Performance Evaluation of Proactive Queues

Raymond A. Marie

Performance Evaluation of Proactive Queues

Raymond A. Marie

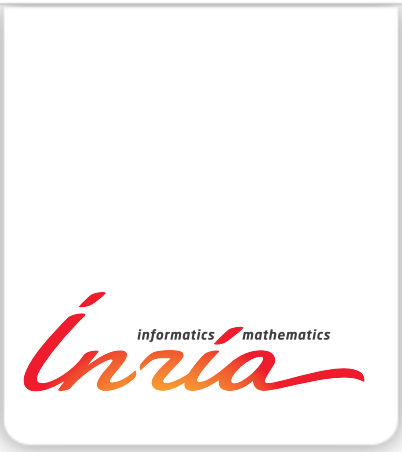
RESEARCH REPORT

N° 9496

January 2023

Project-Team Ermine

ISSN 0249-6399 ISRN INRIA/RR--9496--FR+ENG



Performance Evaluation of Proactive Queues

Raymond A. Marie

Project-Team Ermine

Research Report n 9496 — January 2023 — 65 pages

Abstract: Some service systems can be seen as proactive systems that anticipate future service demand. In this report, we present some basic proactive systems from the perspective of the performance criteria used for classical queues. The resources may anticipate services up to a given limit. This limit is a key factor with respect to the efficiency of the system under study. The performance variables are mainly service time, waiting time and response time; the metrics are mainly the probability distributions and the expectations of those, and also conditional distributions and expectations. According to the relative values of the limit of anticipation and of the number of servers, the results are more or less simple. We start this study with the case of the proactive single-server queue (M/M/1/<C>); the request arrival process is Poisson and the independent service times follow an exponential distribution; C denotes the limit of anticipation. Then, we study the proactive M/M/r/<C> queue; we have to consider two different situations, the easy case being when r is lower than $(C+1)$. Finally, we generalize the previous study by considering the case where the arrival rates and service rates depend on the state of the queue.

Key-words: Performance evaluation, Analytical models, Stochastic modeling, Queueing models, proactive queue.

R. A. Marie is with University of Rennes /Irisa. E-mail: Raymond.Marie@irisa.fr.

RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

Evaluation de performance de files proactives

Résumé : Certains systèmes de service peuvent être considérés comme des systèmes proactifs qui anticipent la demande de service future. Dans ce rapport, nous présentons quelques systèmes proactifs de base du point de vue des critères de performance utilisés pour les files d'attente classiques. Les ressources peuvent anticiper les services jusqu'à une limite donnée. Cette limite est un facteur clé en ce qui concerne l'efficacité du système étudié. Les variables de performance sont principalement le temps de service, le temps d'attente et le temps de réponse ; les métriques sont principalement les distributions de probabilité et les espérances de celles-ci, ainsi que les distributions et les espérances conditionnelles. En fonction des valeurs relatives de la limite d'anticipation et du nombre de serveurs, les résultats sont plus ou moins simples. Nous commençons cette étude par le cas de la file d'attente proactive à serveur unique (M/M/1/<C>) ; le processus d'arrivée des demandes est poissonien et les temps de service indépendants suivent une distribution exponentielle ; C dénote la limite d'anticipation. Ensuite, nous étudions la file d'attente proactive M/M/r/<C> ; nous devons considérer deux situations différentes, le cas facile étant celui où r est inférieur à $(C + 1)$. Enfin, nous généralisons l'étude précédente en considérant le cas où les taux d'arrivée et les taux de service dépendent de l'état de la file d'attente.

Mots-clés : Evaluation de performance, Modèles analytiques, Modélisation stochastique, Files d'attente, Files proactives.

1 Introduction

Some service systems can be seen as proactive systems that anticipate future service demand. Perhaps the simplest paradigm is that of the pizza vendor who may decide to start production of one or more pizzas when he has no more orders pending. In a more professional context, let's consider an ad-hoc network of mini-sensors sending data in the form of messages transmitted through the network to an operating station. The energy is provided by a battery recharged by photovoltaic cells ([1]). The transmission of a message requires a quantum of energy and the fully charged battery contains a limited number of quanta. In case of strong activity of the sensor or in case of bad weather conditions, the battery discharges and it is necessary to wait to recover a quantum of energy to send the possible pending messages. In this example, the server is the photovoltaic unit that provides the quanta and the users are the messages to be transmitted.

In an industrial production context, a cell included in a Kanban chain can also be seen as providing a proactive service. The cell under consideration can contain several machines in parallel, implemented to satisfy the demand as well as possible, without exceeding a maximum level of anticipation C. The machines are the resources (the servers) and the clients are the requests of the cell located downstream. By persevering in this way, we can include storage systems; thus, even if he does not manufacture anything, the car dealer anticipates future demand by ordering cars whose future owners he does not yet know.

In the context of integrated logistical support (ILS), let us consider the example of a maintenance workshop where a sub-assembly of a given piece of equipment is repaired, for example the engine of an aircraft. Two policies (noted respectively A and B) are possible. With policy A, the engine taken in hand by a repair person will be reassembled on the original aircraft. With policy B, the engine is disconnected from the aircraft and the stripped aircrafts are placed in a discipline queue (first in, first out) and the first engine to be repaired is reassembled on the aircraft at the head of the queue. When technically practicable, this policy B makes sense if the maintenance center is initially allocated a stock of C engines in good condition. Note that these two policies do not provide the same performance in terms of aircraft unavailability. With policy A, the engine repair time corresponds to the service time of the customer (here the aircraft). If there are r repairers (or r repair teams), the whole set behaves like a classical $M/M/r$ queue. The aircraft downtime is the sum of a possible waiting time for lack of available repairers and the engine repair time. With policy B, the set is proactive and the service seen by the customers (the airplanes) can be very short, or even almost zero if the time for the engine to be disengaged from the airplane is negligible.

The purpose of this research report is to present some basic proactive systems from the perspective of the performance criteria used for classical queues. We start by studying the case of the single-server queue (next Section); the request arrival process is Poissonian and the independent service times follow an exponential distribution. In Section 3, we study the proactive $M/M/r$ queue. Then, in the section 4, we generalize the study of the previous section to the queue by considering the case where the arrival rates and service rates depend on the state of the queue. Finally, Section 5 expose our conclusions.

2 Single-server Queue M/M/1/<C>

2.1 Modelization

When the client queue is empty, the proactive server can produce up to C items/services. C is the value of the allowed anticipation window. The diagram in Figure 1 illustrates the operation of the queue. The queue has an operator of type "*Join*".

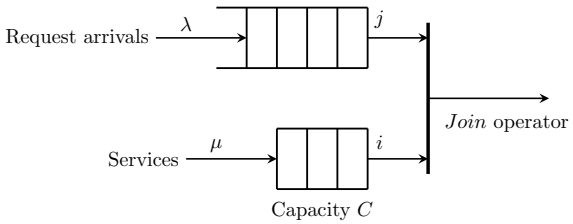


Figure 1: Schematization of the proactive single-server queue.

The arrivals of the requests follow a Poissonian process of rate λ and the service times are independent and identically distributed (*iid*) according to an exponential distribution of rate μ. The parameter C corresponds to the maximum number of anticipated services. The variables i and j denote respectively the number of services waiting for a customer and the number of customers waiting for a service. Note that, by construction, the product of the values taken by the variables i and j is zero for any state of the queue. Both variables are simultaneously zero when the server anticipates a first service. If a client shows up before the end of this anticipated service, the client must wait until the end of the service. If j clients are present in the queue, they are served according to the discipline "first come - first served" (PAPS). These assumptions will allow us to consider a Markov model of the queue.

The variable k defined by the equality

$$k \stackrel{\triangle}{=} (C - i + j)\,, \tag{1}$$

makes it possible to determine in an ordered form the possible states of this waiting system. The values of k less than C corresponding to the existence of anticipated services and the values of k greater than C corresponding to customers waiting (or in service for the lead customer). The variable k equals zero when all the anticipable services are available ($i = C$). It is equal to C when both no customer is present ($j = 0$) and, on the other hand, no advance service is yet available ($i = 0$). For a given value k, the associated values i and j are characterized by the relations

$$i = (C - k)^+ \tag{2}$$
$$j = (k - C)^+ \tag{3}$$

where notation $(x)^+$ means $\max(0, x)$.

2.2 Performances Determination

Let us denote p_k the asymptotic probability of state k, $k = 0, 1, 2, \ldots$. First we determine these asymptotic state probabilities (*cf.* Figure 2).

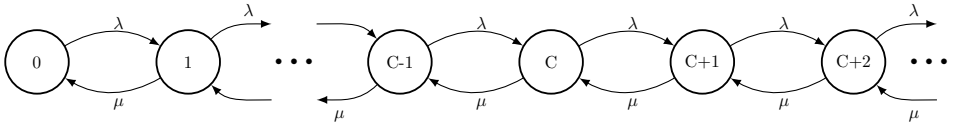


Figure 2: CMTC associated to the proactive single-server queue.

It is easy to check, after having put $\rho = \lambda/\mu$, that the steady state state probabilities are written

$$p_k = (1-\rho)\rho^k, \qquad k = 0, 1, 2, \ldots$$

under the necessary ergodic condition $\rho < 1$.

In order to determine the expectation of the response time, let's start by determining the expectation of the number of requests present in the system:

$$\begin{aligned}
\mathbb{E}[N_C] &= \sum_{j=1}^{\infty} j p_{C+j} = \sum_{j=1}^{\infty} j(1-\rho)\rho^{C+j}\,, \\
&= (1-\rho)\rho^{C+1} \sum_{j=1}^{\infty} j\rho^{j-1} = (1-\rho)\rho^{C+1} \left(\sum_{j=1}^{\infty} \rho^j \right)' \,, \\
&= (1-\rho)\rho^{C+1} \left(\frac{1}{(1-\rho)} \right)' = (1-\rho)\rho^{C+1} \frac{1}{(1-\rho)^2}\,, \\
&= \frac{\rho^{C+1}}{(1-\rho)}\,.
\end{aligned}$$

Remembering that for a file $M/M/1$, the expectation $\mathbb{E}[N]$ is equal to $\rho/(1-\rho)$, we notice that this last expectation is related to $\mathbb{E}[N_C]$ via the relation:

$$\mathbb{E}[N_C] = \rho^C \mathbb{E}[N_C]|_{C=0} = \rho^C \mathbb{E}[N]\,.$$

Let's use Little's formula to deduce the expectation of the response time:

$$\mathbb{E}[R_C] = \frac{\mathbb{E}[N_C]}{\lambda} = \frac{1}{\mu} \frac{\rho^C}{(1-\rho)} = \rho^C \mathbb{E}[R_C]|_{C=0}\,. \tag{4}$$

With this queue, a customer will receive an immediate service if, upon arrival, i is strictly positive. If i and j are both zero, it will be enough for him to wait until the end of the (previously anticipated) service to exit the system. If the value of the variable j is strictly positive when it arrives, the client will have to wait for the server to finish serving

the customer (s) already present before getting the server for himself. The probability that the wait for server availability is strictly positive is the probability that the client finds the system in a k state strictly greater than C; By complementarity, if W_C denotes the waiting time to obtain server availability, we have:

$$\begin{aligned}\mathbb{P}(W_C = 0) &= \sum_{j=0}^{C} p_j \,, \\ &= (1-\rho)\sum_{j=0}^{C} \rho^j = (1-\rho)\frac{1-\rho^{C+1}}{(1-\rho)} \,, \\ &= (1-\rho^{C+1}) \,. \end{aligned} \tag{5}$$

The probability $\mathbb{P}(S_C = 0)$ that the response time is itself zero must be less than the probability $\mathbb{P}(W_C = 0)$ since the event $[S_C = 0]$ is included in the event $[W_C = 0]$. The sought probability is that the client finds the system in a state k such that the variable i is strictly positive:

$$\begin{aligned}\mathbb{P}(S_C = 0) &= \sum_{j=0}^{C-1} p_j \,, \\ &= (1-\rho)\sum_{j=0}^{C-1} \rho^j = (1-\rho)\frac{1-\rho^{C}}{(1-\rho)} \,, \\ &= (1-\rho^{C}) \,. \end{aligned} \tag{6}$$

In Figure 3, we have represented the probabilities that a service or waiting time is zero as a function of C, for C = 0. 1, 2, ..., and 5, when $\rho = 0.75$ and $\mu = 0.5$. These two probabilities are increasing according to the C level of anticipation. It can be verified, as the previous results have shown, that $\mathbb{P}(S_{C+1} = 0) = \mathbb{P}(W_C = 0)$.

Given that the service time is exponentially distrbuted with rate μ, the expectation of the service time perceived by the customers is given by

$$\mathbb{E}[S_C] = \frac{1}{\mu}\mathbb{P}(S_C > 0) = \frac{1}{\mu}(1-(1-\rho^C)) = \frac{\rho^C}{\mu} = \rho^C \mathbb{E}[S_C]|_{C=0} \,. \tag{7}$$

Since the response time remains equal to the sum of the waiting time for the server (possibly zero time) and the service time perceived by the client (possibly zero time), we can determine the expectation of waiting $\mathbb{E}[W_C]$ subtracting $\mathbb{E}[S_C]$ to $\mathbb{E}[R_C]$:

$$\begin{aligned}\mathbb{E}[W_C] &= \mathbb{E}[R_C] - \mathbb{E}[S_C] \,, \\ &= \frac{\rho^C}{\mu(1-\rho)} - \frac{\rho^C}{\mu} = \frac{\rho^C}{\mu}\left(\frac{\rho}{(1-\rho)} - 1\right) \,, \end{aligned}$$

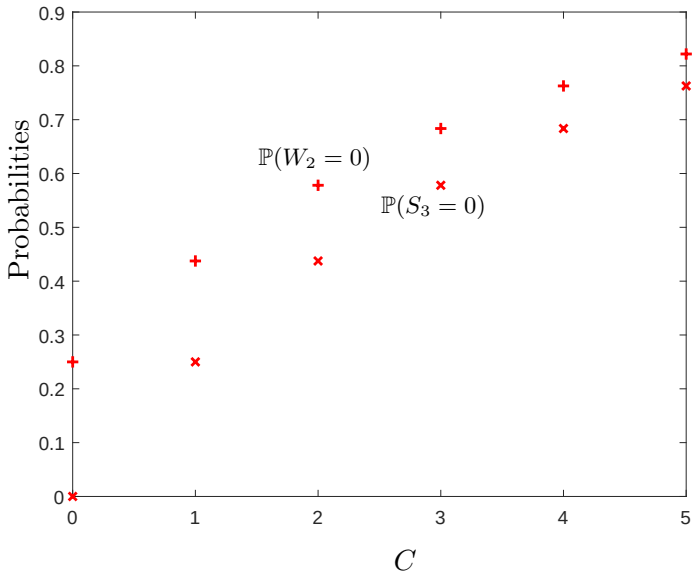


Figure 3: Probabilities of null service times (x) or of null waiting times (+) as a function of C, for C=0, 1, 2, ..., 5; with $\rho = 0.75$ and $\mu = 0.5$.

$$\begin{aligned} &= \rho^C \frac{\rho}{\mu(1-\rho)}\,, \\ &= \rho^C \mathbb{E}[W_C]|_{C=0}\,. \end{aligned} \tag{8}$$

In this single-server queue, the ρ parameter is necessarily less than 1. This causes ρ^C to decrease as the C capacity increases. Thus, the expectations $\mathbb{E}[N_C]$, $\mathbb{E}[R_C]$, $\mathbb{E}[W_C]$ and $\mathbb{E}[S_C]$ decreases according to the same proportion when increasing the C capacity.

In Figure 4, we have represented the evolution of the expectations $\mathbb{E}[W_C]$ and $\mathbb{E}[R_C]$ as a function of C , for C = 0. 1, 2, ..., and 5, when $\rho = 0.75$ and $\mu = 0.5$. As we could expect, these two expectations are decreasing according to the C level of anticipation. For a given value C, the difference $(\mathbb{E}[R_C] - \mathbb{E}[W_C])$ is equal to the expectation of the service time $\mathbb{E}[S_C]$ that decreases when C increases (*cf.* the result 7) ; for $C = 0$, the difference is equal to $1/\mu = 2$. It can also be verified, as the previous results have shown, that $\mathbb{E}[R_{C+1}] = \mathbb{E}[W_C]$.

Let's now look at the conditional expectation of waiting time knowing that the client did not find the server available when he arrived. For that, let's start from the following relation:

$$\mathbb{E}[W_C] = \mathbb{E}[W_C|W_C > 0]\mathbb{P}(W_C > 0) + \mathbb{E}[W_C|W_C = 0]\mathbb{P}(W_C = 0)\,,$$

which, since the conditional expectation $\mathbb{E}[W_C|W_C = 0]$ is null, allows to write:

$$\mathbb{E}[W_C|W_C > 0] = \frac{\mathbb{E}[W_C]}{\mathbb{P}(W_C > 0)}\,.$$

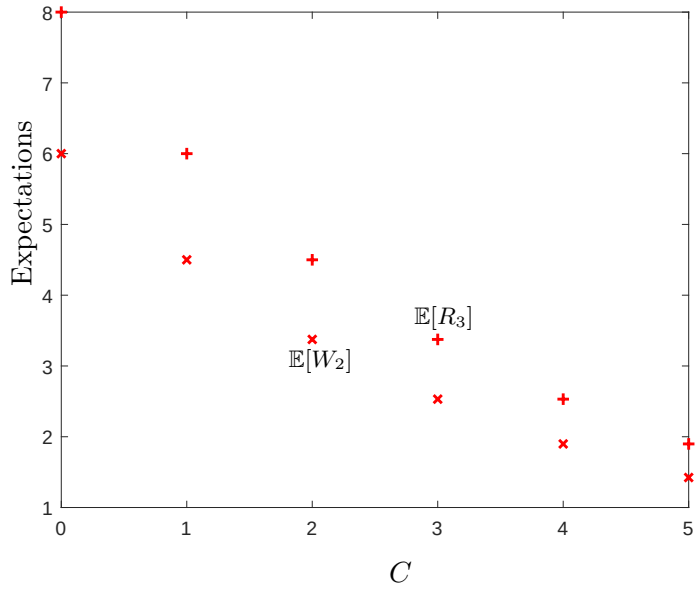


Figure 4: Expectations of waiting times (x) and of response times (+) as a function of C, for C=0, 1, 2, ..., 5; with $\rho = 0.75$ and $\mu = 0.5$.

In the case of the $M/M/1$ queue, it is not difficult to find that this conditional expectation is $\mathbb{E}[W|W > 0] = 1/\mu(1-\rho)$. For the queue with a proactive server, we get:

$$\begin{aligned}
\mathbb{E}[W_C|W_C > 0] &= \frac{\mathbb{E}[W_C]}{\mathbb{P}(W_C > 0)}\,, \\
&= \rho^C \frac{\rho}{\mu(1-\rho)} \frac{1}{\rho^{C+1}}\,, \\
&= \frac{1}{\mu(1-\rho)}\,, \\
&= \mathbb{E}[W_C|W_C > 0]|_{C=0} = \mathbb{E}[W|W > 0]\,.
\end{aligned}$$

We observe that this conditional expectation is identical to that of the classical $M/M/1$ queue. This result is not surprising because, if the client finds one or more clients when they arrive, everything happens as if he was in a non-proactive $M/M/1$ queue.

In the same way, since the probability $\mathbb{P}(R_C = 0)$ must be equal to the probability $\mathbb{P}(S_C = 0)$, we obtain from expressions 4 and 6 the expression of conditional expectation $\mathbb{E}[R_C|R_C > 0]$:

$$\begin{aligned}
\mathbb{E}[R_C|R_C > 0] &= \frac{\mathbb{E}[R_C]}{\mathbb{P}(R_C > 0)} = \frac{\mathbb{E}[R_C]}{\mathbb{P}(S_C > 0)}\,, \\
&= \frac{\rho^C}{\mu(1-\rho)} \frac{1}{\rho^C} = \frac{1}{\mu(1-\rho)}\,, \\
&= \mathbb{E}[R_C|R_C > 0]|_{C=0} = \mathbb{E}[R_C]|_{C=0}\,,
\end{aligned}$$

$$= \quad \mathbb{E}[R] = \mathbb{E}[W_C | W_C > 0]\,.$$

Note that this conditional expectation is identical to that of the classical $M/M/1$ queue for which the response time is positive with probability 1. It is also equal to the conditional expectation of the waiting time, for any value of C. Recall that, for the classical $M/M/1$ queue, the unconditional expectation $\mathbb{E}[W]$ is related to the conditional expectation of the waiting time by the relation $\mathbb{E}[W] = \rho\,\mathbb{E}[W|W > 0]$. Let's go further and look for the probability distribution of W_C. We already know that it is a mixed distribution with a mass at zero, equal to $(1-\rho^{C+1})$. We will consider any customer but to clarify the situation, we will give a name to this customer chosen purely randomly: August. Since arrivals follow a Poisson process, the probability that August finds j clients in front of him on arrival is equal, when j is strictly positive, to $(1-\rho)\rho^{C+j}$. Indeed, Poisson arrivals find the system in equilibrium (this is the property PASTA : *Poisson Arrivals See Time Averages*). Knowing that he finds j clients in front of him, the waiting time experimented by August is distributed according to the Erlang distribution of order j and of parameter μ whose density is written:

$$f_X(t) = \frac{\mu^j t^{j-1}}{(j-1)!} e^{-\mu t}, \qquad t \geq 0\,.$$

Which makes it possible to write the unconditional distribution:

$$\begin{aligned}
dF_{W_C}(t) &= (1-\rho^{C+1})\delta(t) + \sum_{j=1}^{\infty}(1-\rho)\rho^{C+j}\frac{\mu^j t^{j-1}}{(j-1)!}e^{-\mu t}\,,\\
&= (1-\rho^{C+1})\delta(t) + (1-\rho)\rho^C e^{-\mu t}\sum_{j=1}^{\infty}\frac{\lambda^j t^{j-1}}{(j-1)!}\,,\\
&= (1-\rho^{C+1})\delta(t) + (1-\rho)\rho^C e^{-\mu t}\lambda\sum_{i=0}^{\infty}\frac{(\lambda t)^i}{i\,!}\,,\\
&= (1-\rho^{C+1})\delta(t) + \frac{(\mu-\lambda)}{\mu}\rho^C\lambda e^{-(\mu-\lambda)t}\,,\\
&= (1-\rho^{C+1})\delta(t) + \rho^{C+1}(\mu-\lambda)e^{-(\mu-\lambda)t}\,.
\end{aligned}$$

where $\delta(t)$ denotes the Dirac function.

Thus, the probability distribution of W_C is a mixed probability distribution with a mass at zero, equal to $(1-\rho^{C+1})$, and a continuous part over the domain $]0, +\infty]$ corresponding to the exponential with rate $(\mu-\lambda)$ weighted by the cœfficient ρ^{C+1}.

Note that if $C = 0$, we find the result known for the non-proactive $M/M/1$ queue. The conditional probability distribution of the waiting time knowing that the client has waited is therefore the exponential distribution with rate $(\mu-\lambda)$ as well for the proactive queue as for the $M/M/1$ standard queue.

the cumulative probability distribution of W_C can be written:

$$F_{W_C}(t) = (1-\rho^{C+1})u(t) + \rho^{C+1}(1-e^{-(\mu-\lambda)t})\,. \tag{9}$$

In the same way, we can determine the probability distribution of the response time R_C. This mixed probability distribution also has a mass in zero, equal to $(1-\rho^C)$, and the probability that August finds j clients in front of him on his arrival remains equal to $(1-\rho)\rho^{C+j}$. Given its own service, the response time is distributed according to the Erlang distribution of order $(j+1)$ and of parameter μ. Note that if $j=0$, the response time corresponds to the end of the service started by the server before the August arrival time; but the service time being distributed according to the exponential distribution of rate μ, it is the same for this residual service time. We thus have:

$$\begin{aligned}
dF_{R_C}(t) &= (1-\rho^C)\delta(t) + \sum_{j=0}^{\infty}(1-\rho)\rho^{C+j}\frac{\mu^{j+1}t^j}{j!}e^{-\mu t}, \\
&= (1-\rho^C)\delta(t) + \sum_{i=1}^{\infty}(1-\rho)\rho^{C-1+i}\frac{\mu^i t^{i-1}}{(i-1)!}e^{-\mu t}, \\
&= (1-\rho^C)\delta(t) + (1-\rho)\rho^{C-1}e^{-\mu t}\sum_{i=1}^{\infty}\frac{\lambda^i t^{i-1}}{(i-1)!} \\
&= (1-\rho^C)\delta(t) + (1-\rho)\rho^{C-1}e^{-\mu t}\lambda\sum_{k=0}^{\infty}\frac{(\lambda t)^k}{k\,!} \\
&= (1-\rho^C)\delta(t) + \frac{(\mu-\lambda)}{\mu}\rho^{C-1}\lambda e^{-(\mu-\lambda)t} \\
&= (1-\rho^C)\delta(t) + \rho^C(\mu-\lambda)e^{-(\mu-\lambda)t}\,,
\end{aligned}$$

Again, the continuous part of this mixed probability distribution corresponds to the exponential distribution of rate $(\mu-\lambda)$, weighted here by the factor ρ^C. Let's note that by incrementing the allowed anticipation window from C to $(C+1)$, the probability distribution of the response time R_{C+1} is that of the waiting time W_C. Thus the cumulative distribution function of R_C can be written:

$$F_{R_C}(t) = (1-\rho^C)u(t) + \rho^C(1-e^{-(\mu-\lambda)t})\,. \tag{10}$$

This cumulative distribution function is represented in Figure 5, as a function of C, C= 0, 1, 2, 3, and 4; with the same parameter values as in the previous figure: $\rho = 0.75$ and $\mu = 0.5$. The complementary cumulative distribution function $\bar{F}_{R_C}(t) = \mathbb{P}(R_C > t)$ is equal to the decreasing function $\rho^C e^{-(\mu-\lambda)t}$. For an arbitrary value t, we have $\bar{F}_{R_{C+1}}(t) = \rho\bar{F}_{R_C}(t)$ and as ρ is necessarily less than one, the gap between the curves is reduced when C increases.

It follows from the relations 9 and 10 that $F_{W_C}(t) = F_{R_{C+1}}(t)$. Thus, Figure 5 also represents the distribution function $F_{W_C}(t)$ for values $C = 0, 1, 2$ and 3 ; This explains why we did not previously use a figure in order to illustrate the distribution function $F_{W_C}(t)$.

3 Multi-server Queue M/M/r/<C>

In the presence of several servers, the study of the queue becomes more complex but we will see that it is still possible to determine many performance metrics. The queue that we

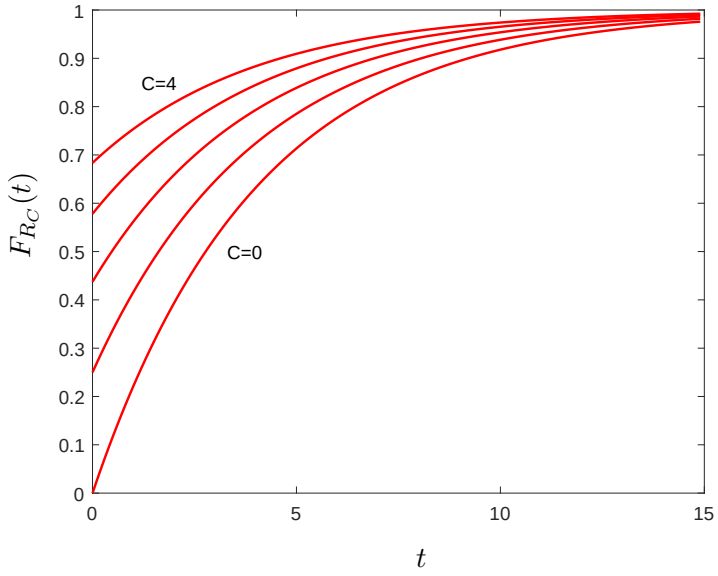


Figure 5: Cumulative distribution function of R_C, $F_{R_C}(t)$,as a function of C, for C=0. 1, 2, 3, and 4; with $\rho = 0.75$ and $\mu = 0.5$.

characterize by the quadruplet M/M/r/$< C >$ where r denotes the number of servers and the last element still denotes the maximum number of anticipatable services, corresponds to a Markovian model and will be easier to study if r is not greater than $(C + 1)$.

3.1 Modelization

We model this new queue by defining the variables i, j and k in a similar way to the previous case (see Figure 1 (page 12) as well as the relation 1) (page 4)). As there are several servers, it is necessary to specify the service procedure when the variable i is strictly positive, so that the maximum number of anticipated services does not exceed the limit C. We can realize the procedure thanks to C tokens representing as many authorizations to anticipate a service, the token being released when a client who presents himself can benefit from the anticipated service. Note that such a procedure is easily modeled using a timed stochastic Petri net. For a given value k, the associated values i and j are characterized by the same relations as in the mono-server case:

$$i = (C - k)^+ \quad (11)$$
$$j = (k - C)^+ \quad (12)$$

where again notation $(x)^+$ means $\max(0, x)$.

The diagram in Figure 6 illustrates the operation of the queue in this multiserver case where the number r of servers may be less than or greater than C. In order to further explore this alternative, let us note n_a the number of active servers (busy with a proactive

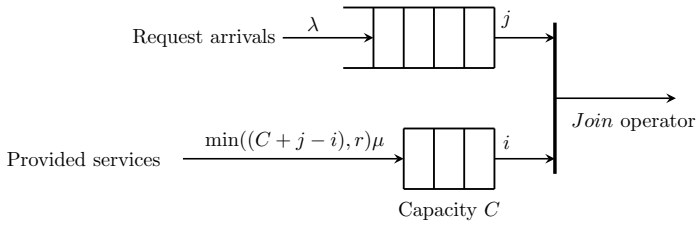


Figure 6: Schematization of the proactive multi-server queue.

task or not) and n_{pa} the number of proactive servers. Figure 7 illustrates, in the case where C is greater than or equal to $(r-1)$, the values of variables i, j, n_a and n_{pa} based on k. The dashed curve represents the number n_a of active servers; the number of proactive servers n_{pa} corresponds to the curve in small dots. Figure 8 illustrates the equivalent results in the case where C is less than $(r-1)$. In both cases, the number n_a of active servers remains equal to $min(k, r)$. The general expression of the number of servers occupied with a proactive task, more complex, can be written:

$$n_{pa} = \mathbf{1}_{(k \leq \max(r,C))} \inf(k, \min(r, C)) + \mathbf{1}_{(\max(r,C)<k<(r+C))}(r + C - k)\,,$$

knowing that a server is no longer occupied with a proactive task if at least r clients are present (The task may have changed from the "proactive " status to the "active "). Note that if C is greater than or equal to $(r-1)$, all the servers are active as soon as j is positive, i.e. also as soon as k is greater than C.

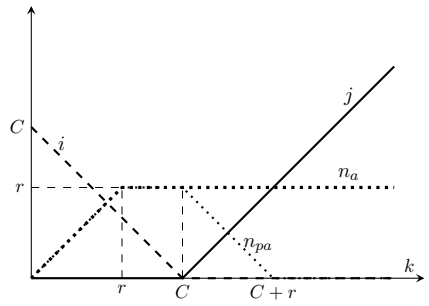


Figure 7: Case $C \geq (r-1)$. Illustration of relations between i, j, n_a and n_{pa} as a function of k.

Let's now consider the j clients that, at an instant, are in the buffer. These customers are positioned according to their order of arrival : the customer located in p-th position arrived after (resp. before) the customer located in $(p-1)$-th (resp. $(p+1)$-th) position. When j is greater than r, the customer located in p-th position is on hold if p is greater than r. If p is less than or equal to r, the query of this client is already in progress. When r is greater than 1, we need to specify the intended service protocol: servers perform identical

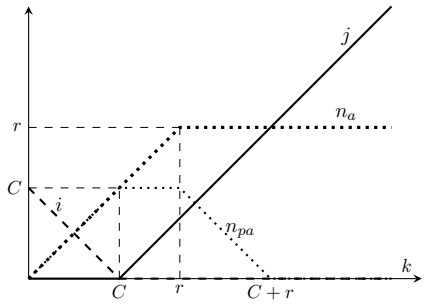


Figure 8: Case $C < (r-1)$. Illustration of relations between i, j, n_a and, n_{pa} as a function of k.

requests that satisfy the client at the top of the buffer. The probability distribution of the service of a client will depend on the state of the file seen by this client but also on relative values of the parameters r and C.

Our model is again a CTMC associated with the possible values of k, $k = 0, 1, 2, ...$ Its transition graph is first represented in Figure 9 in the case where the anticipation C is greater than or equal to $(r-1)$ and then represented in Figure 10 in the case where the anticipation C is lower than $(r-1)$. Because of these ordered states, these particular CTMC are also called birth-death-processes (BDP).

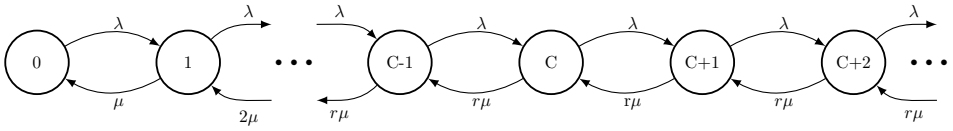


Figure 9: CTMC associated to the proactive multi-serveur queue (case $C \geq (r-1)$).

If $C \geq (r-1)$ (condition that can also be written $r \leq (C+1)$), the r servers are active as soon as $k \geq r$ and in particular the r servers are active when one or more clients are present. On the other hand, if $C < (r-1)$ (a condition that can also be written $r > (C+1)$), the r servers are all active only when k is greater than or equal to r, i.e. when the number j of present clients is greater than or equal to $(r-C)^+$, in order to avoid violating the constraint on the maximum number of anticipated services.

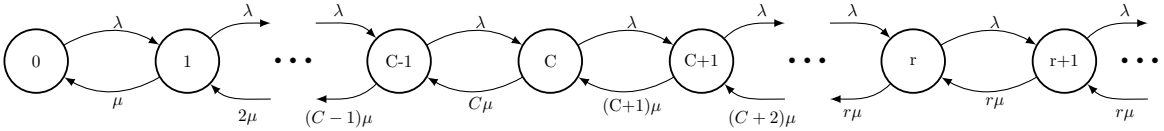


Figure 10: CTMC associated to the proactive multi-serveur queue (case $C < (r-1)$).

Note that in the case where C is greater than or equal to $(r-1)$, the r servers are all active when j is strictly positive, unlike when C is less than $(r-1)$. As we will see later, some performance metrics will have a simpler expression in the first case than in the second.

3.2 Performance Determination

Since now r is eventually greater than 1, let redefine ρ as $\rho = \dfrac{\lambda}{r\mu}$. So, ρ represent the loading rate of this proactive multi-serveur queue. It is easy to verify that the state probabilities of the stationary regime satisfy the relations:

$$p_k = \begin{cases} \rho \dfrac{r}{k} p_{k-1} & \text{if } k \leq r \ , \\ \rho p_{k-1} & \text{if } k \geq r \ . \end{cases}$$

This allows us to express all probabilities in terms of the probability p_0:

$$p_k = \begin{cases} \rho^k \dfrac{r^k}{k!} p_0 & \text{if } k \leq r \ , \\ \rho^k \dfrac{r^r}{r\,!} p_0 & \text{if } k \geq r \ . \end{cases}$$

Subject to satisfying the necessary and sufficient condition of ergodicity $\rho < 1$, the probability p_0 is obtained through the normalizing equation:

$$p_0 = \left[\sum_{i=0}^{r-1} \rho^i \frac{r^i}{i!} + \frac{r^r}{r!} \frac{\rho^r}{(1-\rho)} \right]^{-1} \ , \tag{13}$$

Note that these state probabilities p_k, $k = 0, 1, ...$, **do not depend** on C. Of course, this is not the case for the performances we are looking for.

3.2.1 Case $C \geq (r-1)$

First, we consider the situation where C equals or exceeds $(r-1)$; this is the simplest situation. Let's place ourselves at the moment of arrival of an arbitrary client (to whom we have already given the nice name of August in order to distinguish him), finding the queue in the state (i, j). Several cases are possible:

- case $i \geq 1$ (and therefore $j = 0$): the service is instantaneous (the service distribution is the Dirac's distribution $\delta(0)$),

- case $i = 0$ and $j = 0$: August takes the leading position in the buffer and his service time follows an exponential distribution of rate $r\mu$.

- case $1 \leq j < r$ (and $i = 0$): August takes the position $(j+1)$ and his service time follows an Erlang distribution of order $(j+1)$ and of parameter $r\mu$.

- case $j \geq r$ (and $i = 0$): August takes the position $(j+1)$ and starts his wait which ends when he reaches the position r; his service time will follow an Erlang distribution of order r and parameter $r\mu$.

Let us first determine the expectation of the number of requests waiting in the queue:

$$\begin{aligned}
\mathbb{E}[\nu_C] &= \sum_{j=r+1}^{\infty} (j-r)p_{C+j}\,, \\
&= \sum_{k=C+r+1}^{\infty} (k-C-r)p_k\,, \\
&= \sum_{k=C+r+1}^{\infty} (k-C-r)\frac{\rho^k r^r}{r\,!}p_0 = p_0\frac{r^r}{r\,!}\sum_{i=1}^{\infty} i\rho^{i+C+r}\,, \\
&= \rho^{C+r+1}p_0\frac{r^r}{r\,!}\sum_{i=1}^{\infty} i\rho^{i-1} = \rho^{C+r+1}p_0\frac{r^r}{r\,!}\frac{1}{(1-\rho)^2}\,, \\
&= \rho^C p_0\frac{r^r}{r\,!}\frac{\rho^{r+1}}{(1-\rho)^2}\,, \qquad (14) \\
&= \rho^C \mathbb{E}[\nu_C]|_{C=0}\,. \qquad (15)
\end{aligned}$$

Note that, given the initial expression of $\mathbb{E}[\nu_C]$, this result is valid regardless of the value of C (that C is less than, equal to or greater than $(r-1)$). Thanks to Little's formula, we immediately deduce from this result the expectation of the waiting time:

$$\begin{aligned}
\mathbb{E}[W_C] &= \frac{\mathbb{E}[\nu_C]}{\lambda}\,, \\
&= \rho^C p_0\frac{r^r}{r\,!}\frac{\rho^r}{r\mu(1-\rho)^2}\,, \\
&= \rho^C \mathbb{E}[W_C]|_{C=0}\,, \qquad (16)
\end{aligned}$$

a result also valid regardless of the value of C.

Let's now determine the expectation of the number of requests present in the system (in the case where $C \geq (r-1)$):

$$\begin{aligned}
\mathbb{E}[N_C] &= \sum_{j=1}^{\infty} j p_{C+j}\,, \\
&= \sum_{k=C+1}^{\infty} (k-C)p_k = \sum_{k=C+1}^{\infty} (k-C)\frac{\rho^k r^r}{r\,!}p_0\,, \\
&= p_0\frac{r^r}{r\,!}\sum_{i=1}^{\infty} i\rho^{i+C} = \rho^{C+1}p_0\frac{r^r}{r\,!}\sum_{i=1}^{\infty} i\rho^{i-1}\,, \\
&= \rho^{C+1}p_0\frac{r^r}{r\,!}\frac{1}{(1-\rho)^2}\,. \qquad (17)
\end{aligned}$$

Note that this expression is simpler than the one obtained for the non-proactive queue. It will be quite the opposite in the second case where C will be less than $(r-1)$... unfortunately !

We deduce the expectation of the response time using Little's formula:

$$\begin{aligned}\mathbb{E}[R_C] &= \frac{\mathbb{E}[\mathbb{N}_C]}{\lambda}\,,\\ &= \rho^C p_0 \frac{r^r}{r\,!}\frac{1}{r\mu(1-\rho)^2}\,.\end{aligned} \tag{18}$$

Comparing expressions 14 and 17, we notice that, when $C \geq r$, $\mathbb{E}[N_C]$ can express itself according to $\mathbb{E}[\nu_C]$:

$$\mathbb{E}[N_C] = \mathbb{E}[\nu_{C-r}] \quad C \geq r\,. \tag{19}$$

From this, it is immediately deduced that $\mathbb{E}[R_C]$ can express itself according to $\mathbb{E}[W_C]$:

$$\mathbb{E}[R_C] = \mathbb{E}[W_{C-r}] \quad C \geq r\,. \tag{20}$$

In the boundary case where $C = r-1$, we obtain $\mathbb{E}[N_{r-1}] = \rho^{-1}\mathbb{E}[\nu_0]$ and $\mathbb{E}[R_{r-1}] = \rho^{-1}\mathbb{E}[W_0]$.

Let's now determine the expectation $\mathbb{E}[NS_C]$ of the number of in-service requests using results 14 and 17:

$$\begin{aligned}\mathbb{E}[NS_C] &= \mathbb{E}[N_C] - \mathbb{E}[\nu_C]\,,\\ &= \rho^C p_0 \frac{r^r}{r\,!}\frac{1}{(1-\rho)^2}\left[\rho - \rho^{r+1}\right]\,,\\ &= \rho^{C+1} p_0 \frac{r^r}{r\,!}\frac{(1-\rho^r)}{(1-\rho)^2}\,.\end{aligned} \tag{21}$$

Let's use Little's formula to obtain the expectation of service time:

$$\begin{aligned}\mathbb{E}[S_C] &= \frac{\mathbb{E}[NS_C]}{\lambda}\,,\\ &= \frac{1}{r\,\mu}\rho^C p_0 \frac{r^r}{r\,!}\frac{(1-\rho^r)}{(1-\rho)^2}\,.\end{aligned} \tag{22}$$

It should be noted that this unconditional expectation of service time takes into account customers whose service time is zero. Let us recall here that these last two results giving the expressions of $E[NS_C]$ and $E[S_C]$ are valid only if C is equal to or greater than $(r-1)$ (and thus not valid when $C = 0$ if $r > 1$!).

Let us determine the probability $\mathbb{P}(S_C = 0)$ that the service time of a customer is zero. This probability is the one that the client finds the system in a state k such that the variable i is strictly positive.

$$\begin{aligned}\mathbb{P}(S_C = 0) &= \sum_{j=0}^{C-1} p_j\,,\\ &= \sum_{j=0}^{\infty} p_j - \sum_{j=C}^{\infty} p_j = 1 - p_0 \frac{r^r}{r!}\sum_{i=C}^{\infty}\rho^i\,,\end{aligned}$$

$$= 1 - p_0 \rho^C \frac{r^r}{r!} \sum_{j=0}^{\infty} \rho^j \,,$$

$$= 1 - \rho^C \frac{r^r}{r!} \frac{p_0}{(1-\rho)} \,. \tag{23}$$

Note that this result is also valid if there is only one server since the probability p_0 is then equal to $(1-\rho)$ (cf the expression 6). By associating this last result 23 with the expression of $E[N_C]$ (*cf.* 17), it is easy to show that:

$$\mathbb{E}[N_C] = \frac{\rho}{(1-\rho)} \mathbb{P}(S_C > 0) \,. \tag{24}$$

This expression for the expectation $\mathbb{E}[N_C]$ is worth commenting on. First note that the probability $\mathbb{P}(S_C > 0)$ represents the percentage of customers who cannot pass through the station without stopping.

If $Z(t)$ denotes the state of the CMTC at time t (the one shown in Figure 9), let us respectively note $B(t)$ and $I(t)$ the two processes defined using the $Z(t)$ process as:

$$B(t) = \begin{cases} 1 & \text{if } Z(t) > C \,, \\ 0 & \text{else} \,. \end{cases}$$

$$I(t) = \begin{cases} 1 & \text{if } Z(t) < C \,, \\ 0 & \text{else} \,. \end{cases}$$

Thus, $B(t)$ is 1 if at least one client is present in the queue ($j > 0$ and $i = 0$) and $I(t)$ is 1 if at least one anticipated service is available in the queue ($i > 0$ and $j = 0$).

Figure 11 represents possible partial trajectories over a time interval between t_0 and at t_f. At time t_0, we see that $B(t_0) = 1$ and $I(t_0) = 0$; that means that at least one client is present in the queue.

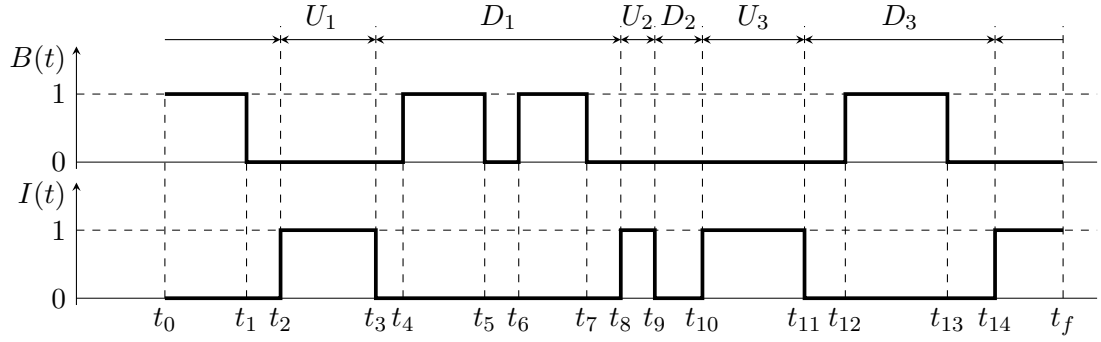


Figure 11: Partial trajectories of $B(t)$ and of $I(t)$ over a time interval between t_0 and t_f.

We see that there are 14 changes of state that we can interpret in the following way:

- At time t_1, the service of the last customer in the queue ends: $B(t_1^+) = 0$ and $I(t_1) = 0$; the next event will be either a new arrival or an early service provisioning.

- At time t_2, an anticipated service occurs: $B(t_2) = 0$ and $I(t_2^+) = 1$.

- At time t_3, a customer arrival ends a possible series of instantaneous services performed on the time interval of lenght $U_1 = (t_3 - t_2)$: $B(t_3) = 0$ and $I(t_3^+) = 0$; the next event can be either a new arrival or an anticipated service provisioning.

- At time t_4, a customer arrival occurs (before the provision of an anticipated service): $B(t_4^+) = 1$ and $I(t_4) = 0$.

- At time t_5, the last of the customers who arrived on the interval $[t_4, t_5[$ leaves the queue, his service terminated: $B(t_5^+) = 0$ and $I(t_5) = 0$; the next event can be either a new arrival or a provision of an anticipated service.

- At time t_6, a customer arrival occurs (before the provision of an anticipated service): $B(t_6^+) = 1$ and $I(t_6) = 0$.

- At time t_7, the service of the last customer in the queue ends: $B(t_7^+) = 0$ and $I(t_7) = 0$; the next event can be either a new arrival or an early service provision.

- At time t_8, an early service provisioning occurs: $B(t_8) = 0$ and $I(t_8^+) = 1$,

and so on...

As can be seen in the figure, time can be decomposed into a sequence of intervals U_i and D_i, $i = 1, 2, 3, ...$ During each interval U_i, $i = 1, 2, 3, ...$, the queue is empty because the services are instantaneous.

The state $k = C$ of the CMTC is a regenerative point for the process $Z(t)$ (and thus for the process $N(t)$, defined as the number of clients present in the queue (i.e. $N(t) = \max(0, (Z(t) - C))$) as well as for the processes $B(t)$ and $I(t)$. This means that the random variables U_i, $i = 1, 2, 3, ...$, are *iid.* The same is true for the random variables D_i, $i = 1, 2, 3, ...$ Moreover, the latter are independent of the former.

During each interval U_i, $i = 1, 2, 3, ...$, the client queue is empty because the services are instantaneous. While during each interval D_i, $i = 1, 2, 3, ...$, there is no anticipated service and the queue behaves like a $M/M/1$ queue with arrival rate $r\lambda$ and service rate $r\mu$; , i.e. with a load rate equal to ρ (because $C \geq (r - 1)$). Considering the queue in stationary regime, this allows us to write the expressions of the following conditonal expectations:

$$\mathbb{E}[N_C | I(t) = 1] = 0\,, \tag{25}$$

$$\mathbb{E}[N_C | I(t) = 0] = \frac{\rho}{(1-\rho)}\,, \tag{26}$$

where t is any instant of the queue in steady state. By deconditioning, we retrieve the relation 24:

$$\begin{aligned} \mathbb{E}[N_C] &= 0 + \frac{\rho}{(1-\rho)} \times \frac{\mathbb{E}[D_i]}{\mathbb{E}[U_i] + \mathbb{E}[D_i]}\,, \\ &= \frac{\rho}{(1-\rho)} \mathbb{P}(S_C > 0)\,. \end{aligned} \tag{27}$$

We can now determine the conditional expectation of service time knowing that it is not zero:

$$\begin{aligned}
\mathbb{E}[S_C|S_C>0] &= \frac{\mathbb{E}[S_C]}{\mathbb{P}(S_C>0)}\,,\\
&= \frac{1}{r\,\mu}\rho^C p_0 \frac{r^r}{r\,!}\frac{(1-\rho^r)}{(1-\rho)^2}\,\frac{r!}{\rho^C\,r^r}\frac{(1-\rho)}{p_0}\,,\\
&= \frac{(1-\rho^r)}{r\mu(1-\rho)}\,. \qquad (28)
\end{aligned}$$

We can find this result without knowing the unconditional expectation $\mathbb{E}[S_C]$ noticing that, knowing that the newcomer finds j clients present, his service time expectation will be equal to $\dfrac{(j+1)}{r\,\mu}$ for $j=0,1,...(r-2)$ and to $\dfrac{r}{r\,\mu}$ for $j\geq(r-1)$:

$$\begin{aligned}
\mathbb{P}(S_C>0)\mathbb{E}[S_C|S_C>0] &= \sum_{j=0}^{r-2}\frac{(j+1)}{r\,\mu}p_{C+j} + \sum_{j=r-1}^{\infty}\frac{r}{r\,\mu}p_{C+j}\,,\\
&= \sum_{j=1}^{r-1}\frac{j}{r\,\mu}p_{C+j-1} + \frac{r}{r\,\mu}\sum_{i=0}^{\infty}p_{C+r-1+i}\,,\\
&= p_0\frac{\rho^{C-1}}{r\,\mu}\frac{r^r}{r!}\left\{\sum_{j=1}^{r-1}j\rho^j + r\rho^r\sum_{i=0}^{\infty}\rho^j\right\}\,,\\
&= p_0\frac{\rho^{C-1}}{r\,\mu}\frac{r^r}{r!}\left\{\frac{((r-1)\rho^{(r+1)}-r\rho^r+\rho)}{(1-\rho)^2}+\frac{r\rho^r}{(1-\rho)}\right\}\,,\\
&= p_0\frac{\rho^{C-1}}{r\,\mu}\frac{r^r}{r!}\,\frac{\rho(1-\rho^r)}{(1-\rho)^2}\,,\\
&= p_0\rho^C\frac{r^r}{r!}\,\frac{(1-\rho^r)}{r\mu(1-\rho)^2}\,.
\end{aligned}$$

Which, by dividing by the probability $\mathbb{P}(S_C>0)$ gives us back the expression 28.

The wait of a customer is null if, on his arrival, the service which he will benefit is already started (possibly already available). The probability that a customer's wait is zero is then expressed as:

$$\begin{aligned}
\mathbb{P}(W_C=0) &= \sum_{j=0}^{C-1+r}p_j\,,\\
&= \sum_{j=0}^{\infty}p_j - \sum_{j=C+r}^{\infty}p_j = 1-p_0\frac{r^r}{r!}\sum_{i=C+r}^{\infty}\rho^i\,,\\
&= 1-p_0\rho^{(C+r)}\,\frac{r^r}{r!}\sum_{j=0}^{\infty}\rho^j\,,\\
&= 1-\rho^{(C+r)}\,\frac{r^r}{r!}\frac{p_0}{(1-\rho)}\,. \qquad (29)
\end{aligned}$$

Note that this formula is valid when $r = 1$ since the probability p_0 is then equal to $(1 - \rho)$ (*cf.* expression 5). This result makes it possible to write the complementary probability:

$$\mathbb{P}(W_C > 0) = \rho^{(C+r)} \, \frac{r^r}{r!} \frac{p_0}{(1-\rho)} \, . \tag{30}$$

For the non-proactive $M/M/r$ queue, it's easy to show that:

$$\mathbb{P}(W > 0) = \rho^r \, \frac{r^r}{r!} \frac{p_0}{(1-\rho)} \, . \tag{31}$$

This allows to write a relation between these probabilities in the form:

$$\mathbb{P}(W_C > 0) = \rho^C \times \mathbb{P}(W_C > 0)|_{C=0} = \rho^C \mathbb{P}(W > 0) \, , \tag{32}$$

as well as:

$$\mathbb{P}(W_{C+1} > 0) = \rho \, \mathbb{P}(W_C > 0) \, . \tag{33}$$

Note also, thanks to the relations (30) and (23), that

$$\mathbb{P}(W_C > 0) = \rho^r \, \mathbb{P}(S_C > 0) = \mathbb{P}(S_{C+r} > 0) \, . \tag{34}$$

We can now determine the conditional expectation of waiting knowing that it occurs:

$$\begin{aligned} \mathbb{E}[W_C | W_C > 0] &= \frac{\mathbb{E}[W_C]}{\mathbb{P}(W_C > 0)} \, , \\ &= \rho^C p_0 \frac{r^r}{r \, !} \frac{\rho^r}{r\mu(1-\rho)^2} \, \frac{r!}{\rho^{(C+r)} \, r^r} \frac{(1-\rho)}{p_0} \, , \\ &= \frac{1}{r\mu(1-\rho)} \, . \end{aligned} \tag{35}$$

As in the case where the server is unique, we verify here that this conditional expectation of waiting for the proactive queue is identical to that of the non-proactive queue.

Recall here that if the probability p_0 does not depend on the variable C, it depends on the variable r. It is therefore not a good idea to compare the results obtained by varying the number of servers without taking it into account.

Let's now look for the probability distribution of W_C, always when $C \geq (r-1)$. We already know that W_C is a mixed random variable with a mass at zero, equal to $\mathbb{P}(W_C = 0)$ (given by the expression 29). Since arrivals follow a Poisson process, the probability that August finds j clients in front of him on arrival is equal, when j is strictly positive, to $p_{(C+j)}$ (property PASTA). August suffers a wait if j is greater than $(r-1)$. In this case, its waiting time is distributed according to the Erlang probability distribution of order $(j - r + 1)$ and of parameter $r\mu$ whose density is written as:

$$f_X(t) = \frac{(r\mu)^{(j-r+1)} t^{j-r}}{(j-r)!} e^{-r\mu t}, \qquad t \geq 0 \, .$$

What, through a deconditioning, allows to write:

$$\begin{aligned}
dF_{W_C}(t) &= \mathbb{P}(W_C=0)\delta(t) + \sum_{j=r}^{\infty} p_{C+j} \frac{(r\mu)^{(j-r+1)} t^{j-r}}{(j-r)!} e^{-r\mu t}, \\
&= \mathbb{P}(W_C=0)\delta(t) + p_0 \rho^{C+r} \frac{r^r}{r!} r\mu e^{-r\mu t} \sum_{j=r}^{\infty} \rho^{j-r} \frac{(r\mu)^{(j-r)} t^{j-r}}{(j-r)!}, \\
&= \mathbb{P}(W_C=0)\delta(t) + p_0 \rho^{C+r} \frac{r^r}{r!} r\mu e^{-r\mu t} \sum_{j=r}^{\infty} \frac{\lambda^{(j-r)} t^{j-r}}{(j-r)!}, \\
&= \mathbb{P}(W_C=0)\delta(t) + p_0 \rho^{C+r} \frac{r^r}{r!} r\mu e^{-r\mu t} \sum_{i=0}^{\infty} \frac{(\lambda t)^i}{i\,!}, \\
&= \mathbb{P}(W_C=0)\delta(t) + p_0 \rho^{C+r} \frac{r^r}{r!} r\mu e^{-r\mu t} e^{\lambda t}, \\
&= \mathbb{P}(W_C=0)\delta(t) + \frac{p_0}{(1-\rho)} \rho^{C+r} \frac{r^r}{r!} (1-\rho) r\mu e^{-(r\mu-\lambda)t}, \\
&= \mathbb{P}(W_C=0)\delta(t) + \frac{p_0}{(1-\rho)} \rho^{C+r} \frac{r^r}{r!} (r\mu-\lambda) e^{-(r\mu-\lambda)t}, \\
&= \mathbb{P}(W_C=0)\delta(t) + (1-\mathbb{P}(W_C=0))(r\mu-\lambda) e^{-(r\mu-\lambda)t}. \qquad (36)
\end{aligned}$$

where $\delta(t)$ denotes the Dirac's function. The probability distribution of W_C is a mixed one with a mass in zero, equal to the probability $\mathbb{P}(W_C = 0)$ whose expression is given by the relation 29, and a continuous part on the domain $]0, +\infty]$ corresponding to the exponential distribution of rate $(r\mu - \lambda)$ weighted by the factor $(1 - \mathbb{P}(W_C = 0))$.

Let's now look for the probability distribution of R_C, again when $C \geq (r-1)$. We already know that it is a mixed distribution with a mass in zero, equal to $\mathbb{P}(S_C = 0)$ (given by the expression 23). If on arrival, August finds less than r clients present while $i = 0$, he is taken care of immediately and his service follows an Erlang distribution of order $(j+1)$ (j being the number of clients present at the time of its arrival) and with parameter $r\mu$ because all the servers are active (we are in the case $C \geq (r-1)$). If $j \geq r$, August must wait until he finds himself in the r^{th} position; his waiting time follows the Erlang distribution of order $(j-r+1)$ and of parameter $r\mu$. Then the client goes on his service following the Erlang distribution of r order r and with parameter $r\mu$. Waiting times and service being independent, the probability distribution of the conditional response time knowing j (and $i = 0$) is the Erlang distribution of order $(j+1)$ and of parameter $r\mu$, result identical to the previous case where j was less than r.

So we have:

$$\begin{aligned}
dF_{R_C}(t) &= \mathbb{P}(R_C=0)\delta(t) + \sum_{j=0}^{\infty} p_{C+j} \frac{(r\mu)^{(j+1)} t^{j}}{(j\,!)} e^{-r\mu t}, \\
&= \mathbb{P}(R_C=0)\delta(t) + p_0 \rho^{C} \frac{r^r}{r!} r\mu e^{-r\mu t} \sum_{j=0}^{\infty} \rho^{j} \frac{(r\mu)^{j} t^{j}}{j\,!},
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{P}(R_C = 0)\delta(t) + p_0 \rho^C \frac{r^r}{r!} r\mu e^{-r\mu t} \sum_{j=0}^{\infty} \frac{(\lambda t)^j}{j\,!}, \\
&= \mathbb{P}(R_C = 0)\delta(t) + p_0 \rho^C \frac{r^r}{r!} r\mu e^{-r\mu t} e^{\lambda t}, \\
&= \mathbb{P}(R_C = 0)\delta(t) + \frac{p_0}{(1-\rho)} \rho^C \frac{r^r}{r!} (1-\rho) r\mu e^{-(r\mu-\lambda)t}\,, \\
&= \mathbb{P}(R_C = 0)\delta(t) + (1 - \mathbb{P}(R_C = 0))(r\mu - \lambda) e^{-(r\mu-\lambda)t}\,, \qquad (37)
\end{aligned}
$$

Again, the continuous part of this mixed probability distribution corresponds to the exponential distribution of rate $(r\mu - \lambda)$, weighted here by the probability $\mathbb{P}(R_C > 0)$; the latter being equal to the probability $\mathbb{P}(S_C > 0)$ that the service is not zero.

3.2.2 Cas $C < (r-1)$

Let's now examine the situation when r is greater than $(C+1)$. This situation is more complex than the previous one because now the number n_a of activated servers is limited to the quantity $n_a = \min((C+j), r)$s where j corresponds to the number of clients present in the buffer. Indeed, one should not anticipate more than C services. However, some previous results remain valid when r is greater than $(C+1)$. This is the case of expectations $\mathbb{E}[\nu_C]$ (*cf.* (14) and (15)) et $\mathbb{E}[W_C]$ (*cf.* (16)). But this is not the case for the expectation $\mathbb{E}[N_C]$ whose expression will become more complex. This is because even though this last expectation can always be written as when C was greater than or equal to $(r+1)$, in the form:

$$\mathbb{E}[N_C] = \sum_{k=C+1}^{\infty} (k - C) p_k\,,$$

the expressions of the first values of p_k as a function of p_0 do not respect the general form obtained when k becomes greater than $(r-1)$.

Here, considering what we already know the expression of $\mathbb{E}[\nu_C]$, it seemed preferable to calculate first the expectation $\mathbb{E}[NS_C]$ of the number of requests during service (in the present case where $C < (r-1)$):

We show in appendix A that $\mathbb{E}[NS_C]$ can be written as:

$$\mathbb{E}[NS_C] = p_0 \sum_{i=1}^{r-C} i \frac{(r\rho)^{i+C}}{(i+C)!} + p_0 \frac{r^r}{r\,!} \rho^{r+1} \left\{ \frac{(1-\rho^C)}{(1-\rho)^2} + \frac{(r-C)}{(1-\rho)} \right\}. \qquad (38)$$

Note that for $C = 0$, this expression 38 gives the expectation of the number of busy servers in queue $M/M/r$ (*i.e.,* the value $r\rho$):

$$\mathbb{E}[NS_C]|_{C=0} = p_0 \sum_{i=1}^{r} i \frac{(r\rho)^i}{(i)!} + r p_0 \frac{r^r}{r\,!} \frac{\rho^{r+1}}{(1-\rho)}\,,$$

$$
\begin{aligned}
&= r\rho\, p_0 \left[\sum_{j=0}^{r-1} \frac{(r\rho)^j}{j\,!} + \frac{r^r}{r\,!} \frac{\rho^r}{(1-\rho)} \right] , \\
&= r\rho\, p_0 \left[p_0 \right]^{-1} , \\
&= r\rho\,. \qquad (39)
\end{aligned}
$$

We can still rewrite the right part of the expression 38 in the following form,

$$
\begin{aligned}
\mathbb{E}[NS_C] &= p_0 \sum_{i=1}^{r-C} i \frac{(r\rho)^{i+C}}{(i+C)!} + p_0 \frac{r^r}{r\,!} \rho^{r+1} \left\{ \frac{1}{(1-\rho)^2} + \frac{(r-C)}{(1-\rho)} \right\} - p_0 \frac{r^r}{r\,!} \frac{\rho^{C+r+1}}{(1-\rho)^2} \\
&= p_0 \sum_{i=1}^{r-C} i \frac{(r\rho)^{i+C}}{(i+C)!} + \mathbb{E}[\nu_C]|_{C=0} \left[1 + (r-C)(1-\rho) \right] - \mathbb{E}[\nu_C] \qquad (40)
\end{aligned}
$$

Thus we get an expression of the expectation of N_C:

$$
\mathbb{E}[N_C] = p_0 \sum_{i=1}^{r-C} i \frac{(r\rho)^{i+C}}{(i+C)!} + \mathbb{E}[\nu_C]|_{C=0} \left[1 + (r-C)(1-\rho) \right] . \qquad (41)
$$

Recall that the expression of $\mathbb{E}[\nu_C]$ is given by the formula 14. Note also that this last expression 41 also gives the good result if we give the zero value to C:

$$
\mathbb{E}[N_C]|_{C=0} = r\rho + \mathbb{E}[\nu_C]|_{C=0}\,. \qquad (42)
$$

Similarly, we obtain a new expression for the probability that the response time is zero valid for the case $C < (r-1)$:

$$
\begin{aligned}
\mathbb{P}(S_C = 0) &= \sum_{j=0}^{C-1} p_j \\
&= p_0 \sum_{i=0}^{C-1} \frac{(r\rho)^i}{i\,!}\,. \qquad (43)
\end{aligned}
$$

This last expression is not as nice as the previous formula 23 obtained in the case $C \geq (r-1)$. Note nevertheless that we can express the probability $\mathbb{P}(S_C = 0)$ in a slightly different and potentially useful form if the number of servers r is important and if the parameter C is close to r:

$$
\begin{aligned}
\mathbb{P}(S_C = 0) &= \sum_{j=0}^{C-1} p_j\,, \\
&= \sum_{j=0}^{r-1} p_j - \sum_{j=C}^{r-1} p_j\,,
\end{aligned}
$$

$$= \mathbb{P}(S_C = 0)|_{C=r} - p_0 \sum_{j=C}^{r-1} \frac{(r\rho)^j}{j!}\,, \tag{44}$$

$$= 1 - \rho^r \frac{r^r}{r!}\frac{p_0}{(1-\rho)} - p_0 \sum_{j=C}^{r-1} \frac{(r\rho)^j}{j!}\,. \tag{45}$$

It seems interesting to find the expression of $\mathbb{E}[N_C]$ (41) using an approach similar to that used to justify the formula 24 (page 17). For this we consider the BDP obtained from the one on Figure 10 by eliminating the states for which $I(t) = 1$. This gives us the BDP represented on Figure 12.

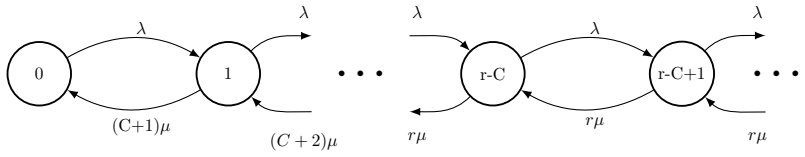


Figure 12: BDP associated with the proactive multi-server pseudo-file (case $C < (r-1)$).

Let us note q_k the asymptotic probability of the state k of this new BDP, $k = 0, 1, ...$ These probabilities are classically expressed in terms of the asymptotic probability q_0:

$$q_k = \begin{cases} \rho^k \dfrac{r^k C\,!}{(C+k)\,!} q_0 & \text{si } k \le r - C\,, \\ \\ \rho^k \dfrac{C\,!\; r^{r-C}}{r\,!} q_0 & \text{si } k \ge r - C\,. \end{cases}$$

Note that the transition rates retained in this new BDP have not been changed. Which allows us to state that

$$\frac{q_i}{q_j} = \frac{p_{C+i}}{p_{C+j}}\,,\ \forall i, j\,.$$

This means that the distribution $q_0, q_1, q_2, ...$ is a homothety of the partial distribution $p_C, p_{C+1}, p_{C+2}, ...$

Noting furthermore that (*cf.* 43):

$$\sum_{j=0}^{\infty} q_j = 1 \qquad \text{and } \sum_{j=0}^{\infty} p_{C+j} = 1 - \sum_{k=0}^{C-1} p_k = \mathbb{P}(S_C > 0)\,,$$

we deduce that

$$p_{C+j} = q_j(1 - \sum_{k=0}^{C-1} p_k) = q_j \mathbb{P}(S_C > 0)\,,\ \forall j\,.$$

Let us note $\mathbb{E}[\widetilde{N_C}]$ the expectation of the number of customers present in this pseudo queue. We have

$$\begin{aligned}
\mathbb{E}[\widetilde{N_C}] &= \sum_{j=1}^{\infty} j q_j \,, \\
&= \sum_{j=1}^{r-C} j q_j + \sum_{j=r-C+1}^{\infty} j q_j \,, \\
&= C\,!\, q_0 \sum_{j=1}^{r-C} j \frac{(r\rho)^j}{(C+j)\,!} + \frac{C\,!\, r^{r-C}}{r\,!} q_0 \sum_{j=r-C+1}^{\infty} j\rho^j \,, \\
&= C\,!\, q_0 \sum_{j=1}^{r-C} j \frac{(r\rho)^j}{(C+j)\,!} + \frac{C\,!\, r^{r-C}}{r\,!} q_0 \sum_{i=1}^{\infty} (i+r-C)\rho^{(i+r-C)} \,, \\
&= C\,!\, q_0 \sum_{j=1}^{r-C} j \frac{(r\rho)^j}{(C+j)\,!} + \frac{C\,!\, r^{r-C}}{r\,!} q_0 \left\{ \rho^{(r-C+1)} \sum_{i=1}^{\infty} i\rho^{i-1} + (r-C)\rho^{(r-C+1)} \sum_{i=1}^{\infty} \rho^{i-1} \right\} , \\
&= C\,!\, q_0 \sum_{j=1}^{r-C} j \frac{(r\rho)^j}{(C+j)\,!} + \frac{C\,!\, (r\rho)^{r-C}}{r\,!} q_0 \left\{ \frac{\rho}{(1-\rho)^2} + \frac{\rho(r-C)}{(1-\rho)} \right\} .
\end{aligned} \tag{46}$$

Multiplying $\mathbb{E}[\widetilde{N_C}]$ by $\mathbb{P}(S_C > 0)$, we obtain:

$$\begin{aligned}
&\mathbb{E}[\widetilde{N_C}] \times \mathbb{P}(S_C > 0) \\
&= \left[C\,!\, q_0 \sum_{j=1}^{r-C} j \frac{(r\rho)^j}{(C+j)\,!} + \frac{C\,!\, (r\rho)^{r-C}}{r\,!} q_0 \left\{ \frac{\rho}{(1-\rho)^2} + \frac{\rho(r-C)}{(1-\rho)} \right\} \right] (1 - \sum_{k=0}^{C-1} p_k) \,.
\end{aligned} \tag{47}$$

Since $q_0(1 - \sum_{k=0}^{C-1} p_k)$ is equal to p_C and that $p_C = \rho^C \frac{r^C}{C!} p_0$, we get

$$\begin{aligned}
\mathbb{E}[\widetilde{N_C}] \times \mathbb{P}(S_C > 0) &= p_0 \sum_{j=1}^{r-C} j \frac{(r\rho)^{j+C}}{(j+C)\,!} + \frac{(r\rho)^r}{r\,!} p_0 \left\{ \frac{\rho}{(1-\rho)^2} + \frac{\rho(r-C)}{(1-\rho)} \right\} , \\
&= p_0 \sum_{j=1}^{r-C} j \frac{(r\rho)^{j+C}}{(j+C)\,!} + p_0 \frac{r^r}{r\,!} \frac{\rho^{r+1}}{(1-\rho)^2} \left\{ 1 + (r-C)(1-\rho) \right\} , \\
&= p_0 \sum_{j=1}^{r-C} j \frac{(r\rho)^{j+C}}{(j+C)\,!} + \mathbb{E}[\nu_C]|_{C=0} \left\{ 1 + (r-C)(1-\rho) \right\} , \\
&= \mathbb{E}[N_C].
\end{aligned} \tag{48}$$

This confirms that $\mathbb{E}[N_C]$ is indeed equal to the product $\mathbb{E}[\widetilde{N_C}] \times \mathbb{P}(S_C > 0)$...

We obtain the value of the expectation of the response time $\mathbb{E}[R_C]$ by dividing $\mathbb{E}[N_C]$ by the arrival rate λ.

$$\mathbb{E}[R_C] = \frac{\mathbb{E}[N_C]}{\lambda} = \frac{1}{\mu} p_0 \sum_{i=1}^{r-C} i \frac{(r\rho)^{i+C-1}}{(i+C)!} + \mathbb{E}[W_C]|_{C=0} \left[1 + (r-C)(1-\rho)\right] . \tag{49}$$

In figure 13, we have represented the evolution of the expectations $\mathbb{E}[W_C]$ and $\mathbb{E}[R_C]$ of a queue with 3 servers as a function of C, for C = 0. 1, 2, ..., 5, when $\rho = 0.75$ and $\mu = 0.5$. As we could expect, these two expectations are decreasing according to the C level of anticipation. The expectation $\mathbb{E}[W_C]$ is given by the formula 16 whereas the expectation $\mathbb{E}[R_C]$ is given by the formula 49 if $C < (r-1)$ and by the formula 18 if $C \geq (r-1)$. We observe that the difference $(\mathbb{E}[R_C] - \mathbb{E}[W_C])$, which is equal to the expectation of the service time $\mathbb{E}[S_C]$, decreases when C increases. For $C = 0$, the difference is equal at $1/\mu = 2$.

We can also verify, as the previous results have shown, that for a given value $C \geq r$, $\mathbb{E}[R_C] = \mathbb{E}[W_{C-r}]$.

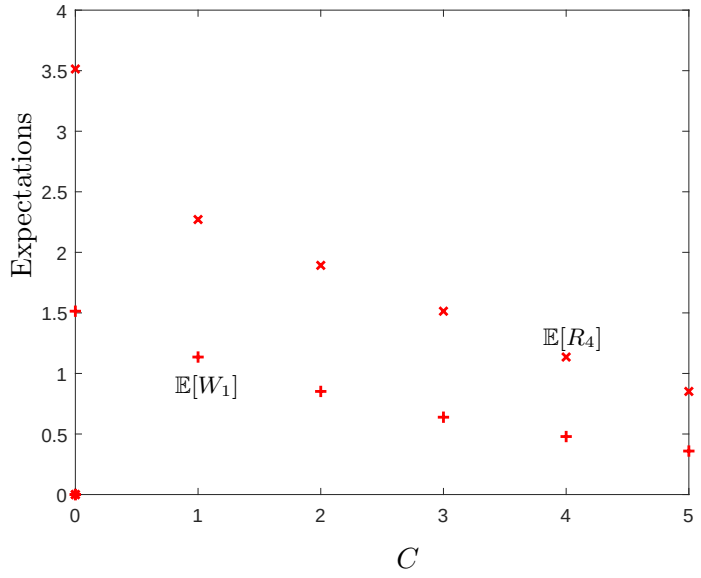


Figure 13: Expectations of waiting times (+) and of response times (x) as a function of C, for $C = 0, 1, 2, ..., 5$. Case with 3 servers, when $\rho = 0.75$ and $\mu = 0.5$.

By dividing the expectation $\mathbb{E}[NS_C]$ (given by the expression 38) by the arrival rate λ, we obtain, if necessary, the expectation $\mathbb{E}[S_C]$ of service time when r is greater than C.

Concerning the probability $\mathbb{P}(W_C = 0)$ that the waiting time is zero, the conditions of its determination are here the same as in the case $C \geq (r-1)$; we must therefore find the

same result as the one obtained page 19:

$$\mathbb{P}(W_C = 0) = 1 - \rho^{(C+r)} \frac{r^r}{r!} \frac{p_0}{(1-\rho)} . \tag{50}$$

It is the same for the conditional expectation of waiting knowing that it exists:

$$\mathbb{E}[W_C | W_C > 0] = \frac{1}{r\mu(1-\rho)} . \tag{51}$$

Again, this conditional expectation of waiting for the proactive queue is identical to that of the non-proactive queue.

On Figure 14, we have represented the probabilities that a service or waiting time is zero as a function of C, $C = 0.1, 2, ..., 8$. The number of servers is 3 while $\rho = 0.75$ and $\mu = 0.5$. These two probabilities are increasing according to the C level of anticipation. The probabilities $\mathbb{P}(W_C = 0)$ are given by the formula 50 whereas the probabilities $\mathbb{P}(S_C = 0)$ are given respectively by the formula 43 if $C < (r-1)$ and by the formula 23 if $C \geq (r-1)$. We can verify, as the previous results have shown, that $\mathbb{P}(S_C = 0) = \mathbb{P}(W_{C-r} = 0)$ when C is greater than or equal to r (here equal to 3).

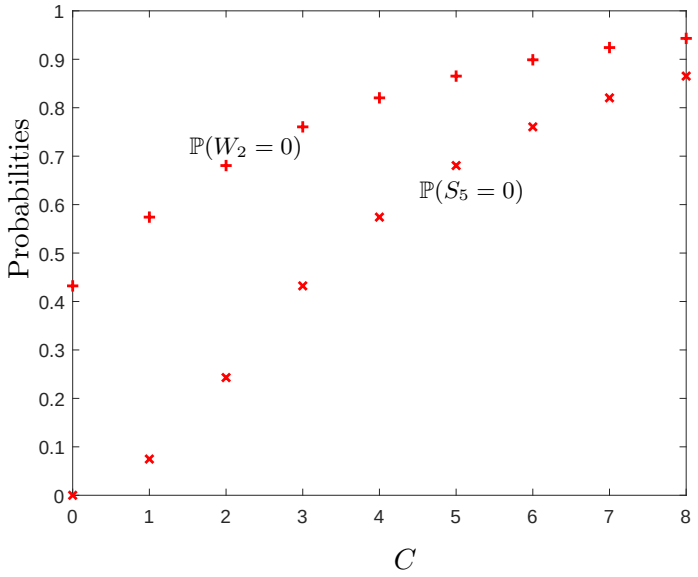


Figure 14: Probabilities of null service times (x) and of waiting times (+) as a function of C, for $C = 0, 1, 2, ..., 8$. Case with 3 servers, when $\rho = 0.75$ and $\mu = 0.5$.

Thanks to Little's formula, we determine the expectation of service time:

$$\mathbb{E}[S_C] = \frac{\mathbb{E}[NS_C]}{\lambda} . \tag{52}$$

Remember that this unconditional expectation of service time takes into account customers whose service time is zero.

The determination of the conditional expectation of service time knowing that it is not zero is very easily done using the relation:

$$\mathbb{E}[S_C|S_C > 0] = \frac{\mathbb{E}[S_C]}{(1 - \mathbb{P}(S_C = 0))}\,, \tag{53}$$

and values obtained from the expressions 43 and 52.

On reading the above, we find that the expressions of several performance indicators become more complex when we consider the case where the number of servers is greater than $(C+1)$. Unfortunately, this will be even more true when we consider the distribution of the sojourn time of any customer (that we will continue to call August). However, if we only consider the distribution of August's waiting time, the expression remains simple. Indeed, as long as at least one client is waiting, the overall service rate of the queue is equal to $r\mu$, regardless of the value of C and therefore, knowing that August found j clients in front of him on arrival, j higher than $(r-1)$, his waiting is distributed according to the Erlang distribution of order $(j-r+1)$ and of parameter $r\mu$; of which the density of this Erlang distribution is written:

$$f_X(t) = \frac{(r\mu)^{(j-r+1)}t^{j-r}}{(j-r)!}e^{-r\mu t}, \qquad t \geq 0\,. \tag{54}$$

As in the case where C was greater than or equal to $(r-1)$, the distribution of W_C is a mixed distribution with a mass $\mathbb{P}(W_C = 0)$ when $t = 0$. The computation of the distribution $dF_{W_C}(t)$ is similar to that made page 21, and provides the result:

$$dF_{W_C}(t) = \mathbb{P}(W_C = 0)\delta(t) + (1 - \mathbb{P}(W_C = 0))(r\mu - \lambda)e^{-(r\mu-\lambda)t}\,. \tag{55}$$

where $\delta(t)$ denotes the Dirac's function. The only difference with the case where C was greater than $(r-1)$ is that, at constant value of the r number of servers, the mass $\mathbb{P}(W_C = 0)$ is here weaker, in the sense that if $C_1 \geq (r-1)$ and $C_2 < (r-1)$, then $\mathbb{P}(W_{C_2} = 0) < \mathbb{P}(W_{C_1} = 0)$ (cf the equality 50).

Now let's take a look at August's customer service distribution. If, on his arrival $i \geq 1$ (and so $j = 0$), August's waiting is null and his service is instantaneous (the distribution of service is Dirac's distribution $\delta(0)$). In a second step, let's examine the conditional service distribution knowing that upon arrival, August finds the system in the state $i = 0$ and $j = 0$, state in which C of the r servers are active. Since r is here strictly greater than $(C+1)$, the arrival of August causes the activation of an additional server, the number n_a of active servers becomes equal to $(C+1)$ and the service rate becomes $(C+1)\mu$ because August will benefit of the first service completed. The probability that August's service ends before a new client arrives is $(C+1)\mu/((C+1)\mu + \lambda)$. If none of the $(C+1)$ active servers completes the execution of its request before the arrival of another client, a new server is activated (r is here greater than $(C+1)$ per hypothesis) and the overall service

rate that August will benefit becomes equal to $(C+2)\mu$. The probability that August's service ends before the arrival of a second new client is $(C+2)\mu/((C+2)\mu+\lambda)$. There is no limit to the number of customers that August can see arriving before he leaves, but after $(r-C-1)$ customer arrivals, August's service rate will equal $r\mu$ and will not change anymore. August's service distribution is here a particular phase-type distribution: it is the Coxian distribution of order $(r-C)$. A schematic representation of this distribution is given on Figure 15. Recall that a Cox distribution can be seen as the distribution of a random sum of independent random variables distributed according to exponential distributions. The diagram of this figure 15 makes it easy to enter the semantics. Noting $\Lambda_i = \lambda + i\mu$, the service starts with a first phase whose duration is distributed according to an exponential distribution with rate $\Lambda_{(C+1)}$. Then, eather the service is terminated (with probability $(C+1)\mu/\Lambda_{(C+1)}$ equal to $(1-\lambda/\Lambda_{(C+1)})$), or it continues (with probability $\lambda/\Lambda_{(C+1)}$) by a second phase distributed according to the exponential distribution with rate $\Lambda_{(C+2)}$; and so more ... The sequence will have at most $(r-C)$ phases; the last possible phase is of rate $r\mu$. It is therefore a distribution of order $(r-C)$ for the case $C < (r-1)$.

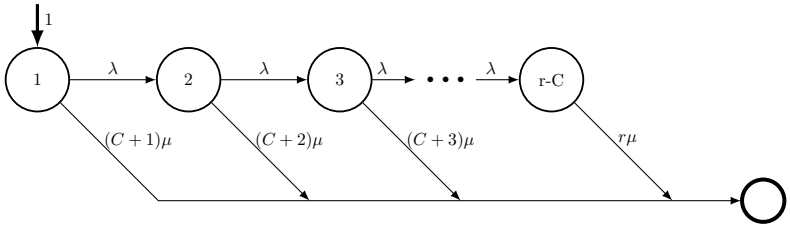


Figure 15: Coxian distribution of order $(r-C)$.

Noting successively:

$$a_i \stackrel{\triangle}{=} \frac{\lambda}{\Lambda_{(C+i)}}, \qquad i = 1, 2, ..., r-C-1, \text{ and } a_{r-C} = 0,$$

$$A_i \stackrel{\triangle}{=} \prod_{j=1}^{i-1} a_j, \qquad j = 2, 3, ..., r-C, \text{ and } A_1 = 1,$$

$$B_i \stackrel{\triangle}{=} A_i(1-a_i), \qquad j = 1, 2, ..., r-C,$$

it is possible to show (*cf.* [3], appendix A.15) that the expectation of the RV S_C following this Coxian probability distribution can be written:

$$\mathbb{E}[S_C] = \sum_{i=1}^{r-C} \frac{A_i}{\Lambda_{(C+i)}} \tag{56}$$

whereas the Laplace transform of the distribution of this RV S_C can be written:

$$L_{S_C}(s) = \sum_{i=1}^{r-C} B_i \left(\prod_{j=1}^{i} \frac{\Lambda_{(C+j)}}{(\Lambda_{(C+j)} + s)} \right) .$$

It remains however to decompose these rational fractions into several simple elements to find the original, *i.e.,* the density function of this Coxian distribution of order k.

Now consider the case where, upon arrival, August finds himself in the p position, such that $1 < p \leq r - C$, he finds $(C + p - 1)$ active servers and his arrival causes the activation for an additional server, the number n_a of active servers becomes equal to $(C + p)$ and the service rate becomes equal to $(C + p)\mu$. Before he can be served and thus leave the queue, August must wait until the $(p - 1)$ customers who precede him are themselves served. During all this time, the overall service rate of the queue may decrease following departures or increase to the maximum rate $r\mu$ due to arrivals.

The August's service distribution is here a phase-type distribution of which we denote the states by the pairs (n, m) where n denotes August's position in the queue and where m denotes: eather the number of clients in the queue behind August when m is less than $(r - C - 1)$, or a number of clients in the queue behind August greater than or equal to $(r - C - 1)$ when $m = (r - C - 1)$.

Indeed, as soon as at least $(r - C - 1)$ customers are present in the queue behind August, the service rate of the queue will remain equal to $r\mu$ at least until the end of August's service. The initial state of this phase-type distribution is the state $(p, 0)$. A schematic representation of this distribution is given in Figure 16.

The semi-formal expression of this phase-type distribution can be obtained by using the algorithm ACE (*cf.* [2])[1]. It is also possible to obtain the formal expression of this phase law using a formal calculation software (for a reasonable number of states of the phase distribution). As an example, consider the special case where $r = 6$, $C = 3$, when August is in $p = 2$ position on arrival (*cf.* Figure 17). For didactic purposes, we will look for the formal expression of the phase-type distribution without using a formal calculation software.

After renumbering the states, we obtain the acyclic CTMC of the Figure 18. This enumeration of states has been chosen so that the infinitesimal generator $A = (a_{ij})$ of the Markov chain is upper triangular. This acyclic CTMC has 7 states of which the first six are transitory states and the seventh is the absorbing state.

Let us note $P_i(t)$ the probability that the CTMC is in state i at time t, $i = 1, 2, ..., 7$. The distribution function of the desired service distribution is equal to the probability $P_7(t)$. Let us recall that the diagonal elements a_{ii} of A are defined by $a_{ii} = -\sum_{j \neq i} a_{ij}$. Note $a_i = -a_{ii}$, $i = 1, 2, ..., 7$. a_i is the total output rate of the state i. We use the integral form of the Kolmogorov equation which, knowing that the initial state is the state 1 with probability 1, is written:

[1] A reminder of this method is presented in appendix C

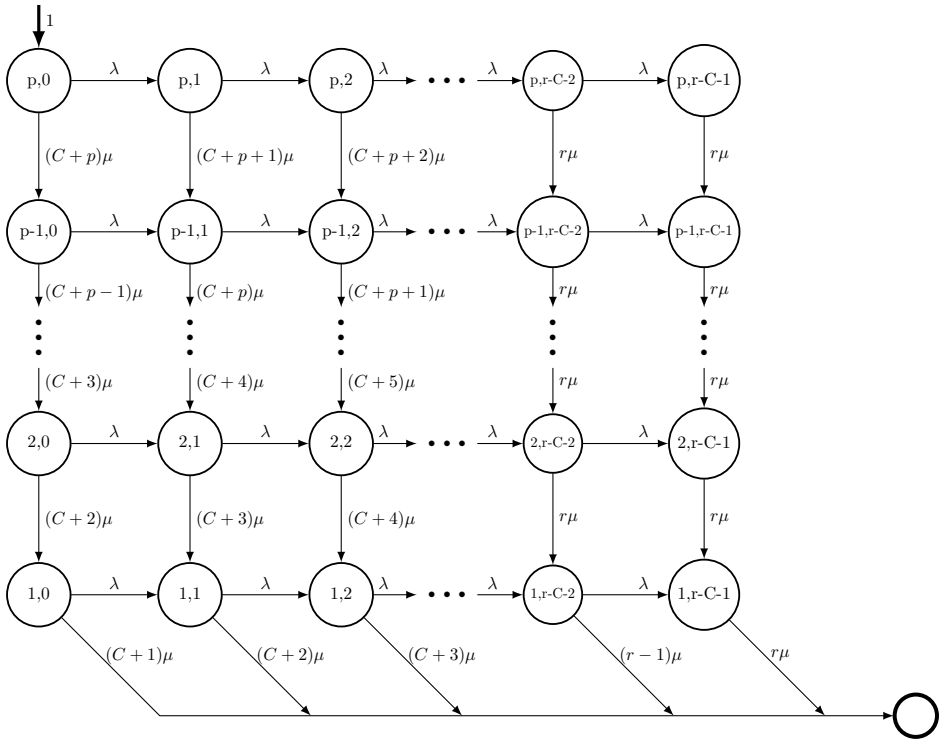


Figure 16: Phase-type service time distribution, case $1 < p \leq r - C$.

$$P_i(t) = \mathbf{1}_{\{i=1\}} \exp\left(-\int_0^t a_1 du\right) + \int_0^t \left[\sum_{k\neq i} P_k(x) a_{ki} \exp\left(-\int_x^t a_i du\right)\right] dx\,. \tag{57}$$

Using again the notations $\Lambda_i = \lambda + i\mu$, for $i = 4, 5, 6$, we get sucessively:

$$P_1(t) = e^{-\Lambda_5 t}\,. \tag{58}$$

$$\begin{aligned}
P_2(t) &= \int_0^t P_1(x)\lambda e^{-\Lambda_6(t-x)} dx\,, \\
&= \lambda e^{-\Lambda_6 t} \int_0^t e^{-\Lambda_5 x} e^{+\Lambda_6 x} dx = e^{-\Lambda_6 t} \frac{\lambda}{(\Lambda_5 - \Lambda_6)} \left[-e^{-(\Lambda_5-\Lambda_6)x}\right]_0^t\,, \\
&= e^{-\Lambda_6 t} \frac{\lambda}{(\Lambda_5 - \Lambda_6)} \left(1 - e^{-(\Lambda_5-\Lambda_6)t}\right) = -\frac{\lambda}{\mu} e^{-\Lambda_6 t} \left(1 - e^{-\mu t}\right)\,, \\
&= \frac{\lambda}{\mu} \left(e^{-\Lambda_5 t} - e^{-\Lambda_6 t}\right)\,.
\end{aligned}$$

The determination of the formal expressions of the probabilities $P_i(t)$, $i = 3, 4, 5, 6$, is presented in the appendix B (page 57). These formal expressions can be written:

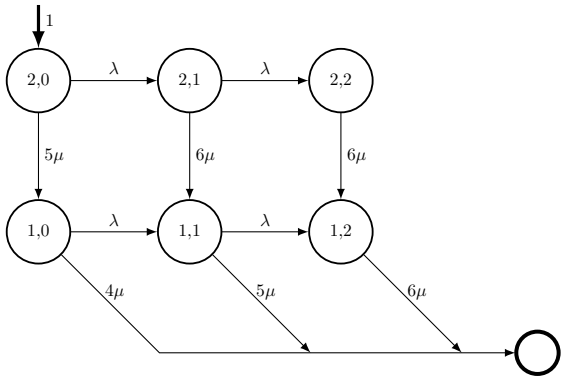


Figure 17: Phase-type service time distribution in the particular case where $p = 2$, $r = 6$ and $C = 3$.

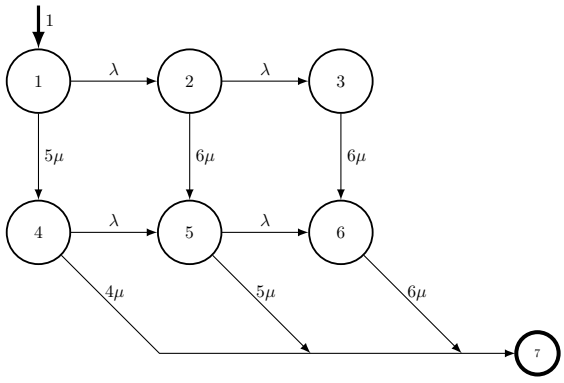


Figure 18: New denomination of the states of the studied acyclic CTMC.

$$P_3(t) = \frac{\lambda}{(\lambda-\mu)}e^{-6\mu t} - \frac{\lambda^2}{\mu(\lambda-\mu)}e^{-\Lambda_5 t} + \frac{\lambda}{\mu}e^{-\Lambda_6 t}\,.$$

$$P_4(t) = 5\left(e^{-\Lambda_4 t} - e^{-\Lambda_5 t}\right)\,.$$

$$P_5(t) = 5\frac{\lambda}{\mu}e^{-\Lambda_4 t} - \left(11\frac{\lambda}{\mu} - \lambda t\right)e^{-\Lambda_5 t} + 6\frac{\lambda}{\mu}e^{-\Lambda_6 t}\,.$$

$$\begin{aligned} P_6(t) = {} & \left(\frac{6\mu\lambda}{(\lambda-\mu)}t - \frac{5\lambda^2}{(\lambda-\mu)^2} + 6 + \frac{5\lambda^2}{\mu(\lambda-2\mu)} - \frac{11\lambda^2}{\mu(\lambda-\mu)} + \frac{6\lambda}{\mu}\right)e^{-6\mu t} + \\ & - \frac{5\lambda^2}{\mu(\lambda-2\mu)}e^{-\Lambda_4 t} + \left(\frac{5\lambda^2}{(\lambda-\mu)^2} + \frac{11\lambda^2}{\mu(\lambda-\mu)} - \frac{\lambda^2 t}{(\lambda-\mu)}\right)e^{-\Lambda_5 t} - 6\left(1+\frac{\lambda}{\mu}\right)e^{-\Lambda_6 t}\,. \end{aligned}$$

By not factoring more in the expression of the coefficients, it is quite easy to check that the state probability $P_6(t)$ is zero at the time origin.

The probability $P_7(t)$ which is none other than August's distribution of service time can be determined using the integral form of the Kolmogorov equation which, knowing that the state 7 is absorbing, is written:

$$P_7(t) = + \int_0^t \left[\sum_{k=4}^{6} P_k(x) a_{k7} \right] dx \,, \tag{59}$$

or, taking into account that at any time the CTMC is in a single state:

$$P_7(t) = 1 - \sum_{k=1}^{6} P_k(t) \,. \tag{60}$$

The beginning of August's service time distribution function is represented in Figure 19, when $t \in [0, 1]$. The different transient probabilities ($\mathbb{P}_i(t)$, i=1, 2, .., 6) of the states of the phase-type distribution are also represented; these states are the transient states of the CTMC while state 7 is the absorbing state. The small time interval chosen for this figure ($t \in [0, 1]$) makes it possible to identify the various functions $\mathbb{P}_i(t)$. The Figure 20 giving the answers when $t \in [0, 3]$ shows that the distribution function of the service time tends to 1 while the probabilities of transient states tend to zero. Note that beyond an example of this size, it seems reasonable to use some software to calculate this distribution function of service time...

Let us now consider the case where, when he arrives, August is in the p position such as $r - C < p \leq r$, he finds the r servers already active. He will take advantage of one of the services already in progress. The overall service rate remains, at least initially, equal to $r\mu$. The distribution function of August's service time is still a phase-type distribution whose states are characterized by the pairs (n, m) as above. The initial state of this phase-type distribution is the state $(p, 0)$, $r - C < p \leq r$. From a state (n, m) such that $n > 1$ and $m < (r - C - 1)$, two transitions are possible: one transition to the state $(n - 1, m)$ with the rate $\min(r, (n + m + C))\mu$ and another one to the state $(n, m + 1)$ with the rate λ. When $n = 1$, the transition with the rate $(1 + m + C)\mu$ leads to the absorbing state signifying the end of the service and the departure of August. To know the distribution function of the service time, it is necessary to study this new acyclic CTMC to deduce the formal expression of this distribution function (as we did in the previous example) or the semi-formal expression using software. For example, taking the previous file where $r = 6$ and $C = 3$ and assuming this time that at his arrival, our August is in position $p = 4$, the diagram defining the phase service distribution is that in Figure 21. In this new situation, it is clear that the expectation of August s' service time is higher than the previous one. Let us now specify that, in this situation where p is greater than $(r - C)$, an initial simplification of the phase distribution is possible. This simplification is presented hereafter with the study of the case where p is greater than r.

Now, if upon arrival August finds himself in the p position such that $p > r$, he suffers a wait until the $(p - r)$ customers at the front of the queue are served. Only then

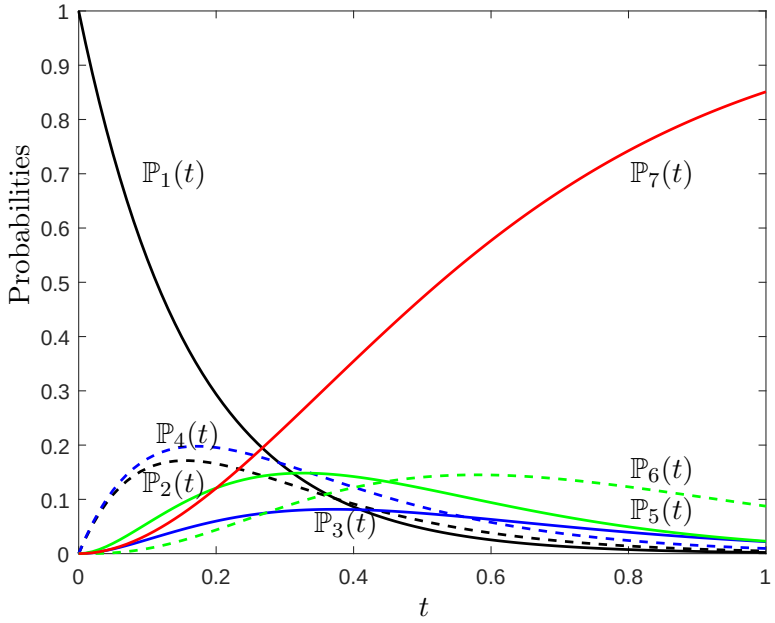


Figure 19: Transient state probabilities of the states of the CTMC as a function of time t, $t \in [0,1]$, for $p = 2$, $r = 6$ and $C = 3$, with $\rho = 0.8$, $\lambda = 3$ and so $\mu = 0.625$.

will August's service begin. But while he was waiting, some customers were able to arrive and stand behind him in the queue. These arrivals will allow for a longer overall maximum service rate because August's service will start in a state (r, m) such that $m > 0$. To identify the distribution function of the service time, one must know what is the probability that August begins his service in a state (r, m), $m = 0, 1, ..., r - C - 1$. For this purpose, we consider the CTMC consisting of states of the type (n, m) such that $\{(n, m)|n = r+1, r+2, ..., p \quad \text{et } m = 0, 1, ..., r-C-1.\}$ and $(r-C)$ absorbing states a_m, $m = 0, 1, ..., r-C-1$, whose transition graph is represented in Figure 22. This CTMC is not irreducible, it contains many transient classes and $(r - C)$ permanent classes corresponding to the absorbing states. The asymptotic state probability of an absorbing state a_m is the probability that August's service begins in the state (r, m). In view of the number of transient classes, such a calculation would be very tedious; but we can circumvent this difficulty by considering an irreducible and ergodic CMTC which reproduces indefinitely the execution of the distribution function of August's waiting (*cf.* the graph of transition represented in Figure 23). Since the overall service rates are the same whatever the value of m when August leaves the $(r + 1)$ position to end up in r position and start serving, we need only to know the relative values of the asymptotic state probabilities $p_{r+1,m}$ of the positive recurent states $(r + 1, m)$, $m = 0, 1, ..., r - C - 1,$. Indeed, if we know these asymptotic probabilities $p_{r+1,m}$ up to a multiplicative constant K, then the probability α_m that August's service starts after a transition from state $(r + 1, m)$ to state (r, m) will

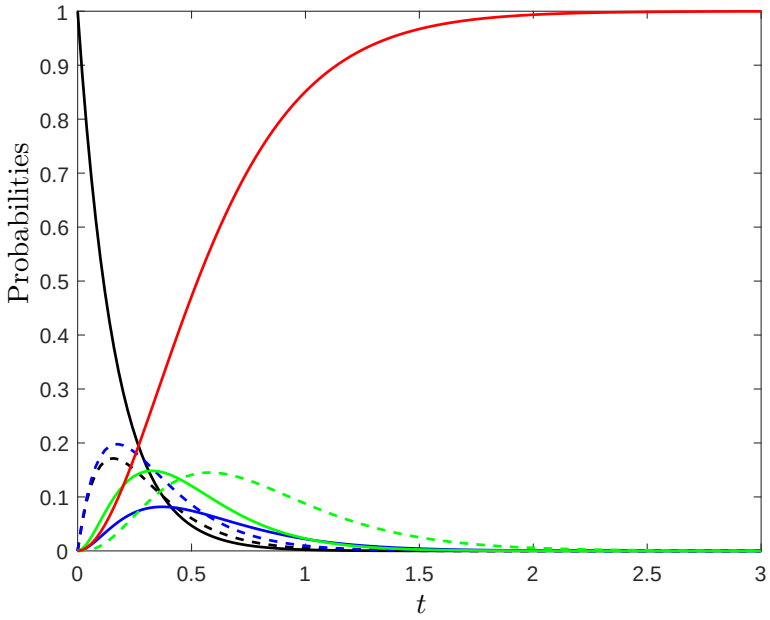


Figure 20: Transient state probabilities of the states of the CTMC as a function of time t, $t \in [0, 3]$, for $p = 2$, $r = 6$ and $C = 3$, with $\rho = 0.8$, $\lambda = 3$ and so $\mu = 0.625$.

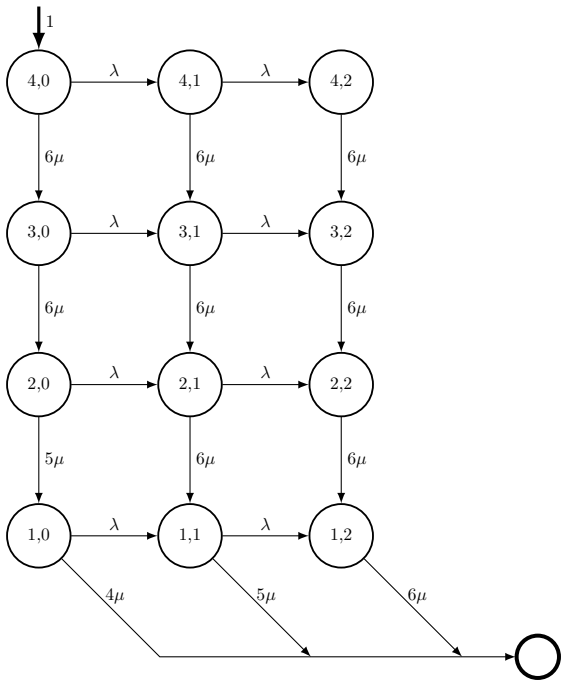


Figure 21: Phase-type service time distribution when $r - C < p \leq r$, in the particular case where $p = 4$, $r = 6$ and $C = 3$. None simplified representation.

be equal to:

$$\alpha_m = \frac{p_{r+1,m}}{\sum_{l=0}^{r-C-1} p_{r+1,l}} = \frac{Kp_{r+1,m}}{\sum_{l=0}^{r-C-1} Kp_{r+1,l}} \,. \tag{61}$$

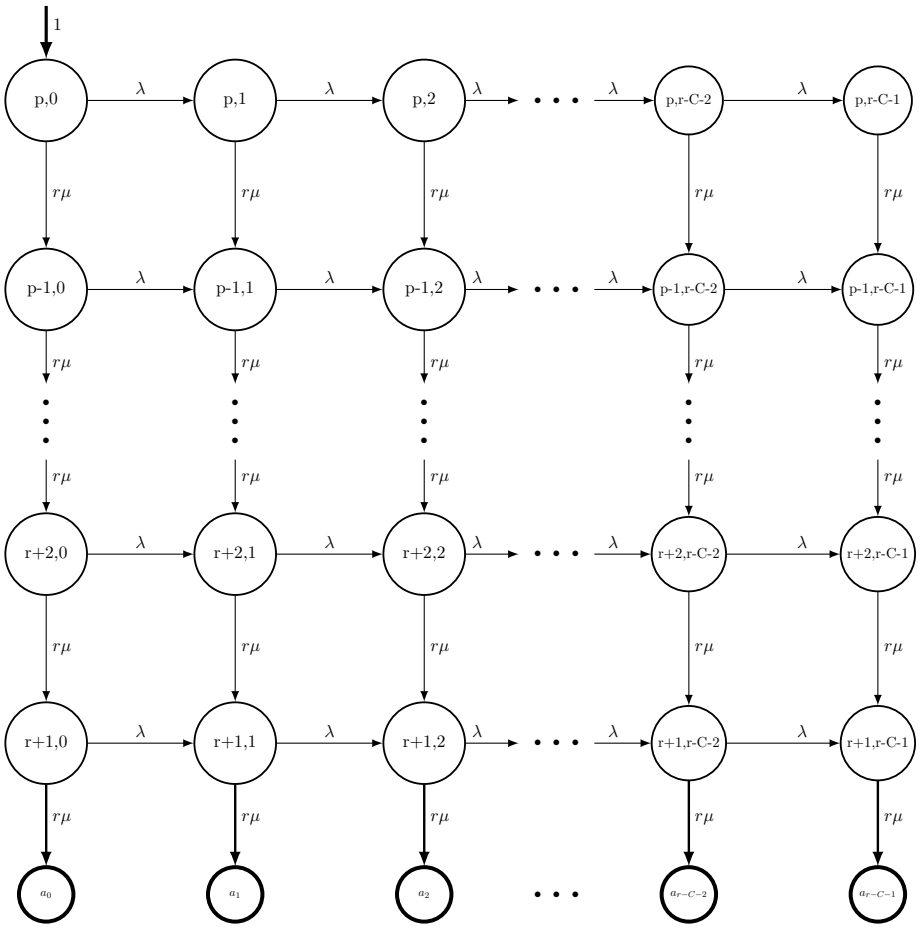


Figure 22: Non irreducible CTMC modeling August's waiting time distribution function.

Let us consider the partition of the state space of the CMTC represented in figure 23 such that:

$$E_i = \{(r+i, m) | m = 0, 1, ..., C-1\} \quad i = 1, 2, ..., (p-r) \,. \tag{62}$$

Thus, we partition the states into $(p-r)$ classes such that each class contains the states corresponding to a given position of August.

Let us introduce the $(p-r)$ state CMTC $Z(t)$ associated to the previous CMTC and such that $Z(t)$ goes from state i to $(i-1)$ when the initial CMTC transits from a state of class E_i to a state of class E_{i-1}, if $i = 2, ..., (p-r)$ and transitions from state 1 to state

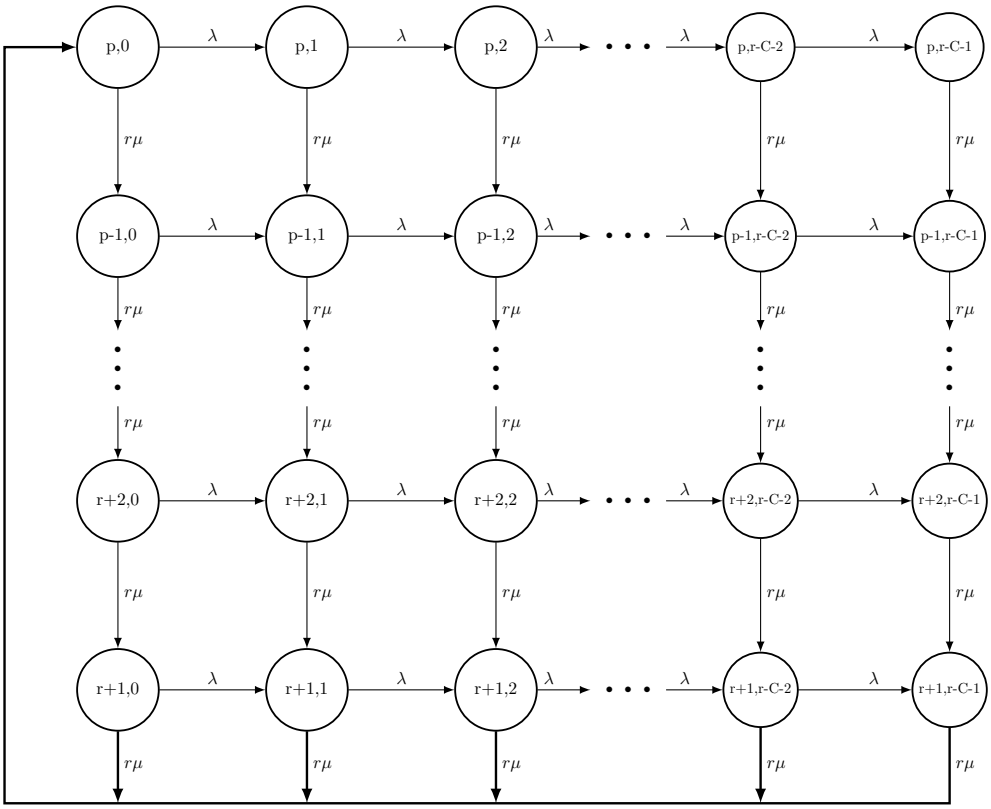


Figure 23: Irreducible CTMC modeling the execution of an infinity of independent realizations of the August's waiting time distribution function.

$(p-r)$ when $i=1$. Since all transition rates from states of class E_i to states not belonging to class E_i cause the CMTC to pass into a state of class E_{i-1} and are all equal to $r\mu$, the CMTC $Z(t)$ transits from state i to state $(i-1)$ with the rate $r\mu$, for $i=2,....,(p-r)$. The same is true for the transition from state 1 to state $(p-r)$. The fact that all the transition rates taking the initial chain from one class E_i to E_{i-1} are all equal allows to obtain an exact aggregation. Examination of the transition graph of the $Z(t)$ allows us to deduce that the stationary state probabilities π_i of this irreducible chain with $(p-r)$ states are all equal to $(p-r)^{-1}$. This leads, at the level of the CMTC of the Figure 23, that the sum of probabilities $\sum_{l=0}^{r-C-1} p_{i,l}$ is such that:

$$\sum_{l=0}^{r-C-1} p_{i,l} = (p-r)^{-1} \qquad i=1,2,...,(p-r)\ . \tag{63}$$

At the level of the CMTC $Z(t)$, the duration of a cycle between two entries in state 1 (or between two entries in any given state) follows an Erlang distribution of order $(p-r)$ and rate $r\mu$. Note that we already know that the distribution of August's waiting knowing

that he starts his waiting at position p, follows such a distribution (*cf.* the relation 54). But the use of the present approach concerns the determination of the vector $\boldsymbol{\alpha}$ in order to determine the initial probabilities of the service distribution of August, knowing that he starts his waiting at position p.

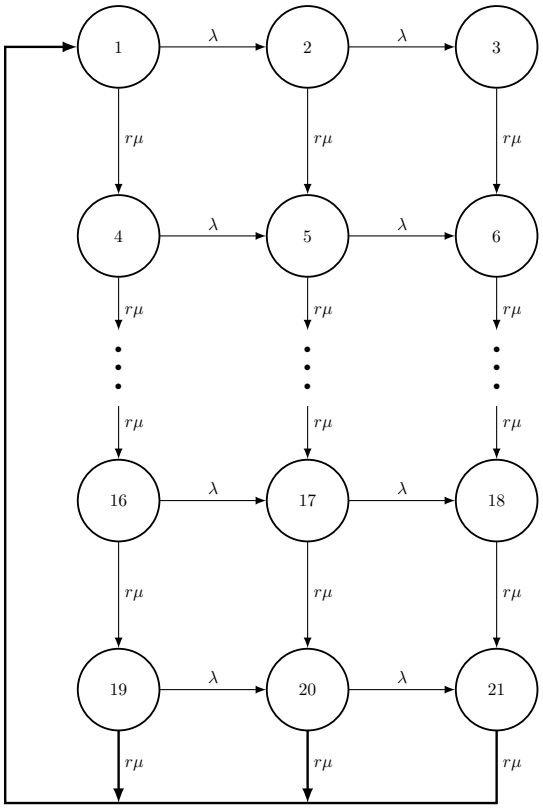


Figure 24: Irreducible CMTC; case $p = r + 7$.

Let's develop the approach in the particular case where, as before, $r - C = 3$ and where $p = r + 7$. Rename the states as noted on the representation of the graph of transition in Figure 24. Note p_h the asymptotic probability of the state h, h, $h = 1, 2, ..., 21$. We consider the cœfficients c_h equal to a multiplicative constant close to asymptotic probabilities. The c_h cœfficients also verify the Chapmann-Kolmogorov equations. For example, we have:

$$\begin{aligned}
\Lambda_r c_2 &= \lambda c_1 \,, \\
r\mu c_3 &= \lambda c_2 \,, \\
\Lambda_r c_4 &= r\mu c_1 \,, \\
\Lambda_r c_5 &= r\mu c_2 + \lambda c_4 \,, \\
r\mu c_6 &= r\mu c_3 + \lambda c_5 \,,
\end{aligned}$$

and so on. Note $\mathbf{c}_j$, $j = 0, 1, ..., 6$, the line vectors

$$\mathbf{c}_j = (c_{1+3j}, c_{2+3j}, c_{3+3j}) \,, \quad j = 0, 1, ..., 6 \,. \tag{64}$$

Recall that ρ is equal to $\lambda/r\mu$ and note that

$$\begin{aligned} c_4 &= \frac{1}{(1+\rho)} c_1 \,, \\ c_5 &= \frac{1}{(1+\rho)} c_2 + \frac{\rho}{(1+\rho)} c_4 = \frac{\rho}{(1+\rho)^2} c_1 + \frac{1}{(1+\rho)} c_2 \,, \\ c_6 &= c_3 + \rho c_5 = \frac{\rho^2}{(1+\rho)^2} c_1 + \frac{\rho}{(1+\rho)} c_2 + c_3 \,. \end{aligned}$$

In matrix form, we get:

$$\mathbf{c}_1 = \mathbf{c}_0 A \,, \tag{65}$$

where

$$A = \begin{bmatrix} \frac{1}{(1+\rho)} & \frac{\rho}{(1+\rho)^2} & \frac{\rho^2}{(1+\rho)^2} \\ 0 & \frac{1}{(1+\rho)} & \frac{\rho}{(1+\rho)} \\ 0 & 0 & 1 \end{bmatrix} \tag{66}$$

Given the repetition of the transition rate pattern on the graph (all rates from a state i to a state $i+1$ are equal to λ and all rates from a state i to a state $i+3$ are equal to $r\mu$), we obtain:

$$\mathbf{c}_j = \mathbf{c}_{j-1} A \,, \quad j = 1, ..., 6 \,. \tag{67}$$

or still

$$\mathbf{c}_j = \mathbf{c}_0 A^j \,, \quad j = 1, ..., 6 \,. \tag{68}$$

Let us introduce the unit column vector $\mathbf{e}$, transpose of the row vector $(1, 1, 1)$ and note that for any row vector (x_1, x_2, x_3), the matrix A has the property:

$$(x_1, x_2, x_3) A \mathbf{e} = (x_1, x_2, x_3) \mathbf{e} = x_1 + x_2 + x_3 \,. \tag{69}$$

This gives rise to the property

$$\mathbf{c}_j \mathbf{e} = \mathbf{c}_0 \mathbf{e} \,, \quad j = 1, ..., 6 \,. \tag{70}$$

This result is not surprising. It follows from the identity expressed by the equation (63).

Moreover the vector $\mathbf{c}_0$ can be written

$$\mathbf{c}_0 = (c_1, c_2, c_3) = c_1 (1, \rho/(1+\rho), \rho^2/(1+\rho)) \,, \quad . \tag{71}$$

This results in $\mathbf{c}_0 \mathbf{e} = c_1 (1 + \rho)$ and more generally:

$$\mathbf{c}_j \mathbf{e} = c_1 (1 + \rho) \,, \quad j = 0, 1, ..., 6 \,. \tag{72}$$

By fixing arbitrarily the constant K equal to the inverse of the probability q_1, the coefficient c_1 becomes equal to 1. Having computed the vector $\mathbf{c}_6 = \mathbf{c}_0 A^6$, we obtain

$$\alpha_0 = \frac{c_{19}}{(c_{19}+c_{20}+c_{21})}\,, \quad \alpha_1 = \frac{c_{20}}{(c_{19}+c_{20}+c_{21})}\,, \quad \alpha_2 = \frac{c_{21}}{(c_{19}+c_{20}+c_{21})}\,. \tag{73}$$

where we already know that the denominator $(c_{19}+c_{20}+c_{21})$ is equal to $(1+\rho)$.
In this case where $p = r+7$, the vector $\boldsymbol{\alpha}$ is equal to:

$$\boldsymbol{\alpha} = (0.0163, \quad 0.0508, \quad 0.9328)\,. \tag{74}$$

We notice that if August arrives in seventh position in the list of customers whose service has not yet started, he has a probability equal to 0.9328 that his service follows with certainty an Erlang distribution of order 6 and rate 6μ. Moreover, being seventh of the waiting customers when he arrives, the probability that August starts his service with at least one customer behind him is greater than 0.98.

More generally, assume that August starts his wait at position p, $p > r$. In this case, the vector $\boldsymbol{\alpha}_p = (\alpha_0, \alpha_1, \alpha_2)$ of the initial probabilities of the service distribution is obtained using the vector $\mathbf{c}_{p-r-1}$ given by the relation

$$\mathbf{c}_{p-r-1} = \mathbf{c}_0 A^{p-r-1}\,, \tag{75}$$

and from the expression of the vector sought as a function of the vector $\mathbf{c}_{p-r-1}$

$$\boldsymbol{\alpha}_p = \frac{1}{(1+\rho)}\mathbf{c}_{p-r-1}\,. \tag{76}$$

Finally, for a given p value, the vector $\boldsymbol{\alpha}_p$ is calculated using the relation:

$$\boldsymbol{\alpha}_p = \frac{1}{(1+\rho)}\mathbf{c}_0 A^{p-r-1}\,. \tag{77}$$

For example, in the case where $p = r+1$ the vector $\boldsymbol{\alpha}_{r+1}$ is collinear with the vector $\mathbf{c}_0$ (because $(p-r-1) = 0$) and takes the value:

$$\boldsymbol{\alpha}_{r+1} = (0.5556, \quad 0.2469, \quad 0.1975)\,, \tag{78}$$

which leads to a slightly higher expectation of service time than the Erlang distribution of the previous case (but in this case, the wait undergone by August will be *a priori* much shorter ...). For $p = r+2$ and $p = r+3$, the vectors $\boldsymbol{\alpha}_{r+2}$ and $\boldsymbol{\alpha}_{r+3}$ take the values:

$$\boldsymbol{\alpha}_{r+2} = (0.3086, \quad 0.2743, \quad 0.4170)\,, \tag{79}$$

$$\boldsymbol{\alpha}_{r+3} = (0.1715, \quad 0.2286, \quad 0.5999)\,, \tag{80}$$

values already very different from that of the vector $\boldsymbol{\alpha}_{r+1}$.

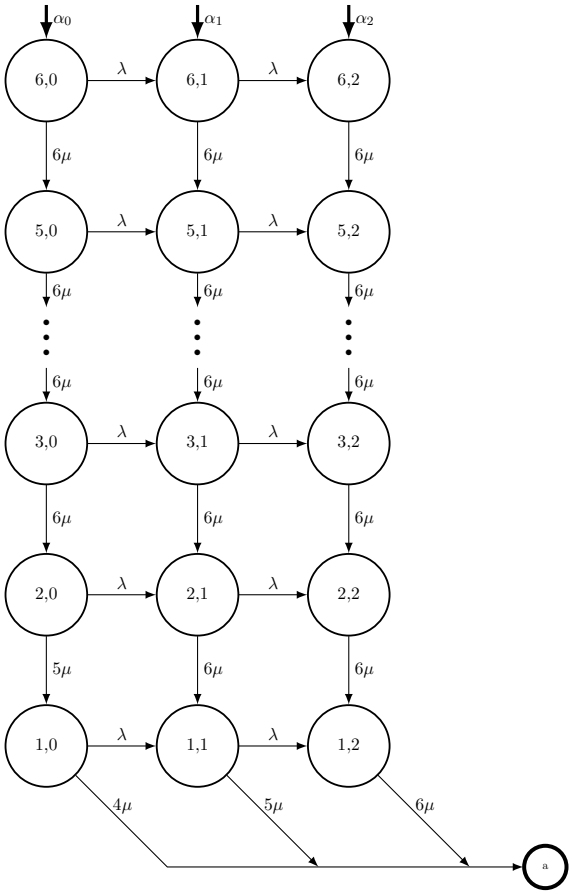


Figure 25: Phase-type service time distribution when $p > r$, for a given p, in the particular case when $r = 6$ and $C = 3$. Before simplification.

At this point, the corresponding distribution is the phase type distribution with three potential initial states shown in Figure 25. The existence of several possible initial states distinguishes this distribution from the one already studied in the case where p was not greater than r and illustrated in Figure 18. But there is still a significant simplification to be made to simplify the calculation of the distribution function of this new distribution (simplification promised on page 33). To do this, let us take the Figure 25 and consider a first part of the states of the phase type distribution corresponding to the subset

$$\{(n, m) | n > r - C - 1 \text{ et } m = 0, 1, ..., C - 1\} \quad . \tag{81}$$

Since the transition rates from any state (n, m) to a state $(n-1, m)$ are all equal to $r\mu$, this subset can be replaced by a linear chain comprising the $(C+1)$ states $r, (r-1), ..., (r-C)$ provided that one can determine the transition probabilities from the state $(r-C)$ to each of the states of the set

$$\{(r-C-1,m)|m=0,1,...,C-1\} \quad , \tag{82}$$

i.e., to be able to determine the new vector $\boldsymbol{\alpha}$ which we will note $\boldsymbol{\alpha}_p^+$. The transition rates from a new state j to a new state $(j-1)$, $j=r,(r-1),...,(r-C+1)$ are obviously equal to $r\mu$. If August arrives at position p in the list of clients whose service has not yet begun, the vector $\boldsymbol{\alpha}_p^+$ we are looking for is obtained by multiplying the previous vector (relative to the beginning of the service) by the matrix A^{C+1}. This gives the new vector $\boldsymbol{\alpha}_p^+$:

$$\boldsymbol{\alpha}_p^+ = \frac{1}{(1+\rho)}\mathbf{c}_0 A^{p-r+C}\,. \tag{83}$$

We thus obtain a phase type distribution with fewer states as shown in Figure 26 associated with the previously studied example. The number of states in the phase type distribution has thus been reduced from 18 to 10 states. Let us notice that in our example, the vectors $\boldsymbol{\alpha}_{r+1}^+$, $\boldsymbol{\alpha}_{r+2}^+$, $\boldsymbol{\alpha}_{r+3}^+$, that take the values:

$$\boldsymbol{\alpha}_{r+1}^+ = (0.0529, \quad 0.1176, \quad 0.8295)\,, \tag{84}$$

$$\boldsymbol{\alpha}_{r+2}^+ = (0.0294, \quad 0.0784, \quad 0.8922)\,, \tag{85}$$

$$\boldsymbol{\alpha}_{r+3}^+ = (0.0163, \quad 0.0508, \quad 0.9328)\,, \tag{86}$$

values largely different from the vectors $\boldsymbol{\alpha}_{r+1}$, $\boldsymbol{\alpha}_{r+2}$ and $\boldsymbol{\alpha}_{r+3}$ calculated previously. This is not surprising if we consider the simple relationship between $\boldsymbol{\alpha}_p^+$ and $\boldsymbol{\alpha}_p$:

$$\boldsymbol{\alpha}_p^+ = \boldsymbol{\alpha}_{p+C+1}\,. \tag{87}$$

As we announced on page 33, this simplification is applicable in the situation where our client August is in position p, such that $r-C<p\leq r$. In this case, the Erlang distribution is of order $(p-r+C+1)$ and the vector $\boldsymbol{\alpha}_p^+$ provides the transition probabilities from the state $(r-C)$ to the states $((r-C-1),l)$, $l=0,1,2$. The relation 87 still allows to determine this vector $\boldsymbol{\alpha}_p^+$. Let us add that if $p=(r-C)$, the arrival of August will cause the activation of the last inactive server and will increase the global service rate to $r\mu$, which allows a small simplification: the Erlang distribution is of order 1 (it is the exponential distribution of rate $r\mu$) and the $\boldsymbol{\alpha}_p^+$ vector is equal to $\mathbf{c}_0/(1+\rho)$.

Note that if $\boldsymbol{\alpha}_p$ does not depend on C (*cf.* 77), $\boldsymbol{\alpha}_p^+$ does depend on C. Depending on the number of states, the distribution function of this new service distribution can be determined according to one of the methods already mentioned.

When p tends to infinity, August's service distribution tends to the limiting distribution corresponding to the situation where at least $(r-C)$ customers are present behind him at the beginning of his service. In this limiting case, the service distribution is the Erlang distribution of order r and rate $r\mu$.

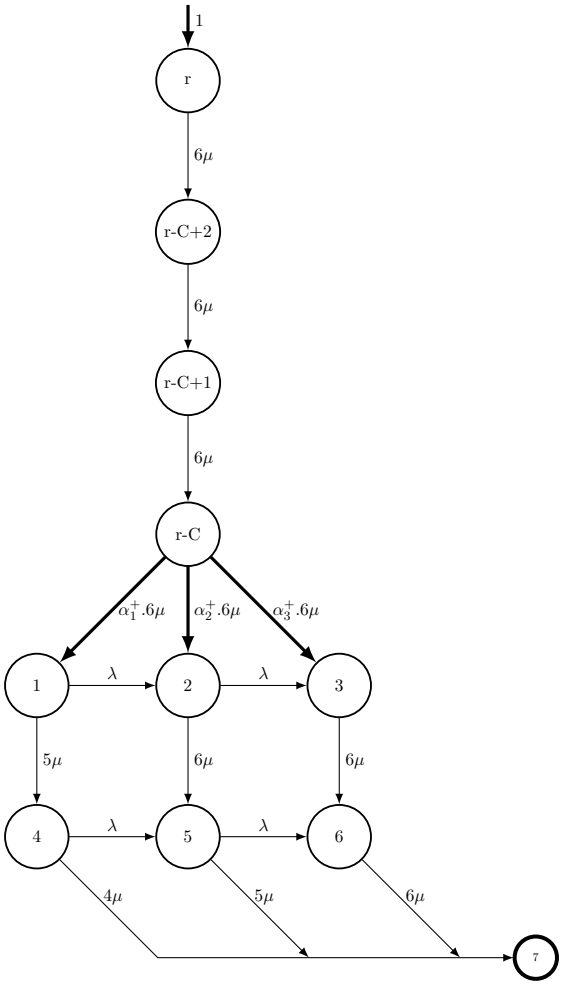


Figure 26: Phase-type service time distribution when $p > r$, for a given p, in the particular case when $r = 6$ and $C = 3$. After simplification.

If we now look for the distribution of August's service time knowing only that he waited (*i.e.*, $p > r$), an event of probability $\mathbb{P}(W > 0)$, we need to determine the vector $\boldsymbol{\alpha}$ conditional on this event. Knowing that the unconditional probability that August is in the queue in p-th position is equal to $p_{(C+p-1)}$, we can give an expression for the vector $\boldsymbol{\alpha}$ conditional on the event $[p > r]$, a vector denoted $\boldsymbol{\alpha}_{att}$:

$$\boldsymbol{\alpha}_{att} = \frac{1}{\mathbb{P}(W > 0)} \frac{1}{(1+\rho)} \sum_{p=r+1}^{\infty} p_{(C+p-1)} \mathbf{c}_0 A^{p-r-1} \,. \tag{88}$$

Or, using the expression of $\mathbb{P}(W>0)$ given in 30:

$$\begin{aligned}
\boldsymbol{\alpha}_{att} &= \frac{1}{\mathbb{P}(W>0)}\frac{1}{(1+\rho)}\sum_{p=r+1}^{\infty}\frac{\rho^{(C+p-1)}r^r}{r\,!}p_0\mathbf{c}_0A^{p-r-1}\,,\\
&= \frac{1}{\mathbb{P}(W>0)}\frac{r^rp_0}{r\,!(1+\rho)}\sum_{p=r+1}^{\infty}\rho^{(C+p-1)}\mathbf{c}_0A^{p-r-1}\,,\\
&= \frac{1}{\mathbb{P}(W>0)}\frac{r^rp_0\rho^{C+r}}{r\,!(1+\rho)}\mathbf{c}_0\sum_{p=r+1}^{\infty}\rho^{p-r-1}A^{p-r-1}\,,\\
&= \frac{(1-\rho)}{(1+\rho)}\mathbf{c}_0\sum_{l=0}^{\infty}(\rho A)^l\,,\\
&= \frac{(1-\rho)}{(1+\rho)}\mathbf{c}_0\sum_{l=0}^{\infty}B^l\,,
\end{aligned}$$

where $B=\rho A$. If the spectral radius of the matrix B is less than 1 (i.e. if the eigenvalues of the matrix B are less than unity in modulus), we can still write

$$\boldsymbol{\alpha}_{att}=\frac{(1-\rho)}{(1+\rho)}\mathbf{c}_0(I-B)^{-1}\,. \tag{89}$$

Let us write matrix B using the version of A given by 66:

$$B=\begin{bmatrix}\dfrac{\rho}{(1+\rho)} & \dfrac{\rho^2}{(1+\rho)^2} & \dfrac{\rho^3}{(1+\rho)^2}\\ 0 & \dfrac{\rho}{(1+\rho)} & \dfrac{\rho^2}{(1+\rho)}\\ 0 & 0 & \rho\end{bmatrix}$$

The determinant of the matrix $(B-\lambda I)$ is equal to $(\rho-\lambda)\left(\dfrac{\rho}{(1+\rho)}-\lambda\right)^2$. The characteristic equation of the matrix B thus admits the simple root $\lambda_1=\rho$ and the double root $\lambda_2=\dfrac{\rho}{(1+\rho)}$. These roots are by definition positive and less than 1 (the spectral radius of the matrix B is equal to ρ). The matrix $(I-B)^{-1}$ exists. We can therefore compute it by starting from:

$$(I-B)=\begin{bmatrix}\dfrac{1}{(1+\rho)} & -\dfrac{\rho^2}{(1+\rho)^2} & -\dfrac{\rho^3}{(1+\rho)^2}\\ 0 & \dfrac{1}{(1+\rho)} & -\dfrac{\rho^2}{(1+\rho)}\\ 0 & 0 & 1-\rho\end{bmatrix}$$

The determinant of the matrix $(I-B)$ is equal to $(1-\rho)\left(\dfrac{1}{(1+\rho)}\right)^2$. We finally obtain:

$$(I-B)^{-1} = \begin{bmatrix} (1+\rho) & \rho^2 & \dfrac{\rho^3}{(1-\rho)} \\ 0 & (1+\rho) & \dfrac{\rho^2}{(1-\rho)} \\ 0 & 0 & \dfrac{1}{(1-\rho)} \end{bmatrix}$$

We can therefore use the relation 89 to determine the vector $\boldsymbol{\alpha}_{att}$, knowing just that August waited. We obtain:

$$\boldsymbol{\alpha}_{att} = ((1-\rho), \quad \rho(1-\rho), \quad \rho^2)\,, \tag{90}$$

If we want to look for the service time distribution of August knowing that he waited, it is better to use the simplified representation and the vector $\boldsymbol{\alpha}^+_{att}$ computable thanks to the relation:

$$\boldsymbol{\alpha}^+_{att} = \boldsymbol{\alpha}_{att} A^{C+1}\,, \tag{91}$$

Recall that if $\boldsymbol{\alpha}_{att}$ does not depend on C (*cf.* 89), of course, $\boldsymbol{\alpha}^+_{att}$ depends on C.

In the numerical example considered above, ρ is equal to 0.8. This gives the vectors:

$$\begin{aligned} \boldsymbol{\alpha}_{att} &= (0.2000, \quad 0.1600, \quad 0.6400) \\ \boldsymbol{\alpha}^+_{att} &= (0.0191, \quad 0.0491, \quad 0.9318)\,. \end{aligned}$$

In the Figures 27 and 28, we represent respectively the beginning and the end side of the distribution function of the service time random variable in the following situations: $p = r + 1$, $W > 0$ (we just know that p is greater than r), as well as the limit when p tends to infinity. We can see that the three corresponding functions are close (hence the use of two distinct figures). The values of the parameters of the queue are those of the previous example $r = 6$ and $C = 3$. These functions were determined using the semi-formal software ACE.

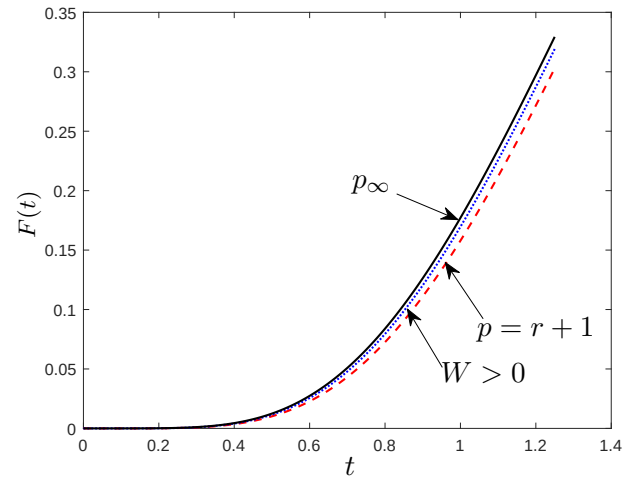


Figure 27: Start of the service time distribution function, $r = 6$, $C = 3$; with $\rho = 0.8$, $\lambda = 3$ and so $\mu = 0.625$.

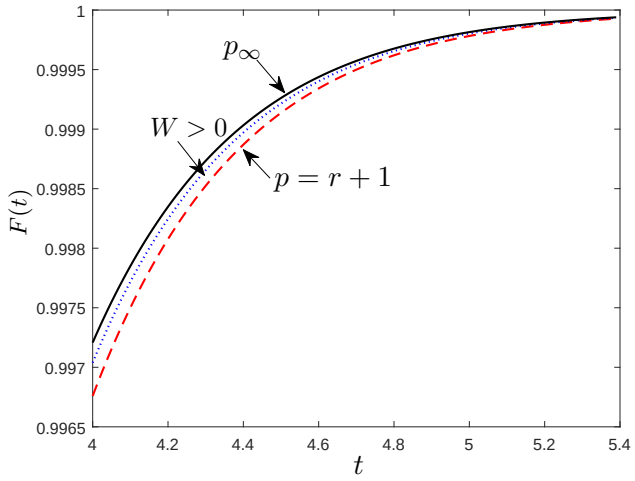


Figure 28: End side of the service time distribution function, $r = 6$, $C = 3$; with $\rho = 0.8$, $\lambda = 3$ and so $\mu = 0.625$.

It is also possible to evoke the response time distribution but, in order to restrict (and simplify) the discourse, we will present the approach by considering only the example already considered, the one where $r = 6$ and $C = 3$.

Let's start by looking for the distribution of the response time knowing that August starts his stay by waiting ($p > r$); we already know (*cf.* relation 55) that the conditional

distribution of the waiting is an exponential distribution of rate $(r\mu - \lambda)$. The conditional distribution of the response time knowing that August starts his stay by waiting can thus be calculated as a phase distribution starting with an initial phase of rate $(r\mu - \lambda)$ and continuing with the first phase of the phase distribution given in Figure 26. The diagram of the phase distribution corresponding to the response time knowing that August begins his stay with a wait (knowing $W > 0$ and thus $p > r$) is represented in Figure 29. The coefficients $\alpha^+_{att,i}$, i=1,2,3, are the components of the vector $\boldsymbol{\alpha}^+_{att}$.

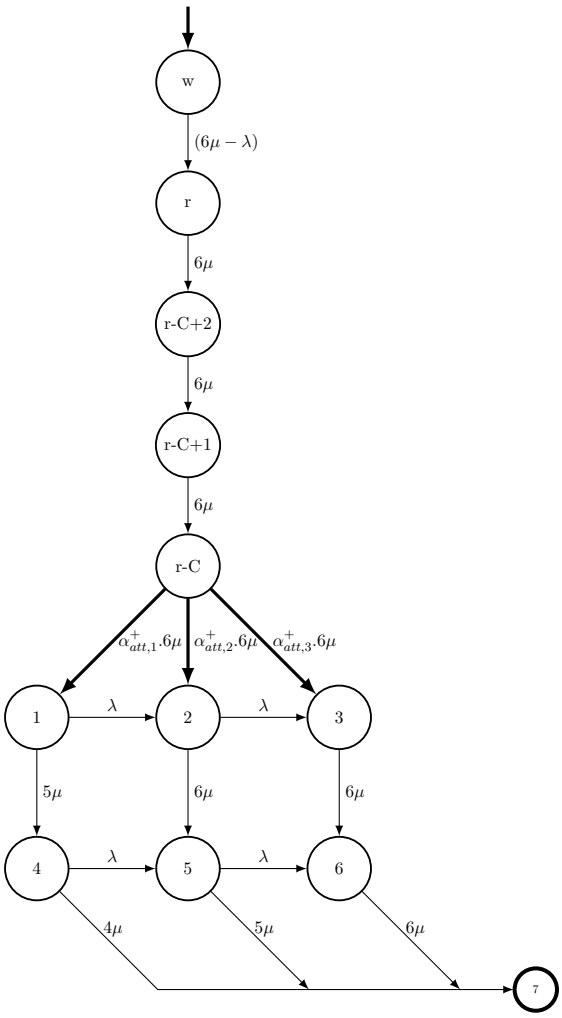


Figure 29: Phase-type response time distribution knowing $W > 0$,
in the particular case where $r = 6$ and $C = 3$.

In the Figure 30, we represent the response time distribution function knowing $W > 0$, with $r = 6$ and $C = 3$. In this figure, we have added (in dotted line) the distribution function of the Erlang limit service distribution of order r and rate $r\mu$ in order to appre-

ciate the experienced wait. Again, these functions were determined using the semi-formal software ACE.

Let us specify that the graphical representation of this distribution function of the residency distribution knowing $W > 0$ in the case $C = 4$ (and all data being equal) would be almost undistinguishable from that where $C = 3$. Indeed, for this conditional distribution, the parameter C only intervenes on the service distribution which, thanks to the value of the vector $\boldsymbol{\alpha}^{+}_{att}$, is already very close to the Erlang-r distribution.

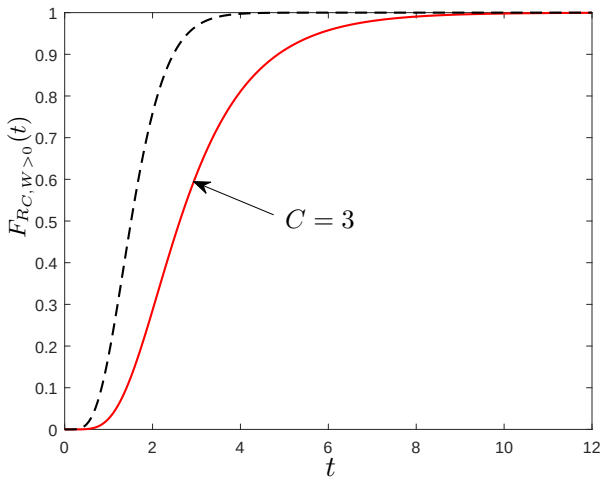


Figure 30: Distribution function $F_{R_{C,W>0}}(t)$ of the response time random variable knowing $W > 0$, when $r = 6$, $C = 3$; with $\rho = 0.8$, $\lambda = 3$ and thus $\mu = 0.625$. The dotted curve represents the distribution function of the Erlang limit service CTMC of order r and rate $r\mu$.

In the present case where $C > (r - 1)$, the representation of the (unconditional) residence distribution $F_{R_C}(t)$ would be tedious but not infeasible. We already know that the distribution of R_C admits a mass at $t = 0$ equal to $\mathbb{P}(S_C = 0)$. When the service is no longer immediate and August finds himself in position p, $p \leq r$, with probability p_{C-1+p}, his stay time is equal to his service time, whose distribution function for any value of p, $p = 1, 2, ..., r$, we have seen previously. Finally, it remains to add to this sum the conditional distribution function $F_{C,W>0}(t)$ that we have just determined (see Fig. 30), weighted by the probability $\mathbb{P}(W > 0)$. This can be summarized as:

$$F_{R_C}(t) = \mathbb{P}(S_C = 0)u(t) + \sum_{q=1}^{r} p_{C-1+q} F_{R_{C,p=q}}(t) + \mathbb{P}(W > 0) F_{R_{C,W>0}}(t)\,. \qquad (92)$$

4 Multi-server Queue with state-dependent rates

In this section, we study proactive queues where the arrival rate $\lambda(i,j)$ and the service rate $\mu(i,j)$ can depend on the variables i and/or j. For these queues, we will use the notation $\lambda(i,j)/\mu(i,j)/ < C >$. The study of their performances requires the writing of algorithms but before presenting them, let us give some concrete examples of this family of models. The first example concerns the study of a finite population of N individuals where the arrival rate $\lambda(i,j)$ depends only on the number j of customers present in the queue:

$$\lambda(j) = (N-j)\lambda_e \quad j \leq N\,. \tag{93}$$

The second example corresponds to the case where the service is not carried out on the place of contact with the customers but is carried out by a distant (and possibly complex) process on request of the site receiving the customers. Without addressing the technique of obtaining it for the moment, we can assume that we are able to determine a sequence of service rates $\{\mu(i,j)\}$ depending only on the number $(C+j-i)$ of requests issued and not satisfied.

The variable k defined by the relation 1 of the page 4 can still be used to characterize the state of the queue since the knowledge of k makes it possible to find the couple (i,j) thanks to the relations 2 and 3 (page 4). In order not to make the expressions too long, we will respectively note $\lambda(k)$ and $\mu(k)$ the rates $\lambda(i(k),j(k))$ and $\mu(i(k),j(k))$ associated to the value of k.

The study of the performance of these queues starts with the determination of the steady state probabilities. These probabilities satisfy the relations:

$$p_k = \frac{\lambda(k-1)}{\mu(k)} p_{k-1} \quad k > 0. \tag{94}$$

Let us set

$$f(k) \stackrel{\triangle}{=} \frac{\lambda(k-1)}{\mu(k)} \qquad \text{and} \qquad H(k) \stackrel{\triangle}{=} \prod_{n=1}^{k} \left(\frac{\lambda(n-1)}{\mu(n)} \right) .$$

This allows us to express all probabilities in terms of the probability p_0.:

$$p_k = H(k)p_0 \quad k > 0. \tag{95}$$

The constraint $k > 0$ can be ignored provided that $H(0) \stackrel{\triangle}{=} 1$.

If the value of the arrival rate $\lambda(k)$ becomes zero for $k \geq K_{lim}$, the number of states in the queue is equal to $(K_{lim}+1)$ and the probability p_0 is obtained thanks to the normalization equation:

$$p_0 = \left[\sum_{k=0}^{K_{lim}} H(k) \right]^{-1} .$$

When j is strictly positive, the value of i is always zero and the arrival rate $\lambda(k)$ can only depend on the number j of customers present in the queue. A priori, the sub-sequence $\{\lambda(k)\}$ is then decreasing, eventually with a null limit value for $j \geq J_{lim}$. On the other hand, the sub-sequence $\{\mu(j)\}$, $j > 0$, is assumed here to be non-decreasing. This assumption is (naturally) satisfied in all real systems modeled by queues of infinite capacity and admitting an asymptotic state probability distribution (in theory, one could also accept a sub-sequence decreasing to a sufficiently high lower bound value to satisfy the ergodicity conditions).

If the limit value of $\lambda(k)$ is not zero and does not have a simple formal expression, things get complicated because it is out of the question to calculate the function $H(k)$ until infinity. We must then accept a slight approximation. We suppose that there exists a value α, $\alpha < 1$, and a value K_α such that

$$\frac{\lambda(k-1)}{\mu(k)} \leq \alpha \quad k \geq K_\alpha \,. \tag{96}$$

Then, we consider a first queue of limited capacity such that $k = 0, 1, ..., K_\alpha$ and a second queue of infinite capacity whose function similar to the function $f(k)$, denoted $\widetilde{f}(k)$, is defined as a function of $f(k)$ as follows

$$\widetilde{f}(k) \stackrel{\triangle}{=} \begin{cases} f(k) & \text{if } k \leq K_\alpha \,, \\ f(K_\alpha) & \text{if } k \geq K_\alpha \,. \end{cases}$$

The previous definition allows us to obtain the expression of the function similar to the function $H(k)$, noted $\widetilde{H}(k)$:

$$\widetilde{H}(k) \stackrel{\triangle}{=} \begin{cases} H(k) & \text{if } k \leq K_\alpha \,, \\ H(K_\alpha) f(K_\alpha)^{k-K_\alpha} & \text{if } k \geq K_\alpha \,. \end{cases}$$

The present approximation relies on the fact that, if the boundary value of $\lambda(k)$ is not zero, the value α is chosen to be sufficiently small so that K_α is sufficiently large and the majoration of $f(k)$ by $f(K_\alpha)$, $k \geq K_\alpha$, does not lead to a significant error on the computed numerical results.

For the queue of limited capacity, the probability, noted $\overline{p_0}$, a upper bound of p_0 is obtained by the normalization equation:

$$\overline{p_0} = \left[\sum_{k=0}^{K_\alpha} H(k) \right]^{-1} .$$

It is easy to see that $\overline{p_0}$ is strictly greater than p_0. Note that to avoid an *overflow* during its computation, this approximate probability $\overline{p_0}$ can be obtained by the recursion

$$\overline{p_0}(n) = \frac{\overline{p_0}(n-1)}{1 + H(n)\overline{p_0}(n-1)} \quad n = 1, 2, ...K_\alpha \quad \text{avec } \overline{p_0}(0) = 1. \tag{97}$$

Note that the sequence $\overline{p_0}(n)$ is strictly decreasing since $(1 + H(n)\overline{p_0}(n-1))$ is strictly greater than 1.

For the approximate queue of infinite capacity, the probability $\underline{p_0}$, a lower bound of p_0, is written:

$$\begin{aligned}
\underline{p_0} &= \left[\sum_{k=0}^{K_\alpha} \widetilde{H}(k) + \sum_{k=K_\alpha+1}^{\infty} \widetilde{H}(k)\right]^{-1}, \\
&= \left[\sum_{k=0}^{K_\alpha} H(k) + \sum_{k=K_\alpha+1}^{\infty} H(K_\alpha)\left(f(K_\alpha)\right)^{k-K_\alpha}\right]^{-1}, \\
&= \left[\sum_{k=0}^{K_\alpha} H(k) + H(K_\alpha)f(K_\alpha) \sum_{k=K_\alpha+1}^{\infty} \left(f(K_\alpha)\right)^{k-(K_\alpha+1)}\right]^{-1}, \\
&= \left[\sum_{k=0}^{K_\alpha} H(k) + \frac{H(K_\alpha)f(K_\alpha)}{(1-f(K_\alpha))}\right]^{-1} \qquad \text{because } f(K_\alpha) < 1.
\end{aligned}$$

Note that if $f(k) = f(K_\alpha)$ for $k \geq K_\alpha$, then $p_0 = \underline{p_0}$. In the opposite case ($\widetilde{H}(k) > H(k)$), the probability p_0 is strictly greater than $\underline{p_0}$. Let us now examine the ratio of $\underline{p_0}$ to $\overline{p_0}$:

$$\begin{aligned}
\frac{\underline{p_0}}{\overline{p_0}} &= \left[\sum_{k=0}^{K_\alpha} H(k) + \frac{H(K_\alpha)f(K_\alpha)}{(1-f(K_\alpha))}\right]^{-1} \times \left[\sum_{k=0}^{K_\alpha} H(k)\right], \\
&= \left[1 + \frac{H(K_\alpha)f(K_\alpha)}{\left(\sum_{k=0}^{K_\alpha} H(k)\right)(1-f(K_\alpha))}\right]^{-1} = \left[1 + \frac{\overline{p_0}H(K_\alpha)f(K_\alpha)}{(1-f(K_\alpha))}\right]^{-1}, \\
&= \frac{(1-f(K_\alpha))}{(1-f(K_\alpha)) + \overline{p_0}H(K_\alpha)f(K_\alpha)}, \\
&= 1 - \frac{\overline{p_0}H(K_\alpha)f(K_\alpha)}{(1-f(K_\alpha)) + \overline{p_0}H(K_\alpha)f(K_\alpha)}.
\end{aligned} \tag{98}$$

Let us put:

$$\varepsilon(K_\alpha) \stackrel{\triangle}{=} \frac{\overline{p_0}H(K_\alpha)f(K_\alpha)}{(1-f(K_\alpha)) + \overline{p_0}H(K_\alpha)f(K_\alpha)} \tag{99}$$

and let us illustrate these approximations on a proactive $M/M/r/ < C >$ queue subjected to a load factor of 0.7 and having 5 servers. For this queue, the function $f(k)$ becomes constant very quickly (for $k \geq r$) and as soon as K_α is chosen to be equal to or greater than r, the minor value is equal to the true value p_0. When $K_\alpha = 40$, we obtain $\overline{p_0} = 0.0259$ and $\varepsilon(40) = 1.0019 \times 10^{-6}$. When $K_\alpha = 50$, we obtain $\overline{p_0} = 0.0259$ and $\varepsilon(50) = 2.8302 \times 10^{-8}$; while $H(40) = 1.6580 \times 10^{-6}$ and $H(50) = 4.6835 \times 10^{-7}$. We notice that the two values of K_α correspond to two very close $\overline{p_0}$ values. The observation of the formula 97 and the very low values of $H(40)$ and $H(50)$ allow to understand the reason of this very strong

proximity of the values. We can say that these approximations provide near-accurate results with low computational cost.

The method of calculation of the upper bound $\overline{p_0}$ consists in calculating simultaneously $\overline{p_0}(n)$, via the recurrence formula 97 and $\varepsilon(n)$ in terms of $\overline{p_0}(n)$ according to the formula

$$\varepsilon(n) \stackrel{\triangle}{=} \frac{\overline{p_0}(n)H(n)f(n)}{(1-f(n)) + \overline{p_0}(n)H(n)f(n)} ,$$

until $\varepsilon(n)$ becomes lower than a chosen bound (*eg,* 10^{-5}).

The relation 98 allows then to obtain the value of the lower bound $\underline{p_0}$ according to $\overline{p_0}$.

Once the probabilities $\overline{p_0}$ and $\underline{p_0}$ of the approximate distributions have been calculated, we can calculate the other probabilities $\overline{p_k}$ and $\underline{p_k}$

$$\overline{p_k} = \widetilde{H}(k)\overline{p_0} , \quad k = 1, 2, ..., K_\alpha \qquad \text{et} \qquad \underline{p_k} = \widetilde{H}(k)\underline{p_0} , \quad k = 1, 2, 3, ...$$

By construction of the probabilities $\overline{p_0}$ and $\underline{p_0}$, we have indeed:

$$\sum_{k=0}^{K_\alpha} \overline{p_k} = 1 , \quad \sum_{k=0}^{\infty} \underline{p_k} = 1 . \tag{100}$$

Let us also note that

$$\frac{p_k}{\overline{p_k}} = \frac{p_0}{\overline{p_0}} , \quad k = 1, 2, ..., K_\alpha \qquad \text{and} \qquad \frac{p_k}{\underline{p_k}} = \frac{p_0}{\underline{p_0}} , \quad k = 1, 2, 3, ... \tag{101}$$

Note that in the very special case where $\widetilde{H}(k) = H(k), k \geq K_{alpha}$, the exact probability distribution $\{p_k\}$is easy to calculate since $p_k = \underline{p_k}$, $k \geq 0$. In the general case where $\widetilde{H}(k) > H(k), k > K_\alpha$, note that there exists a value K_{cross} where the probability $\underline{p_k}$ becomes greater than the probability p_k:

$$K_{cross} = \min_k \{\underline{p_k} > p_k\} .$$

Before turning to performance indicators, let's recall the notion of stochastic dominance.

Definition 4.1. *Let X and Y be two real* RV *admitting $F_X(.)$ and $F_Y(.)$ as distribution function. The* RV *X is stochastically dominated by the* RV *Y if*

$$F_X(u) \geq F_Y(u) , \forall u \in \mathbb{R} . \tag{102}$$

This stochastic dominance (of the first order) is generally noted $X \prec^{st} Y$. This dominance means that small values are more likely for RV X than for RV Y. If the expectations of X and Y exist, it is easy to verify that the first order stochastic dominance leads to:

$$\mathbb{E}[X] \leq \mathbb{E}[Y] \; ; \tag{103}$$

It is sufficient to use the fact that here the expectation $E[X]$, resp. $E[Y]$, is also equal to the integral of the complementary function of its distribution function (*i.e.,* $(1 - F_X(u))$), resp. $(1 - F_Y(u))$)

It is also possible to show that if an X is stochastically dominated to the first order by a RV Y and if g is a **increasing** function, then:

$$\mathbb{E}[g(X)] \leq \mathbb{E}[g(Y)] \; . \tag{104}$$

Let us specify that if the inequality is strict at the level of distribution functions (relation 102), it also becomes strict at the level of expectations (relations 103 and 104).

We will also need to compare expectations $\mathbb{E}[g(.)]$ when g is a decreasing function. In such a case, letting us put $s \stackrel{\triangle}{=} -g$, s is an increasing function and we obtain, if the RV X is stochastically dominated at first order by the RV Y, $E[s(X)] \leq \mathbb{E}[s(Y)]$ and thus $E[-g(X)] \leq \mathbb{E}[-g(Y)]$ or $-\mathbb{E}[g(X)] \leq -\mathbb{E}[g(Y)]$, that is:

$$\mathbb{E}[g(X)] \geq \mathbb{E}[g(Y)] \qquad \text{(case when } g \text{ is } \textbf{decreasing}) \; . \tag{105}$$

Let us return to the relationship 101. In the general case where $\widetilde{H}(k) > H(k), k > K_\alpha$, this one implies that $\underline{p_k} < \overline{p_k}$, $k = 0, 1, 2, ..., K_\alpha$ and that $\underline{p_k} < p_k$, $k = 0, 1, 2, ...$ this implies that:

$$\sum_{j=0}^{k} \overline{p_j} > \sum_{j=0}^{k} \underline{p_j} \; , \quad \forall k \in \mathbb{N} \qquad \text{et} \qquad \sum_{j=0}^{k} p_j > \sum_{j=0}^{k} \underline{p_j} \; , \quad \forall k \in \mathbb{N} \; . \tag{106}$$

In other words, the distribution $\{\underline{p_k}\}_{k\geq 0}$ stochastically dominates (to first order) the distribution $\{\overline{p_k}\}_{k\geq 0}$ as well as the exact (but not calculated) distribution $\{p_k\}_{k\geq 0}$. As it is also easy to verify that the exact distribution stochastically dominates the $\{\overline{p_k}\}_{k\geq 0}$, we finally obtain:

$$\{\overline{p_k}\}_{k\geq 0} \prec^{st} \{p_k\}_{k\geq 0} \prec^{st} \{\underline{p_k}\}_{k\geq 0} \; . \tag{107}$$

Let us insist on the fact that the stochastic dominance of $\{\underline{p_k}\}_{k\geq 0}$ on $\{p_k\}_{k\geq 0}$ exists only in the general case where $\widetilde{H}(k) > H(k), k > K_\alpha$ since in the special case, the distribution $\{\underline{p_k}\}_{k\geq 0}$ corresponds to the exact distribution.

Let's now look at the performance indicators, and among these, first of all those that take the form of:

$$IP_i = \sum_{k=k_{inf}}^{\infty} Y_i(k) p_k \; , \tag{108}$$

where $Y_i(k)$ is an increasing function of k. This is the case for several indicators such as $\mathbb{E}[\nu_C], \mathbb{E}[NS_C], \mathbb{E}[N_C]$ ...

If $IP_{i,\,(\overline{p_0})}$, resp. $IP_{i,\,(\underline{p_0})}$, denotes the performance indicator IP_i calculated with the distribution $\{\overline{p_k}\}_{k\geq 0}$, resp. $\{\underline{p_k}\}_{k\geq 0}$, the relations 104 and 107 allow us to write $IP_{i,\,(\overline{p_0})} \leq IP_i \leq IP_{i,\,(\underline{p_0})}$. But this search for bounds is only useful for the general case where $\widetilde{H}(k) > H(k), k > K_\alpha$. And in this general case, the strict inequalities obtained by the relation 106 allow to write here:

$$IP_{i,\,(\overline{p_0})} < IP_i < IP_{i,\,(\underline{p_0})} \qquad \text{(general case)} \,. \tag{109}$$

Let us apply the method by considering the evaluation of the expectation of the number of pending requests in such a proactive queue.

Let us denote $\mathbb{E}[\nu_C]_{(\overline{p_0})}$ (resp. $\mathbb{E}[\nu_C]_{(\underline{p_0})}$) the expectation of the number of pending requests in the approximate queue with limited (resp. infinite) capacity. The lower bound of $\mathbb{E}[\nu_C]$ is obtained from the relation 14 of the page 15:

$$\mathbb{E}[\nu_C]_{(\overline{p_0})} = \sum_{k=C+r+1}^{K_\alpha} (k-C-r)\overline{p_k} = \overline{p_0} \sum_{k=C+r+1}^{K_\alpha} (k-C-r)H(k)\,, \tag{110}$$

The upper bound of $\mathbb{E}[\nu_C]$ is obtained from the same relation 14:

$$\begin{aligned}
\mathbb{E}[\nu_C]_{(\underline{p_0})} &= \sum_{k=C+r+1}^{\infty} (k-C-r)\,\underline{p_k}\,, \\
&= \sum_{k=C+r+1}^{K_\alpha} (k-C-r)\widetilde{H}(k)\,\underline{p_0} + \sum_{k=K_\alpha+1}^{\infty} (k-C-r)\widetilde{H}(k)\,\underline{p_0}\,, \\
&= \underline{p_0}\left[\sum_{k=C+r+1}^{K_\alpha} (k-C-r)H(k) + \sum_{k=K_\alpha+1}^{\infty} (k-C-r)H(K_\alpha)\,(f(K_\alpha))^{k-K_\alpha}\right], \\
&= \underline{p_0}\left[\frac{1}{\overline{p_0}}\mathbb{E}[\nu_C]_{(\overline{p_0})} + H(K_\alpha)\sum_{k=K_\alpha+1}^{\infty} (k-K_\alpha)\,(f(K_\alpha))^{k-K_\alpha}\right. \\
&\qquad \left. +(K_\alpha-C-r)H(K_\alpha)f(K_\alpha)\sum_{k=K_\alpha+1}^{\infty} (f(K_\alpha))^{k-K_\alpha-1}\right], \\
&= \underline{p_0}\left[\frac{1}{\overline{p_0}}\mathbb{E}[\nu_C]_{(\overline{p_0})} + \frac{H(K_\alpha)f(K_\alpha)}{(1-f(K_\alpha))^2} + (K_\alpha-C-r)\frac{H(K_\alpha)f(K_\alpha)}{(1-f(K_\alpha))}\right], \\
&= \frac{\underline{p_0}}{\overline{p_0}}\mathbb{E}[\nu_C]_{(\overline{p_0})} + \frac{\underline{p_0}H(K_\alpha)f(K_\alpha)}{(1-f(K_\alpha))}\left(\frac{1}{(1-f(K_\alpha))} + (K_\alpha-C-r)\right). \qquad (111)
\end{aligned}$$

Let us illustrate the method by considering the evaluation of the expectation of the number of pending requests in a proactive queue $M/M/10/ < 5 >$ with an arrival rate subject to a discouraging distribution. More precisely, the clients that show up at the entrance of the queue are accepted with a probability depending on the number of clients already present. If j denotes this number of customers, the acceptance probability is given

here by $1/(1+\beta(j-r)^+)$ where β is chosen less than 1. The initial arrival rate is poissonian (noted λ_b) and can be greater than $1/r\mu$. The sequence of arrival rates is thus written:

$$\lambda(k) = \frac{\lambda_b}{1+\beta(k-C-r)^+} \qquad k \geq 0. \tag{112}$$

With the data $r = 10$, $C = 5$, $\mu = 0.2$, $\lambda_b = 3$ and $\beta = 0.2$, we obtain:

$$\begin{aligned} &4,7686 < \mathbb{E}[\nu_C] < 4,7703 \text{ et } K_\alpha = 36 && \text{avec la consigne } \varepsilon \leq 10^{-4}, \\ &4,7701 < \mathbb{E}[\nu_C] < 4,7702 \text{ et } K_\alpha = 39 && \text{avec la consigne } \varepsilon \leq 10^{-5}. \end{aligned} \tag{113}$$

This simple example shows that this approach allows to estimate very precisely the different performance indicators with a low calculation cost.

5 Conclusion

In this paper, just as we know the formulas for determining the performance of non proactive queues, we sought to determine the set of results that would allow us to evaluate the performance of proactive queues. We have seen that in the end, some results are simply obtained while others are more difficult. The fundamental difference between the proactive and non proactive models is that the services of the proactive queues are potentially usable by any client. This window of anticipation of C services allows lucky customers to be served instantly and allows all customers to benefit from shorter response times. Nevertheless, we determined the performance in the more complex case where the number of servers is greater than $(C+1)$. We also proposed an efficient approach to evaluate a queue where the arrival and service rates depend on the number of clients in the queue. We conclude by reminding that proactive models can be used in different industries such as telecommunications, industrial production and integrated logistical support.

A Appendix : Computation of the expectation $\mathbb{E}[NS_C]$ (situation where $1 < p \leq r - C$)

$$
\begin{aligned}
&\mathbb{E}[NS_C] \\
&= \mathbb{E}[N_C] - \mathbb{E}[\nu_C]\,, \\
&= \sum_{j=C+1}^{\infty}(j-C)p_j - \sum_{j=C+r+1}^{\infty}(j-C-r)p_j\,, \\
&= \sum_{j=C+1}^{C+r}(j-C)p_j + \sum_{j=C+r+1}^{\infty}(j-C-(j-C-r))p_j\,, \\
&= \mathbf{1}_{\{r>C+1\}}\sum_{j=C+1}^{r-1}(j-C)p_j + \sum_{j=r}^{C+r}(j-C)p_j + r\sum_{j=C+r+1}^{\infty}p_j\,, \\
&= \mathbf{1}_{\{r>C+1\}}p_0\sum_{j=C+1}^{r-1}(j-C)\frac{(r\rho)^j}{j!} + p_0\frac{r^r}{r\,!}\sum_{j=r}^{C+r}(j-C)\rho^j + rp_0\frac{r^r}{r\,!}\sum_{j=C+r+1}^{\infty}\rho^j\,, \\
&= \mathbf{1}_{\{r>C+1\}}p_0\sum_{i=1}^{r-C-1}i\frac{(r\rho)^{i+C}}{(i+C)!} + p_0\frac{r^r}{r\,!}\left\{\sum_{j=r}^{\infty}(j-C)\rho^j - \sum_{j=C+r+1}^{\infty}(j-C)\rho^j\right\} + \\
&\qquad + rp_0\frac{r^r}{r\,!}\frac{\rho^{C+r+1}}{(1-\rho)}\,, \\
&= \mathbf{1}_{\{r>C+1\}}p_0\sum_{i=1}^{r-C-1}i\frac{(r\rho)^{i+C}}{(i+C)!} + p_0\frac{r^r}{r\,!}\left\{\sum_{i=0}^{\infty}(i+r-C)\rho^{i+r} - \sum_{i=1}^{\infty}(i+r)\rho^{i+C+r}\right\} + \\
&\qquad + rp_0\frac{r^r}{r\,!}\frac{\rho^{C+r+1}}{(1-\rho)}\,, \\
&= \mathbf{1}_{\{r>C+1\}}p_0\sum_{i=1}^{r-C-1}i\frac{(r\rho)^{i+C}}{(i+C)!} + p_0\frac{r^r}{r\,!}\rho^r\left\{\sum_{i=0}^{\infty}(i+r-C)\rho^{i} - \rho^C\sum_{i=1}^{\infty}(i+r)\rho^{i}\right\} + \\
&\qquad + rp_0\frac{r^r}{r\,!}\frac{\rho^{C+r+1}}{(1-\rho)}\,, \\
&= \mathbf{1}_{\{r>C+1\}}p_0\sum_{i=1}^{r-C-1}i\frac{(r\rho)^{i+C}}{(i+C)!} + p_0\frac{r^r}{r\,!}\rho^r\left\{(r-C) + (1-\rho^C)\sum_{i=1}^{\infty}i\rho^i + (r-C-r\rho^C)\sum_{i=1}^{\infty}\rho^i\right\} + \\
&\qquad + rp_0\frac{r^r}{r\,!}\frac{\rho^{C+r+1}}{(1-\rho)}\,, \\
&= \mathbf{1}_{\{r>C\}}p_0\sum_{i=1}^{r-C}i\frac{(r\rho)^{i+C}}{(i+C)!} + p_0\frac{r^r}{r\,!}\rho^{r+1}\left\{\frac{(1-\rho^C)}{(1-\rho)^2} + \frac{(r-C-r\rho^C)}{(1-\rho)}\right\} + rp_0\frac{r^r}{r\,!}\frac{\rho^{C+r+1}}{(1-\rho)}\,, \\
&= p_0\sum_{i=1}^{r-C}i\frac{(r\rho)^{i+C}}{(i+C)!} + p_0\frac{r^r}{r\,!}\rho^{r+1}\left\{\frac{(1-\rho^C)}{(1-\rho)^2} + \frac{(r-C)}{(1-\rho)}\right\}. \qquad (114)
\end{aligned}
$$

B Appendix : Determination of the formal expressions of the probabilities $P_i(t)$, $i = 3, 4, 5, 6$, related to the phase type distribution in the particular case where $p = 2$, $r = 6$ and $C = 3$ (situation where $1 < p \leq r - C$)

Recall the expression of the already calculated probabilities $P_1(t)$ and $P_2(t)$:

$$P_1(t) = e^{-\Lambda_5 t}\,, \tag{115}$$

$$P_2(t) \;=\; \frac{\lambda}{\mu}\left(e^{-\Lambda_5 t} - e^{-\Lambda_6 t}\right)\,.$$

By continuing to use the Kolmogorov equation (*cf.* 57) and the notation $\Lambda_i = \lambda + i\mu$, we obtain successively:

$$\begin{aligned}
P_3(t) &= \int_0^t P_2(x)\lambda e^{-6\mu(t-x)}dx\,,\\
&= \lambda e^{-6\mu t}\int_0^t \frac{\lambda}{\mu}\left(e^{-\Lambda_5 t} - e^{-\Lambda_6 t}\right)e^{+6\mu x}dx\,,\\
&= \frac{\lambda^2}{\mu}e^{-6\mu t}\left\{\int_0^t e^{-(\Lambda_5-6\mu)x}dx - \int_0^t e^{-(\Lambda_6-6\mu)x}dx\right\}\,,\\
&= \frac{\lambda^2}{\mu}e^{-6\mu t}\left\{\left[-\frac{1}{(\lambda-\mu)}e^{-(\Lambda_5-6\mu)x}\right]_0^t - \left[-\frac{1}{\lambda}e^{-(\Lambda_6-6\mu)x}\right]_0^t\right\}\,,\\
&= \frac{\lambda^2}{\mu}e^{-6\mu t}\left\{\frac{1}{(\lambda-\mu)}\left(1-e^{-(\Lambda_5-6\mu)t}\right) - \frac{1}{\lambda}\left(1-e^{-(\Lambda_6-6\mu)t}\right)\right\}\,,\\
&= \frac{\lambda}{(\lambda-\mu)}e^{-6\mu t} - \frac{\lambda^2}{\mu(\lambda-\mu)}e^{-\Lambda_5 t} + \frac{\lambda}{\mu}e^{-\Lambda_6 t}\,.
\end{aligned}$$

$$\begin{aligned}
P_4(t) &= \int_0^t P_1(x)5\mu e^{-\Lambda_4(t-x)}dx\,,\\
&= 5\mu e^{-\Lambda_4 t}\int_0^t e^{-\Lambda_5 x}e^{+\Lambda_4 x}dx = e^{-\Lambda_4 t}\frac{5\mu}{(\Lambda_5-\Lambda_4)}\left[-e^{-(\Lambda_5-\Lambda_4)x}\right]_0^t\,,\\
&= e^{-\Lambda_4 t}\frac{5\mu}{(\Lambda_5-\Lambda_4)}\left(1-e^{-(\Lambda_5-\Lambda_4)t}\right) = 5e^{-\Lambda_4 t}\left(1-e^{-\mu t}\right)\,,\\
&= 5\left(e^{-\Lambda_4 t} - e^{-\Lambda_5 t}\right)\,.
\end{aligned}$$

$$\begin{aligned}
P_5(t) &= \int_0^t P_2(x)6\mu e^{-\Lambda_5(t-x)}dx + \int_0^t P_4(x)\lambda e^{-\Lambda_5(t-x)}dx\,,\\
&= 6\lambda e^{-\Lambda_5 t}\int_0^t \left(e^{-\Lambda_5 x} - e^{-\Lambda_6 x}\right)e^{\Lambda_5 x}dx +\\
&\qquad +5\lambda e^{-\Lambda_5 t}\int_0^t \left(e^{-\Lambda_4 x} - e^{-\Lambda_5 x}\right)e^{\Lambda_5 x}dx\,,
\end{aligned}$$

$$
\begin{aligned}
&= 6\lambda e^{-\Lambda_5 t}\left\{\int_0^t dx - \int_0^t e^{-\mu x}dx\right\} + \\
&\qquad +5\lambda e^{-\Lambda_5 t}\left\{\int_0^t e^{\mu x}dx - \int_0^t dx\right\}, \\
&= 6\lambda t e^{-\Lambda_5 t} - 6\lambda e^{-\Lambda_5 t}\left[\frac{-e^{-\mu x}}{\mu}\right]_0^t + \\
&\qquad +5\lambda t e^{-\Lambda_5 t}\left[\frac{e^{\mu x}}{\mu}\right]_0^t - 5\lambda t e^{-\Lambda_5 t}, \\
&= 6\lambda t e^{-\Lambda_5 t} - 6\frac{\lambda}{\mu}e^{-\Lambda_5 t}\left(1-e^{-\mu t}\right) + \\
&\qquad +5\frac{\lambda}{\mu}e^{-\Lambda_5 t}\left(e^{\mu x}-1\right) - 5\lambda t e^{-\Lambda_5 t}, \\
&= 5\frac{\lambda}{\mu}e^{-\Lambda_4 t} - \left(11\frac{\lambda}{\mu} - \lambda t\right)e^{-\Lambda_5 t} + 6\frac{\lambda}{\mu}e^{-\Lambda_6 t}.
\end{aligned}
$$

$$
\begin{aligned}
P_6(t) &= \int_0^t P_3(x)6\mu e^{-6\mu(t-x)}dx + \int_0^t P_5(x)\lambda e^{-6\mu(t-x)}dx, \\
&= 6\mu e^{-6\mu t}\int_0^t \left(\frac{\lambda}{(\lambda-\mu)}e^{-6\mu x} - \frac{\lambda^2}{\mu(\lambda-\mu)}e^{-\Lambda_5 x} + \frac{\lambda}{\mu}e^{-\Lambda_6 x}\right)e^{+6\mu x}dx + \\
&\qquad +\lambda e^{-6\mu t}\int_0^t \left(5\frac{\lambda}{\mu}e^{-\Lambda_4 x} - \left(11\frac{\lambda}{\mu} - \lambda x\right)e^{-\Lambda_5 x} + 6\frac{\lambda}{\mu}e^{-\Lambda_6 x}\right)e^{+6\mu x}dx, \\
&= 6\mu e^{-6\mu t}\int_0^t \left(\frac{\lambda}{(\lambda-\mu)} - \frac{\lambda^2}{\mu(\lambda-\mu)}e^{-(\lambda-\mu)x} + \frac{\lambda}{\mu}e^{-\lambda x}\right)dx + \\
&\qquad +\lambda e^{-6\mu t}\int_0^t \left(5\frac{\lambda}{\mu}e^{-(\lambda-2\mu)x} - \left(11\frac{\lambda}{\mu} - \lambda x\right)e^{-(\lambda-\mu)x} + 6\frac{\lambda}{\mu}e^{-\lambda x}\right)dx, \\
&= e^{-6\mu t}\left\{\frac{6\mu\lambda}{(\lambda-\mu)}t - \frac{6\lambda^2}{(\lambda-\mu)}\left[\frac{-e^{-(\lambda-\mu)x}}{(\lambda-\mu)}\right]_0^t + 6\lambda\left[\frac{-e^{-\lambda x}}{\lambda}\right]_0^t + \right. \\
&\quad \left. +5\frac{\lambda^2}{\mu}\left[\frac{-e^{-(\lambda-2\mu)x}}{(\lambda-2\mu)}\right]_0^t - \frac{11\lambda^2}{\mu}\left[\frac{-e^{-(\lambda-\mu)x}}{(\lambda-\mu)}\right]_0^t + 6\frac{\lambda^2}{\mu}\left[\frac{-e^{-\lambda x}}{\lambda}\right]_0^t + \lambda^2\int_0^t xe^{-(\lambda-\mu)x}dx\right\}, \\
&= e^{-6\mu t}\left\{\frac{6\mu\lambda}{(\lambda-\mu)}t - \frac{6\lambda^2}{(\lambda-\mu)^2}\left(1-e^{-(\lambda-\mu)t}\right) + 6\left(1-e^{-\lambda t}\right) + \right. \\
&\quad +\frac{5\lambda^2}{\mu(\lambda-2\mu)}\left(1-e^{-(\lambda-2\mu)t}\right) - \frac{11\lambda^2}{\mu(\lambda-\mu)}\left(1-e^{-(\lambda-\mu)x}\right) + \\
&\quad \left. +\frac{6\lambda}{\mu}\left(1-e^{-\lambda t}\right) - \frac{\lambda^2}{(\lambda-\mu)^2}(1+(\lambda-\mu)t)e^{-(\lambda-\mu)t} + \frac{\lambda^2}{(\lambda-\mu)^2}\right\}, \\
&= \left(\frac{6\mu\lambda}{(\lambda-\mu)}t - \frac{5\lambda^2}{(\lambda-\mu)^2} + 6 + \frac{5\lambda^2}{\mu(\lambda-2\mu)} - \frac{11\lambda^2}{\mu(\lambda-\mu)} + \frac{6\lambda}{\mu}\right)e^{-6\mu t} + \\
&\quad -\frac{5\lambda^2}{\mu(\lambda-2\mu)}e^{-\Lambda_4 t} + \left(\frac{5\lambda^2}{(\lambda-\mu)^2} + \frac{11\lambda^2}{\mu(\lambda-\mu)} - \frac{\lambda^2 t}{(\lambda-\mu)}\right)e^{-\Lambda_5 t} - 6\left(1+\frac{\lambda}{\mu}\right)e^{-\Lambda_6 t}.
\end{aligned}
$$

By not further factoring the expression of the coefficients, it is quite easy to verify that the state probability $P_6(t)$ is zero at the origin.

C Appendix : A reminder of the ACE method

Let us denote $Q = \{q_{ij}\}$ the infinitesimal generator of the acyclic CMTC and q_i the global departure rate of a state i. Let us call these global departure rates poles because they correspond to the poles of the Laplace transforms of the expressions of the transient state probabilities of the acyclic CMTC states. If there are N different poles in this chain, the list $(\gamma_1, \gamma_2, ..., \gamma_n, ..., \gamma_N)$ designate these different poles and $\gamma(n)$ designate the value of the γ_n pole. The goal of the presented method is to determine the transient state probability of a state j in the following semi-formal form:

$$P_j(t) = \sum_n e^{\gamma(n)t} \sum_k a(j,n,k)t^k \,, \tag{116}$$

where the coefficients $a(j,n,k)$ are expressed in numerical form. Let us note $\Gamma(j)$ the set of poles present in the expression of the probability $P_j(t)$. We can write more precisely this probability:

$$P_j(t) = \sum_{n=1}^{N} \mathbf{1}_{\{\gamma_n \in \Gamma(j)\}} e^{\gamma(n)t} \sum_{k=0}^{K_j(n)} a(j,n,k)t^k \,, \tag{117}$$

where the integers $K_j(n)$ are assumed to be known.

We assume that the states are ordered so that the matrix Q is upper triangular. Let $J(i)$ be the set of "parent" states of state i. A state j is parent of a state i if $q_{ji} > 0$. Let us note $\Gamma(J(i))$ the set of poles that appear in at least one transient state probability of the "parent" states of state i.

Using the integral form of the Kolmogorov equation, the expression for the probability of state i is written:

$$P_i(t) = P_i(0)e^{\gamma^* t} + \sum_{j \in J(i)} \sum_{n=1}^{N} \mathbf{1}_{\{\gamma_n \in \Gamma(j)\}} \int_0^t e^{\gamma(n)x} \left[\sum_{k=0}^{K_j(n)} a(j,n,k)x^k \right] q_{ji} e^{\gamma^*(t-x)} dx \,, \tag{118}$$

where $\gamma^* = q_i$.

As the state probabilities will *a priori* share more and more common poles, we reverse the double sum:

$$P_i(t) = P_i(0)e^{\gamma^* t} + \sum_{n=1}^{N} \mathbf{1}_{\{\gamma_n \in \Gamma(J(i))\}} \int_0^t e^{\gamma(n)x} \left[\sum_{j \in J(i)} \mathbf{1}_{\{\gamma_n \in \Gamma(j)\}} \sum_{k=0}^{K_j(n)} a(j,n,k)x^k \right] q_{ji} e^{\gamma^*(t-x)} dx \,, \tag{119}$$

Let's define $\tilde{a}(i,n,k)$ as the sum

$$\tilde{a}(i,n,k) = \sum_{j \in J(i)} \mathbf{1}_{\{\gamma_n \in \Gamma(j)\}} \mathbf{1}_{\{k \leq K_j(n)\}} a(j,n,k) q_{ji} \,, \tag{120}$$

for $k = 0, 1, ..., \widetilde{K_i}(n)$, where

$$\widetilde{K_i}(n) = \max\{K_j(n) | j \in J(i)\} \,. \tag{121}$$

This allows us to rewrite the expression of $P_i(t)$ as

$$P_i(t) = P_i(0)e^{\gamma^* t} + \sum_{n=1}^{N} \mathbf{1}_{\{\gamma_n \in \Gamma(J(i))\}} \sum_{k=0}^{\widetilde{K_i}(n)} e^{\gamma^* t} \, \tilde{a}(i,n,k) \int_0^t x^k e^{(\gamma(n)-\gamma^*)x} dx \,, \tag{122}$$

For the calculation of the integral, we must distinguish two possible cases. If $\gamma(n) = \gamma^*$, we immediately obtain:

$$\int_0^t x^k e^{(\gamma(n)-\gamma^*)x} dx = \int_0^t x^k dx = \frac{t^{k+1}}{(k+1)}\,, \tag{123}$$

whereas if $\gamma(n) \neq \gamma^*$, we obtain after having put $\alpha = (\gamma(n) - \gamma^*)$:

$$\int_0^t x^k e^{\alpha x} dx = e^{\alpha t}\left[\sum_{m=0}^{k}(-1)^m \frac{k\,!\,t^{k-m}}{\alpha^{m+1}(k-m)\,!}\right] + (-1)^{k+1}\frac{k\,!}{\alpha^{k+1}}\,, \tag{124}$$

We obtain the new expression for $P_i(t)$:

$$P_i(t) = \sum_{n\neq n^*} \mathbf{1}_{\{\gamma_n\in\Gamma(J(i))\}} \sum_{k=0}^{\widetilde{K_i}(n)} \left[e^{\gamma(n)t}\,\tilde{a}(i,n,k)\left(\sum_{m=0}^{k}(-1)^m \frac{k\,!\,t^{k-m}}{\alpha^{m+1}(k-m)\,!}\right) + \right.$$
$$\left. +(-1)^{k+1} e^{\gamma^* t}\tilde{a}(i,n,k)\frac{k\,!}{\alpha^{k+1}}\right] + \mathbf{1}_{\{\gamma^*\in\Gamma(J(i))\}} \sum_{k=0}^{\tilde{K}_i(n^*)} e^{\gamma^* t}\tilde{a}(i,n^*,k)\frac{t^{k+1}}{(k+1)}\,. \tag{125}$$

To simplify the calculation algorithm, it is useful to note that

$$\begin{aligned}
\sum_{k=0}^{\widetilde{K_i}(n)}\left(\sum_{m=0}^{k}(-1)^m \frac{k\,!\,t^{k-m}}{\alpha^{m+1}(k-m)\,!}\right)\tilde{a}(i,n,k) &= \sum_{k=0}^{\widetilde{K_i}(n)}\left(\sum_{u=0}^{k}(-1)^{k-u}\frac{(k-u)\,!\,t^u}{\alpha^{k-u+1}u\,!}\right)\tilde{a}(i,n,k)\,,\\
&= \sum_{u=0}^{\widetilde{K_i}(n)} t^u \sum_{k=u}^{\widetilde{K_i}(n)}(-1)^{k-u}\frac{k\,!}{\alpha^{k-u+1}u\,!}\tilde{a}(i,n,k)\,,\\
&= \sum_{k=0}^{\tilde{K}_i(n)} t^k \sum_{r=k}^{\widetilde{K_i}(n)}(-1)^{r-k}\frac{r\,!}{\alpha^{r-k+1}k\,!}\tilde{a}(i,n,r)\,.
\end{aligned} \tag{126}$$

This allows us to calculate the new coefficients relative to a pole $\gamma(n) \neq \gamma^*$ in the following way:

$$a(i,n,\widetilde{K_i}(n)) = \frac{1}{\alpha}\tilde{a}(i,n,\widetilde{K_i}(n)) \tag{127}$$

$$a(i,n,k) = \frac{1}{\alpha}\left[\tilde{a}(i,n,k) - (k+1)a(i,n,k+1)\right]\,, \qquad k = (\widetilde{K_i}(n)-1),(\widetilde{K_i}(n)-2),...,0. \tag{128}$$

In order to determine the coefficient $a(i,n^*,0)$ relative to the γ^* pole, note first that, according to the relation 126,

$$a(i,n,0) = \sum_{r=0}^{\widetilde{K_i}(n)}(-1)^r\frac{r\,!}{\alpha^{r+1}}\tilde{a}(i,n,r)\,, \tag{129}$$

and so, that according to the relation 125,

$$a(i,n^*,0) = \sum_{n\neq n^*}\mathbf{1}_{\{\gamma_n\in\Gamma(J(i))\}}\sum_{k=0}^{\widetilde{K_i}(n)}(-1)^{k+1}\frac{k\,!}{\alpha^{k+1}}\tilde{a}(i,n,k)\,. \tag{130}$$

It immediately follows that:

$$a(i,n^*,0) = -\sum_{n\neq n^*}\mathbf{1}_{\{\gamma_n\in\Gamma(J(i))\}}a(i,n,0) + P_i(0)\,. \tag{131}$$

If γ^* is already in the set of poles of the parents of the state i, the order of the polynomial expression associated to the pole increases by one unit (according to the relation 125) and, according to the relation 123, its new coefficients can be obtained by the recurrence:

$$a(i, n^*, k) = \tilde{a}(i, n^*, k-1)/k\,, \qquad k = 1, 2, ..., (\widetilde{K_i}(n) + 1)\,. \tag{132}$$

Note that for poles $\gamma(n)$ different from γ^*, $K_i(n) = \widetilde{K_i}(n)$, while $K_i(n^*) = 0$ si $\gamma^* \notin \Gamma(J(i))$ et $K_i(n^*) = \widetilde{K_i}(n) + 1$ else.

A schema of the algorithm is proposed table 1 and table 2. The presented algorithm uses many input data. First of all, the infinitesimal generator of the acyclic CMTC and the vector of initial state probabilities. The data NS and NP give the number of states and the number of poles respectively. Note that the states are ordered so that the matrix $TQ = \{q_{ij}\}$ is upper triangular. The vector of initial state probabilities is noted $TProbEnt$. The algorithm also uses the vectors $TGamStar$ and $TVPole$. The component $TGamStar(i)$ provides the pole number of the state i while the value of this pole is provided by the component $TVPole(i)$. The algorithm also uses the Boolean matrices $BPole(NS, NP)$, $BPar(NS, NS)$ and $BPolPar(NS, NP)$, matrices defined as follows:

$$BPole(i, n) = \begin{cases} \text{True} & \text{if } n \text{ is a pole of } P_i(t)\ , \\ \text{False} & \text{else}\ . \end{cases}$$

$$BPar(j, i) = \begin{cases} \text{True} & \text{if state } j \text{ is a parent of state } i\ , \\ \text{False} & \text{else}\ . \end{cases}$$

$$BPolPar(i, n) = \begin{cases} \text{True} & \text{if pole } n \text{ figures in the probability of at least one parent of state } i\ , \\ \text{False} & \text{else}\ . \end{cases}$$

A second algorithm allowing to fill the matrices $TGamStar, TVPole, BPole, BPar$ and $BPolPar$ using only the knowledge of the matrix TQ is (fortunately) proposed in table 3.

The matrices $BPole$ (provided by the algorithm of the table 3), $a(i, n, k)$, and $Kmax(i, n)$ (provided by the algorithm of the tables 1 and 2) are used to compute the values of the various state probabilities $P_i(t)$.

Table 1: ACE Algorithm (Beginning)

```
BEGIN
INPUT DATA : NE ; NP; TQ(i,j) ; TProbEnt(i) ; BPoles(i,n) ;BPar(j,i) ; TVPole(n) ;
        BPolPar(i,n) ; TGamStar(i) ; i,j = 1,2,3,...,NE; n = 1,2,...,NP .
OUTPUT DATA : a(i,n,k), Kmax(i,n) .
% Initialization state 1 :
a(1,1,0) := TProbEnt(1)
Kmax(1,1) := 0
% Loop over the states :
For i = 2 : NE  Do
    For n = 1 : NP  Do
        % Search for K~_i(n)
        ValKmax := -1
        For j = 1 : (i - 1)  Do
            If ((BPar(j,i)) and (Kmax(j,n) > ValKmax))
                ValKmax := Kmax(j,n)
            End If
        Done
        If (ValKmax ≥ 0)
            K~_i(n) := ValKmax
        End If
        % End of search for K~_i(n)
        For k = 0 : K~_i(n) Do
            Sum := 0.0
            For j = 1 : (i - 1) Do
                ValKmj := Kmax(j,n)
                If ((BPar(j,i)) and (BPolPar(i,n)) and (k ≤ ValKmj))
                    %   j ∈ J(i), n ∈ Γ(J(i)) and k ≤ Kmax(j,n)
                    Sum := Sum + a(j,n,k) * TQ(j,i)
                End If
            Done
            atil(i,n,k) := Sum
        Done
    Done
    % Ending of the computation of coefficients ã(i,n,k) relative to the node i.
```

Table 2: ACE Algorithm (Continuation and End)

```
    % Start computation of a(i,n,k) relative to the poles of the node i exept pole γ*
    GS := TGamStar(i)
    For n = 1 : NP Do
        If ( (BPolPar(i,n)) and (n ≠ GS) )
            ValKtil := K̃_i(n)
            AlphaP := TVPole(n) − TVPole(GS)
            a(i,n,ValKtil) := atil(i,n,ValKtil)/AlphaP
            If (ValKtil > 0)
                k := ValKtil
                For m = 1 : ValKtil Do
                    k := k − 1
                    a(i,n,k) := (atil(i,n,k) − (k+1) * a(i,n,(k+1)))/AlphaP
                Done
            End If
        End If
    Done
    % Start computation of coefficients relative to the pole γ*
    % Computation of a(i,GS,0)
    Sum := 0.0
    For n = 1 : NP Do
        If ((BPolPar(i,n)) and (n ≠ GS) )
            Sum := Sum + a(i,n,0)
        End If
    Done
    a(i,GS,0) := −Sum + TProbEnt(i)
    If (BPolPar(i,GS) = False)
        Kmax(i,GS) := 0
    Else
        % γ* is also part of Γ(J(i)) (because BPolPar(i,GS) = True)
        For k = 0 : K̃_i(GS) Do
            kp1 := k + 1
            a(i,GS,kp1) := atil(i,GS,k)/kp1
        Done
        Kmax(i,GS) := K̃_i(GS) + 1
    End If
Done
END
```

Table 3: Pre-ACE Algorithm

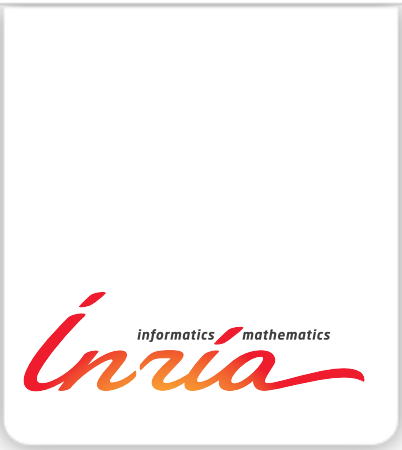
```
BEGIN
INPUT DATA : NE ; TQ(i,j), i,j = 1,2,3,...,NE .
OUTPUT DATA : NP ; TGamStar(i) ;TVPole(n) ;BPoles(i,n) ;BPar(i,j) ;
              BPolPar(i,n) ; i,j = 1,2,3,...,NE; n = 1,2,...,NP .
% Initialization state 1 :
TGamStar(1) := 1
TVPole(1) := TQ(1,1)
BPoles(1,1) := True
Ncourant := 1
% boucle sur les états :
For i = 2 : NE Do
    EssaiPole := TQ(i,i)
    BTrouv := False
    j := 1
    Tantque ( (j < i) and (BTrouv = False) ) Do
        If (TVPole(TGamStar(j)) == EssaiPole)
            BTrouv := True
            TGamStar(i) := TGamStar(j)
        End If
        j:= j +1
    Done
    If (BTrouv = False)
        Ncourant := Ncourant + 1
        TGamStar(i) := Ncourant
        TVPole(Ncourant) := TQ(i,i)
    End If
    For j = 1 : (i - 1) Do
        If (TQ(j,i) > 0.0)
            BPar(j,i) := True
            BPolPar(i,:) := (BPolPar(i,:)) or (BPoles(j,:))
        End If
        BPoles(i,:) := BPolPar(i,:)
        BPoles(i,TGamStar(i)) := True
    Done
Done
NP := Ncourant
END
```

References

[1] J. Doncel and J.-M. Fourneau. Balancing energy consumption and losses with energy packet network models. In *Proceedings of the 2019 IEEE International Conference on Fog Computing (ICFC)*, pages 59–68. IEEE Computer Society, 2019.

[2] R. A. Marie, A. L. Reibman, and K. S. Trivedi. Transient analysis of acyclic Markov chains. *Performance Evaluation*, 7(3):175–194, August 1987.

[3] Raymond A. Marie. *Introduction aux probabilités - Applications en informatique et télécommunications, fiabilité et gestion d'incertitudes*. Ellipses, June 2016.



RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399