Optimizing the Number of Robots for Web Search Engines

J. Talim – Z. Liu – P. Nain
INRIA, B.P. 93, 06902, Sophia Antipolis Cedex, France
{jtalim, liu, nain}@sophia.inria.fr

E. G. Coffman, Jr.
Bell Labs, Lucent Technologies
Murray Hill, NJ 07974, USA
egc@bell-labs.com

Abstract

Robots are deployed by a Web search engine for collecting information from different Web servers in order to maintain the currency of its data base of Web pages. In this paper, we investigate the number of robots to be used by a search engine so as to maximize the currency of the data base without putting an unnecessary load on the network. We use a queueing model to represent the system. The arrivals to the queueing system are Web pages brought by the robots; service corresponds to the indexing of these pages. The objective is to find the number of robots, and thus the arrival rate of the queueing system, such that the indexing queue is neither starved nor saturated. For this, we consider a finite-buffer queueing system and define the cost function to be minimized as a weighted sum of the loss probability and the starvation probability. Under the assumption that arrivals form a Poisson process, and that service times are independent and identically distributed random variables with an exponential distribution, or with a more general service function, we obtain explicit/numerical solutions for the optimal number of robots to deploy.

Keywords: Web search engines; Web robots; Queues.

1 Introduction

The World Wide Web has become a major information publishing and retrieving mechanism on the Internet. The amount of information as well as the number of Web servers has been growing exponentially fast in recent years. In order to help users find useful information on the Web, search engines such as Alta Vista, HotBot, Yahoo, Infoseek, Magellan, Excite and Lycos, etc. are available. These systems consist of four main components: a database that contains web pages (full text or summary), a user interface that deals with queries, an indexing engine that updates the database, and robots that traverse the Web servers and bring Web pages to the indexing engine. Thus, the quality of a search engine depends on many factors, e.g., query response time, completeness, indexing speed, currency, and efficient robot scheduling.

Our interest here focuses on the function served by robots: establishing currency by bringing new pages to be indexed and bringing changed/updated pages for re-indexing. We investigate the problem of choosing the number of robots to meet the conflicting demands of low network traffic and an up-to-date data base. The specific model, illustrated in Figure 1, centers on the indexing engine, which is represented by a finite, single-server queue/buffer, and multiple robots acting as sources of arriving pages. The times between successive page accesses are independent and identically distributed for each robot; the robots themselves are identical and function independently. The indexing (service) times are independent, identically distributed, and independent of the arrival processes.

When a robot arriving with a page for the indexing buffer finds the buffer full, the page being delivered is lost, at least temporarily. In this situation, a potential update or new page has been lost, and network congestion has been created to no benefit. On the other hand, if the buffer is ever empty, and hence the indexing engine is idle, data base updating is at a standstill waiting for the robots to bring more pages. To reduce the probability of the first of these two events, we want to keep the number of robots suitably small, but to reduce the probability of the second, we want to keep the number of robots suitably large. To make the objective concrete, we will formulate a cost function as a weighted sum of the probabilities of an empty buffer and a full buffer. We will then study the problem of finding the number of robots that minimizes this cost function.

There is a large literature on search engines and their components. The search engines themselves may well be their own best source of references; we recommend this entree to the research on any aspect of the subject. In particular, much can be found on the design and control (including distributed control) of robots. However, we have found very little on the modeling and analysis of robot scheduling and the indexing queue. The work in [2] is the only such effort we know about. In [2], the authors propose a natural model of Web-page obsolescence, and study the problem of scheduling a single search engine robot so as to minimize the extent to which the search engine's data base is out-of-date.

Section 2 introduces the probability model, sets notation, and formalizes the optimization problem. Section 3 solves the optimization problem for exponentially distributed service times and presents an explicit computation of the optimal number of robots. The sensitivity of the results to various model parameters is also addressed. Extensions to more general service time distributions are given in Section 4. Section 5 concludes with a brief discussion of ongoing research and open issues.

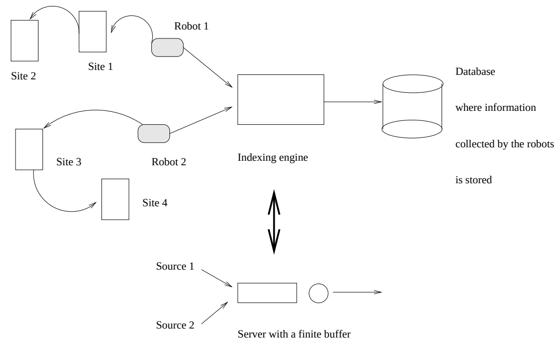


Figure 1: Model of search engine with 2 robots

2 The model

The search engine is modeled as a single server finite capacity queue. The system capacity is $K \geq 2$ (including the position in the server). There are $N \geq 1$ robots: each robot brings new pages to the queue according to a Poisson process with rate $\lambda > 0$. These N Poisson processes are assumed to be mutually independent and independent of the service times. Hence, new pages are generated according to a Poisson process with intensity λN. An incoming page finding a full queue is lost.

We denote by $F(x) = \mathbf{P}\{\sigma \leq x\}$ the probability distribution of the service times (with σ a generic service time) and by $\bar{\sigma} > 0$ the mean service time. Define $\mu = 1/\bar{\sigma}$.

In this notation, we define the cost function as the weighted sum of two terms:

- the probability of an empty buffer $\mathbf{P}\{X = 0\}$, where X is a random variable representing the stationary queue-length in a M/G/1/K queue with arrival rate λN and service time distribution F;
- the probability of losing an arriving page, that is, the probability that the queue length *seen by an arrival* in this M/G/1/K is K, which we denote by $\mathbf{P}^*\{X = K\}$.

With $\rho := N\lambda/\mu > 0$, the cost function is then defined as

$$C(\rho, \gamma, K) := \gamma\, \mathbf{P}\{X = 0\} + \mathbf{P}^*\{X = K\} \qquad (1)$$

where γ is a positive constant.

In the remainder of this paper we will resort to queueing theory to compute $C(\rho, \gamma, K)$ and to optimize this quantity as a function of ρ, or equivalently of N, the number of robots. We first cover the situation where the service times are exponentially distributed.

3 The M/M/1/K search engine model

In this section, our search engine model is the well known M/M/1/K queue. We first address the problem of finding the optimal number of sources (robots) in this model.

3.1 Optimizing the number of robots

The following proposition gives the well known stationary queue-length probabilities at arbitrary epochs for this queue [3]:

Proposition 1 *For any $\rho > 0$*

$$\mathbf{P}(X = i) = \begin{cases} \frac{1-\rho}{1-\rho^{K+1}}\rho^i & \text{for } i = 0, 1, \ldots, K \\ 0 & \text{for } i > K. \end{cases}$$

When $\rho = 1$ the non-zero stationary queue-length probabilities at arbitrary epochs are all equal and given by $Prob(X = i) = 1/(K+1)$ for $i = 0, 1, \ldots, K$. ◇

The expression for the cost function $C(\rho, \gamma, K)$ now flows from Proposition 1 and the PASTA property, which ensures that in the M/M/1/K queue the stationary queue-length probabilities at arbitrary epochs and the stationary queue-length probabilities at arrival epochs are equal (i.e., $\mathbf{P}^*(X = k) = \mathbf{P}(X = k)$). We find that

$$C(\rho, \gamma, K) = \begin{cases} \frac{(1-\rho)(\gamma + \rho^K)}{1-\rho^{K+1}} & \text{for } \rho \neq 1 \\ \frac{\gamma+1}{1+K} & \text{for } \rho = 1. \end{cases} \quad (2)$$

Note that for any $K \geq 2$, $\gamma > 0$, the mapping $\rho \to C(\rho, K, \gamma)$ is continuous and differentiable in $(0, \infty)$, including the point $\rho = 1$.

Lemma 1 *For any $\gamma > 0$, $K \geq 2$, the mapping $\rho \to C(\rho, \gamma, K)$ has a unique minimum in $[0, \infty)$, to be denoted $\rho(\gamma, K)$. Furthermore, $0 < \rho(\gamma, K) < 1$ if $\gamma < 1$, $\rho(1, K) = 1$ and $\rho(\gamma, K) > 1$ if $\gamma > 1$.*

◇

Proof. Fix $K \geq 2$. We have

$$\frac{\partial C(\rho, \gamma, K)}{\partial \rho} = \frac{R(\rho, \gamma, K)}{(1-\rho^{K+1})^2} \quad (3)$$

with

$$\begin{aligned} R(\rho, \gamma, K) &= \rho^{2K} - K\gamma\rho^{K+1} + (\gamma - 1)(K+1)\rho^K \\ &\quad + K\rho^{K-1} - \gamma. \end{aligned}$$

Tedius but elementary algebra show that the polynomial $R(\rho, \gamma, K)$ in the variable ρ

(i) has a zero of multiplicity two (respectively, three) at point $\rho = 1$ when $\gamma \neq 1$ (respectively, $\gamma = 1$);

(ii) has a zero of multiplicity one in $[0, 1)$ and no zero in $(1, \infty)$ when $\gamma < 1$;

(iii) has a zero of multiplicity one in $(1, \infty)$ and no zero in $[0, 1)$ when $\gamma > 1$;

(iv) has no zero other than 1 when $\gamma = 1$.

We deduce from the above that

$$\frac{\partial C(\rho, \gamma, K)}{\partial \rho} = \frac{(1-\rho)^2}{(1-\rho^{K+1})^2} Q(\rho, \gamma, K)$$

where $Q(\rho, \gamma, K)$ is a polynomial in the variable ρ with a single zero $\rho(\gamma, K)$ in $[0, \infty)$ with $\rho(\gamma, K) < 1$ if $\gamma < 1$, $\rho(1, K) = 1$ and $\rho(\gamma, K) > 1$ if $\gamma > 1$. Furthermore, the inequality $Q(0, \gamma, K) = -\gamma < 0$ implies that, for any $\gamma > 0$, we have $Q(\rho, \gamma, K) < 0$ for $0 \leq \rho < \rho(\gamma, K)$ and $Q(\rho, \gamma, K) > 0$ for $\rho > \rho(\gamma, K)$, which proves the lemma. ∎

It is worth observing that the optimum $\rho(\gamma, K)$ does not depend on the buffer size K when $\gamma = 1$. This means that if the same weight is given to the probability of starvation and to the loss probability, then the optimal arrival rate is equal to the service rate, independent of the buffer size.

We now return to the original problem, namely the computation of the number N of robots that minimizes the cost function $C(\rho, \gamma, K)$ with $\rho = \lambda N/\mu$. The answer is found in the next result which is a direct corollary of Lemma 1.

Proposition 2 *For any $\gamma > 0$, $K \geq 2$, let $N(\gamma, K)$ be the optimal number of robots to use.*

Then,

$$N(\gamma, K) = \arg\min_n C(n\lambda/\mu, \gamma, K) \tag{4}$$

with $n \in \{\lfloor \rho(\gamma, K)\mu/\lambda \rfloor, \lceil \rho(\gamma, K)\mu/\lambda \rceil\}$, where for any real number x, $\lfloor x \rfloor$ (respectively $\lceil x \rceil$) denotes the largest (respectively smallest) integer less (respectively greater) than or equal to x. $\diamond$

In the next section we investigate the impact of the parameter γ on the optimal number of robots.

3.2 Impact of γ on the optimal number of robots

Recall that the parameter γ is a positive constant that allows us to stress either the loss probability or the probability of starvation. Part of the impact of γ on $\rho(\gamma, K)$, and therefore on $N(\gamma, K)$, the optimal number of robots, is captured in the following result.

Proposition 3 *For any $K \geq 2$, the mapping $\gamma \to \rho(\gamma, K)$ is nondecreasing in $[0, \infty)$, with $\lim_{\gamma\to\infty} \rho(\gamma, K) = \infty$.*

$\diamond$

Proof. Pick two constants $0 < \gamma_1 < \gamma_2$ and define

$$\begin{aligned} \Delta(\rho, \gamma_1, \gamma_2, K) &:= C(\rho, \gamma_2, K) - C(\rho, \gamma_1, K) \\ &= \frac{1-\rho}{1-\rho^{K+1}}(\gamma_2 - \gamma_1). \end{aligned}$$

We assume that $\rho(\gamma_2, K) < \rho(\gamma_1, K)$ and show that this yields a contradiction.

Under the condition $\gamma_1 < \gamma_2$ the mapping $\rho \to \Delta(\rho, \gamma_1, \gamma_2, K)$ is strictly decreasing in $[0, \infty)$. Therefore,

$$\begin{aligned} 0 &< \Delta(\rho(\gamma_2, K), \gamma_1, \gamma_2, K) \\ &\quad - \Delta(\rho(\gamma_1, K), \gamma_1, \gamma_2, K) \\ &= [C(\rho(\gamma_2, K), \gamma_2, K) - C(\rho(\gamma_1, K), \gamma_2, K)] \\ &\quad + [C(\rho(\gamma_1, K), \gamma_1, K) - C(\rho(\gamma_2, K), \gamma_1, K)] \\ &\leq 0 \end{aligned} \tag{5}$$

where the last inequality follows from the definition of $\rho(\gamma, K)$. Since $\rho \to \Delta(\rho, \gamma_1, \gamma_2, K)$ is strictly decreasing on $[0, \infty)$ we deduce from (5) that $\rho(\gamma_2, K) \geq \rho(\gamma_1, K)$ must hold, which contradicts the assumption that $\rho(\gamma_2, K) < \rho(\gamma_1, K)$. Therefore $\rho(\gamma_2, K) \geq \rho(\gamma_1, K)$ and the mapping $\gamma \to \rho(\gamma, K)$ is nondecreasing in $[0, \infty)$.

On the other hand, one can check that $\partial C(\rho, \gamma, K)/\partial\rho \leq 0$ for $\rho = (\gamma/K)^{1/(K-1)}$ when $\gamma > 0$. Hence, by Lemma 1, we conclude that

$$\rho_0(\gamma, K) := \left(\frac{\gamma}{K}\right)^{1/(K-1)} \leq \rho(\gamma, K), \quad \forall \gamma > 0. \tag{6}$$

Letting γ tend to infinity on both sides of (6) yields the second result of the proposition. ∎

Proposition 3 has a simple physical interpretation. As the parameter γ increases the probability of starvation becomes the main quantity to minimize. The minimization is done by increasing the arrival rate or, equivalently, by increasing the number of robots. Figure 2 provides two numerical examples, illustrating the monotonicity of the optimal number of robots.

The next section focuses on the impact of the buffer size K on the optimal number of robots.

3.3 Impact of K on the optimal number of robots

In this section, we examine the behavior of $\rho(\gamma, K)$ as a function of K. The first result establishes an upper bound on $\rho(\gamma, K)$ that complements the lower bound given in (6).

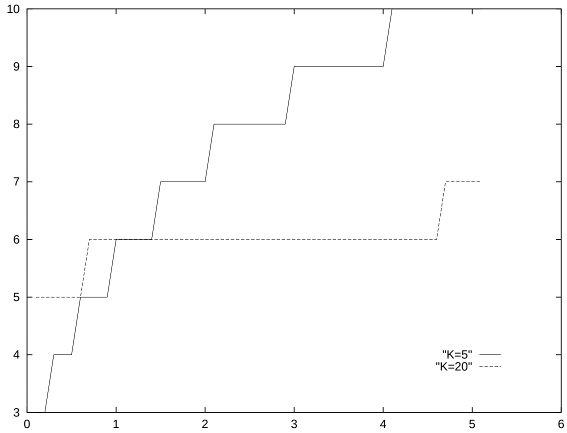


Figure 2: The mapping $\gamma \to N(\gamma, K)$ ($\mu/\lambda = 5.7$)

Lemma 2 *For any* $\gamma \geq 1$, $K \geq 2$

$$\rho(\gamma, K) < ((K+1)\gamma)^{1/(K-1)} := \rho_1(\gamma, K). \quad (7)$$

$\diamond$

Proof. Thanks to Lemma 1, it is enough to show that $\partial C(\rho, \gamma, K)/\partial \rho > 0$ at the point $\rho = \rho_1(\gamma, K)$ or, equivalently from (3), that $R(\rho_1(\gamma, K), \gamma, K) > 0$.

By writing $R(\rho, \gamma, K)$ in the form

$$R(\rho, \gamma, K) = \rho^{K+1}\left(\rho^{K-1} - K\gamma\right) + (\gamma - 1)(K+1)\rho^K + K\rho^{K-1} - \gamma,$$

we find that

$$R(\rho_1(\gamma, K), \gamma, K) = ((K+1)\gamma))^{\frac{K+1}{K-1}}\gamma + (\gamma-1)(K+1)\left((K+1)\gamma\right)^{\frac{K}{K-1}} + \gamma(K(K+1) - 1)$$

which is strictly positive, in particular for $K \geq 2$ and $\gamma \geq 1$. ∎

Using Lemma 1 together with the lower and upper bounds on $\rho(\gamma, K)$ reported in (6) and (7), respectively, we get that

$$\rho_0(\gamma, K) \leq \rho(\gamma, K) < 1, \quad \text{for } 0 < \gamma < 1 \quad (8)$$

and

$$\rho_1(\gamma, K) \geq \rho(\gamma, K), \quad \text{for } \gamma > 1. \quad (9)$$

By combining (8) and (9) with the limits $\lim_{K\uparrow\infty} \rho_0(\gamma, K) = \lim_{K\uparrow\infty} \rho_1(\gamma, K) = 1$ ($\gamma > 0$) and the identity $\rho(1, K) = 1$ (see Lemma 1), we conclude that

$$\lim_{K\to\infty} \rho(\gamma, K) = 1 \quad (10)$$

for any $\gamma > 0$. In other words, we have shown that the optimal arrival rate converges to the service capacity when the buffer size increases to infinity.

The limiting result (10) can be used to find an approximation for the optimal number of robots to be deployed when K is large. Indeed, the relation

$$\lim_{K\to\infty} N(\gamma, K) = \lim_{K\to\infty} \arg \min_{\{n\in\lfloor\mu/\lambda\rfloor, \lceil\mu/\lambda\rceil\}} C(\lambda\, n/\mu, \gamma, K), \quad (11)$$

which follows from (4) and (11), suggests the following approximation, for large K:

$$N(\gamma, K) \sim \begin{cases} \lceil\mu/\lambda\rceil & \text{if } C(\rho_+, \gamma, \infty) \leq C(\rho_-, \gamma, \infty) \\ \lfloor\mu/\lambda\rfloor & \text{if } C(\rho_+, \gamma, \infty) > C(\rho_-, \gamma, \infty) \end{cases} \quad (12)$$

with the notation $C(\rho, \gamma, \infty) := \lim_{K\to\infty} C(\rho, \gamma, K)$.

Since $C(\rho, \gamma, \infty) = \gamma(1 - \rho)$ for $\rho \leq 1$ and $C(\rho, \gamma, \infty) = 1 - 1/\rho$ for $\rho \geq 1$ (use (2)), we may rewrite (12) as

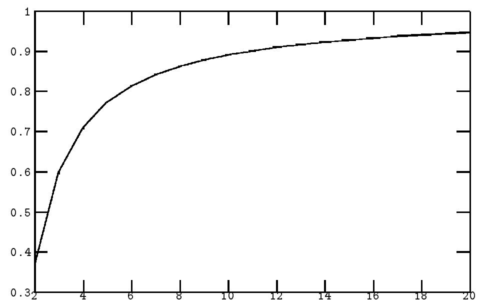
$$N(\gamma, K) \sim \begin{cases} \lceil\mu/\lambda\rceil & \text{if } (\rho_+ - 1)/\rho_+ \leq \gamma\,(1 - \rho_-) \\ \lfloor\mu/\lambda\rfloor & \text{if } (\rho_+ - 1)/\rho_+ \geq \gamma\,(1 - \rho_-) \end{cases} \quad (13)$$

with $\rho_+ := (\lambda/\mu)\,\lceil\mu/\lambda\rceil$ and $\rho_- := (\lambda/\mu)\lfloor\mu/\lambda\rfloor$.

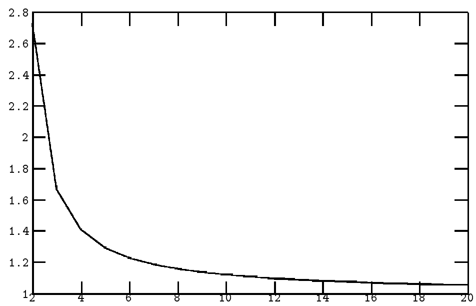
The limiting result (10) may seem counterintuitive at first. Indeed, one may be tempted to argue that the component $P(X = K)$ in the cost function $C(\rho, \gamma, K)$ converges to 0 as the buffer size increases to infinity and to conclude from this that $C(\rho, \gamma, K)$ is minimized when $P(X = 0)$ converges to 0, which occurs when the arrival rate converges to infinity. This interpretation is not correct though, as $\lim_{K\to\infty} P(X = K) = (\rho - 1)/\rho > 0$ when $\rho > 1$ (see Proposition 1).

It is not an easy task to study the behavior of $\rho(\gamma, K)$ as a function of K. We suspect the mapping $K \to \rho(\gamma, K)$ to be increasing when $0 < \gamma < 1$ and decreasing when $\gamma > 1$, but we have not been able to prove it. The conjectured behavior of the mapping $K \to \rho(\gamma, K)$ (resp. $K \to N(\gamma, K)$) is illustrated in Figure 3 (resp. Figure 4). Figure 5 displays the behavior of the optimal number of robots as a function of the ratio μ/λ when the buffer size is infinite. In both curves the parameter γ is held fixed and taken equal to 0.5 and 2, respectively.

We see from Figure 5 that the optimal number of robots $N(\gamma, \infty)$ is equal to 6 when the buffer size is infinite for $\gamma \in \{0.5, 2\}$. Figure 4(b) tells us that for $K \geq 13$ (resp. $K \geq 18$) $N(\gamma, \infty)$ gives the correct value for $N(\gamma, K)$ when $\gamma = 0.5$ (resp. $\gamma = 2$), which seems to indicate that the accuracy of the approximation (13) may be very sensitive to the model parameters.

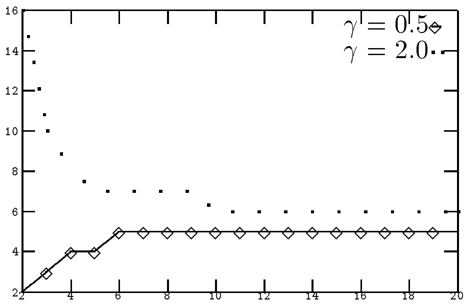


(a) $\gamma = 0.5$

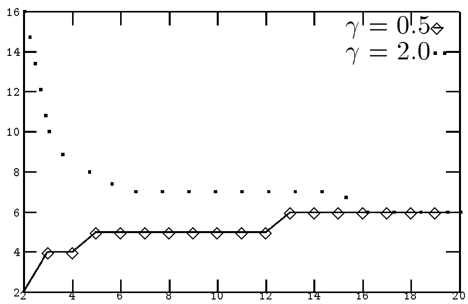


(b) $\gamma = 2$

Figure 3: The mapping $K \to \rho(\gamma, K)$



(a) $\mu/\lambda = 5.7$



(b) $\mu/\lambda = 6$

Figure 4: The mapping $K \to N(\gamma, K)$

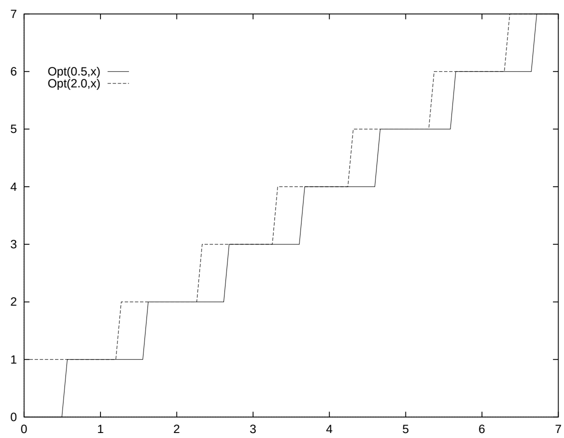


Figure 5: The mapping $\mu/\lambda \to N(\gamma, \infty)$

4 The M/G/1/K queue search engine model

The queueing model in Section 3 is again considered in this section but we now relax the assumption that the service times are exponentially distributed, i.e. we model the search engine as an M/G/1/K queue with service time distribution F.

4.1 Preliminaries

We first introduce additional notation and definitions. For $\Re(\theta) \geq 0$, let $\mathcal{F}(\theta) = \mathbf{E}[\exp(-\theta\sigma)]$ be the Laplace-Stieltjes transform (LST) of the service time distribution.

We denote by $[\omega^n]f$ the coefficient of ω^n in the Taylor series expansion of f. For $\rho \geq 0$, $|\omega| \leq 1$, define

$$\mathcal{G}_\rho(\omega) := \mathcal{F}(\rho\,(1-\omega)/\bar{\sigma}) - \omega. \tag{14}$$

For $\rho \geq 0$, let $\omega_0(\rho)$ be the zero of $\mathcal{G}_\rho(\omega)$ having the smallest modulus, with $\nu(\rho)$ its multiplicity. It is known from Takàcs' lemma [1, pp. 653-654] that $\omega_0(\rho) = 1$ with $\nu(\rho) = 1$, if $\rho < 1$; $\omega_0(\rho) = 1$ with $\nu(1) = 2$; and $\omega_0(\rho) < 1$ with $\nu(\rho) = 1$, if $\rho > 1$.

The cost structure (1) is kept unchanged. As in Section 3, we first calculate (Lemma 3) the unknown probabilities entering the definition of $C(\rho, \gamma, K)$, namely $\mathbf{P}\{X = 0\}$ and $\mathbf{P}^*\{X = K\}$, in the case where X represents the stationary queue length in an M/G/1/K queue with service time distribution F.

Lemma 3 *For* $\rho \geq 0$

$$\mathbf{P}\{X = 0\} = \frac{1}{1 + \rho\,\alpha_K(\rho)} \tag{15}$$

and

$$\mathbf{P}^*\{X = K\} = \mathbf{P}\{X = K\} = \frac{1 + (\rho - 1)\alpha_K(\rho)}{1 + \rho\,\alpha_K(\rho)} \tag{16}$$

with

$$\alpha_K(\rho) := [\omega^{K-2}]\frac{1}{\mathcal{G}_\rho}. \tag{17}$$

◇

Proof. The proof is based on Cohen's analysis of the M/G/1/K queue in [1, Chapter III.6].

Fix $\rho \geq 0$ and introduce

$$B := 1 + \frac{\rho}{2\pi i}\int_{D_r} \frac{1}{\mathcal{G}_\rho(\omega)} \cdot \frac{d\omega}{\omega^{K-1}}$$

with D_r any circle in the complex plane with center 0 and radius strictly less than $\omega_0(\rho)$.

According to [1, p. 575], we have

$$\mathbf{P}\{X = 0\} = \frac{1}{B} \tag{18}$$

and

$$\mathbf{P}\{X = K\} = \frac{1}{2\pi i B}\int_{D_r}\left(\frac{\rho - 1}{\mathcal{G}_\rho(\omega)} + \frac{1}{1-\omega}\right)\frac{d\omega}{\omega^{K-1}}. \tag{19}$$

The integrals in the right-hand sides of (18) and (19) can be evaluated by the theorem of residues [6, p.102] which gives (15) and (16), respectively. The proof is concluded by noting that the first equality in (16) is a consequence of the PASTA property [7]. ∎

By Lemma 3 we can rewrite the cost function (1) as

$$C(\rho, \gamma, K) = 1 + \frac{\gamma - \alpha_K(\rho)}{1 + \rho\,\alpha_K(\rho)} \tag{20}$$

with $\alpha_K(\rho)$ given by (17).

In the next section we devise an algorithm for computing $\alpha_K(\rho)$ in the case where $\mathcal{F}(\theta)$ is rational (e.g. the service time distribution F is phase-type).

Remark 1 *Since* $C(\rho, 1, K) = \mathbf{P}\{X = 0\} + \mathbf{P}\{X = K\} \leq 1$, *we deduce from (20) that* $\alpha_K(\rho) \geq 1$ *for* $\rho \geq 0$. *In particular,* $\alpha_K(0) = 1$.

4.2 Computing the cost function

From now on we assume that $\sigma > 0$ a.s. (service times have no mass at zero) and that the LST of the service times is a *rational function.* More precisely, we consider the situation where

$$\mathcal{F}(\theta) = \frac{B(\theta)}{A(\theta)}$$

with $A(\theta) := \sum_{i=0}^{q} a_i\,\theta^i$ and $B(\theta) := \sum_{i=0}^{r} b_i\,\theta^i$ are relatively prime polynomials (they have no common roots). Since $A(\theta)$ and $B(\theta)$ are relatively prime polynomials the coefficients a_0 and b_0 cannot both be equal to 0 (otherwise $A(\theta)$ and $B(\theta)$ would have the common factor θ); this in turn implies that $b_0 \neq 0$ since $1 = \mathcal{F}(0) \neq 0$. Lastly, we observe that $q > r$ under the condition $\lim_{\theta\to+\infty} \mathcal{F}(\theta) = 0$, which follows from the assumption that $\sigma > 0$ a.s.

Recall the definition of $\mathcal{G}_\rho(\omega)$ (see Lemma 3). The above setting implies that

$$\frac{1}{\mathcal{G}_\rho(\omega)} = \frac{Q(\omega,\rho)}{R(\omega,\rho)} \tag{21}$$

with

$$Q(\omega,\rho) := A(\rho(1-\omega)/\bar{\sigma}) = \sum_{i=0}^{q} a_i(\rho)\,\omega^i \tag{22}$$

$$\begin{aligned} R(\omega,\rho) &:= B(\rho(1-\omega)/\bar{\sigma}) - \omega\,A(\rho(1-\omega)/\bar{\sigma}) \\ &= \sum_{i=0}^{q+1} b_i(\rho)\,\omega^i \end{aligned} \tag{23}$$

where

$$a_i(\rho) := (-1)^i \sum_{j=i}^{q} \binom{j}{i} \frac{a_j}{\bar{\sigma}^j}\,\rho^j, \quad \text{for } 0 \leq i \leq q \tag{24}$$

and

$$\begin{aligned} b_0(\rho) &:= \sum_{j=0}^{r} \frac{b_j}{\bar{\sigma}^j}\,\rho^j \\ b_i(\rho) &:= (-1)^i \left[\sum_{j=i}^{r} \binom{j}{i} \frac{a_j}{\bar{\sigma}^j}\,\rho^j \right. \\ &\qquad \left. + \sum_{j=i-1}^{q} \binom{j}{i-1} \frac{b_j}{\bar{\sigma}^j}\,\rho^j\right], \quad \text{for } 1 \leq i \leq r \\ b_i(\rho) &:= (-1)^{i-1} \sum_{j=i-1}^{q} \binom{j}{i-1} \frac{a_j}{\bar{\sigma}^j}\,\rho^j \\ &\qquad \text{for } r+1 \leq i \leq q+1. \end{aligned} \tag{25}$$

Note that $\{a_i(\rho), b_{i+1}(\rho), 0 \leq i \leq q\}$ are all polynomials of degree q and that $b_0(\rho)$ is a polynomial of degree r.

The next result provides an efficient scheme of convolution type for computing $\alpha_K(\rho)$ and subsequently the cost $C(\rho,\gamma,K)$ in (20).

Lemma 4 *For $K \geq 2$, $\rho \geq 0$,*

(i) $b_0(\rho) \neq 0$;

(ii) $\alpha_K(\rho)$ *can be computed by means of the following recursion:*

$$\alpha_K(\rho) = \begin{cases} \dfrac{a_0(\rho)}{b_0(\rho)} & \text{for } K = 2 \\[2ex] \dfrac{a_{K-2}(\rho)}{b_0(\rho)} - \displaystyle\sum_{i=1}^{K-2} \frac{b_i(\rho)}{b_0(\rho)}\,\alpha_{K-i}(\rho) & \\ \qquad \text{for } 3 \leq K \leq q+2 & \\[2ex] -\displaystyle\sum_{i=1}^{q+1} \frac{b_i(\rho)}{b_0(\rho)}\,\alpha_{K-i}(\rho) & \\ \qquad \text{for } K \geq q+3. & \end{cases} \tag{26}$$

◇

Proof. From (25) and the definition of the polynomial $B(\theta)$ we see that $b_0(\rho) = B(\rho/\bar{\sigma})$. Therefore, $b_0(\rho) = 0$ would imply that $\mathcal{F}(\rho/\bar{\sigma}) = 0$ since $A(\rho\bar{\sigma}) \neq 0$ (as A and B have no common roots), which would yield a contradiction, since $\mathcal{F}(\rho/\bar{\sigma}) = \mathbf{E}[\exp(-\rho\sigma/\bar{\sigma})] > 0$. Consequently, $b_0(\rho) \neq 0$ for all $\rho \geq 0$ and $K \geq 2$.

The proof of (ii) is an easy induction on K using $\alpha_K(\rho) = (1/(K-2)!)\lim_{\omega\downarrow 0} d^{K-2}\mathcal{G}_\rho(\omega)^{-1}/d\omega^{K-2}$. It is omitted for the sake of conciseness. ■

The complexity of computing $\{\alpha_K(\rho),\, K \geq 2\}$ can be reduced even further. Indeed, we observe from (26) that for $K \geq q+3$, $\alpha_K(\rho)$ satisfies a linear relation of order $q+1$, namely,

$$\begin{aligned} \alpha_K(\rho) &= c_1\alpha_{K-1}(\rho) + c_2\alpha_{K-2}(\rho) \\ &\quad + \cdots + c_{q+1}\alpha_{K-q-1}(\rho) \end{aligned} \tag{27}$$

for $K \geq q+3$, with $c_n := -b_n(\rho)/b_0(\rho)$.

We can then invoke the theory of linear relations [5, Theorem 2.2, p. 48] to conclude from the above that

$$\alpha_K(\rho) = \sum_{i=1}^{q+1} \lambda_i^K(\rho) \sum_{j=1}^{\nu_i} K^{j-1} d_{i,j}(\rho), \quad K \geq 2, \tag{28}$$

where $\{\lambda_i(\rho), 1 \leq i \leq m\}$, $1 \leq m \leq q+1$, are the distinct roots of the (characteristic) polynomial

$$P(x) = x^{q+1} - c_1 x^q - c_2 x^{q-1} - \cdots - c_{q+1},$$

with ν_i the multiplicity of λ_i.

The unknown coefficients $\{d_{i,j}(\rho), 1 \leq j \leq \nu_i, 1 \leq i \leq m\}$ in (28) are computed from the initial conditions on $\alpha_2(\rho), \ldots, \alpha_{q+2}(\rho)$ that can themselves be obtained from (26).

Remark 2 *Since $\sum_{i=0}^{q+1} b_i(\rho)$ by definition of $R(\omega, \rho)$ (take $\omega = 1$ in (23)), the linear relation of order $q+1$ in (27) can be reduced to a linear relation of order q. This procedure is illustrated in Section 4.4.*

4.3 Asymptotic behavior of the optimal number of robots

Unlike in the M/M/1/K case, an explicit computation for the optimal number of robots to deploy is out of reach. Even the task of showing the uniqueness of this optimum is non-trivial. In this section, we will content ourselves with the derivation of basic asymptotics.

We first show the existence of a finite optimum.

Lemma 5 *For any $\gamma > 0$, $K \geq 2$, there exist finite real numbers $0 \leq \rho_1(\gamma, K) < \rho_2(\gamma, K) < \cdots < \rho_n(\gamma, K)$, $1 \leq n < \infty$, that minimize the cost $C(\rho, \gamma, K)$ over $[0, \infty)$.* ◇

Proof. It is easily shown by an induction on K using (26) that $\alpha_K(\rho)$ is a rational function in the variable ρ, namely, $\alpha_K(\rho) = f(\rho)/g(\rho)$ with $f(\rho)$ and $g(\rho)$ polynomials of degree q^{K-1} and r^{K-1}, respectively ($g(\rho) = b_0(\rho)^{K-1}$). Since $q > r$ this implies that $\alpha_K(\rho)$ has a limit which is infinite as $\rho \to \infty$; since $\alpha_K(\rho) \geq 1$ for all $\rho \geq 0$ and $K \geq 2$ as pointed out in Remark 1, we deduce that, necessarily,

$$\lim_{\rho \to \infty} \alpha_K(\rho) = +\infty, \tag{29}$$

which together with (20) implies that

$$\lim_{\rho \to \infty} C(\rho, \gamma, K) = 1.$$

On the other hand, (29) also implies that there exists $\rho_0 \geq 0$ such that $\alpha_K(\rho_0) > \gamma$, which in turn implies that $C(\rho_0, \gamma, K) < 1$ from (20). This shows that the mapping $\rho \to C(\rho, \gamma, K)$ reaches is minimum in $[0, \infty)$. The number of points in $[0, \infty)$ where $C(\rho_0, \gamma, K)$ is minimum is finite as a consequence of the fact that $\rho \to C(\rho, \gamma, K)$ is a rational function because $\alpha_K(\rho)$ is. This concludes the proof. ∎

The next result shows that the optimal number of robots increases to infinity as the coefficient γ increases to infinity.

Proposition 4 *For $K \geq 2$,*

$$\lim_{\gamma \to \infty} \rho_1(\gamma, K) = +\infty.$$

◇

Proof. Throughout the proof, $K \geq 2$ is held fixed. We know (see the proof of Lemma 5) that $\alpha_K(\rho) = f(\rho)/b_0(\rho)^{K-1}$, with f a polynomial of degree q^{K-1}. Since $b_0(\rho) \neq 0$ for $\rho \geq 0$ (see Lemma 4), we deduce that the mapping $\rho \to \alpha_K(\rho)$ is continuous on $[0, \infty)$, which in turn implies from (20) and the condition $\alpha_K(\rho) \geq 1$ for all $\rho \geq 0$ (see Remark 1) that, for any $\gamma > 0$, the mapping $\rho \to C(\rho, \gamma, K)$ is also continuous on $[0, \infty)$.

Assume that $\liminf_{\gamma \to \infty} \rho_1(\gamma, K) = L < \infty$. Then there exists a sequence $\{\gamma_n\}_n$ with $\lim_{n \to \infty} \gamma_n = +\infty$ such that $\lim_{n \to \infty} \rho_1(\gamma_n, K) = L$. With the continuity of the mappings $\rho \to \alpha_K(\rho)$ and $\rho \to C(\rho, \gamma, K)$ on $[0, \infty)$, this implies that (see (20))

$$\lim_{n \to \infty} C(\rho_1(\gamma_n, K), \gamma_n, K) = +\infty.$$

To complete the proof we need to show that $\lim_{\gamma \to \infty} C(\rho_1(\gamma, K), \gamma, K) < \infty$ when $\lim_{\gamma \to \infty} \rho_1(\gamma, K) = +\infty$.

We see from (29) that, for any $\gamma > 0$, one can find $\rho = h(\gamma, K)$ such that

$$\mathbf{P}\{X = 0\} = \frac{1}{1 + \rho\,\alpha_K(\rho)} < \frac{1}{\gamma}$$

By definition (2) of the cost $C(\rho, \gamma, K)$ this implies that $C((g(\gamma, K), \gamma, K) \leq 2$ for any $\gamma > 0$, thereby showing that $\lim_{\gamma\to\infty} \rho_1(\gamma, K) = +\infty$ must hold. ∎

We now turn our attention to the analysis of the behavior of the optimal number of robots as the buffer size increases to infinity. Lemma 6 below gives the asymptotics of $\alpha_K(\rho)$ as K gets large. From this result we will deduce the asymptotic behavior of $C(\rho, \gamma, K)$ (Proposition 5) and then the optimal number of robots (Proposition 6) as K goes to infinity.

Lemma 6 *For $\rho \geq 0$,*

$$\alpha_K(\rho) \sim \frac{D(\rho) K^{\nu(\rho)-1}}{\omega_0(\rho)^K} \quad (K \to \infty)$$

where $D(\rho)$ is given by

$$D(\rho) := \begin{cases} -\dfrac{Q(1,\rho)}{R^{(1)}(1,\rho)} & \text{for } 0 \leq \rho < 1 \\[2ex] 2\,\dfrac{Q(1,1)}{R^{(2)}(1,1)} & \text{for } \rho = 1 \\[2ex] -\dfrac{\omega_0(\rho)\, Q(\omega_0(\rho),\rho)}{R^{(1)}(\omega_0(\rho),\rho)} & \text{for } \rho > 1 \end{cases} \tag{30}$$

with $R^{(i)}(\omega,\rho) := \partial^i R(\omega,\rho)/\partial\omega^i$ for $i = 1, 2$. ◇

Lemma 6 follows directly from the definition of $\alpha_K(\rho)$, the coefficient of ω^{K-2} in the Taylor series expansion of $1/\mathcal{G}_\rho(\omega)$ $(K \geq 2)$, the definition of $\omega_0(\rho)$ and $\nu(\rho)$ (see the beginning of Section 4.1), and [5, Theorem 4.1, p. 159].

Proposition 5 *For any $\gamma > 0$*

$$C(\rho,\gamma,K) \sim \begin{cases} (1-\rho)\gamma & \text{for } 0 \leq \rho < 1 \\ 0 & \text{for } \rho = 1 \\ \dfrac{\rho - 1}{\rho} & \text{for } \rho > 1 \end{cases}$$

as $K \to \infty$. ◇

Proof. Definition (20) of $C(\rho, \gamma, K)$ and Lemma 6 already imply that

$$C(\rho,\gamma,K) \sim \begin{cases} 1 + \dfrac{\gamma - D(\rho)}{1 + \rho D(\rho)} & \text{for } 0 \leq \rho < 1 \\ 0 & \text{for } \rho = 1 \\ \dfrac{\rho - 1}{\rho} & \text{for } \rho > 1 \end{cases}$$

as $K \to \infty$. It remains to evaluate $D(\rho)$ for $0 \leq \rho < 1$.

We have (cf. (22), (23), (30))

$$D(\rho) = \frac{1}{1 - \rho H}$$

with $H := (A'(0) - B'(0))/(\bar{\sigma}\,A(0)) = -\mathcal{F}'(0)/\bar{\sigma} = 1$ (hint: $A(0) = B(0)$). Hence, $D(\rho) = 1/(1-\rho)$ and

$$C(\rho,\gamma,K) \sim (1-\rho)\gamma \quad (K \to \infty)$$

for $0 \leq \rho < 1$. ∎

From Proposition 5 we get the following

Proposition 6 *For any $\gamma > 0$, the cost $C(\rho, \gamma, K)$ is minimized at $\rho = 1$ when $K \to \infty$.* ◇

In direct analogy with the M/M/1/K case (see (12)) Proposition 6 suggests the following approximation for the optimal number of robots $N(\gamma, K)$ when K is large:

$$N(\gamma,K) \sim \begin{cases} \lceil \mu/\lambda \rceil & \text{if } C(\rho_+,\gamma,\infty) \leq C(\rho_-,\gamma,\infty) \\[1ex] \lfloor \mu/\lambda \rfloor & \text{if } C(\rho_+,\gamma,\infty) \geq C(\rho_-,\gamma,\infty) \end{cases}$$

or equivalently from Proposition 5

$$N(\gamma,K) \sim \begin{cases} \lceil \mu/\lambda \rceil & \text{if } \dfrac{\rho_+ - 1}{\rho_+} \leq (1-\rho_-)\gamma \\[2ex] \lfloor \mu/\lambda \rfloor & \text{if } \dfrac{\rho_+ - 1}{\rho_+} \geq (1-\rho_-)\gamma. \end{cases}$$

We observe that the above approximation is the same as the one found in the M/M/1/K case (13), thus suggesting that it may not be very accurate in general for moderate values of K. Indeed, we would expect the optimal number of robots to depend on the distribution of the service times and not just its mean.

4.4 Example

In this example the server must complete two tasks: it first checks whether the information contained in a page has changed; if it has, the server updates the data base accordingly.

We assume that the times required to perform the tasks have exponential distributions, with rate μ_1 for the reading task and μ_2 for the updating task. These durations are further assumed to be independent, as well as independent from page to page. Let $0 \leq p \leq 1$ be the probability that an update has to be performed. Note that if $p = 0$ this model reduces to the M/M/1/K queue studied in Section 3.

In this setting the LST of the service time distribution is

$$\begin{aligned}\mathcal{F}(\theta) &= p\,\frac{\mu_1\mu_2}{\mu_1-\mu_2}\left(\frac{1}{\mu_2+\theta}-\frac{1}{\mu_1+\theta}\right)\\ &\quad +(1-p)\,\frac{\mu_1}{\mu_1+\theta}\end{aligned}$$

and the mean service time is

$$\bar{\sigma} = \frac{1}{\mu_1} + \frac{p}{\mu_2}.$$

From the above we readily deduce that (see (14))

$$\frac{1}{\mathcal{G}_\rho(\omega)} = \Big[p\mu_1\mu_2 + (1-p)\mu_1(\mu_2 + f(\omega,\rho)) - \omega(\mu_1 + f(\omega,\rho))(\mu_2 + f(\omega,\rho))\Big] / (\mu_1 + f(\omega,\rho))(\mu_2 + f(\omega,\rho))$$

for $|\omega| < 1$, $\rho > 0$, with $f(\omega,\rho) := \rho(1-\omega)/\bar{\sigma}$.

The coefficients of the polynomials Q and R in (21) are easily identified; we find that

- $a_0(\rho) = (\rho/\bar{\sigma})^2 + (\mu_1+\mu_2)\rho/\bar{\sigma} + \mu_1\mu_2$
- $a_1(\rho) = -2(\rho/\bar{\sigma})^2 - (\mu_1+\mu_2)\rho/\bar{\sigma}$
- $a_2(\rho) = (\rho/\bar{\sigma})^2$

and

- $b_0(\rho) = ((1-p)\mu_1)\rho/\bar{\sigma} + \mu_1\mu_2$
- $b_1(\rho) = -(\rho/\bar{\sigma})^2 + ((1-p)\mu_1 + \mu_1 + \mu_2)\,\rho/\bar{\sigma} + \mu_1\mu_2$
- $b_2(\rho) = 2(\rho/\bar{\sigma})^2 + (\mu_1+\mu_2)\rho/\bar{\sigma}$
- $b_3(\rho) = -(\rho/\bar{\sigma})^2$.

Applying Lemma 4 we see that the unknown quantity $\alpha_K(\rho)$ in (20) is given by

$$\alpha_2(\rho) = \frac{a_0(\rho)}{b_0(\rho)} \tag{31}$$

$$\alpha_3(\rho) = \frac{a_1(\rho)}{b_0(\rho)} - \frac{a_0(\rho)b_1(\rho)}{b_0(\rho)^2} \tag{32}$$

$$\begin{aligned}\alpha_4(\rho) &= \frac{a_2(\rho)}{b_0(\rho)} - \frac{a_1(\rho)b_1(\rho)}{b_0(\rho)^2}\\ &\quad + \frac{a_0(\rho)b_1(\rho)^2}{b_0(\rho)^3} - \frac{a_0(\rho)b_2(\rho)}{b_0(\rho)^2}\end{aligned} \tag{33}$$

and

$$\begin{aligned}\alpha_K(\rho) &= -\frac{b_1(\rho)}{b_0(\rho)}\,\alpha_{K-1}(\rho) - \frac{b_2(\rho)}{b_0(\rho)}\,\alpha_{K-2}(\rho)\\ &\quad -\frac{b_3(\rho)}{b_0(\rho)}\,\alpha_{K-3}(\rho) \quad \text{for } K \geq 5.\end{aligned} \tag{34}$$

As pointed out in Remark 2, the identity $\sum_{i=0}^{3} b_i(\rho) = 0$ may be used to reduce the order of the linear relation (34). Define

$$u_K(\rho) := \alpha_K(\rho) - \alpha_{K-1}(\rho) \tag{35}$$

for $K \geq 3$. Then (34) becomes

$$\begin{aligned}u_K(\rho) &= -\left(1 + \frac{b_1(\rho)}{b_0(\rho)}\right) u_{K-1}(\rho)\\ &\quad + \frac{b_3(\rho)}{b_0(\rho)}\,u_{K-2}(\rho) \quad \text{for } K \geq 5.\end{aligned}$$

The solution to this linear relation of order 2 is given by [5, Theorem 2.2, p. 48]

$$u_K(\rho) = \lambda_1(\rho)^K\,d_1(\rho) + \lambda_2(\rho)^K\,d_2(\rho) \quad \text{for } K \geq 3, \tag{36}$$

with $\lambda_1(\rho)$ and $\lambda_2(\rho)$ the (distinct) roots of the polynomial $b_0(\rho)x^2 + (b_0(\rho) + b_1(\rho))x - b_3(\rho)$, namely,

$$\lambda_i(\rho) = \frac{-(b_0(\rho) + b_1(\rho)) \pm \sqrt{\Delta(\rho)}}{2\, b_0(\rho)}, \quad i = 1, 2,$$

with $\Delta(\rho) := (b_0(\rho) + b_1(\rho))^2 + 4\, b_0(\rho) b_3(\rho)$. (The proof that $\Delta(\rho) > 0$ for all $\rho > 0$ is left to the reader.) The coefficients $d_i(\rho)$ in (36) are computed from the initial conditions $u_2(\rho)$ and $u_3(\rho)$. We find

$$\begin{aligned} d_1(\rho) &= \frac{\lambda_2(\rho)\,(\alpha_3(\rho) - \alpha_2(\rho)) - (\alpha_4(\rho) - \alpha_3(\rho))}{\lambda_1(\rho)^3\,(\lambda_2(\rho) - \lambda_1(\rho))} \\ d_2(\rho) &= \frac{\alpha_4(\rho) - \alpha_3(\rho) - \lambda_1(\rho)\,(\alpha_3(\rho) - \alpha_2(\rho))}{\lambda_2(\rho)^3\,(\lambda_2(\rho) - \lambda_1(\rho))} \end{aligned}$$

where $\alpha_i(\rho)$ $(i = 2, 3, 4)$ are given in (31)-(33).

Combining (35) and (36), we finally get

$$\begin{aligned} \alpha_K(\rho) &= \alpha_2(\rho) + d_1(\rho) \sum_{i=2}^{K} \lambda_1(\rho)^i \\ &\quad + d_2(\rho) \sum_{i=2}^{K} \lambda_2(\rho)^i, \quad K \geq 2. \end{aligned}$$

Figure 6 represents what we have found to be typical behavior of the cost function $C(\rho, \gamma, K)$ as a function of ρ. One can observe that the minimum is unique and obtained, say, at $\rho = \rho(\gamma, K)$. We have computed $\rho(\gamma, K)$ for various values of the parameters γ, K and the probability p that a page has to be updated.

Figure 7 displays the mapping $K \to \rho(\gamma, K)$ for two values of the probability p ($p = 1$ and $p = 0.5$) and $\gamma = 1$; Figure 8 displays the mapping $K \to \rho(\gamma, K)$ for two values of γ ($\gamma = 0.5$ and $\gamma = 2$) and $p = 1$. As in the M/M/1/K case, we observe in Figure 8 that $\rho(\gamma, K)$ is increasing when $\gamma < 1$ and decreasing when $\gamma > 1$.

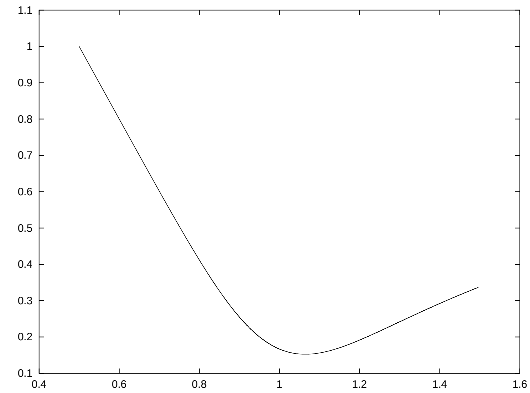
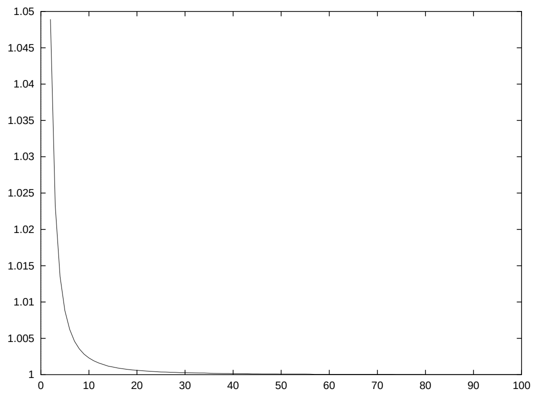
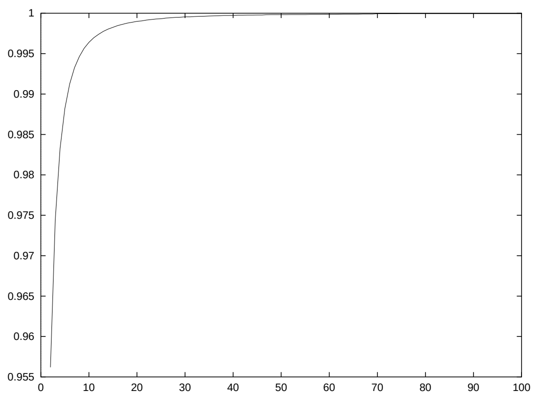


Figure 6: The mapping $\rho \to C(\rho, \gamma, K)$ with $\mu_1 = 0.5$, $\mu_2 = 0.1$, $p = 1$,$\gamma = 2$ and $K = 15$



$p = 1$



$p = 0.5$

Figure 7: The mapping $K \to \rho(\gamma, K)$ with $\gamma = 1$

5 Concluding remarks

Simple queueing models (the M/M/1/K and M/G/1/K queues) of search engines have been proposed, analyzed, and optimized in order to find the optimal number of robots to use. The cost function is a weighted sum of the loss probability and the starvation probability.

Extensions of these models to dynamic models where the number of active robots may change over time as a function of the workload in the queue have been proposed in a companion paper [4].

Several interesting, open issues remain, including the situation where the robots are not homogeneous and/or are allocated to different parts of the network. For instance, one may wish to determine the optimal number of robots to be allocated to a given area.

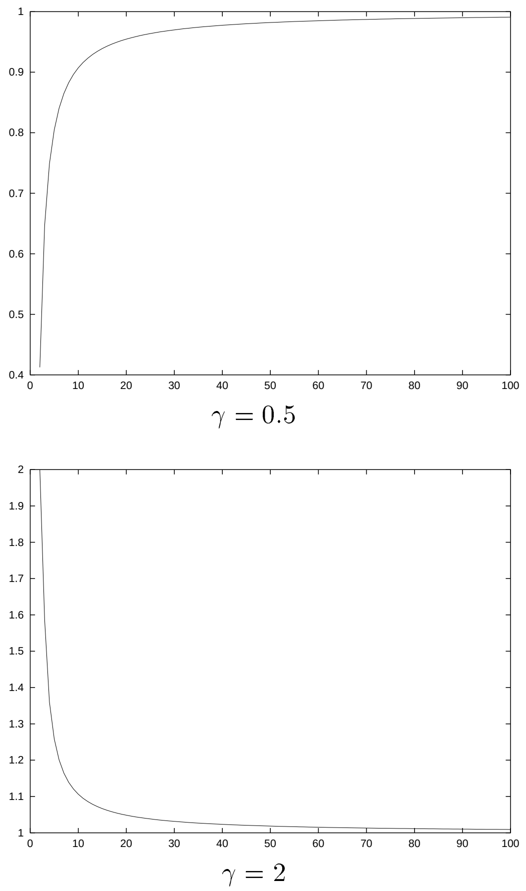


Figure 8: The mapping $K \rightarrow \rho(\gamma, K)$ with $p = 1$

References

[1] Cohen, J. W ., *The Single Server Queue*, North-Holland Publishing Company, 1982.

[2] Coffman Jr., E. G, Liu, Z. and Weber, R. R., “Optimal robot scheduling for web search engines”, *J. Scheduling*, **1**, pp. 14-22, 1998.

[3] Kleinrock, L., *Queueing Systems, Vol. I*, Wiley & Sons, New York, 1975.

[4] Talim, J., Liu, Z., Nain, P. and Coffman, Jr., E. G., “Optimal number of robots for search engines: the dynamic case.” *Preprint*, Oct. 1997.

[5] Sedgewick, R. and Flajolet, P., *Analysis of algorithms*, Addison-Wesley Publishing Company, 1996.

[6] Titchmarsh, E. C., *The Theory of Functions*, Oxford University Press, 2nd Edition, 1939.

[7] Wolff, R. L., “Poisson Arrivals See Time Averages,” *Opns. Res.*, **30**, pp. 223-231, 1982.