International Journal of Operational Research

ISSN online: 1745-7653 - ISSN print: 1745-7645
https://www.inderscience.com/ijor

Distribution of occupied resources on a fractional resource sharing in a queueing system

T.B. Ravaliminoarimalalason, M. Rakotomalala, F. Randimbindrainibe

DOI: 10.1504/IJOR.2021.10039107

Article History:

Received:	04 March 2021
Accepted:	11 May 2021
Published online:	12 January 2024

Distribution of occupied resources on a fractional resource sharing in a queueing system

T.B. Ravaliminoarimalalason*,
M. Rakotomalala and F. Randimbindrainibe

Ecole Doctorale en Sciences et Techniques de
l'Ingénierie et de l'Innovation,
Ecole Supérieure Polytechnique d'Antananarivo,
University of Antananarivo,
Madagascar, East Africa
Email: tokybaz@gmail.com
Email: mrakotom_diego@yahoo.fr
Email: falimanana@mail.ru
*Corresponding author

Abstract: Many server systems can share their resources by fractional way not discrete. It can be found in major cases of communication systems sharing power, spectrum or bandwidth resources for example. The objective of this work is to build analytical expressions of the amount of occupied resources in a structure modelled as queueing system. The queue server shares its resources to customers that request services to him. Both infinite and finite capacity are highlighted and the requested resources can be fractional. The amount of occupied resources as real-valued random variable is characterised by its distribution functions that we proposed in this paper. They are validated by simulations, and then can be used to predict the performances of such system or to dimension the appropriate needed capacity. Impacts of system load factor and system capacity has been also analysed.

Keywords: queueing; dimensioning; load factor; resource occupation; resource sharing.

Reference to this paper should be made as follows: Ravaliminoarimalalason, T.B., Rakotomalala, M. and Randimbindrainibe, F. (2024) 'Distribution of occupied resources on a fractional resource sharing in a queueing system', *Int. J. Operational Research*, Vol. 49, No. 1, pp.131–145.

Biographical notes: T.B. Ravaliminoarimalalason is a PhD, telecommunication engineer, and Lecturer-Researcher at the Department of Telecommunication and at the Ecole Doctorale en Sciences et Techniques de l'Ingénierie et de l'Innovation, Ecole Supérieure Polytechnique d'Antananarivo, University of Antananarivo, Madagascar.

M. Rakotomalala is a PhD, telecommunication engineer, and Lecturer-Researcher at the Institut Supérieur de Technologie Antananarivo and at the Ecole Doctorale en Sciences et Techniques de l'Ingénierie et de l'Innovation, Ecole Supérieure Polytechnique d'Antananarivo, University of Antananarivo, Madagascar.

F. Randimbindrainibe is a Professor, and Lecturer-Researcher at the Ecole Doctorale en Sciences et Techniques de l'Ingénierie et de l'Innovation, Ecole Supérieure Polytechnique d'Antananarivo, University of Antananarivo, Madagascar.

1 Introduction

Many systems share their resources to their users. They can be modelled as queueing systems having a single server and multiple resources, or multiple servers each having their own resources. Romm and Skitovich (1971) started from a generalisation of Erlang theorem by studying the allocation of a random resource to each arrival to a queue. He is focused on the stationary distribution of the queueing system operating in a discrete time. Green (1980) studied a queueing system in which the customers can ask to be served by random number of servers. If the queue has more than one server, or the server has more one resource, they can be shared to the customers. It is called a processor-sharing queue that Yashkov (1987) reviewed and proposed techniques to analyse it. Tikhonenko (2010) studied the processor-sharing queue with limited resources, exponential time distribution. Sopin and Vikhrova (1995) evaluated the probability characteristics in a queueing system with random requirements, and Lisovskaya et al. (2017) analysed the customer capacity in the case of multi-server queue. Pagano (2020) extended the work on non-Poisson arrivals and non-exponential service time. He considered a parallel service with double customer and a renewal arrival process. Vikhrova (2017) combined a study of queueing system with random requirements and limited resources. Analytical model of the distribution of the amount of occupied resources in a discrete resource sharing is already built (Ravaliminoarimalalason and Randimbindrainibe, 2021). It corresponds to a $M^x/M/m$. The current work is an extension on a continuous domain: We consider a fractional resource sharing where the requested quantities from customers are not discrete but continuous, distributed as real positive random variables. The main objective is to find analytical models that can be used to predict the behaviour of the system and also to dimension it.

We called by fractional resources the resources in continuous domain, in opposite of discrete one. This case can be found in many communication systems (Vishnevski and Dudin, 2017) such as bandwidths resources (Massoulie and Roberts, 2010; Krishnan et al., 2011), spectrums (Ye et al., 2016), powers (Zhang et al., 2019) for example, even in the latest standard of mobile communication (Yang et al., 2016).

2 Analytical expressions

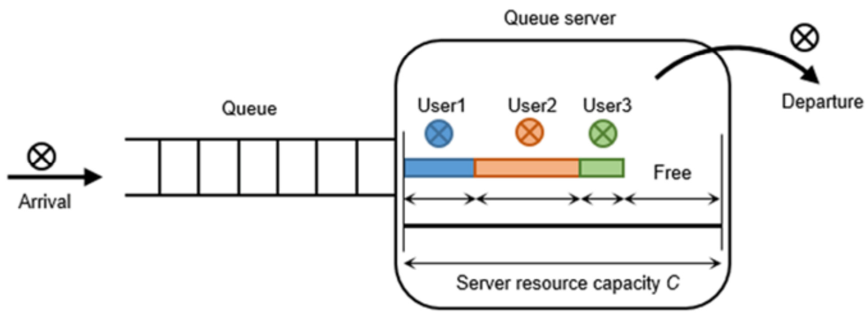
2.1 Description

We recall the results we obtained in the case of discrete resource sharing. Given an $M/M/1$ queue with discrete resources C, $C \in \mathbb{N} \cup \{+\infty\}$ to allocate to its customers. Each customer asks to be served by a random amount r of resources. The queue server can allocate this amount of resources to this customer if they are available, otherwise the customer must wait until they become available. Thus, as described in Figure 1, the server can serve simultaneously several customers depending on the availability of resources it has. The classic notion of an $M/M/1$ queue that can only receive one customer at a time is no longer relevant. It is a sharing of the resources available at the queue server level through the customers that come there according to their needs.

The following theorems can be cited. We denote by:

- $P(R = r)$ the probability distribution of random variable R that indicates the quantity of occupied resources in the server. It can have a general distribution.
- $P(R_k = r)$ the probability that k customers are using r resources.

Figure 1 System description (see online version for colours)



Theorem 1: For the case of infinite resources in the queue server, the probability distribution $P(R = r)$ is:

$$P(R=r)=\exp(-\rho)\cdot\sum_{k=1}^{+\infty}\frac{\rho^k}{k!}P(R_k=r),\ r\in\mathbb{N} \tag{1}$$

where $\rho = \lambda/\mu$ is the queue load factor, λ the customer arrival process intensity, and $1/\mu$ the average service time requested by the customers.

Theorem 2: For the case of finite resources C, $C\in\mathbb{N}^*$, in the queue server, the probability distribution $P(R = r)$ is:

$$P(R=r)=\sum_{s=1}^{C}\left(\frac{\left(\sum_{k=1}^{s}\frac{\rho^k}{k!}P(R_k=r)+\sum_{k=s+1}^{+\infty}\frac{\rho^k}{s!s^{k-s}}P(R_s=r)\right)\cdot P_s}{\sum_{k=1}^{s}\frac{\rho^k}{k!}+\sum_{k=s+1}^{+\infty}\frac{\rho^k}{s!s^{k-s}}}\right)\text{ for } r=0,\ldots,C \tag{2}$$

where $P_s=\sum_{r=1}^{C}P(R_s=r)\cdot P(R_1>C-r)$ is the probability that the queue cannot receive more than s customers at the same time.

Proofs of these Theorems 1 and 2 are explicitly described in our paper (Ravaliminoarimalalason and Randimbindrainibe, 2021).

2.2 *Continuous infinite resource sharing*

This time, we are going to look at the case of resources whose customer needs are positive real numbers, with values in $\mathbb{R}$ or part of $\mathbb{R}$. The amount of available resources in the queue server is no longer a natural number, but can be also any positive real number. We therefore use a real-valued random variable for the amount of occupied resources.

2.2.1 General case

Theorem 3: We can extend Theorem 1 in the case of a real-valued random variable. The probability $P(R \le r)$ that the quantity R of occupied resources in the server is less or equal than r is then:

$$P(R \le r) = \exp(-\rho) \cdot \sum_{k=1}^{+\infty} \frac{\rho^k}{k!} P\left(R_k \le r\right) \tag{3}$$

In other words, the cumulative distribution function $F(x)$ of the quantity of occupied resources is:

$$F(x) = \exp(-\rho) \cdot \sum_{k=1}^{+\infty} \frac{\rho^k}{k!} F_k(x) \tag{4}$$

where $F_k(x)$ is the cumulative distribution function of the quantity of resources occupied by k customers. The quantity of resources occupied by k customers is obtained by the sum of k independent and identically distributed random variables of the occupation by one customer.

Proof: The demonstration is done by analogy to that of discrete resources. If the queue has $N = k$ customers, they are all served simultaneously. By denoting r_i the quantity of resources used by the i^{th} customer, we can write:

$$P(R \le r) = \sum_{k=1}^{+\infty} P(N = k) \cdot P\left(r_1 + \ldots + r_k \le r\right)$$

$P(N = k)$ is the k^{th} element π_k of the stationary distribution π of the queueing system: $P(N = k) = \frac{\rho^k}{k!} \exp(-\rho)$. And $P(r_1 + \ldots + r_k \le r)$ is the probability that the sum of used resources by k customers is less or equal to r, it can be denoted $P(R_k \le r)$.

This probability $P(R \le r)$ is the cumulative distribution function of the real-valued random variable R, which will be noted $F(r)$.

□

Theorem 4: The probability density function of the quantity of occupied resources is:

$$f(x) = \exp(-\rho) \cdot \sum_{k=1}^{+\infty} \frac{\rho^k}{k!} f_k(x) \tag{5}$$

where $f_k(x)$ denotes the probability density function of the quantity of resources occupied by k customers. It should be noted that the density function $f_k(x)$ is obtained by the convolution product of s probability density functions $f_1(x)$ of the quantity of resources occupied by a single customer: $f_k(x) = f_1 * \ldots * f_1(x)$. Indeed, it is due to the sum of real-valued random variable of the amount of resources occupied by each customer.

Proof: Theorem 4 is deduced by derivation of the distribution function in equation (4).

□

2.2.2 Exponential distribution probability of single usage

We will take the particular case of exponential distribution probability of single usage with parameter μ_R, i.e., a customer uses an amount $R_1 = r$ of resources following the exponential distribution $\mathcal{E}(\mu_R)$. An elementary property of exponential distribution (Asmussen, 2003; Kleinrock, 1975) is given in the following Lemma 1 that we will be used in below theorem.

Lemma 1: The sum of k independent exponential random variables with the same rate parameters μ_R is following the gamma distribution probability with parameters k and μ_R which we denote $\Gamma(k, \mu_R)$.

Theorem 5: The density probability function of the amount of occupied resources is:

$$f(x) = \exp\left(-\rho - \mu_R x\right) \cdot \sum_{k=1}^{+\infty} \frac{k\left(\rho\mu_R\right)^k}{k!^2} x^{k-1} \tag{6}$$

Its cumulative distribution function is:

$$F(x) = P(R \leq x) = \exp(-\rho) \cdot \sum_{k=1}^{+\infty} \frac{k\rho^k}{k!^2} \gamma\left(k, \mu_R x\right) \tag{7}$$

where $\gamma(k, x) = \int_0^x t^{k-1} \exp(-t)dt$ is the incomplete gamma function.

Proof: By virtue of Lemma 1 and that the gamma $\Gamma(k, \mu_R)$ probability density function is $f_k(x) = \frac{x^{k-1}}{(k-1)!} \mu_R^k . \exp(-\mu_R x),\ x \geq 0,$ so:

$$\begin{aligned} f(x) &= \exp(-\rho) \cdot \sum_{k=1}^{+\infty} \frac{\rho^k}{k!} \frac{x^{k-1}}{(k-1)!} \mu_R^k . \exp\left(-\mu_R x\right) \\ &= \exp(-\rho) . \exp\left(-\mu_R x\right) \cdot \sum_{k=1}^{+\infty} \frac{\rho^k \mu_R^k}{k!} \frac{x^{k-1}}{(k-1)!} \\ &= \exp\left(-\rho - \mu_R x\right) \cdot \sum_{k=1}^{+\infty} \frac{\left(\rho\mu_R\right)^k}{k!} \frac{x^{k-1}}{(k-1)!} \\ &= \exp\left(-\rho - \mu_R x\right) \cdot \sum_{k=1}^{+\infty} \frac{k\left(\rho\mu_R\right)^k}{k!^2} x^{k-1},\ x \geq 0 \end{aligned}$$

And the cumulative distribution function of $\Gamma(k, \mu_R)$ is $F_k(x) = \frac{\gamma(k, \mu_R x)}{(k-1)!},\ x \geq 0$ where $\gamma(k, x)$ is the incomplete gamma function $\gamma(k, x) = \int_0^x t^{k-1} \exp(-t)dt,$ then:

$$F(x) = \exp(-\rho).\sum_{k=1}^{+\infty} \frac{\rho^k}{k!} \frac{\gamma(k, \mu_R x)}{(k-1)!} = \exp(-\rho) \cdot \sum_{k=1}^{+\infty} \frac{k\rho^k}{k!^2} \gamma(k, \mu_R x), \; x \ge 0$$

□

2.3 Continuous finite resource sharing

Now, let's put a finite amount of available resources in the queue server.

2.3.1 General case

Theorem 6: The cumulative distribution function $F(x)$ of occupied resources is:

$$F(x) = P(R \le x) = \sum_{s=1}^{+\infty} \left(\frac{\left(\sum_{k=1}^{s} \frac{\rho^k}{k!} F_k(x) + \sum_{k=s+1}^{+\infty} \frac{\rho^k}{s! s^{k-s}} F_s(x) \right) \cdot P_s}{\sum_{k=1}^{s} \frac{\rho^k}{k!} + \sum_{k=s+1}^{+\infty} \frac{\rho^k}{s! s^{k-s}}} \right) \text{ for } 0 \le x \tag{8}$$

where P_s is the probability that the capacity of the server in terms of number of customers is equal to s. The load factor ρ is $\rho = \lambda/\mu$ in this expression.

Proof: We use the nomenclature $\pi_{h,s}$ used to denote the element of the stationary distribution of the queueing system (Gaver et al., 1984) indicating the probability that the system with capacity s in terms of customers is hosting h customers (whether served or awaiting). We denote by $h = k + w$ the hosted customers where k is the number of served customers and w the number of customers awaiting service.

$$\pi_{h,s} = \begin{cases} \dfrac{\frac{\rho^h}{h!}}{\sum_{i=0}^{s} \frac{\rho^i}{i!} + \sum_{i=1}^{+\infty} \frac{\rho^{s+i}}{s! s^i}}, & \text{for } 0 \le h \le s \\[2ex] \dfrac{\frac{\rho^h}{s! s^{h-s}}}{\sum_{i=0}^{s} \frac{\rho^i}{i!} + \sum_{i=1}^{+\infty} \frac{\rho^{s+i}}{s! s^i}}, & \text{for } h > s \end{cases}$$

If the server's customer capacity is $s = 1$, the use of resources less than r at this server is limited to that single customer. Then we have the probability:

$$\begin{aligned} &\pi_{1,s} \cdot P(R_1 \le r) + \pi_{2,s} \cdot P(R_1 \le r) + \pi_{3,s} \cdot P(R_1 \le r) + \ldots \\ &= \pi_{1,s} \cdot P(R_1 \le r) + \sum_{k=2}^{+\infty} \pi_{k,s} \cdot P(R_1 \le r) \end{aligned}$$

If the server's customer capacity is $s = 2$, the use of resources less than r at this server is limited to those two customers at most. Then we have also the probability:

$$\pi_{1,s} \cdot P(R_1 \le r) + \pi_{2,s} \cdot P(R_2 \le r) + \pi_{3,s} \cdot P(R_2 \le r) + \ldots$$
$$= \sum_{k=1}^{2} \pi_{k,s} \cdot P(R_k \le r) + \sum_{k=3}^{+\infty} \pi_{k,s} \cdot P(R_2 \le r)$$

And so on… The probability of use of less than r resources is obtained for all possible values of s from 1 to $+\infty$ $(s \in \mathbb{N}^*)$. So:

$$P(R \le r) = \sum_{s=1}^{+\infty} \left(\left(\sum_{h=1}^{s} \pi_{h,s} \cdot P(R_h \le r) + \sum_{h=s+1}^{+\infty} \pi_{h,s} \cdot P(R_s \le r) \right) \cdot P_s \right)$$

We deduce the cumulative distribution function:

$$F(x) = P(R \le x) = \sum_{s=1}^{+\infty} \left(\left(\sum_{h=1}^{s} \pi_{h,s} \cdot F_h(x) + \sum_{h=s+1}^{+\infty} \pi_{h,s} \cdot F_s(x) \right) \cdot P_s \right)$$

□

Remark: Till now, we based the calculation on the resources R_k requested by k customers. It is possible that $R_k > C$, and in this case, all those k customers cannot be served simultaneously, then we are talking about a blocked system. Anyway, we can consider fictive occupation: an occupation of resources if the k customers' requests were accepted by the server. $F(x)$ mentions this fictive occupation, and the probability of blocking is $F^c(C) = 1 - P(R \le C) = P(R > C)$.

Real occupation is related to the real amount of resources occupied by the customers. It can be obtained by considering $F(x)$ between $x = 0$ and $x = C$. In this case, a normalisation factor should be used to have $F(C) = 1$.

It remains to determine the expression of P_s, the probability that the system has a capacity s in terms of number of customers.

Theorem 7: The probability distribution P_s of the server capacity in terms of number of customers is:

$$P_s = \int_0^C \int_{C-x}^{+\infty} f_s(x) f_1(t)\,dt\,dx \tag{9}$$

where $f_s(x)$ is the density probability function of the amount of resources used by s customers: $f_s(x) = f_1 * \ldots * f_1(x)$, s consecutive convolution product of $f_1(x)$.

Proof: For the case of discrete resources of step 1, $P_s = \sum_{r=1}^{C} P(R_s = r) \cdot P(R_1 > C - r)$. In fact, the current use by s customers is r and if another one wants to use more than $C - r$ resources, the used resources would exceed the finite capacity C. The queueing system cannot therefore serve more than s customers in this case, and this, for all the possible values of r from 1 to C and in steps of 1.

Now, let's consider a step of $\varepsilon > 0$. $P_s = \sum_{r=\varepsilon}^{C} P(R_s = r) \cdot P(R_1 > C - r)$ with $r = \varepsilon$, 2ε, 3ε, …, $n\varepsilon = C$.

$$P_s = \sum_{\varepsilon=1}^{C/\varepsilon} P(R_s = n\varepsilon) \cdot P(R_1 > C - n\varepsilon)$$

However, for any continuous random variable X, the density function is the derivate of the cumulative distribution function:

$$\lim_{\varepsilon \to 0^+} \frac{F_X(t+\varepsilon) - F_X(t)}{\varepsilon} = f_X(t)$$

So, for ε infinitely small, we can write that $F_X(t + \varepsilon) - F_X(t) = f_X(t)\varepsilon + o(\varepsilon)$ where $o(\varepsilon)$ is the little-o function of ε.

Then we have: $F_X(t + \varepsilon) - F_X(t) = f_X(t)\varepsilon + o(\varepsilon) = P(t \le X \le t + \varepsilon)$ for ε infinitely small.

The continuous case is obtained from the discrete one when the step ε of used resources tends to 0^+, and $\lim_{\varepsilon \to 0^+} P(t \le X \le t + \varepsilon) = P(X = t)$, so for $\varepsilon \to 0^+$,

$$P(X = t) = f_X(t)\varepsilon.$$

For the random variable R_s, $P(R_s = t) = f_s(t)\varepsilon$.

So, we can write P_s as:

$$P_s = \sum_{n\varepsilon=\varepsilon}^{C} \varepsilon \cdot f_s(n\varepsilon) \cdot P(R_1 > C - n\varepsilon)$$

And by applying the definition of Riemann integral in this expression of P_s, we have:

$$P_s = \sum_{n\varepsilon=\varepsilon}^{C} \varepsilon \cdot f_s(n\varepsilon) \cdot P(R_1 > C - n\varepsilon) = \int_0^C f_S(x) \cdot P(R_1 > C - x)\, dx$$

where $P(R_1 > C - x) = \int_{C-x}^{+\infty} f_1(t)dt$ is the complementary cumulative distribution function of R_1 at point $C - x$.

Then:

$$\begin{aligned} P_s &= \int_0^C f_S(x) \cdot P(R_1 > C - x)\, dx = \int_0^C f_S(x) \cdot \left(\int_{C-x}^{+\infty} f_1(t)dt \right) dx \\ &= \int_0^C \int_{C-x}^{+\infty} f_S(x) f_1(t) dt dx \end{aligned}$$

□

2.3.2 Exponential distribution probability of single usage

Let's take again the particular case of exponential distribution probability of single usage with parameter μ_R. As the amount of available resources in the queue server is finite C, to avoid that the system is blocked eternally, we should have $P(R_1 > C)$ very small.

Theorem 8: The server's capacity in terms of number of customers follows the Poisson distribution with parameter $\mu_R C$.

Proof: Using the expression of P_s in equation (9), we have:

$$
\begin{aligned}
P_s &= \int_0^C f_S(x)\cdot P\left(R_1 > C-x\right)dx \\
&= \int_0^C \frac{x^{s-1}}{(s-1)!}\mu_R^s \cdot \exp\left(-\mu_R x\right)\cdot \exp\left(-\mu_R (C-x)\right)\cdot dx \\
&= \int_0^C \frac{x^{s-1}}{(s-1)!}\mu_R^s \cdot \exp\left(-\mu_R C\right)\cdot dx = \mu_R^s \exp\left(-\mu_R C\right)\cdot \int_0^C \frac{x^{s-1}}{(s-1)!}dx \\
&= \mu_R^s \exp\left(-\mu_R C\right)\left[\frac{x^s}{s!}\right]_0^C = \frac{\mu_R^s C^s}{s!}\exp\left(-\mu_R C\right) = \frac{\left(\mu_R C\right)^s}{s!}\exp\left(-\mu_R C\right)
\end{aligned}
$$

We obtained the probability distribution of a Poisson random variable with parameter $\mu_R C$.

□

Theorem 9: The cumulative distribution function $F(x)$ of the amount of occupied resources is:

$$
F(x) = \\
\sum_{s=1}^{+\infty}\left(\frac{\left(\sum_{k=1}^{s}\frac{k\rho^k}{k!^2}\gamma\left(k,\mu_R x\right)+\sum_{k=s+1}^{+\infty}\frac{s\rho^k}{s!^2\, s^{k-s}}\gamma\left(s,\mu_R x\right)\right)\cdot\frac{\left(\mu_R C\right)^s}{s!}e^{-\mu_R C}}{\sum_{k=1}^{s}\frac{\rho^k}{k!}+\sum_{k=s+1}^{+\infty}\frac{\rho^k}{s!s^{k-s}}}\right) \text{ for } 0 \le x \le C
$$

Proof: Just to replace $F_k(x)$ by its expression $F_k(x) = \dfrac{\gamma(k,\mu_R x)}{(k-1)!}$ for gamma distribution, and P_s by its expression from Theorem 8.

$$
F(x) = \sum_{s=1}^{+\infty}\left(\frac{\left(\sum_{k=1}^{s}\frac{\rho^k}{k!}\frac{\gamma\left(k,\mu_R x\right)}{(k-1)!}+\sum_{k=s+1}^{+\infty}\frac{\rho^k}{s!s^{k-s}}\frac{\gamma\left(s,\mu_R x\right)}{(s-1)!}\right).\frac{\left(\mu_R C\right)^s}{s!}e^{-\mu_R C}}{\sum_{k=1}^{s}\frac{\rho^k}{k!}+\sum_{k=s+1}^{+\infty}\frac{\rho^k}{s!s^{k-s}}}\right)
$$

$$
\sum_{s=1}^{+\infty}\left(\frac{\left(\sum_{k=1}^{s}\frac{k\rho^k}{k!^2}\gamma\left(k,\mu_R x\right)+\sum_{k=s+1}^{+\infty}\frac{s\rho^k}{s!^2\, s^{k-s}}\gamma\left(s,\mu_R x\right)\right)\cdot\frac{\left(\mu_R C\right)^s}{s!}e^{-\mu_R C}}{\sum_{k=1}^{s}\frac{\rho^k}{k!}+\sum_{k=s+1}^{+\infty}\frac{\rho^k}{s!s^{k-s}}}\right)
$$

□

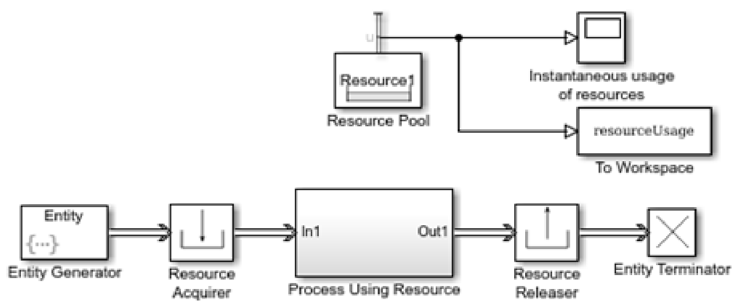
3 Results and discussions

3.1 Validation of the analytical expressions

3.1.1 Infinite resources sharing

To validate our analytical results, we compare the results of simulations obtained with MATLAB and Simulink. Figure 2 shows the Simulink model we used for the comparison. We consider a single server queue in the process using resource block with an infinite amount of resources. We have carried out several scenarios, but we present below one case. Entities arrived in this queue according to a Poisson process of intensity λ = 1.2 arrivals per second. They asked to be served by a quantity of resources communicated to the resource acquirer block. The amount of requested resources follows an exponential distribution of average $1/\mu_R$ = 5 resources. If the requested amount is available, the service starts for a duration following an exponential distribution of average $1/\mu$ = 0.8 seconds. Used resources will be released once the service finished. The server can host simultaneously more than one entity if resources are available to serve them.

Figure 2 Resource sharing Simulink model



The codes to generate the entities are given below, where entity.durService and entity.reqResource are attributes of the created entities indicating respectively the required service time and the amount of requested resources. pLambda, pMu and pMuR are mask parameters of Entity Generator block representing respectively the arrival process intensity, exponential distribution parameter of service time and exponential distribution parameter of requested resources (see online version for colours).

Entity generation:

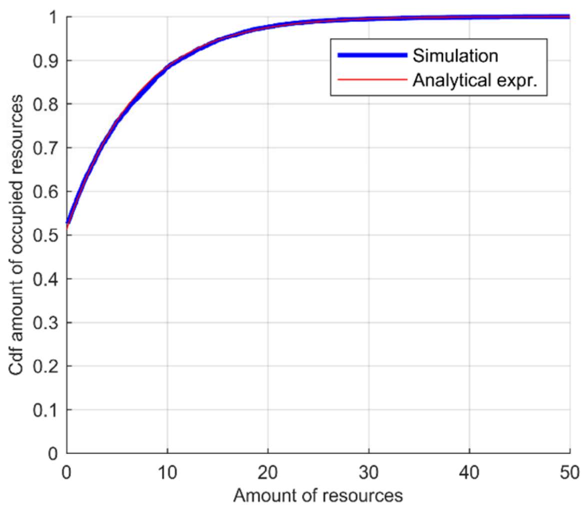
```
arrivalRate = pLambda;
dt = random('exp', 1/arrivalRate);
```

Event action:

```
avgDuration = 1/pMu;
entity.durService = random('exp', avgDuration);
avgResource = 1/pMuR;
entity.reqResource = random('exp', avgResource);
```

In Figure 3, red curve indicates the cumulative distribution function of the amount of occupied resources that we have just determined the expression. Blue curve indicates the increasing cumulative frequency of the number of occupied resources observed on the simulation over a period of 5,000 seconds. We note superposition of the two curves indicating that the analytical expression of the cumulative distribution function that we found coincides well with the increasing cumulative frequency obtained during the simulation, and this, with all the simulations that we have carried out. The analytical expression could therefore be used to predict the outcome of the system, to dimension the amount of required resources, …

Figure 3 Cdf amount of occupied resources – case infinite resources (see online version for colours)



3.1.2 Finite resources sharing

For the case of finite resources sharing, we consider a queue with a single server in Figure 2 but now having C = 30 resources. The customer's arrivals form a Poisson process of intensity $\lambda = 1.2$ arrivals per second. The service time requested by customers still follows an exponential distribution with average $1/\mu = 8$ seconds. Customers request an amount of resources following an exponential distribution with average $1/\mu_R = 5$ resources. With it, the probability that a customer will need more than 30 resources is around 2.5×10^{-3}.

In Figure 4, red curve is from the analytical expression of the distribution function of the amount of occupied resources (in the current work) and blue curve indicates the increasing cumulative frequency of the amount of occupied resources observed on the simulation for a period of 5,000 seconds.

There is also superposition of the two curves indicating that the analytical expression of the distribution function coincides well with the cumulative frequency obtained during the simulation. All simulations that we have carried out confirm this result. So, we can also use the analytical expression in the current work to predict the outcome of the system, to dimension the amount of required resources, …

3.2 Impact of the arrival rate and service time on the probability of blocking

The main objective of this paper is to find analytical model to ease the dimensioning of the server resource capacity. Using the validated analytical expressions, we can build the following results. We have utilised the exponential distribution of single usage model; however, the general case can be used for further practice case.

For a fixed capacity $C = 30$ resources, and average usage $1/\mu_R = 5$ resources by a single customer, we have the impact of arrival rate, or service time in Figure 5. Note that the load factor ρ increases if the service time or the arrival rate increase also.

Figure 4 Cdf amount of occupied resources – case finite resources $C = 30$ (see online version for colours)

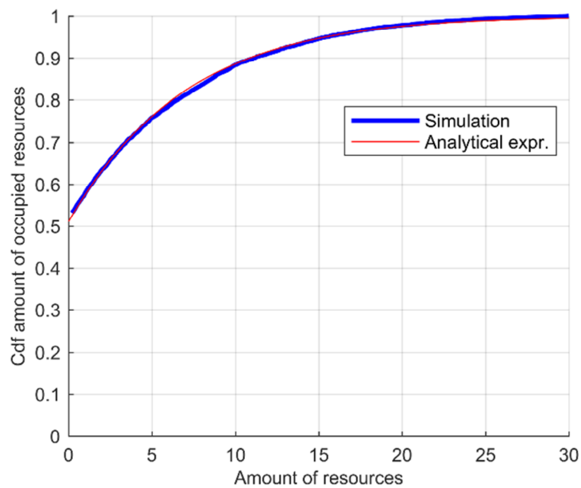
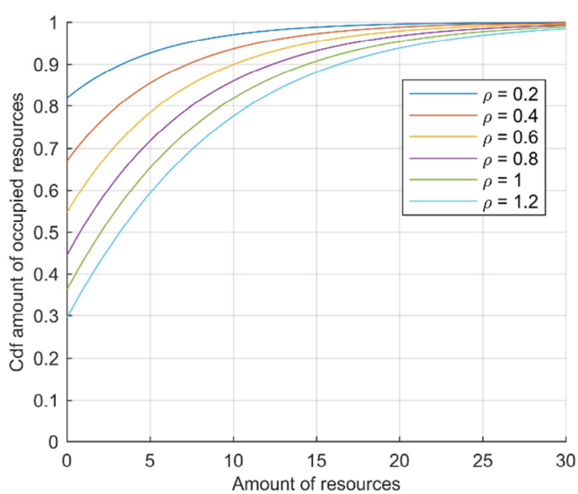


Figure 5 Cdf for fixed resources capacity (see online version for colours)



In Figure 5, considering a value of amount of resources, less the load factor is, less the probability of certain range of resource occupation is. The system has a high probability to be idle (0 occupied resource) for low value of ρ, i.e., low arrival rate, or low service time.

For small value of ρ, the cdf approaches 1 at a lower amount of resources. It means that $P(R \leq 17)$, $P(R \leq 25)$ and $P(R \leq 30)$ are very similar, we do not need to deploy more than 17 resources in this example. We have to take an admissible threshold ε to read in the figure that $P(R > D) \leq \varepsilon$, D is the quantity of resources to deploy. It is also applicable for large value of ρ and the result will be higher quantity of resources.

From the value of cdf at $C = 30$ in Figure 5, we can have the probability of blocking. Table 1 lists the values of such probability for each ρ. It increases with the load factor. It means that the arrival rate, or the service time impacts on the probability of blocking of the system. Higher is ρ, the system has higher probability to be at a blocked state.

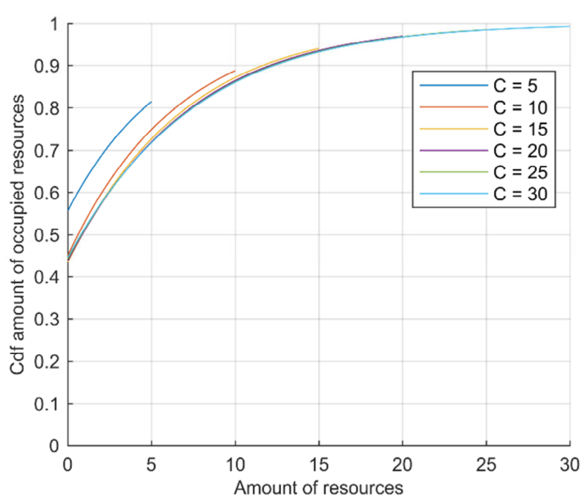
Table 1 Cdf amount of occupied resources at point C for each ρ

ρ	$\lim_{x \to C^-} F_R(x)$	*Blocking probability*
0.2	0.9992	0.0008
0.4	0.9979	0.0021
0.6	0.9958	0.0042
0.8	0.9929	0.0071
1	0.9892	0.0108
1.2	0.9846	0.0154

3.3 Impact of the server resource capacity versus blocking point

Let's vary the capacity C of available resources, and fix the average usage $1/\mu = 5$ resources by a single customer, and a load factor $\rho = 0.8$. Related cumulative distribution function of the fictive occupation of resources (cut at the amount C of resources) are shown in Figure 6.

Figure 6 Cdf by varying resources capacity (see online version for colours)



We have different probabilities of blocking for each C as listed in Table 2. The probability decreases when the capacity increases. So, to have less frequent blocked system, we need to add more resources.

Table 2 Cdf amount of fictive occupied resources at point C for each C

C	$\lim_{x\to C^-} F_R(x)$	*Blocking probability*
5	0.8148	0.1852
10	0.8873	0.1127
15	0.9404	0.0596
20	0.9699	0.0301
25	0.9853	0.0146
30	0.9929	0.0071

We found that from $C = 15$, the cumulative distribution functions are similar. They differ only on the probability of blocking that should be seen at point C of the amount of resources. So, we can use the graph of $C = 30$ to do the following dimensioning rule mentioned in paragraph 3.2: take an admissible threshold ε to read in the figure that $P(R > D) \leq \varepsilon$, D is the quantity of resources to deploy. For example, we can read in Figure 6 that, with $\varepsilon = 0.05$ (5% of blocking), we should have 17 resources to share.

Thus, the best way is to use the expression on infinite resources, that is more simple.

4 Conclusions

From the above results and conclusion, we conclude that our analytical expressions can be used for predict or dimension a resources sharing modelled as a queueing system. They are useful to evaluate the performances such amount of occupied resources, amount of needed capacity, blocking probability. It is straightforward that the load system factor impacts the resource usage by the arrival rate or the service time but from this paper we can evaluate how much it effects the number of resources. The adaptation of these results with multi-class customers could be a future work.

References

Asmussen, S. (2003) *Applied Probability and Queues*, Springer-Verlag, New York.

Gaver, D., Jacobs, P. and Latouche, G. (1984) 'Finite birth-and-death models in randomly changing environments', *Adv. Appl. Probab.*, Vol. 16, No. 4, pp.715–731.

Green, L. (1980) 'A queueing system in which customers require a random number of servers', *Operations Research*, Vol. 28, No. 6, pp.1335–1346.

Kleinrock, L. (1975) *Queueing Systems, Theory*, Vol. 1, Wiley, New York.

Krishnan, N., Yates, R.D., Mandayam, N.B. and Panchal, J.S. (2011) 'Bandwidth sharing for relaying in cellular systems', *IEEE Transactions on Wireless Communications*, Vol. 11, No. 1, pp.117–129.

Lisovskaya, E., Moiseeva, S. and Pagano, M. (2017) 'On the total customers' capacity in multi-server queues', *International Conference on Information Technologies and Mathematical Modelling*, Springer, Cham, pp.56–67.

Massoulie, L. and Roberts, J.W. (2000) 'Bandwidth sharing and admission control for elastic traffic', *Telecommunication Systems*, Vol. 15, No. 1, pp.185–201.

Pagano, M. (2020) 'Analysis of a resource-based queue with the parallel service and renewal arrivals', Lecture notes from international conference of *Distributed Computer and Communication Networks (DCCN 2020)*, pp.335–349, Vol. 12563, pp.335–349, Springer, Cham.

Ravaliminoarimalalason, T.B. and Randimbindrainibe, F. (2021) 'Distribution of occupied resources on a discrete resources sharing in a queueing system', *International Journal of Engineering Research and Technology*, Vol. 10, No. 1, pp.638–643.

Romm, E. and Skitovich, V. (1971) 'On certain generalization of problem of Erlang', *Autom. Remote Control*, Vol. 32, No. 6, pp.1000–1003.

Sopin, E.S. and Vikhrova, O.G. (1995) 'Probability characteristics evaluation in queueing system with random requirements', *CEUR Workshop Proceedings*, pp.85–92.

Tikhonenko, O. (2010) 'Queueing system with processor sharing and limited resources', *Autom. Remote Control*, Vol. 75, No. 5, pp.803–815.

Vikhrova, O.G. (2017) 'About probability characteristics evaluation in queuing system with limited resources and random requirement', *Discrete and Continuous Models and Applied Computational Science*, Vol. 25, No. 3, pp.209–216.

Vishnevski, V.M. and Dudin, A.N. (2017) 'Queueing systems with correlated arrival flows and their applications to modeling telecommunication networks', *Autom. Remote Control*, Vol. 78, No. 8, pp.1361–1403.

Yang, C., Li, J., Guizani, M., Anpalagan, A. and Elkashlan, M. (2016) 'Advanced spectrum sharing in 5G cognitive heterogeneous networks', *IEEE Wireless Communications*, Vol. 23, No. 2, pp.94–101.

Yashkov, S. (1987) 'Processor-sharing queues: some progress in analysis', *Queueing Syst*., Vol. 2, No. 1, pp.1–17.

Ye, Y., Wu, D., Shu, Z. and Qian, Y. (2016) 'Overview of LTE spectrum sharing technologies', *IEEE Access*, Vol. 4, pp.8105–8115.

Zhang, J., Chen, S., Guo, X., Shi, J. and Hanzo, L. (2019) 'Boosting fronthaul capacity: Global optimization of power sharing for centralized radio access network', *IEEE Transactions on Vehicular Technology*, Vol. 68, No. 2, pp.1916–1929.