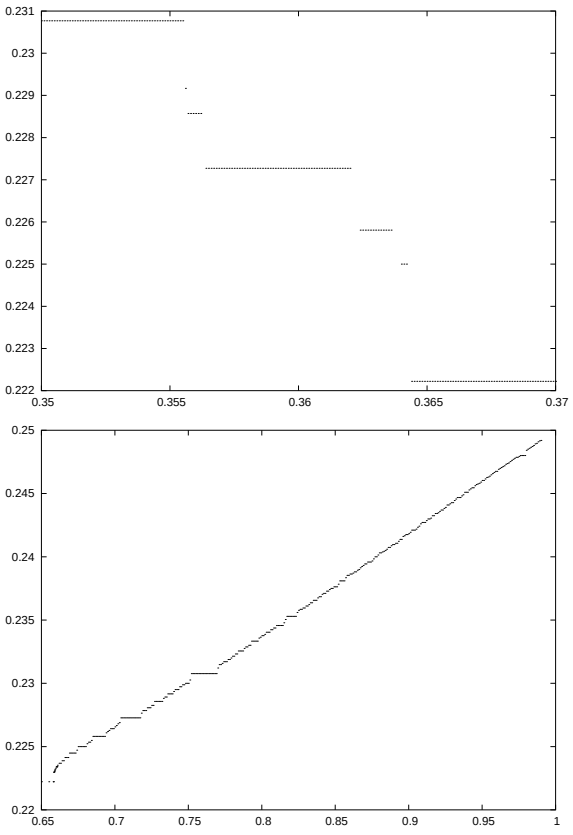

Figure 5: Zooms of the curve $\rho_s \mapsto \alpha_{opt}$

Behavior of α_{opt} when λ is fixed and μ_1 and μ_2 vary During this set of experiments (presented in Figure 6), the arrival rate is set to $\lambda = 1$, and μ_1 vary (on the x-axis) and μ_2 vary (on the y-axis) while restricting our investigations to the stability domain : I_s. The precision is chosen to be $1/500$. Figure 6 displays areas where the values of α_{opt} remain constant : each white cell represents the set of parameters that yield an identical optimal ratio.

The larger cell represents the zone where the optimal policy is the *Round Robin* policy ($\alpha_{opt} = 1/2$). Surprisingly, this zone is larger than the well-known case where the two services are equal (see [14]).

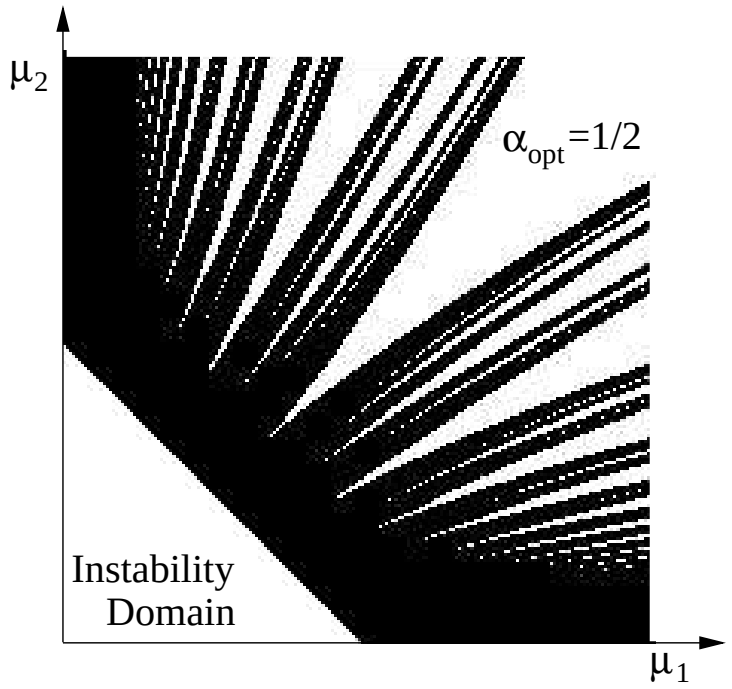

Figure 6: Curve of α_{opt} when the intensity of the services vary

A fractal appearance of the curve has to be guessed. This curve does not allow us to conclude about the continuity issue. Indeed, the discontinuity of α_{opt} should find a concrete translation on the figure by the join of two cells which is not the case. On the other hand, the distances between the cells seem to be smaller when the load decreases.

4.3.3 Effective computations for the average sojourn time

We present now computations of the optimal ratio for the minimization of the average sojourn time of the customers in the system. From a theoritical point of view, Theorem 23 can be applied for the average sojourn time and the dominant subset is still the set of Sturmian words. The cost function is now (with Little's Formula)

$$\mathbf{R}(\overline{m}_\alpha) = \mathbb{E}N_1(\overline{m}_\alpha) + \mathbb{E}N_2(\overline{m}_{1-\alpha}).$$

where $\mathbb{E}N_i(m)$ represents the average number of customers in the system in queue i when the sample is made according m. The value of $\mathbb{E}N_i(m)$ is given by Equation (9).

Moreover, properties exhibited by Lemma 24 and Proposition 25 can be extended to the average sojourn time. Hence, the algorithm described in Section 4.3.1 can be used to compute the optimal ratio.

It could be noticed that this problem is equivalent to the problem of the minimization of the expected number of customers in the system.

On Figure 7 we represent the variation of α_{opt} for the minimization of the average sojourn time in function of the total load of the system. The values of parameters are these ones of Figure 4 and the precision is fixed to be equal to $1/2000$.

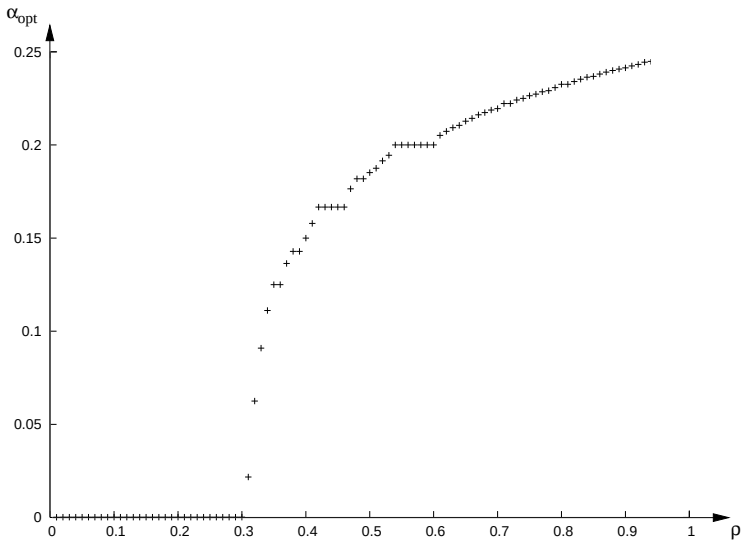

Figure 7: Curve of α_{opt} for sojourn time when the intensity of the services vary

As for the waiting time, the optimal ratio remains constant over long intervals.

4.4 Comparisons with other open-loop routing algorithms

As shown above, computing the optimal routing policy is quite involved and requires a lot of computer power. In this section, we show some heuristics that provide good solutions very fast.

Bernoulli Routing Here is the first way to approximate α_{opt}.

1. Compute the optimal Bernoulli routing policy. This provides α_B, the probability that one packet is sent to Q_1.
2. Use α_B as the routing density for a Sturmian routing.

We assume that when they arrive the customers are randomly assigned in one of the two queues. The probability to be affected in Q_1 follow a Bernoulli distribution of parameter α. The goal is then to compute the parameter α which minimizes the mean waiting time.

Here, the input process in each queue is a Poisson process with intensity $\alpha\lambda$ in Q_1 and $(1-\alpha)\lambda$ in Q_2. The objective function $g_B(\alpha)$ is now equal to

$$g_B(\alpha) = \alpha\left(\frac{\alpha\lambda}{\mu_1(\mu_1-\alpha\lambda)}\right) + (1-\alpha)\left(\frac{(1-\alpha)}{\mu_2(\mu_2-(1-\alpha)\lambda)}\right), \tag{20}$$

using classical properties of the expected waiting time in a $M/M/1$ queue ([8]).

This gives the optimal α_B as the root inside I_s of $dg_B(\alpha)/d\alpha = 0$ which can be computed readily.

Gamma approximation Another way to find an approximation of α_{opt} is by assuming that the input process in Q_1 has a Gamma distribution with parameters $(1/\alpha, \lambda)$ and the input process in Q_2 has a Gamma distribution with parameters $(1/1-\alpha, \lambda)$. Then we compute the expected waiting times in both queues and we minimize the weighted sum, to obtain the best α, denoted α_G.

Of course no routing policy can make the arrival process in both Q_1 and Q_2 Gamma-distributed, since $1/\alpha$ and $1/1-\alpha$ cannot be both integers (unless $\alpha = 2$). However, as we will see next, the optimal parameter α_G is a very good approximation of α_{opt}.

A similar approach has been taken in [5]. However, in that paper the authors did not compute the optimal policy neither α_{opt}.

To compute α_G, we start from the distribution function with density:

$$\frac{\lambda e^{-\lambda t}(\lambda t)^{\frac{1}{\alpha}-1}}{\Gamma(\alpha^{-1})}dt,$$

and we use the classical formula for the expected waiting time for a $\Gamma\backslash M\backslash 1$ queue (see [8]). The waiting time in one queue is equal to $W = \frac{\eta}{\mu(1-\eta)}$, where η is the unique solution in $]0,1[$ of the equation $\mathcal{L}(\mu(1-\eta)) = \eta$, $\mathcal{L}$ denoting the Laplace transform of the random variable of the inter-arrivals (here $\mathcal{L}(x) = (\lambda/(\lambda+x))^{1/\alpha}$). Then, we apply an optimization procedure which is very similar to the one used for the optimal policy except that each step is much faster (almost constant time in our computations).

Erlang mixture approximation This third way to approximate α_{opt} is based on the following property of mechanical words [6]. The mechanical word $\overline{m}_\alpha$ can be factorized only using the two words u and v :

$$u = 1\overbrace{0\dots0}^{s} \text{ and } v = 1\overbrace{0\dots0}^{s-1},$$

with $s = \lfloor \alpha^{-1} \rfloor$. In addition, if we denote by $|\overline{m}_\alpha|_u$ the number of u in $\overline{m}_\alpha$, and by α_1 the first partial quotient of the continued fraction, *i.e.* $\alpha = 1/(s+\alpha_1)$, then

$$|\overline{m}_\alpha|_u/(|\overline{m}_\alpha|_u + |\overline{m}_\alpha|_v) = \alpha_1.$$

We consider now the input process made of i.i.d. inter arrivals distributed according to a mixture of two Erlang distributions. This mixture is composed by an Erlang with parameters (s,λ) weighted by $1-\alpha_1$ and an Erlang with parameters $(s+1,\lambda)$ weighted by α_1. Such an input process presents some similar performance features than the mechanical sampling as illustrated in [12].

Thus, by assuming that the input process in Q_1 is the mixture of Erlang built from α and that the input process in Q_2 is the mixture of Erlang built from $1-\alpha$, we are able to compute the expected performances and to optimize the parameter α. The optimal α is denoted by α_E.

Also here, no routing policy can realize the mixture input process in both queues, nevertheless α_E is a good approximation of α_{opt} as we will see later.

The computation of α_E is similar to that of α_G except that the Laplace transform is now $\mathcal{L}(x) = (1-\alpha_1)\,(\lambda/(\lambda+x))^s + \alpha_1\,(\lambda/(\lambda+x))^{s+1}$. Computations are still in an almost constant time and substantial improvement of the running time can be achieved using formal inversion of the generating function of the waiting time. This allows us to get α_E in almost constant time over the whole range of parameters.

4.4.1 Numerical Experiments

We give now the numerical results of the comparisons.

Figure 8, displays the variations of the optimal ratios, obtained by the three methods, in function of the traffic intensity ρ_s for the average waiting time, while Figure 9 displays the optimal ratios for the average sojourn time. The parameters are similar as these of Figure 4, it means that the rates of services are $\mu_1 = 7/16$ and $\mu_2 = 21/16$ and that λ varies. All the ratios are computed with a precision of 10^{-9} while α_{opt} is computed with a precision of 10^{-4} on Figure 8 and $1/2000$ on Figure 9.

On Figures 8 and 9, one can observe that the ratios obtained by the Bernoulli routing method are always smaller than the exact values of α_{opt}. An intuitive explanation for that behavior is that sending two consecutive

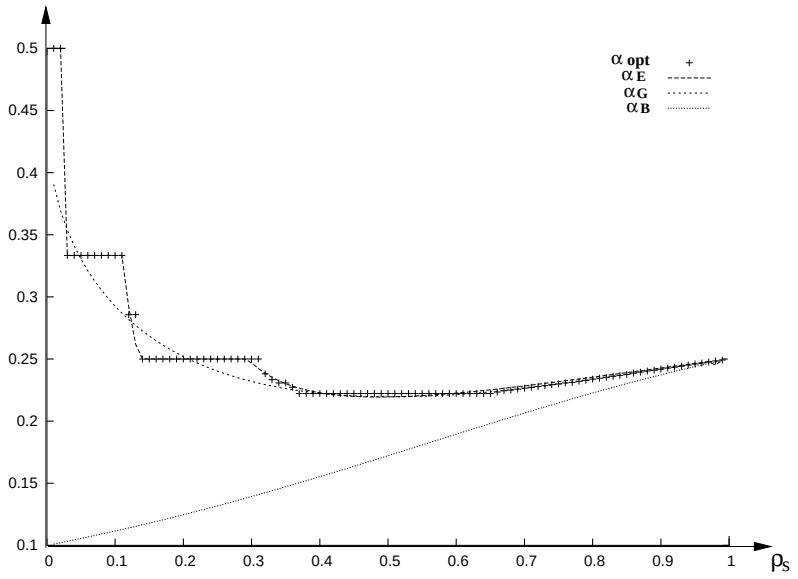

Figure 8: Comparisons for average waiting times

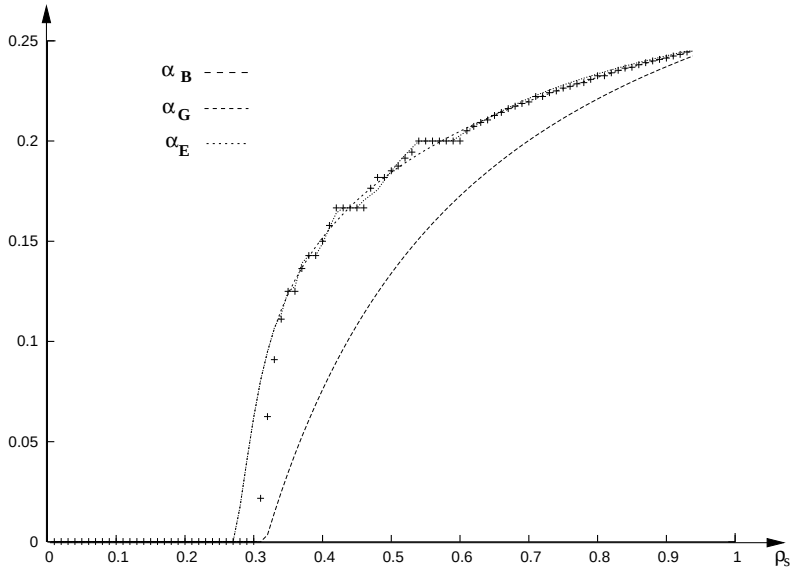

Figure 9: Comparison with heuristics for average sojourn time

packets in the slowest queue has a high impact on the waiting time and a rather large probability. Hence, the Bernoulli routing tends to send less packets in the slowest queue, on average.

One can also observe that the ratios α_G obtained using the Gamma approximation and the ratios α_E computed by the mixture of Erlangs approximation are remarkably close to α_{opt}. For loads larger than 0.8, the differences with the exact optimal ratio $|\alpha_{opt} - \alpha_G|$ and $|\alpha_{opt} - \alpha_E|$ are smaller than 10^{-4} which is the best precision of the computations of α_{opt} used in these Figures.

One additional noticeable point appears for α_E which remains constant over long interval of times as the optimal ratio α_{opt} does.

Comparison of the performances Figure 10 displays the relative error between the performances of the heuristics and α_{opt}. More precisely, the performances are computed by a mechanical sampling of slope α_B, α_G and α_E respectively and are compared with the optimal policy $\overline{m}_{\alpha_{opt}}$. The values of these experiments are displayed in Figure 8.

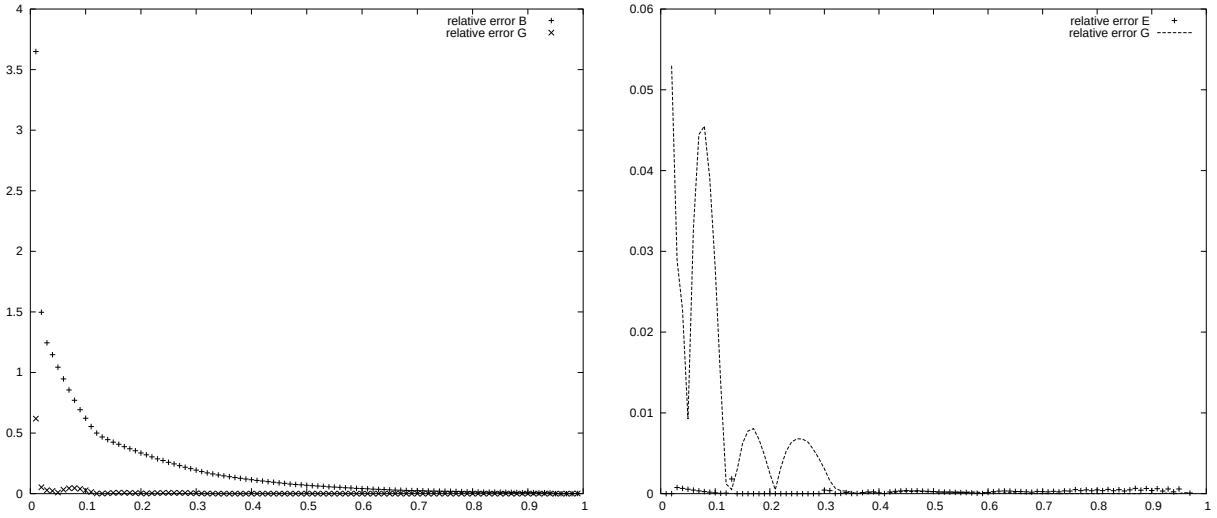

Figure 10: Relative errors for average sojourn times

Using α_B or α_G or α_E instead of α_{opt} for the routing policy always gives good performances, furthermore α_G or α_E are remarkably close approximations of α_{opt}. Indeed, on Figure 10, the average relative difference for the average waiting time is equal to 23% with the use of α_B, equal to 1% with the use of α_G and 0.1% with α_E.

5 Conclusion

In this paper, we present the computation of the optimal policy of open-loop routing in two exponential queues when the arrivals in the system follow a Poisson process. This requires a lot of computer power and the use of specific tricks to make it work efficiently. Some heuristics are used to approximate the policy which perform very well with a computational running time which can be very small.

However, some qualitative issues such as continuity are still not settled and the periodicity (or not) of the optimal policy remains open.

References

[1] E. Altman. *Markov Decision Processes, Models, Methods, Directions, and Open Problems*, chapter Applications of Markov Decision Processes in Communication Networks : a Survey, pages 488–536. Kluwer, 2001.

[2] E. Altman, B. Gaujal, and A. Hordijk. Balanced sequences and optimal routing. *J. Assoc. Comput. Mach.*, 47:752–775, 2000.

[3] E. Altman, B. Gaujal, and A. Hordijk. Multimodularity, convexity and optimization properties. *Mathematics of Operation Research*, 25:324–347, May 2000.

[4] F. Baccelli and P. Bremaud. *Elements of queueing theory*. Springer, 1992.

[5] M.B. Combe and O. Boxma. Optimization of static traffic allocation policies. *Theoritical Computer Science*, 125:17–43, 1994.

[6] B. Gaujal and E. Hyon. Optimal routing policies in two deterministic queues. *Réseaux et systèmes répartis - Calculateurs Parallèles*, 13(6):601–633, 2001.

[7] B. Gaujal and E. Hyon. Optimal routing policies in deterministic queues in tandem. In *WODES*, pages 251–257. IEEE, 2002.

[8] D. Gross and C.M. Harris. *Fundamentals of Queueing theory*. Wiley, 2nd edition edition, 1985.

[9] B. Hajek. Optimal control of two interacting service stations. *IEEE Trans. Aut. Cont.*, 29:491–499, 1984.

[10] B. Hajek. Extremal splittings of point processes. *Mathematics of Operation Research*, 10(4):543–556, 1985.

[11] A. Hordijk, G.M. Koole, and J.A. Loeve. Analysis of a customer assignment model with no state information. *Probability in the Engineering and Informational Sciences*, 8:419–429, 1994.

[12] E. Hyon. *Contrôle d'admission en boucle ouverte dans les réseaux*. PhD thesis, INPL, 2002. in French.

[13] G. Latouche and V. Ramaswami. A logarithmic reduction algoritm for quasi-birth-death processes. *Journal of Applied Probability*, 30:650–674, 1993.

[14] Z. Liu and R. Righter. Optimal load balancing on distributed homogenous unreliable processors. *Journal of Operation Research*, 46:563–573, 1998.

[15] J. A. Loeve. *Markov Decision Chains with Partial Information*. PhD thesis, Leiden University, 1995.

[16] M. Lothaire. *Algebraic Combinatorics on Words*, chapter Sturmian Words. Cambridge University Press, 2002.

[17] R.A. Milito and E. Fernandez-Gaucherand. Open-loop routing of n arrivals to m parallel queues. *IEEE Transactions on Automatic Control*, 40:2108–2114, 1995.

[18] M.F. Neuts. *Matrix-Geometric Solutions in stochastic Models An Algoritmic Approach*. John Hopkins University Press, 1981.

[19] M.F. Neuts. *Structured stochastic matrices of M/G/1 type and their applications*. Marcel Dekker, 1989.

[20] Dinard van der Laan. *The structure and performance of optimal routing sequences*. PhD thesis, Leiden University, 2003.

[21] Q. Ye. On latouche-ramaswami's logarithmic reduction algorithm for quasi-birth-and-death processes. *Stochastics models*, 18:449–467, 2002.

Unité de recherche INRIA Lorraine
LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399