On the Total Customers' Capacity in Multi-server Queues

Ekaterina Lisovskaya[1(✉)], Svetlana Moiseeva[1], and Michele Pagano[2]

[1] Tomsk State University, Tomsk, Russia
{ekaterina_lisovs,smoiseeva}@mail.ru
[2] University of Pisa, Pisa, Italy
m.pagano@iet.unipi.it

Abstract. In this paper we consider a generalization of M/GI/N/∞ queues, in which customer capacity is an additional parameter of the system and it is independent of the service time. In more detail we focus on the distributions of the total capacity of customers in the different elements of the queue (waiting line, service and entire system) and provide approximate expressions for the corresponding characteristic functions. To verify the goodness of the proposed approximation, several sets of simulations have been carried out, considering discrete and continuous distributions of the customer capacity and using the Kolmogorov distance as a measure of similarity.

Keywords: N-server queuing system · Customer with random capacity · Approximation of the probability distribution

1 Introduction

Queuing theory is one of the most relevant branches of probability theory and applied mathematics [3,6,12,13]. Indeed, queuing systems represent a powerful mathematical tool for performance analysis of a wide variety of real-life systems, including, for instance, telecommunication networks, financial markets, supply chain management and airplane traffic control.

In many application customers are simply characterized in terms of arrival and service processes [1,8,9]. For instance, in computer networks it is typically assumed that the customer volume (i.e., the packet length) is proportional to the service time (namely, the time needed to transmit the packet itself). In this work, we consider a more general model and assume that customer volume and service time are described by independent random variables with arbitrary distributions. As highlighted in the next section, customer capacity plays a relevant role in modeling new network architectures.

In more detail, we consider a queuing system with Poisson arrivals, N servers and unlimited capacity (such assumption is widely used in modeling for sake of analytical tractability). Extending the approach developed by some of the authors in [4] (in which an approximate expression for the distribution of the

A. Dudin et al. (Eds.): ITMM 2017, CCIS 800, pp. 56–67, 2017.
DOI: 10.1007/978-3-319-68069-9_5

number of customers was derived), we will be able to find an explicit approximation for the distribution of the customers' total capacity in the queuing system as well as in its elements (waiting line and service).

The rest of the paper is organized as follows. In Sect. 2, we review the most relevant works on queuing systems with random capacity of customers and highlight the novelty of our contribution. Section 3 properly defines the analyzed queuing model and recalls an approximation for the probability distribution of the number of customers in the system, while Sect. 4 presents the original contribution of the paper, i.e. provides a general expression for the characteristic function of the total customers' capacity. Then, in Sect. 5 the goodness of the approximation (in terms of Kolmogorov distance) is verified through discrete-event simulations for different values of the system parameters. Finally, Sect. 6 concludes the paper with some final remarks.

2 Related Work

In recent years queuing systems with random customer capacity have attracted the interest of researchers for their applicability in different fields, mainly in the framework of computer networks. In this section some of the most relevant contributions are discussed.

In the paper [2] an efficient analytical model that evaluates the behavior of the downlink LTE (Long-Term Evolution) channel with CLA (Cross-Layer Adaptation) is presented. Since video traffic is resource–intensive, it is a challenging issue to stream video over low bandwidth networks, whereas video communication over LTE becomes an open research topic nowadays due to LTEs high throughput capabilities.

The paper [11] deals with a model of a multi-server queuing system with losses caused by lack of resources necessary to service claims. A claim accepted for servicing occupies a random amount of resources of several types with given distribution functions. Random vectors that define the requirements of claims for resources are independent of the processes of customer arrivals and servicing, mutually independent, and identically distributed. Under the assumptions of a Poisson arrival process and exponential service times, the authors analytically find the joint distribution of the number of customers in the system and the vector of amounts of resources occupied by them. Moreover, sample computations are presented to illustrate an application of the model to analyzing the characteristics of a videoconferencing service in an LTE wireless network.

In [10] the authors consider queuing systems, in which customers occupy some resources that are released after customer departure. Arriving customers are lost if there are not enough free resources required for their servicing. In such systems for each customer it is necessary to record the vector of occupied resources until its departure.

Multi-server queuing systems with AQM-type (Active Queue Management) mechanisms are considered in [16,17]. In more detail, in the first work M/M/n-type ($n \geq 1$) queuing systems with a bounded total volume and finite queue

size are considered. It is assumed that the volumes of the arriving packets are generally distributed random variables. Moreover, an AQM-type (Active Queue Management) mechanism is used to control the actual buffer state: each of the arriving packets is dropped with a probability depending on its volume and the occupied volume of the system at the pre-arrival epoch. The stationary queue-size distribution and the loss probability are derived, and numerical examples illustrating theoretical results are also provided. Then, in [17] the analysis is extended to the case of arbitrary distribution of the service time.

The main aim of the paper [14] is to develop a simulation model for queuing systems with non-priority cyclic service RR (round robin) discipline and to compare, in terms of queuing performance, such service discipline with traditional FCFS (first come-first served).

Finally, the paper [15] investigates single server queuing systems with batch Poisson arrivals and without demands losses under assumption that each demand has some random capacity (generally, each demand is characterized by an l-dimensional indication vector). Service time of the demand arbitrary depends on its capacity (indications). The Laplace-Stieltjes transform of total capacities (random vector of sum of indications) of demands that were served during a busy period of the system is determined.

The main novelty of our approach is that it deals with systems without losses and, in this way, permits to dimension the system resources (in terms of buffer space) in order to have loss probabilities below any given threshold (as well-known in the literature, the complementary probability provides an upper bound for the loss probability in the corresponding finite-buffer system). Moreover, our approach is quite general and may be applied to any distribution (discrete or continuous) of the customer capacity, provided that its characteristic function is well-defined. Finally, we also provide the distribution of the overall capacity for the customers in the different components of the queue (waiting line and buffer); such information may be useful to dimension the different elements of the real system under analysis.

3 Approximation of Probability Distribution of the Customers' Number in the System

We consider the M/GI/N/∞ queue. The arrival process is distributed by Poisson law with rate λ. The system has N servers and service times on each server are i.i.d. with distribution function $A(x)$. The arriving customer occupies any free server or goes to the queue in case of all servers are busy. Let each customer have some random capacity $v > 0$ with distribution function $G(y)$. Customers' capacities and service times are mutually independent and do not dependent on the epochs of customers' arrivals.

Denote by $i(t)$ and $V(t) = \sum_{i=1}^{i(t)} v_i$ the number of customers in the system at time t and their total capacity, respectively.

Let $P(i) = P\{i(t) = i\}$ be the stationary probability distribution of the number of customers in the system. We denote by π_i an approximation of $P(i)$, which is defined as a composite distribution [4]:

$$\pi_i = \begin{cases} C_1 P_1(i), 0 \le i \le N, \\ C_2 P_2(i - N + 1), i \ge N. \end{cases} \quad (1)$$

Note that the equality of the two expression for $i = N$ provides an additional condition to determine the constants C_1 and C_2.

The probabilities $P_1(i)$, where $0 \le i \le N$, are the probabilities of the number of occupied servers in an N-server M/GI/N/0 queue with customer losses when all servers are busy. Hence, they can be determined by the Erlang B formula:

$$P_1(i) = \frac{(\lambda a)^i}{i!} \left(\sum_{k=0}^{N} \frac{(\lambda a)^k}{k!} \right)^{-1}$$

where a is the mean service time.

The probabilities $P_2(i)$ refers to states in which all servers are busy. In this case, the block of occupied servers is considered as a single one, characterized by an equivalent service time distribution $B(x)$ and an equivalent mean service time b. Therefore, the probabilities $P_2(i)$, where $i \ge 1$, are defined as the probabilities of having i customers in a single-server queuing system with waiting (i.e., the classical M/GI/1 queue). Hence, they can be determined by the Pollaczek-Khinchin formula [4] and we can write

$$P_2(i) = (1 - \lambda b) \sum_{k=0}^{i} \alpha_k b_{i-k},$$

where the coefficients of the expansion are given by

$$\alpha_0 = \frac{1}{\beta_0}, \ \alpha_n = \frac{1}{\beta_0} \left[\alpha_{n-1} - \sum_{k=0}^{n-1} \alpha_k \beta_{n-k} \right],$$

$$b_0 = \beta_0, \ b_n = \beta_n - \beta_{n-1},$$

$$\beta_n = \int_0^\infty e^{-\lambda z} \frac{(\lambda z)^n}{n!} dB(z),$$

and the distribution function $B(x)$ has the form

$$B(x) = 1 - (1 - A(x)) \left(1 - \frac{1}{a} \int_0^x (1 - A(z))\, dz \right)^{N-1}.$$

The constants C_1 and C_2 in (1) can be found from the normalization condition and the conditions of "stitching" [4]. So the expression (1) becomes:

$$\pi_i = \begin{cases} \dfrac{P_2(1)}{P_2(1) + P_1(N)\,(1 - (P_2(0) + P_2(N)))} P_1(i), 0 \le i \le N, \\ \dfrac{P_1(N)}{P_2(1) + P_1(N)\,(1 - (P_2(0) + P_2(N)))} P_2(i - N + 1), i \ge N. \end{cases} \quad (2)$$

4 Characteristic Function for the Total Capacity

Starting from the definition of conditional expectation, we can write the characteristic function of the total capacity in the form

$$h(u) = M\left\{e^{juV(t)}\right\} = M\left\{M\left\{e^{ju\sum_{k=1}^{i} v_k}\middle| i(t) = i\right\}\right\}$$

$$= \sum_{i=0}^{\infty} M\left\{e^{ju\sum_{k=1}^{i} v_k}\right\} P\{i(t) = i\} = \sum_{i=0}^{\infty}\left(M\left\{e^{juv}\right\}\right)^i P\{i(t) = i\},$$

where we took into account that for $i = 0$ the queue is empty and the sum at the exponent is 0.

Then, using approximation (2), the characteristic function can be rewritten as

$$h(u) = \sum_{i=0}^{\infty}\left(M\left\{e^{juv}\right\}\right)^i \pi_i,$$

and, taking the inverse Fourier transform, we obtain an approximation of the density function of the customers' total capacity in the M/GI/N/∞ queue:

$$f_V(x) = \int_{-\infty}^{\infty} e^{-jux} h(u) du. \tag{3}$$

Similarly, we can obtain the characteristic functions of the total capacity of the customers in the service and in the waiting line. These results have practical relevance since the customers in each element of the queue typically require specific resources (for instance, in routers there is a physical separation between input buffers and output ports).

In more detail, for the customers in the service we obtain:

$$h_{serv}(u) = \sum_{i=0}^{N}\left(M\left\{e^{juv}\right\}\right)^i \frac{P_2(1)P_1(i)}{P_2(1) + P_1(N)\left(1 - (P_2(0) + P_2(N))\right)},$$

and for the customers in the waiting queue:

$$h_{wait}(u) = \sum_{i=0}^{\infty}\left(M\left\{e^{juv}\right\}\right)^{i+N} \frac{P_1(N)P_2(i+1)}{P_2(1) + P_1(N)\left(1 - (P_2(0) + P_2(N))\right)}.$$

5 Simulation and Numerical Examples

Several simulation experiments, performed in the same way as [5], have been carried out to estimate the distribution function of the customers number and the total customers capacity and verify the goodness of the proposed approximation. To this aim, it was also necessary to calculate numerically the approximations (2) and (3) since a close-form solution is, in general, not available.

As a measure of the similarity between simulation and approximation results, we consider the Kolmogorov distance

$$\Delta = \sup_x |F(x) - D(x)| .$$

Here $F(x)$ represents the approximation based on (2) or (3), respectively for $i(t)$ and $V(t)$, and $D(x)$ is the cumulative distribution function built on the basis of the simulation results (in order to reduce the variance of the estimates 10^{10} arrivals have been generated). As typically done in the literature [7], we suppose that an approximation is applicable if its Kolmogorov distance is less than 0.03.

In the following we present the result for three different scenarios, in order to highlight the applicability of our approximation in different settings. Note that the parameters for the arrival and service processes were selected in such a way that the condition for the stationary regime existence is always met ($N > \lambda a$).

Example 1. Let us consider the following parameters for the queue:

- arrival rate $\lambda = 25$
- number of servers $N = 10$
- exponential distribution (with parameter μ) of the service time, i.e.

$$A(x) = \begin{cases} 1 - e^{-\mu x}, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

- uniform distribution (in the interval $[a, b]$) of customers' capacity, i.e.

$$G(y) = \begin{cases} 0, & y < a, \\ \dfrac{y-a}{b-a}, & a \leq y \leq b, \\ 1, & y > b. \end{cases}$$

Furthermore, we used the following numerical values: $\mu = 5$, $a = 0$ and $b = 1$. It is easy to verify that the distributions are very similar both for the customer numbers and the total capacity, as highlighted by Figs. 1 and 2; indeed, we obtain that $\Delta_i = 0.007$ and $\Delta_V = 0.012$, respectively for $i(t)$ and $V(t)$.

Example 2. In the second set of simulation we changed the distribution of the service time. In more detail, the parameters of the queuing system are as follows:

- arrival rate $\lambda = 25$
- number of servers $N = 10$

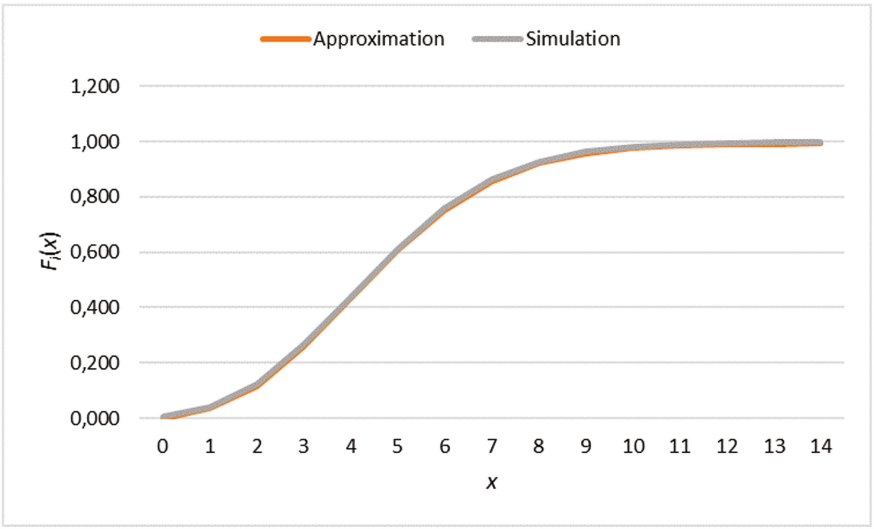

Fig. 1. Example 1 – Distributions of the customers number

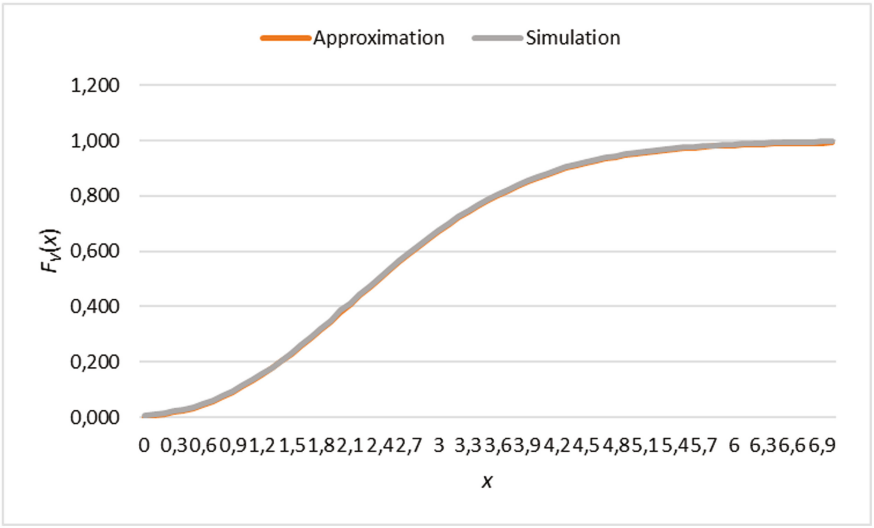

Fig. 2. Example 1 – Distributions of the total capacity

– gamma distribution (with parameters α and β) of the service time, i.e.

$$A(x) = \begin{cases} \dfrac{\gamma(\alpha, \beta x)}{\Gamma(\alpha)}, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

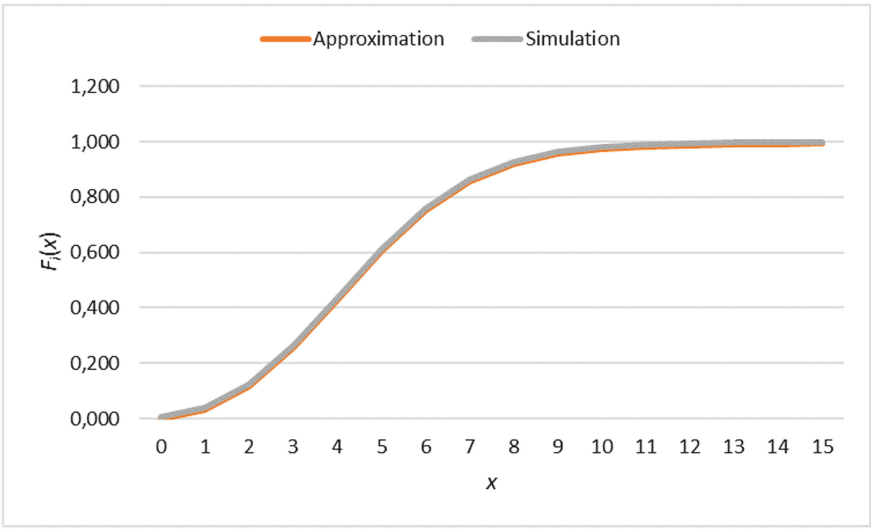

Fig. 3. Example 2 – Distributions of the customers number

– uniform distribution (in the interval $[a, b]$) of customers' capacity, i.e.

$$G(y) = \begin{cases} 0, & y < a, \\ \dfrac{y-a}{b-a}, & a \le y \le b, \\ 1, & y > b. \end{cases}$$

In this case (with $\alpha = 0.5$, $\beta = 2.5$ and, as before, $a = 0$, $b = 1$), the approximation is even closer since $\Delta_i = 0.009$ and $\Delta_V = 0.007$ (see Figs. 3 and 4).

Example 3. In the third set of simulations we verified the goodness of the approximation in case of discrete distribution of the customer capacity. In more detail, we considered the following set of parameters:

– arrival rate $\lambda = 45$
– number of servers $N = 6, 7, 8$
– gamma distribution (with parameters α and β) of the service time, i.e.

$$A(x) = \begin{cases} \dfrac{\gamma(\alpha, \beta x)}{\Gamma(\alpha)}, & x \ge 0, \\ 0, & x < 0, \end{cases}$$

– geometric distribution (in the form representing the number of failures before the first success, with parameter p) of customers' capacity:

$$G(y) = P\{v = y\} = p\,(1-p)^y\,.$$

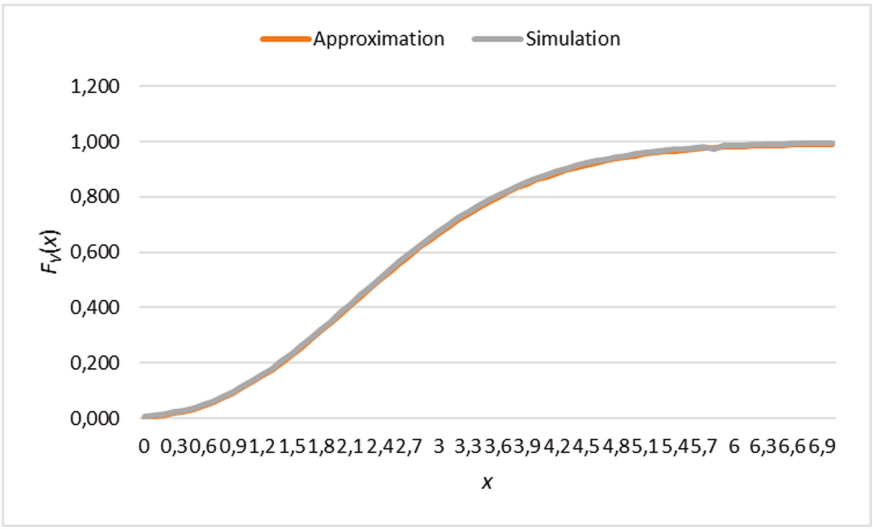

Fig. 4. Example 2 – Distributions of the total capacity

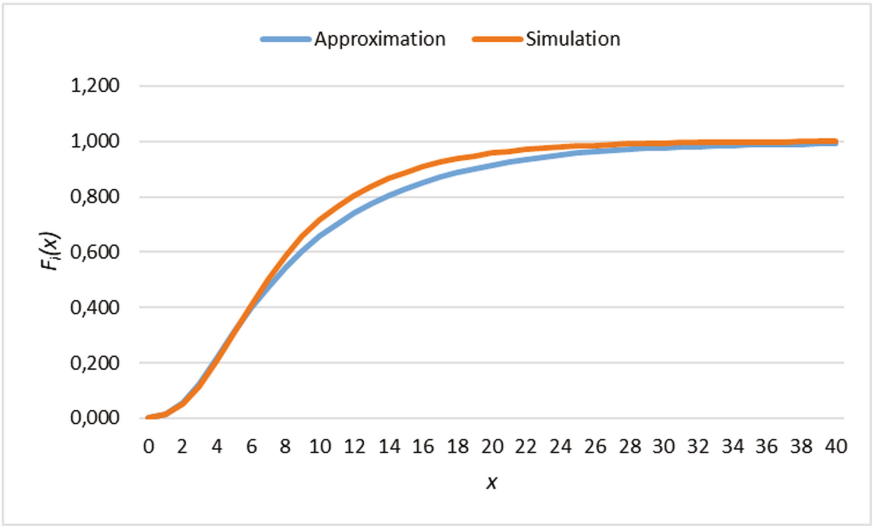

Fig. 5. Example 3 ($N = 6$) – Distributions of the customers number

In all the scenarios we assumed $\alpha = 3.5$, $\beta = 29.7$, $p = 0.4$ and checked how the value of N influences the goodness of the approximation. Figures 5 and 6 points out that the approximation is rather poor for $N = 6$ (indeed, in this case the values of the Kolmogorov distance are $\Delta_i = 0.064$ and $\Delta_V = 0.048$), while it improves when the number of servers is increased, as highlighted by the corresponding values of the Kolmogorov distance (namely $\Delta_i = 0.029$ and

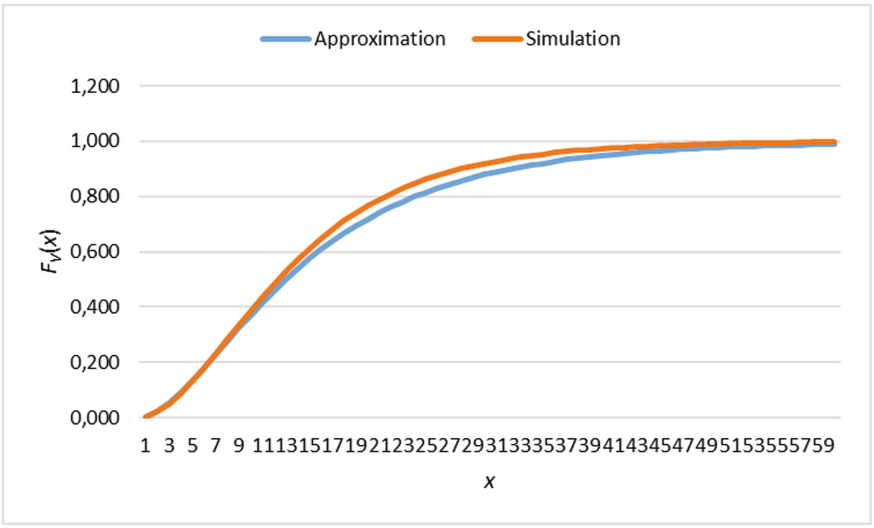

Fig. 6. Example 3 ($N = 6$) – Distributions of the total capacity

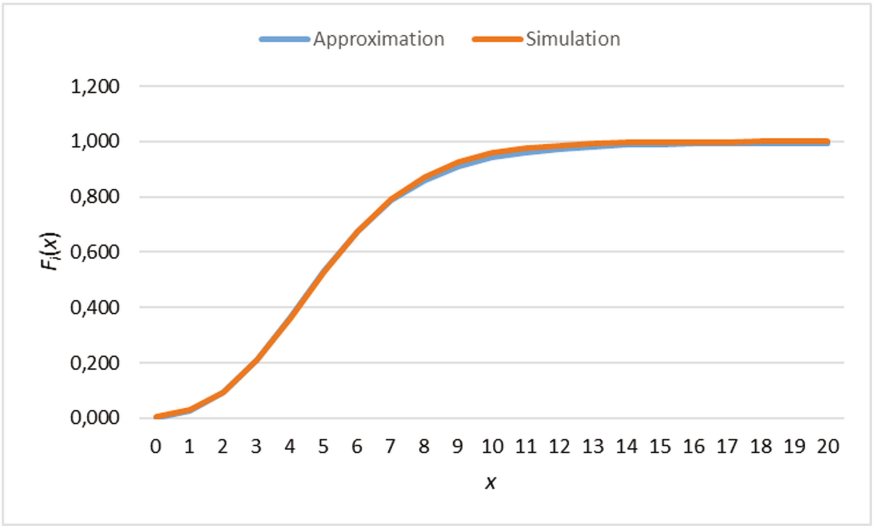

Fig. 7. Example 3 ($N = 8$) – Distributions of the customers number

$\Delta_V = 0.016$ for $N = 7$, $\Delta_i = 0.017$ and $\Delta_V = 0.005$ for $N = 8$) that are clearly below the admissibility threshold. Finally, a visual evidence of the goodness of the proposed approximation is provided by Figs. 7 and 8, referring to $N = 8$ (for sake of brevity, the graphs for $N = 7$ are omitted).

We can conclude that the accuracy of the total capacity approximation is suitable over a wide range of system parameters and improves with the increase of the number of servers in the system.

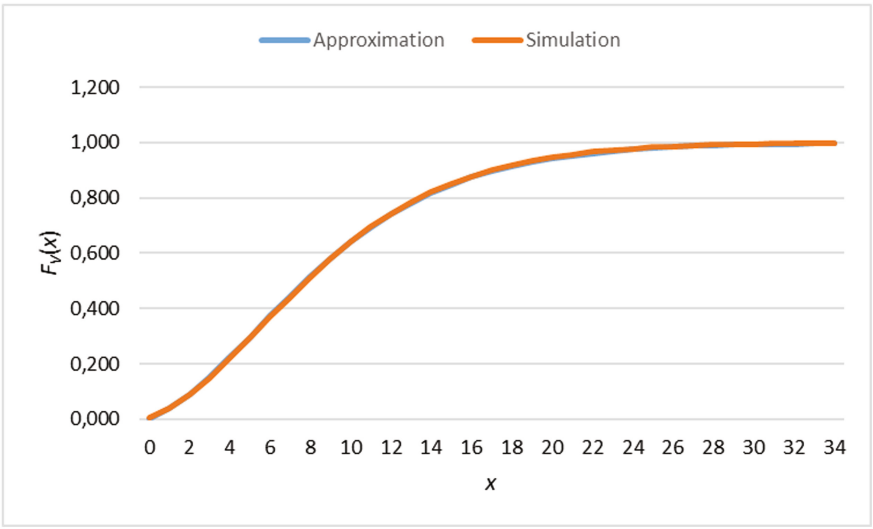

Fig. 8. Example 3 ($N = 8$) – Distributions of the total capacity

6 Conclusions

In this paper we analyzed a generalization of M/GI/N/∞ queues with customers of random capacity. Such models present not only theoretical interest, but also practical relevance in modeling new network architectures (eg., CLA in LTE) and AQM mechanisms in queues.

In more detail we considered the distribution of the total capacity of customers in the system and, starting from our previous results in [4] and the definition of conditional expectation, derived an approximate expression for its characteristic function. Then, we extended the proposed methodology to the total capacity of the customers in the waiting line and in the service, providing the general expressions of the corresponding characteristic functions.

Finally, the goodness of the proposed approximation was verified (in terms of Kolmogorov distance) through several sets of simulations, considering continuous as well as discrete distributions of the customer capacity.

References

1. Apachidi, X.N., Katsman, Y.: Development of a queuing system with dynamic priorities. Key Eng. Mater. **685**, 934–938 (2016)
2. Efimushkina, T., Gabbouj, M., Samuylov, K.: Analytical model in discrete time for cross-layer video communication over LTE. Autom. Control Comput. Sci. **48**(6), 345–357 (2014)
3. Fedorova, E.: The second order asymptotic analysis under heavy load condition for retrial queueing system MMPP/M/1. In: Dudin, A., Nazarov, A., Yakupov, R. (eds.) ITMM 2015. CCIS, vol. 564, pp. 344–357. Springer, Cham (2015). doi:10.1007/978-3-319-25861-4_29

4. Lisovskaya, E., Moiseeva, S.: Study of the Queuing Systems M/GI/N/∞. Commun. Comput. Inf. Sci. **564**, 175–184 (2015)
5. Lisovskaya, E., Pagano, M.: Imitacionnoe modelirovanie sistemy massovogo obsluzhivaniya trebovanij sluchajnogo ob"ema. Problemy optimizacii slozhnyh sistem: Trudy 12-j Mezhdunarodnoj Aziatskoj shkoly-seminara, 352–357 (in Russian)(2016)
6. Moiseev, A., Nazarov, A.: Queueing network MAP/$(\mathrm{GI}/\infty)^K$ with high-rate arrivals. Eur. J. Oper. Res. **254**(2), 161–168 (2016)
7. Moiseev, A., Sinyakov, M.: Razrabotka ob'ektno-orientirovannoj modeli sistemy imitacionnogo modelirovaniya processov massovogo obsluzhivaniya. Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naya tekhnika i informatika 1, 89–93 (In Russian)(2010)
8. Moiseev, A.: Asymptotic Analysis of the Queueing Network $SM/(GI/\infty)^K$. Commun. Comput. Inf. Sci. **564**, 73–84 (2015)
9. Moiseeva, S., Zadiranova, L.: Feedback in infinite-server queuing systems. In: Vishnevsky, V., Kozyrev, D. (eds.) DCCN 2015. CCIS, vol. 601, pp. 370–377. Springer, Cham (2016). doi:10.1007/978-3-319-30843-2_38
10. Naumov, V.A., Samuilov, K.E.: On Modeling Queueing Systems with Multiple Resources. Vestn. Ross. Univ. Druzhby Narodov, Ser. Mat. Informatika. Fiz. 3, 60–64 (2014)
11. Naumov, V.A., Samuilov, K.E., Samuilov, A.K.: On the total amount of resources occupied by serviced customers. Autom. Remote Control **77**(8), 1419–1427 (2016)
12. Nazarov, A., Broner, V.: Inventory management system with Erlang distribution of batch sizes. In: Dudin, A., Gortsev, A., Nazarov, A., Yakupov, R. (eds.) ITMM 2016. CCIS, vol. 638, pp. 273–280. Springer, Cham (2016). doi:10.1007/978-3-319-44615-8_24
13. Pankratova, E., Moiseeva, S.: Queueing system GI/GI/∞ with n types of customers. Commun. Comput. Inf. Sci. **564**, 216–225 (2015)
14. Raspopov, A., Katsman, Y.Y.: Resource allocation algorithm modeling in queuing system based on quantization. Key Eng. Mater. **685**, 886–891 (2016)
15. Tikhonenko, O., Kawecka, M.: Busy period characteristics for single server queue with random capacity demands. In: Kwiecień, A., Gaj, P., Stera, P. (eds.) CN 2012. CCIS, vol. 291, pp. 393–400. Springer, Heidelberg (2012). doi:10.1007/978-3-642-31217-5_41
16. Tikhonenko, O., Kempa, W.M.: On the queue-size distribution in the multi-server system with bounded capacity and packet dropping. Kybernetika **49**(6), 855–867 (2013)
17. Tikhonenko, O., Kempa, W.M.: Performance evaluation of an M/G/n-type queue with bounded capacity and packet dropping. Int. J. Appl. Math. Comput. Sci. **26**(4), 841–854 (2016)