On the Application of Dynamic Screening Method to Resource Queueing System with Infinite Servers

Michele Pagano and Ekaterina Lisovskaya

Abstract Infinite-server queues are a widely used modelling tool thanks to their analytical tractability and their ability to provide conservative upper bounds for the corresponding multi-server queueing systems. A relatively new research field is represented by resource queues, in which every customer requires some volume of resources during her staying in the queue and frees it only at the end of the service. In a nutshell, in this paper the joint distribution of the processes describing the number of busy servers and the total volume of occupied resources is derived and the parameters of the corresponding bidimensional Gaussian distribution are explicitly calculated as a function of the arrival process characteristics and the service time and customers capacity distributions. The aim of this paper is twofold: on one side it summarizes in a *ready-to-be-used way* the main results for different arrival processes (namely, Poisson processes, renewal processes, MAP, and MMPP), on the other it provides a detailed description of the employed methodology, presenting the key ideas at the basis of powerful analysis tools (dynamic screening and asymptotic analysis methods), developed in the last two decades by Tomsk researchers.

Keywords Resource queuing systems · Dynamic screening method · Asymptotic analysis method · Renewal processes · MMPP · MAP

M. Pagano (✉)
Department of Information Engineering, University of Pisa, Pisa, Italy
e-mail: michele.pagano@iet.unipi.it

E. Lisovskaya
Tomsk State University, Tomsk, Russian Federation
e-mail: ekaterina_lisovs@mail.ru

V. C. Joshua et al. (eds.), *Applied Probability and Stochastic Processes*,
Infosys Science Foundation Series, https://doi.org/10.1007/978-981-15-5951-8_12

1 Introduction

Infinite-server queues play a relevant role in queueing theory and in performance analysis. Indeed, different issues can be modelled in such a way, that the number of *servers* is really infinite or so big, that in practice there are always free servers. A typical example is represented by economical models, in which there is no reason to limit the number of contracts that can be signed between credit organizations and clients. Although in real systems physical resources are always finite, these models can be applied to the analysis of computing clusters and multi-core supercomputers, as well as to high-capacity routers (see [1] and references therein).

Moreover, infinite-server queues have a higher analytical tractability than the corresponding multi-server systems: for several classes of arrival processes not only mean values of the performance indexes are available, but it is also possible to determine the corresponding probability distributions, at least under some asymptotic conditions. For instance, “heavy traffic” scenarios are often encountered in computer networks and the knowledge of the steady-state distribution of the number of busy servers can provide conservative upper bounds for the correct dimensioning of the system (e.g., output capacity of a router).

Traditionally, in network modelling the service was associated to packets transmission or calls duration, but this assumption is getting less and less true in modern network architectures. Indeed, issues related to virtual machine allocation in cloud environments or performance of LTE (Long Term Evolution) networks require new queueing models, in which the customers ask for *some resources* (CPU/memory and radio resources, respectively) that are released at the end of the service. Such models are known in the literature as *resource queueing systems*. For instance, in [2] they are applied to the analysis of M2M traffic characteristics in a LTE network cell, while [3] presents an overview of the resource queuing systems used for modeling of a wide class of real systems with limited resources, focusing on wireless networks with exponentially distributed service time. Resource queues in connection with AQM (Active Queue Management) mechanisms are investigated in [4] under the processor sharing discipline, but the analysis is limited to Poisson arrivals. Finally, analytical results for systems with finite resources are given in [5], where M/M/n/m queues are considered and the service time is assumed to be proportional to the customer capacity.

All the above-mentioned works deal with finite resource queueing systems, and analytical results are obtained under stringent condition for the arrival process and the service time distribution. However, the inadequacy of the Poisson process as arrival model is well-known in the literature [6, 7] and more realistic traffic models have been proposed in the literature, such as MAPs (Markov Arrival Processes) and MMPPs (Markov Modulated Poisson Processes). In case of infinite-server resource queues the previous limitations disappear and such models can be used to calculate conservative bounds on system performance under realistic traffic conditions and general distributions of the service time and the customer capacity.

The aim of the paper is to present the principal results in the field of infinite-server resource queueing systems, most of them derived and tested by the authors in the last few years and gathered in [8], where complete proofs and simulation results are also reported. However, to the best of our knowledge, this paper is the first attempt (at least in the English scientific literature) to collect the main results for such systems in a review work and provide for networking specialists the possibility of choosing the most suitable model and finding the relevant performance indexes. It is worth noticing that, above all in case of correlated arrivals, the derivation of the Gaussian approximation is quite cumbersome, so we mainly focus on the methodological elements and present a complete analysis only for Poisson arrivals. Indeed, we also aim to popularize powerful tools developed in the last decades by the "Tomsk queueing theory School," such as the alternative description of MAPs, the dynamic screening method, and the asymptotic analysis method.

The rest of the paper is organized as follows. In Sect. 2 we provide a thorough description of the analyzed queueing systems and recall some background results and definitions, while the following section details the application of the proposed methodology to the case of Poisson arrivals. Then, in Sect. 4 we generalize the analysis to renewal processes and MAPs, highlighting the differences with the Poisson case and summarizing the key results. Finally, the main contributions of the paper are pointed out in the Conclusions, together with future research directions.

2 Reference Model and Theoretical Background

In this section we describe the system under analysis, introducing the notation and the mathematical apparatus used in the rest of the paper. In describing the different arrival processes we detail the definition of MAPs, since our notation is slightly different (although equivalent) from the one most widely used in the western literature. Finally, we briefly recall the dynamic screening method for the study of non-Markovian queueing systems.

2.1 Infinite-Server Resource Queueing System

Let us consider an infinite-server queueing system with infinite resources (so no customer will be rejected) as shown in Fig. 1. An arriving customer can occupy any free server for a random service time $\xi \geq 0$, characterized by a distribution function $B(\tau) = P\{\xi < \tau\}$ with finite first moment (roughly speaking, it is just required that the mean service time is finite and no assumptions are made on its variance). As already mentioned in the introduction, the customer requires during his service also some resource of random volume ν, described by a distribution function $G(y) = P\{\nu < y\}$ with finite first and second moments. When the service is completed, the customer leaves the system and frees the resource. Moreover, service times $\{\tau\}$

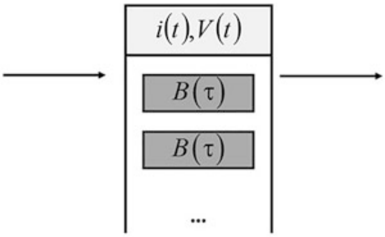

Fig. 1 Infinite-server resource queueing system

and customer capacities $\{\nu\}$ are assumed to be mutually independent and do not dependent on the epochs of customers' arrivals.

Let us fix an initial moment t_0 (to get the steady-state regime it will be enough to consider $t_0 \to -\infty$) and let the system be empty at time t_0.

Denote by $i(t)$ the number of customers in the system at time t; then, the total volume of occupied resources (i.e., the total customers capacity) is given by

$$V(t) = \sum_{i=1}^{i(t)} \nu_i$$

and the bidimensional process $\{i(t), V(t)\}$ unambiguously characterizes the state of the considered queueing system.

Due to the independence of the two components, it is easy to find a relation among them. Indeed, the characteristic function of the total customers capacity can be rewritten as

$$\begin{aligned} h(v) &= M\left\{e^{jvV(t)}\right\} = M\left\{M\left\{e^{jv\sum_{k=1}^{i}\nu_k}|i(t)=i\right\}\right\} \\ &= \sum_{i=0}^{\infty} M\left\{e^{jv\sum_{k=1}^{i}\nu_k}\right\} P\{i(t)=i\} = \sum_{i=0}^{\infty}\left(M\left\{e^{jv\nu}\right\}\right)^i P\{i(t)=i\} \end{aligned}$$

and, taking into account that

$$M\left\{e^{jv\nu}\right\} = \int_0^{\infty} e^{jvy} dG(y) = G^*(v)\,,$$

we get the link between traditional (the number of busy servers does not depend on the occupied resources) and resource queueing systems:

$$h(v) = \sum_{i=0}^{\infty}\left(G^*(v)\right)^i P\{i(t)=i\}\,. \tag{1}$$

However, this elegant result does not solve our problem. Indeed, the distribution of the number of busy servers is known only for a limited set of systems (see

Sect. 3.1 for Poisson arrivals); even in that case, quite often the analytical expression of the distribution of the total customers capacity is not available and only numerical approximations can be found.

The process $\{i(t), V(t)\}$ is, in general, non-Markovian; among the different approaches (for instance, the analysis of the embedded Markov chain and the method of supplementary variables) proposed in the literature, we consider the dynamic screening method that provides a unified framework for the analysis of infinite-server queueing systems (including tandem queues, queueing networks, and resource systems) and is described in Sect. 2.3.

2.2 Arrival Process

The arrival process plays a major role in determining not only the queueing behavior, but also the analytical tractability of the system. Indeed, analytical results can be obtained only for Poisson arrivals, but it is well-known that the distribution of inter-arrival times is typically quite far from the exponential one and, above all, the arrivals are correlated [6]. To cope with these issues, we will consider two different classes of traffic models, widely used in the literature: renewal processes and MAPs (which include MMPPs as a special case). Unfortunately, for both classes closed-form results are not available and only asymptotic approximations can be obtained under heavy traffic conditions. In this paper we introduce a scale parameter $N \to \infty$ (high intensity parameter) and focus on the case of "infinitely growing arrival rate" (for the other regime, known in the literature as "infinitely growing service time," see for instance [9]).

In more detail, renewal processes are characterized by the sequence of inter-arrival times $\{\zeta_n\}$, which are independent identically distributed random variables with common distribution $A(z) = P\{\zeta < z\}$; in our analysis only the existence of finite mean and variance is assumed. Hence, the asymptotic condition simply corresponds to a scaled distribution $A(zN)$ and the mean interarrival time goes to 0 as $1/N$ when $N \to \infty$.

As far as MAPs are concerned, we make use of the characterization developed by Tomsk researchers on the basis of the theory of doubly stochastic processes, which includes the following components [1]:

- $k(t)$: a continuous time ergodic Markov chain with K states and infinitesimal generator matrix $\boldsymbol{Q} = \|q_{k\nu}\|,\ k, \nu = 1, \ldots, K$
- $\lambda_k \geq 0,\ k = 1, \ldots, K$: the conditional arrival rate for each state of the underlying Markov chain $k(t)$, typically denoted through the diagonal matrix $\boldsymbol{\Lambda} = \operatorname{diag}\{\lambda_k\},\ k = 1, \ldots, K$
- $d_{k\nu}\ k, \nu = 1, \ldots, K$: the conditional probabilities that there is an arrival when the Markov chain $k(t)$ changes its state from k to ν (it is assumed that $d_{kk} = 0$), grouped in the matrix $\boldsymbol{D} = \|d_{k\nu}\|,\ k, \nu = 1, \ldots, K$

In this way, unlike the *classical* notation [10], the model parameters have a clear physical interpretation and MMPPs can be easily obtained by setting $\boldsymbol{D} = \boldsymbol{0}$ since state transitions of the underlying Markov chain just imply a rate change, but arrivals are not generated. Moreover, the asymptotic condition can be taken into account multiplying all the coefficient of the matrices $\boldsymbol{Q}$ and $\boldsymbol{\Lambda}$ by the high intensity parameter $N \to \infty$.

It is worth mentioning that this notation is equivalent to the *classical* one, based on matrices $\boldsymbol{D_0}$ and $\boldsymbol{D_1}$. Indeed, it is possible to show that

$$\begin{aligned}\boldsymbol{D_0} &= \boldsymbol{Q} - [\boldsymbol{\Lambda} + \boldsymbol{D} \circ \boldsymbol{Q}] \\ \boldsymbol{D_1} &= \boldsymbol{\Lambda} + \boldsymbol{D} \circ \boldsymbol{Q}\end{aligned}$$

where $\circ$ denotes the Hadamard product (or entrywise product).

2.3 Dynamic Screening Method

In a nutshell, the dynamic screening method is based on the construction of a suitable *screened process* and its *markovization* by the addition of a suitable component, depending in general on the arrival process.

Let us consider two time axes (see Fig. 2): the first one displays the arrival times of all customers, while the other one corresponds to the screened customers. For any $t \geq t_0$ let us define a continuous function $S(t)$ that assumes values in the interval [0, 1]; then, a customer arriving at time t is screened on the second axis (i.e., generates an event on it) with probability $S(t)$. Since the screening probability depends on the arrival time t, the method is called dynamic.

In more detail, for an infinite-server queue we assume that the system is empty at the initial time t_0, fix an arbitrary moment $T > t_0$ and put

$$S(t) = 1 - B(T - t)$$

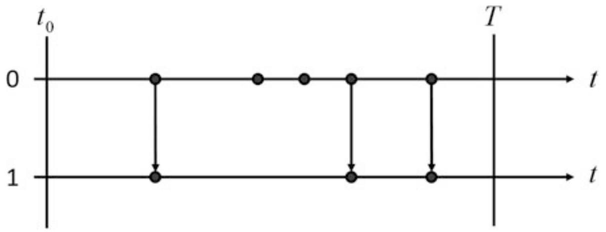

Fig. 2 Screening of the customers' arrivals

i.e., $S(t)$ represents the probability that a customer, arrived at time $t < T$, is not served by time T and so it still occupies some resources in the queue. Instead, with probability $1 - S(t)$ the customer has left the system and it is not screened on the second axis.

Let us denote by $\{n(t)\}$ and $\{W(t)\}$ the counting process representing the number of screened event in the interval $[t_0, t)$ and their total capacity, respectively. The process $\{n(t)\}$ is, in general, non-Markovian (except the case of Poisson arrivals), but it can be *markovized* by adding a suitable component:

- the residual time before the next arrival $z(t)$ in case of renewal processes ⇒ the process $\{n(t), z(t)\}$ is Markovian;
- the state $k(t)$ of the modulating Markov chain in case of MAPs ⇒ the process $\{n(t), k(t)\}$ is Markovian;
- the residual time $z(t)$ and the state $l(t)$ of the embedded Markov chain in case of semi-Markov processes ⇒ the process $\{n(t), z(t), l(t)\}$ is Markovian.

Moreover, the probability distributions of the number of customers in the system $\{i(t)\}$ and the number of screened arrivals on the second axis $\{n(t)\}$ coincide at time T:

$$P\{i(T) = m\} \;=\; P\{n(T) = m\} \quad \forall m = 0, 1, 2, \ldots \tag{2}$$

The latter result, known as *the fundamental equation of the dynamic screening method*, can be easily verified starting from the equality of the corresponding conditional probabilities (given a sequence of L arrivals at times $t_1, t_2, \ldots t_L$)

$$P\{i(T) = m | t_1, t_2, \ldots t_L\} \;=\; P\{n(T) = m | t_1, t_2, \ldots t_L\} \quad \forall m = 0, 1, 2, \ldots$$

for any number of arrivals L and any sequence of arrival times $t_1, t_2, \ldots t_L$, which is a direct consequence of the chosen $S(t)$ as can be verified by direct calculation. Since the distributions of the multidimensional random variable $(L, t_1, t_2, \ldots t_L)$ are the same in the two cases, also the distributions of the random variables $i(T)$ and $n(T)$ (i.e., of the values of the processes $\{i(t)\}$ and $\{n(t)\}$ at time T) coincide. It is easy to prove the same property for the extended process $\{i(t), V(t)\}$:

$$\begin{aligned} P\{i(T) = m, V(T) < z\} \;&=\; P\{n(T) = m, W(T) < z\} \\ &\qquad \forall m = 0, 1, 2, \ldots \;\text{ and } z \geq 0 \end{aligned} \tag{3}$$

that, by analogy with (2), represents *the fundamental equation of the dynamic screening method for resource queueing systems.*

To summarize, the essence of the dynamic screening method consists in the following steps:

1. Choose a suitable screening function $S(t)$ and build the corresponding screened process $\{n(t)\}$;
2. Markovize the process $\{n(t), W(t)\}$, by adding the suitable component $\varkappa(t)$;

3. Determine the probabilistic characteristics of the extended process

$$\{n(t), W(t), \varkappa(t)\};$$

4. Derive the joint distribution of the process $\{n(t), W(t)\}$ (and, in case, the marginal distributions if relevant);
5. Set $t = T$ and, according to (3), get the distribution of the process $\{i(t), V(t)\}$ at time $t = T$.

Finally, note that T was chosen arbitrarily (the only condition is $T > t_0$) and so we can calculate the probability distribution of the joint process *at any time*; in particular, letting $t_0 \to -\infty$, we can get the steady-state distribution, which is typically the parameter of interest in the study of queueing systems.

3 Analysis of Infinite-Server Resource Queueing System: Poisson Arrivals

Let us assume that the arrival process is Poissonian with rate λ and denote by M^v/GI/∞ the corresponding resource queueing system to highlight that customers are characterized by their capacity v. Although in this special case the analysis can be carried out in different ways, we will take advantage of the analytical simplicity of the input process to better illustrate our general methodology. In more detail, at first in Sect. 3.1 we derive the Kolmogorov equation for the characteristic function of the bidimensional process $\{i(t), V(t)\}$ and find the corresponding analytical solution that is possible thanks to the special structure of the arrival process. Then, in Sect. 3.2 we present the *general* approach that provides first- and second-order approximations of the characteristic function.

3.1 Direct Solution of Kolmogorov Equations

Let us define the screened process as described in Sect. 2; thanks to the memoryless property of the exponential distribution, now the bidimensional stochastic process $\{n(t), W(t)\}$ is Markovian and no additional component is required. To visually simplify the analysis, let us introduce the following notation:

$$P\{n(T) = n, W(T) < w\} \overset{\Delta}{=} P(n, w, t) \quad \forall n = 0, 1, 2, \dots \text{ and } w > 0\,,$$

and assume that $P(n, w, t) = 0$ for negative values of n and w. According to the formula of total probability the following equality holds

$$\begin{aligned} P(n, w, t + \Delta t) &= P(n, w, t)(1 - \lambda\Delta t) + P(n, w, t)\lambda\Delta t\,(1 - S(t)) \\ &\quad + \lambda\Delta t S(t)\int_0^\infty P(n-1, w-y, t)dG(y) + o(\Delta t)\,, \end{aligned}$$

from which the set of Kolmogorov differential equations can be easily derived:

$$\frac{\partial P(n, w, t)}{\partial t} = \lambda S(t)\left[\int_0^\infty P(n-1, w-y, t)dG(y) - P(n, w, t)\right] \tag{4}$$

for $n = 0, 1, 2, \ldots$ and $w > 0$, with initial conditions

$$P(n, w, t_0) = \begin{cases} 1 & n = w = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

To solve the Kolmogorov differential equations, let us introduce the characteristic function

$$h(u, v, t) \overset{\Delta}{=} M\left\{\exp\left(jun(t) + jvW(t)\right)\right\} = \sum_{n=0}^\infty e^{jun}\int_0^\infty e^{jvw}P(n, dw, t). \tag{6}$$

Taking into account that

$$\begin{aligned} &\sum_{n=0}^\infty e^{jun}\int_0^\infty e^{jvw}\int_0^w P(n-1, d(w-y), t)dG(y) \\ &= e^{ju}\sum_{n=0}^\infty e^{ju(n-1)}\int_0^\infty e^{jvy}e^{jv(w-y)}\int_0^w P(n-1, d(w-y), t)dG(y) \\ &= e^{ju}\int_0^\infty e^{jvy}\left[\sum_{n=0}^\infty e^{ju(n-1)}\int_0^w e^{jv(w-y)}P(n-1, d(w-y), t)\right]dG(y) \\ &= e^{ju}\int_0^\infty e^{jvy}h(u, v, t)dG(y) = e^{ju}h(u, v, t)\int_0^\infty e^{jvy}dG(y) \\ &= e^{ju}G^*(v)h(u, v, t), \end{aligned}$$

where

$$G^*(v) \overset{\Delta}{=} \int_0^\infty e^{jvy}dG(y)\,,$$

Eq. (4) can be rewritten as

$$\frac{\partial h(u,v,t)}{\partial t} = \lambda S(t) h(u,v,t)\left[e^{ju}G^*(v)-1\right] \tag{7}$$

with the initial condition

$$h(u,v,t_0) = 1 \tag{8}$$

and its solution is given by

$$h(u,v,t) = \exp\left\{\lambda\left[e^{ju}G^*(v)-1\right]\int_{t_0}^{t} S(\tau)d\tau\right\}. \tag{9}$$

For $t = T$ and $t_0 \to -\infty$, by virtue of (3) we obtain the characteristic function of the bidimensional process describing the number of busy servers and the total customers capacity in steady-state conditions:

$$h(u,v) = \exp\left\{\lambda b\left[e^{ju}G^*(v)-1\right]\right\}, \tag{10}$$

where

$$b \overset{\Delta}{=} \int_0^{\infty} (1-B(\tau))\,d\tau\,.$$

Putting $v = 0$ in (10), we get the characteristic function for the number of busy servers in steady-state conditions

$$h(u) \overset{\Delta}{=} h(u,v)|_{v=0} = \exp\left\{\lambda b\left[e^{ju}-1\right]\right\} \tag{11}$$

that coincides with the characteristic function of the Poisson distribution with parameter λb, in agreement with the well-known classical results for the M/GI/∞ queueing systems.

In a similar way the characteristic function for the total customers capacity is

$$h(v) \overset{\Delta}{=} h(u,v)|_{u=0} = \exp\left\{\lambda b\left[G^*(v)-1\right]\right\} \tag{12}$$

in accordance with the results obtained by Oleg Tikhonenko [5] and with Eq. (1). Indeed, as shown by (11), in M/GI/∞ the number of busy servers has Poisson distribution with parameter λb and by direct substitution into (1) we get

$$\begin{aligned} h(v) &= \sum_{i=0}^{\infty}\left(G^*(v)\right)^i P\left\{i(t)=i\right\} = \sum_{i=0}^{\infty}\left(G^*(v)\right)^i \frac{(\lambda b)^i}{i!}e^{-\lambda b} \\ &= e^{-\lambda b}e^{\lambda b G^*(v)} = \exp\left\{\lambda b\left[G^*(v)-1\right]\right\}. \end{aligned}$$

3.2 The Asymptotic Analysis Method

The asymptotic analysis method in queueing systems aims at determining their characteristics under some limit condition [11]. In the following we will consider its application to the differential Eq. (7) in case of "infinitely growing arrival rate" and we look for its approximate solutions with different order of accuracy, namely "first-order asymptotic" $h(u, v, t) \approx h_1(u, v, t)$ and "second-order asymptotic" $h(u, v, t) \approx h_1(u, v, t)h_2(u, v, t)$, also known as Gaussian approximation. Note that it is possible to derive higher order asymptotics, but in that case the inversion of the characteristic function is, in general, possible only by numerical methods and, as stated in [1], at least for "traditional" queueing systems the gain is not significant in case of heavy traffic.

3.2.1 First-Order Asymptotic Analysis

By performing the substitutions

$$\varepsilon = \frac{1}{\lambda},\ u = \varepsilon x,\ v = \varepsilon y,\ h(u, v, t) = f_1(x, y, t, \varepsilon) \tag{13}$$

in Eq. (7), we obtain the following Cauchy problem:

$$\begin{cases} \varepsilon \dfrac{\partial f_1(x, y, t, \varepsilon)}{\partial t} = S(t) f_1(x, y, t, \varepsilon) \left(e^{j\varepsilon x} G^*(\varepsilon y) - 1\right) \\ f_1(x, y, t_0, \varepsilon) = 1. \end{cases} \tag{14}$$

For $\varepsilon \to 0$, taking into account the *first-order* Taylor–Maclaurin expansion

$$e^{j\varepsilon x} = 1 + j\varepsilon x + O\left(\varepsilon^2\right)$$

the limit function $f_1(x, y, t) = \lim\limits_{\varepsilon \to 0} f_1(x, y, t, \varepsilon)$ satisfies the following differential equation:

$$\frac{\partial f_1(x, y, t)}{\partial t} = S(t) f_1(x, y, t) \left(jx + jya_1\right),$$

where a_1 is the average customer capacity, i.e.,

$$a_1 = \int_0^\infty y dG\left(y\right).$$

Taking into account the initial condition $f_1(x, y, t_0) = 1$, we get

$$f_1(x, y, t) \; = \; \exp\left\{(jx + jya_1)\int_{t_0}^{t} S(\tau)d\tau\right\}$$

and, after performing the substitutions inverse to (13), the first-order approximation of $h(u, v, t)$, i.e.,

$$h(u, v, t) \; \approx \; \exp\left\{\lambda\,(ju + jva_1)\int_{t_0}^{t} S(\tau)d\tau\right\}. \tag{15}$$

3.2.2 Second-Order Asymptotic Analysis

The second-order asymptotic provides the bidimensional Gaussian approximation of the process $\{i(t), V(t)\}$. Rewriting the corresponding characteristic function as

$$h(u, v, t) \; = \; h_2(u, v, t)\exp\left\{(jx + jya_1)\int_{t_0}^{t} S(\tau)d\tau\right\}$$

the differential Kolmogorov equation (7) becomes

$$\frac{\partial h_2(u, v, t)}{\partial t} + \lambda(ju + jva_1)S(t)h_2(u, v, t) \; = \; h_2(u, v, t)\lambda S(t)\left(e^{j\varepsilon u}G^*(v) - 1\right)$$

and, after performing the substitutions

$$\varepsilon^2 = \frac{1}{\lambda}, \; u = \varepsilon x, \; v = \varepsilon y, \; h_2(u, v, t) = f_2(x, y, t, \varepsilon), \tag{16}$$

we obtain the following differential equation:

$$\begin{aligned}\varepsilon^2\frac{\partial f_2(x, y, t, \varepsilon)}{\partial t} &+ (j\varepsilon x + j\varepsilon ya_1)S(t)f_2(x, y, t, \varepsilon)\\ &= S(t)f_2(x, y, t, \varepsilon)\left(e^{j\varepsilon x}G^*(\varepsilon y) - 1\right)\end{aligned} \tag{17}$$

with the initial condition

$$f_2(x, y, t_0, \varepsilon) = 1. \tag{18}$$

As before we consider the limit as $\varepsilon \to 0$ and then use the *second-order* Taylor–Maclaurin expansion

$$e^{j\varepsilon x} \; = \; 1 + j\varepsilon x + \frac{(j\varepsilon x)^2}{2} + O\left(\varepsilon^3\right).$$

Then, the limit function $f_2(x, y, t) = \lim_{\varepsilon \to 0} f_2(x, y, t, \varepsilon)$ satisfies the following differential equation:

$$\frac{\partial f_2(x, y, t)}{\partial t} = S(t) f_2(x, y, t) \left(\frac{(jx)^2}{2} + \frac{(jy)^2}{2} a_2 + jxjya_1 \right), \qquad (19)$$

where a_2 is the second moment of the random variable describing the customer capacity, i.e.,

$$a_2 = \int_0^\infty y^2 dG\,(y)\,.$$

The solution of (19), with the initial condition $f_2(x, y, t_0) = 1$, is

$$f_2(x, y, t) = \exp \left\{ \left(\frac{(jx)^2}{2} + \frac{(jy)^2}{2} a_2 + jxjya_1 \right) \int_{t_0}^{t} S(\tau) d\tau \right\}$$

and, performing the substitutions inverse to (16), we get the second-order approximation of $h(u, v, t)$, i.e.,

$$h(u, v, t) \approx \exp \left\{ \lambda \left(ju + jva_1 + \frac{(ju)^2}{2} + \frac{(jv)^2}{2} a_2 + jujva_1 \right) \int_{t_0}^{t} S(\tau) d\tau \right\}. \qquad (20)$$

Finally, for $t = T$ and $t_0 \to -\infty$, by virtue of (3) we obtain the second-order asymptotic for the characteristic function of the steady-state distribution of the bidimensional process $\{i(t), V(t)\}$

$$h(u, v) \approx \exp \left\{ ju\lambda b + jv\lambda a_1 b + \frac{(ju)^2}{2} \lambda b + \frac{(jv)^2}{2} \lambda a_2 b + jujv\lambda a_1 b \right\} \qquad (21)$$

that corresponds to the characteristic function of a bivariate Gaussian process with correlated components. This result has a much wider validity, not limited to Poisson arrival, as shown in the next section.

4 Asymptotic Analysis of Infinite-Server Resource Queueing System

Dynamic screening and asymptotic analysis can be applied to a great variety of arrival processes and queueing systems. For instance, as shown in this section, the proposed methodology can be easily extended to renewal processes and MMPPs, a special case of MAPs widely used in teletraffic [10, 12]. For sake of brevity, we will just sketch the procedure, highlighting the additional complexity due to the change in the input process as well as the general validity of the Gaussian approximation and providing references with the detailed proof of the results.

4.1 The MMPP$^{(\nu)}$/GI/∞ Queue

As already stated in Sect. 2.2, an MMPP is characterized by the two matrices $\boldsymbol{Q}$ and $\boldsymbol{\Lambda}$ and the evolution of the queue depends on the state of the modulating Markov chain $k(t)$. Therefore, it is now necessary to work with the tridimensional Markovian process $\{k(t), n(t), W(t)\}$. Denoting the probability distribution of this process by

$$P(k, n, w, t) = P\left\{k(t) = k, n(t) = n, W(t) < w\right\},$$

and applying the formula of total probability as in the Poisson case, we get

$$\begin{aligned} P(k, n, w, t + \Delta t) = {} & P(k, n, w, t)(1 - \lambda_k \Delta t)(1 + q_{kk}\Delta t) \\ & + P(k, n, w, t)\lambda_k \Delta t(1 - S(t)) \\ & + \lambda_k \Delta t S(t) \int_0^w P(k, n-1, w-y, t) dG(y) \\ & + \sum_{\nu \neq k} q_{\nu k} \Delta t P(\nu, n, w, t) + o(\Delta t), \end{aligned} \tag{22}$$

for $k = 1, \ldots, K$, $n = 0, 1, 2, \ldots$ and $w > 0$.

From (22), we obtain the system of Kolmogorov differential equations

$$\begin{aligned} \frac{\partial P(k, n, w, t)}{\partial t} = {} & \lambda_k S(t) \left[\int_0^w P(k, n-1, w-y, t) dG(y) - P(k, n, w, t) \right] \\ & + \sum_{\nu} q_{\nu k} P(\nu, n, w, t), \end{aligned} \tag{23}$$

with initial conditions

$$P(k, n, w, t_0) = \begin{cases} r(k) & n = w = 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $\{r(k)\}$, $k = 1, \ldots, K$ are the stationary state probabilities of the modulating Markov chain $k(t)$. Note that the first term on the right-hand side of (23) is similar to the one in (4), while the other one takes into account the state transitions in the modulating Markov chain.

Introducing the *partial* characteristic function

$$\begin{aligned} h(k,u,v,t) &= M\left\{\exp\left(jun(t)+jvW(t)\right)\right\} \\ &= \sum_{n=0}^{\infty} e^{jun}\int_0^{\infty} e^{jvw}P(k,n,dw,t), \end{aligned}$$

we can write the following system of equations:

$$\frac{\partial h(k,u,v,t)}{\partial t} = \lambda_k S(t)h(k,u,v,t)\left[e^{ju}G^*(v)-1\right]+\sum_{\nu} h(\nu,u,v,t)q_{\nu k}$$

with the initial condition

$$h(k,u,v,t_0) = r(k) \quad \text{for } k=1,\ldots,K\,,$$

or in matrix form:

$$\frac{\partial \mathbf{h}(u,v,t)}{\partial t} = \mathbf{h}(u,v,t)\left[\mathbf{\Lambda}S(t)(e^{ju}G^*(v)-1)+\mathbf{Q}\right], \tag{24}$$

with the initial condition

$$\mathbf{h}(u,v,t_0)=\mathbf{r}\,,$$

where

$$\mathbf{h}(u,v,t)=[h(1,u,v,t),h(2,u,v,t),\ldots,h(K,u,v,t)]$$

and

$$\mathbf{r}=[r(1),r(2),\ldots,r(K)]$$

is the row-vector of the stationary distribution of the modulating Markov chain:

$$\begin{cases} \mathbf{rQ}=\mathbf{0} \\ \mathbf{re}=1\,, \end{cases}$$

$\mathbf{e}$ being a column-vector with all entries equal to 1.

To the matrix differential equation (24) we apply the asymptotic analysis method to get asymptotic results under the condition of “infinitely growing arrival rate.” Denoting by N the scaling parameter, we consider the family of MMPP processes with $\mathbf{\Lambda} = N\tilde{\mathbf{\Lambda}}$ and $\mathbf{Q} = N\tilde{\mathbf{Q}}$ as $N \to \infty$. Calculations are more cumbersome since now we need to work with a matrix (and not scalar) equation, but, as shown in [13], the procedure is analogous to the Poisson case: the first- and second-order approximations are derived and then, by setting $t = T$ and $t_0 \to -\infty$, we obtain the characteristic function of the process $\{i(t), V(t)\}$ at steady state:

$$h(u, v) \approx \exp\left\{N\lambda(ju + jva_1)b_1 + \frac{(ju)^2}{2}(N\lambda b_1 + N\kappa b_2)\right.$$
$$\left. + \frac{(jv)^2}{2}(N\lambda a_2 b_1 + Na_1^2\kappa b_2) + jujv(N\lambda a_1 b_1 + N\kappa a_1 b_2)\right\}, \qquad (25)$$

where a_1 and a_2 are the first and the second moments of the random variable describing the customer capacity,

$$b_1 = \int_0^\infty (1 - B(\tau))d\tau, \quad b_2 = \int_0^\infty (1 - B(\tau))^2 d\tau$$

and

$$\lambda = \mathbf{r}\tilde{\mathbf{\Lambda}}\mathbf{e}, \quad \kappa = 2\mathbf{g}\left(\tilde{\mathbf{\Lambda}} - \lambda\mathbf{I}\right)\mathbf{e},$$

where the row-vector $\mathbf{g}$ satisfies the linear matrix system

$$\begin{cases} \mathbf{g}\tilde{\mathbf{Q}} = \mathbf{r}\left(\lambda\mathbf{I} - \tilde{\mathbf{\Lambda}}\right) \\ \mathbf{g}\mathbf{e} = 1. \end{cases}$$

The form of the characteristic function (25) implies that the bidimensional process $\{i(t), V(t)\}$ is asymptotically Gaussian with the vector of mathematical expectations

$$\mathbf{a} = N\left[\lambda b_1 \quad \lambda a_1 b_1\right]$$

and the covariance matrix

$$\mathbf{K} = N\begin{bmatrix} \lambda b_1 + \kappa b_2 & \lambda a_1 b_1 + \kappa a_1 b_2 \\ \lambda a_1 b_1 + \kappa a_1 b_2 & \lambda a_2 b_1 + \kappa a_1^2 b_2 \end{bmatrix}.$$

In the general case of MAPs [14], the procedure is exactly the same, only equality (22) slightly changes since a transition of the modulating Markov chain $k(t)$ from state ν to state k (with $k \neq \nu$) can now generate an arrival with probablity $d_{\nu k}$. This corresponds to substitute the matrix $\boldsymbol{\Lambda}$ with $\boldsymbol{\Lambda} + \boldsymbol{Q} \circ \boldsymbol{D}$, leaving unchanged all the rest. Apart from the value of λ and κ, equality (25) still holds for the steady-state characteristic function and hence the previous considerations about Gaussianity can be extended to MAP$^{(\nu)}$/GI/∞ resource queues.

4.2 The $GI^{(\nu)}/GI/\infty$ Queue

Let us consider as input flow a renewal process and assume that the inter-arrival time, characterized by the distribution $A(z)$, has finite mean and variance, i.e.,

$$a = \frac{1}{\lambda} = \int_0^{\infty} (1 - A(z))\, dz \quad \text{and} \quad \sigma^2 = \int_0^{\infty} (z - a)\, dA(z)\,.$$

In this case the memoryless property does not hold, hence it is necessary to take into account the residual time $z(t)$ to obtain a Markovian process $\{z(t), n(t), W(t)\}$. Denoting its probability distribution by

$$P(z, n, w, t) \;=\; P\left\{z(t) < z, n(t) = n, W(t) < w\right\},$$

the formula of total probability leads to the following equality (for $n = 0, 1, 2, \ldots$, and $z, w > 0$):

$$\begin{aligned} P(z, n, w, t + \Delta t) = {} & [P(z + \Delta t, n, w, t) - P(\Delta t, n, w, t)] \\ & + P(\Delta t, n, w, t)(1 - S(t))A(z) \\ & + A(z)S(t) \int_0^w P(\Delta t, n - 1, w - y, t) dG(y) + o(\Delta t), \end{aligned}$$

from which the Kolmogorov differential equation is easily derived:

$$\begin{aligned} \frac{\partial P(z, n, w, t)}{\partial t} = {} & \frac{\partial P(z, n, w, t)}{\partial z} + \frac{\partial P(0, n, w, t)}{\partial z}(A(z) - 1) \\ & + S(t)A(z)\left[\int_0^w \frac{\partial P(0, n - 1, w - y, t)}{\partial z} dG(y) - \frac{\partial P(0, n, w, t)}{\partial z}\right], \end{aligned} \tag{26}$$

with initial condition

$$P(z, n, w, t_0) \;=\; \begin{cases} R(z) & n = w = 0 \\ 0 & \text{otherwise}\,, \end{cases}$$

where

$$R(z) = \frac{1}{a} \int_0^z (1 - A(u)) du$$

is the stationary distribution of the renewal arrival process. Also in this case it is useful to rewrite the Kolmogorov equation in terms of the *partial* characteristic function

$$\begin{aligned} h(z,u,v,t) &= M\left\{\exp\left(jun(t)+jvW(t)\right)\right\} \\ &= \sum_{n=0}^{\infty} e^{jun} \int_0^{\infty} e^{jvw} P(z,n,dw,t) \end{aligned}$$

and we obtain the following equation:

$$\begin{aligned} \frac{\partial h(z,u,v,t)}{\partial t} &= \frac{\partial h(z,u,v,t)}{\partial z} \\ &\quad + \frac{\partial h(0,u,v,t)}{\partial z}\left[A(z)-1+A(z)S(t)\left(e^{ju}G^*(v)-1\right)\right], \end{aligned} \tag{27}$$

with the initial condition

$$h(z,u,v,t_0) = R(z)\,. \tag{28}$$

Since the exact solution of (27) is, in general, not available, we apply the asymptotic analysis method under the condition of "infinitely growing arrival rate," rewriting the distribution function as $A(Nz)$ with $N \to \infty$ as in Sect. 4.1. Following our *usual* approach, we get the second-order approximation of $h(z,u,v,t)$ and, setting $z \to \infty, t = T, t_0 \to -\infty$, we obtain the characteristic function of the process $\{i(t), V(t)\}$ in the steady-state regime (see [15] for the detailed proof):

$$\begin{aligned} h(u,v) \;\approx\; \exp\Bigg\{ & N\lambda(ju+jva_1)b_1 + \frac{(ju)^2}{2}(N\lambda b_1 + N\kappa b_2) \\ & + \frac{(jv)^2}{2}(N\lambda a_2 b_1 + N a_1^2 \kappa b_2) + jujv(N\lambda a_1 b_1 + N\kappa a_1 b_2)\Bigg\}, \end{aligned} \tag{29}$$

where a_1, a_2, b_1, and b_2 are the same as in (25), while the expression of κ has changed:

$$\kappa = \lambda^3\left(\sigma^2 - a^2\right).$$

In complete analogy with the result in Sect. 4.1, the bidimensional process $\{i(t), V(t)\}$ is asymptotically Gaussian with the vector of mathematical expectations

$$\mathbf{a} = N\left[\lambda b_1 \quad \lambda a_1 b_1\right]$$

and the covariance matrix

$$\mathbf{K} = N \begin{bmatrix} \lambda b_1 + \kappa b_2 & \lambda a_1 b_1 + \kappa a_1 b_2 \\ \lambda a_1 b_1 + \kappa a_1 b_2 & \lambda a_2 b_1 + \kappa a_1^2 b_2 \end{bmatrix}$$

that have exactly the same expression (apart from the definition of κ and λ) as in the MMPP case.

5 Conclusions

In this work we analyzed infinite-server resource queueing systems, collecting in a review paper the most relevant results we obtained in the last few years. To the best of our knowledge it is the first attempt in the English literature to describe a general analysis methodology for such systems and provide a list of *ready-to-be-used* formulas for different arrival processes (namely, Poisson processes, renewal processes, MAP, and MMPP). The proposed approach is based on the application at first of the dynamic screening method (for markovization purposes) and then of the asymptotic analysis method (to find at least an asymptotic solution for the corresponding Kolmogorov equations). In a nutshell, the paper highlights that, under the condition of "infinitely growing arrival rate," the joint distribution of the processes describing the number of busy servers and the total volume of occupied resources is bivariate Gaussian and provides analytical expressions for its parameters (mean vector and covariance matrix) as a function of the arrival process characteristics, the distribution of the service time and the first and second moments of the customers capacity distribution.

Finally, it is worth mentioning that the proposed methodology is much more general and can be applied to other arrival processes (e.g., semi-Markov processes), heterogeneous customers/servers, multi-resource customers as well as to more complex resource systems, including tandem queues and queueing networks.

Acknowledgments The publication has been prepared with the support of the University of Pisa PRA 2018–2019 Research Project "CONCEPT—COmmunication and Networking for vehicular CybEr-Physical sysTems."

References

1. Nazarov, A., Moiseev, A.: Infinite-Server Queueing System and Networks (in Russian). Publishing House STL, Tomsk (2015)
2. Sopin, E.S., Ageev, K.A., Markova, E.V., Vikhrova, O.G., Gaidamaka, Y.V.: Performance analysis of M2M traffic in LTE network using queuing systems with random resource requirements. Autom. Control Comput. Sci. **52**(5), 345–353 (2018)

3. Gorbunova, A.V., Naumov, V.A., Gaidamaka, Y.V., Samouylov, K.E.: Resource queuing systems as models of wireless communication systems (in Russian). Inform. Primen **12**(3), 48–55 (2018)
4. Tikhonenko, O., Kempa, W.: Queueing system with processor sharing and limited memory under control of the AQM mechanism. Autom. Remote Control **76**(10), 1784–1796 (2015)
5. Tikhonenko, O., Kawecka, M.: Total volume distribution for multiserver queueing systems with random capacity demands. In: Kwiecień, A., Gaj, P., Stera, P. (eds.) Computer Networks, pp. 394–405. Springer, Berlin (2013)
6. Pagano, M., Rykov, V., Yuri, K.: Teletraffic Models (in Russian). Publishing House Infra-M, Moscow (2018)
7. Paxson, V., Floyd, S.: Wide area traffic: the failure of Poisson modeling. IEEE/ACM Trans. Netw. **3**(3), 226–244 (1995)
8. Lisovskaya, E.: Asymptotic methods for the analysis of resource queueing systems with non-Poissonian arrival flows. Ph.D. Thesis, Tomsk State University (2018). Candidate of physical and mathematical Sciences
9. Lisovskaya, E., Moiseeva, S., Pagano, M., Potatueva, V.: Study of the MMPP/GI/∞ queueing system with random customers' capacities. Inf. Appl. **11**(4), 109–117 (2017)
10. Heffes, H., Lucantoni, D.M.: A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. IEEE J. Sel. Areas Commun. **4**, 856–868 (1986)
11. Nazarov, A., Moiseeva, S.: The Asymptotic Analysis Method in Queueing Theory (in Russian). Publishing House STL, Tomsk (2006)
12. Heyman, D.P., Lucantoni, D.: Modeling multiple IP traffic streams with rate limits. IEEE/ACM Trans. Netw. **11**(6), 948–958 (2003)
13. Lisovskaya, E., Moiseeva, S., Pagano, M.: The total capacity of customers in the infinite-server queue with MMPP arrivals. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds.) Distributed Computer and Communication Networks, pp. 110–120. Springer, Cham (2016)
14. Kononov, I., Lisovskaya, E.: Analysis of infinite-server queues with arrivals of random volume (in Russian). In: Proceedings of the XV International Conference Named After A. F. Terpugov, vol. 1, pp. 67–71. Publishing House TSU, Tomsk (2016)
15. Lisovskaya, E., Moiseeva, S.: Asymptotic analysis of non-Markovian infinite-server queueing with renewal arrivals of random volume customers (in Russian). Tomsk State University J. Control Comput. Sci. **39**, 30–38 (2017)