QUEUEING SYSTEMS WITH LIMITED ACCESS TO SERVICE STATION

WOJCIECH M. KEMPA
SILESIAN UNIVERSITY OF TECHNOLOGY,
FACULTY OF APPLIED MATHEMATICS,
INSTITUTE OF MATHEMATICS, GLIWICE, POLAND

ABSTRACT: *Queueing models with different-type limitations in the access to the service station are discussed. Such systems are used in modelling of different-type phenomena occurring in telecommunication and computer networks, production management, transport and logistics. Exemplary models are described and characterized by analytical results. In particular, results for queueing systems with single and multiple server vacation policies and the threshold-type N-policy are given. Moreover, the Active Queue Management (AQM) is described as the set of techniques for reducing the risk of buffer overflows.*

KEYWORDS: *Active Queue Management (AQM), N-policy, queueing system, single/multiple vacation policy.*

1 Introduction

As it seems, queueing theory is nowadays one of the most rapidly developing scientific disciplines of the applied probability area. Fast technological progress in the design of computer and telecommunication networks, in particular relating to different aspects of wireless communication, causes that the range of problems and phenomena which can be investigated using the methods of queueing theory is essentially increasing. As one can observe, many analytical (theoretical) results in queueing theory are obtained, as one can say, "on the occassion" in the process of finding solutions of many practical problems under different-type grants executed by consortia including scientific and industrial individuals. Indeed, queueing theory is particularly applicable in the process of development and performance evaluation of packet-oriented computer networks. In this case the operation of a node of the network (like e.g. IP router) can be modelled by using an appropriate queueing system. Indeed, Leonard Kleinrock, who is the author of classical handbooks in queueing theory, can be deservedly called "a father of the Internet". In his laboratory, in Boelter Hall at University of California Los Angeles, one of two first nodes of the ARPANET network was built (this network can be named the "ancestor" of todays Internet). Obviously, applications of queueing theory reach much deeper and concern the production management and organization, transport problems, logistics etc.

Queueing theory, which was primarily considered as a branch of applied probability, now is in fact an independent scientific discipline. The result of its dynamic development and increasing applicability in solving different-type real-life problems is a huge number of textbooks, monographs, journal articles and scientific conferences devoted to theoretical aspects and practical applications of queueing theory. Among many items, let us mention about classical theoretical works by Bocharov et al. [2], Borovkov [3], Cohen [7], Kleinrock [19], Takács [24] or Takagi [25].

2 Models with service limitations

Queueing models with different-type limitations in the access to the service station are being intensively studied currently. Indeed, in practice one can very often meet a situation in which the service process for some reason is suspended or delayed. The limitation in free access to the service can be associated with the temporary deactivation of the service station (server vacation), during which the service process is completely blocked, or may be connected with the imposibility for a customer (a job, a packet etc.) to join the waiting room (a buffer) due to its finite capacity or some other mechanisms "filtering" the input flow. A detailed discussion on modelling of different-type practical issues leading to queues with limited access to the service station can be found e.g. in an interesting survey by Doshi, 1986 [8].

Indeed, the following practical situations and motivations can be mentioned here:

- machine failures – they may occur randomly, independently on the status of the queue and the number of customers present. Repair times can be considered as server vacations, during which the processing is completely blocked;
- maintenances – every time when the machine processing jobs becomes idle, it undergoes preventive maintenance of random (usually) duration. For example, processors in computer and communication systems execute testing and maintenance besides executing their primary tasks (like processing jobs, receiving and transmitting data etc.). The maintenance and testing is needed to preserve high quality of service and reliability of the system mainly.
- energy saving – a typical problem e.g. in production lines and wireless network communication. In this case each busy period of the appropriate queueing system may be considered as an active period (active mode) e.g. in the functioning of the wireless network node (e.g. wireless sensor network) while, similarly, each idle time can be treated as a power saving period (sleep mode).

In the literature different-types mechanisms of limitation of the access to the service station are proposed. The most important solutions are the following ones:

- single/multiple vacation policy (SV/MV) – in the case of SV policy, every time when the system becomes idle, the service station takes on a (single) vacation (usually of random duration) during which the processing of jobs is suspended. After closing the vacation, if there are jobs waiting for service, the processing restarts immediately; otherwise, the server waits in readiness for the first arrival. In the MV policy successive independent vacations are taken on as far as at least one job waiting for service is detected at the completion epoch of one of them;
- N-policy – after the idle time the processing (a new busy period) is being initialized simultaneously with arriving of the Nth job;
- T-policy – in this case the server is turned on only for a time of a fixed length T every time when at least one job waits for service (this solution can be used in modelling of TDM (Time Division Multiplexing));
- setup/closedown times – may occur if the server is switched off being idle: after the last service the service station needs some time to be deactivated safely; similarly, the first processing in a new busy period is preceded by a setup time during which the server becomes

full ready for job processing. During setup and closedown times the service process is completely blocked;

- server breakdowns (failures) – occur randomly if the service station is busy with processing. After the breakdown occurrence the server needs some time (repair period) to restart the service process;

- gated service discipline – in this mechanism a single processing period is divided into two subperiods. During the first one the arriving jobs are allowed to join the server; at the beginning of the second service subperiod, the service "gate" is being closed up and all incoming jobs must accumulate in the buffer and wait for the opening of the "gate". The gated service discipline is usually used in modelling of the operation of transportation networks (buses, trains, ferries etc.). However, such a phenomenon may also occur in parallel computing and the Internet traffic.

Evidently, due to vast literature which is still growing, it is impossible to present and discuss main results for each of the above mentioned queueing models. In the next sections we present results for some chosen models, both for transient and stationary stochastic characteristics.

3 Single and multiple vacation policies

In these two queueing models the limitation in the access to the service station is connected with the so called "server vacation". More precisely speaking, every time when the server becomes idle it initializes a vacation period during which the processing of accumulated and incoming jobs is completely blocked. In the single vacation policy the service station takes exactly one vacation (usually of random duration) at the end of each busy period. If, at the end of the vacation, the system is still empty, the server waits for the first arrival and the processing starts immediately. Otherwise, the completion epoch of the vacation coincides with the start of the service process. In the multiple vacation policy the service station keeps on taking vacations until, returning from a vacation, at least one customer is accumulated in the buffer queue and waits for service.

Miller in 1964 [23] was the first who investigated a queueing model of the $M/G/1$ type in that the service station is unavailable during some random time (referred to as vacation). Next Levy and Yechiali (1975, [21]) introduced several generalized models of the classical $M/G/1$ queue with temporarily unavailable server.

Queueing models with different-type vacation policies are nowadays intensively studied due to their applications e.g. in modelling of different phenomena occurring in telecommunications and computer networks. In particular, vacation queueing models are used in modelling and perfomance evaluation of the energy saving periods (sleep modes) occuring in wireless telecommunication (like LTE or WiMAX 802.16e). In a wireless network users need the continuous availability of a dedicated wideband data channel. To provide such an undisturbed networking it is necessary to exchange control packets frequently, even in the case of no data to be exchange using the network. As a consequence a large amount of energy is consumed to control such a high-speed connection (see e.g. Mancuso and Alouf, 2012 [22] for discussion on this topic). The power saving can be achieved, however, by using the so called sleep mode operations.

Let us take into consideration an infinite-buffer $M/G/1$ type model in which customers arrive according to a Poisson process with rate λ and are being served individually with a CDF $F(\cdot)$ of the service time with LST $f(\cdot)$. Moreover, let (V_n), $n \geq 1$, be a sequence of independent positive random variables with a common CDF $G(\cdot)$, with LST $g(\cdot)$, where V_n corresponds to the nth server

vacation. We assume that at time $t = 0$ a customer enters the empty system with the server being on vacation, and that V_1 denotes the residual duration of the first vacation. Besides, assume that the sequence (V_n) is independent on the arrival and service processes.

Investigate the models with single and multiple vacations simultaneously. Let us take into consideration only moments of service and vacation completions (we treat them as one sequence). Let $(X_n^\star, I_n^\star)$ denote the state of the system at the nth such moment, where $X_n^\star$ stands for the number of customers present and

$$I_n^\star = \begin{cases} 0 & \text{if the } n\text{th moment is a vacation termination instant,} \\ 1 & \text{if the } n\text{th moment is a service completion instant.} \end{cases} \tag{1}$$

Let us note that for the system operating under single vacation policy we have (see [8])

$$(X_{n+1}^\star, I_{n+1}^\star) = \begin{cases} (X_n^\star + \eta_S - 1, 1) & \text{if} \quad X_n^\star \geq 1, \\ (\eta_S, 1) & \text{if} \quad (X_n^\star, I_n^\star) = (0,0), \\ (\eta_V, 0) & \text{if} \quad (X_n^\star, I_n^\star) = (0,1), \end{cases} \tag{2}$$

where η_S and η_V stand, respectively, for the number of customers arriving during one service time and during one vacation.

For the model with multiple vacation policy we obtain, similarly,

$$(X_{n+1}^\star, I_{n+1}^\star) = \begin{cases} (X_n^\star + \eta_S - 1, 1) & \text{if} \quad X_n^\star \geq 1, \\ (\eta_V, 0) & \text{if} \quad X_n^\star = 0. \end{cases} \tag{3}$$

Let $X^\star$ and $I^\star$ be the limitting random variables for $X_n^\star$ and $I_n^\star$ as $n \to \infty$, so $(X_n^\star, I_n^\star) \xrightarrow{d} (X^\star, I^\star)$. As it was noted in [8], the number of customers $\widehat{X}$ "seen" by an arbitrary service completion is just the random variable $X^\star$ conditioned by $I^\star = 1$. Using this conditioning it can be proved (see [8] or [21]) that for the queueing system with single server vacations the following representation is true:

$$\begin{aligned} \widehat{P}^{SV}(z) &\stackrel{def}{=} \mathbf{E}\{z^{\widehat{X}}\} = \mathbf{E}\{z^{X^\star} \,|\, I^\star = 1\} = \frac{(1-\rho)(1-z)f(\lambda - \lambda z)}{f(\lambda - \lambda z) - z} \cdot \frac{1 - g(\lambda - \lambda z) + (1-z)f(\lambda)}{[f(\lambda) + \lambda \mathbf{E}(V)](1-z)} \\ &= \widehat{P}(z) \cdot \frac{1 - g(\lambda - \lambda z) + (1-z)f(\lambda)}{[f(\lambda) + \lambda \mathbf{E}(V)](1-z)}, \end{aligned} \tag{4}$$

where $\rho = \lambda \int_0^\infty t dF(t) < 1$ is the occupation rate in the "usual" $M/G/1$-type system (without vacations) and $\mathbf{E}(V)$ stands for the mean vacation duration.

The first factor on the right side of (4) is the PGF of the number of customers at a service completion epoch in the "usual" $M/G/1$-type queue (a well-known Pollaczek-Khintchine formula). In consequence, $\widehat{X}$ can be represented as a sum of two independent random variables, one of them is the number of customers at the service completion epoch in the corresponding "usual" $M/G/1$ system without vacations. This conclusion is known as the so called decomposition property. As it turns out, such a property can be proved not only for a model with single vacation.

For the $M/G/1$-type queue operating under multiple vacation policy we have, similarly,

$$\widehat{P}^{MV}(z) = \widehat{P}(z) \cdot \frac{1 - g(\lambda - \lambda z)}{\lambda \mathbf{E}(V)(1-z)}. \tag{5}$$

Moreover, in 1986 in [10] Fuhrmann and Cooper showed that the stationary distribution of the queue size (number of jobs) in the $M/G/1$-type queueing system with generalized server

vacation is a convolution of the distribution functions of two independent positive random variables. One of them is the stationary queue-size distribution of the number of jobs in the ordinary $M/G/1$-type queueing system (without server vacations). Some other results for the stationary state of queueing models with server vacations can be found e.g. in [4] and [5].

Let us investigate now a finite-buffer $M/G/1/K$-type queue with Poisson arrivals with rate λ and generally distributed service times with CDF $F(\cdot)$ with LST $f(\cdot)$. The system size is assumed to be K, i.e. we have a buffer with $K-1$ places and one place in service facility. Assume, moreover, that the multiple vacation policy is implemented in that one (single) server vacation has general distribution with CDF $G(\cdot)$ with LST $g(\cdot)$. Denote by $X(t)$ the number of packets present in the system at time t. As one can note, in the literature most of results obtained for different queueing systems relate to the stationary state of the system (as $t \to \infty$). However, transient analysis (for fixed time t) is often desired, e.g. just after the start of the system operation or after application of a new control mechanism, or in the case of low traffic intensity (in this case the convergence rate of transient characteristics to the stationary may be slow).

Introduce the the following notation:

$$\widehat{Q}_n(s,m) \stackrel{def}{=} \int_0^\infty e^{-st}\mathbf{P}\{X(t)=m\,|\,X(0)=K-n\}dt, \quad s>0,\, 0\le m,n\le K. \tag{6}$$

It can be proved (see Kempa, 2015 [16]) that the following system of equations is true:

$$\sum_{i=-1}^{n} a_{i+1}(s)\widehat{Q}_{n-i}(s,m) - \widehat{Q}_n(s,m) = \psi_n(s,m), \quad 0\le n\le K-1, \tag{7}$$

and

$$\widehat{Q}_K(s,m) = \sum_{i=1}^{K-1} b_i(s)\widehat{Q}_{K-i}(s,m) + \widehat{Q}_0(s,m)\sum_{i=K}^{\infty} b_i(s) + d(s,m) + \frac{\delta_{m,0}}{s+\lambda}, \tag{8}$$

where

$$a_k(s) \stackrel{def}{=} \int_0^\infty e^{-(s+\lambda)y}\frac{(\lambda y)^k}{k!}dF(y), \tag{9}$$

$$b_k(s) \stackrel{def}{=} \big(1-g(s+\lambda)\big)^{-1}\int_0^\infty e^{-(s+\lambda)y}\frac{(\lambda y)^k}{k!}dG(y), \tag{10}$$

$$d(s,m) \stackrel{def}{=} \big(1-g(s+\lambda)\big)^{-1}\Big(I\{1\le m\le K-1\}\varphi_{G,m}(s) + \delta_{m,K}\sum_{i=K-1}^{\infty}\varphi_{G,i+1}\Big)(s), \tag{11}$$

$$\psi_n(s,m) \stackrel{def}{=} a_{n+1}(s)\widehat{Q}_0(s,m) - \widehat{Q}_1(s,m)\sum_{k=n+1}^{\infty} a_k(s) - h_{K-n}(s,m), \tag{12}$$

$$h_k(s,m) \stackrel{def}{=} I\{k\le m\le K-1\}\varphi_{F,m-k}(s) + \delta_{m,K}\sum_{i=K-k}^{\infty}\varphi_{F,i}(s), \tag{13}$$

where for arbitrary CDF $H(\cdot)$

$$\varphi_{H,k}(s) \stackrel{def}{=} \int_0^\infty e^{-(s+\lambda)t}\frac{(\lambda t)^k}{k!}[1-H(t)]dt. \tag{14}$$

There is proved in Korolyuk, 1975 [20] that each solution of the infinite-sized system of type (7), written for $n \geq 0$, can be stated as

$$\widehat{Q}_n(s,m) = C(s,m)R_{n+1}(s) + \sum_{k=0}^{n} R_{n-k}(s)\psi_k(s,m), \quad n \geq 0, \tag{15}$$

where $C(s,m)$ is independent on n and successive terms of the sequence $(R_k(s))$ (called Korolyuk's potential) can be computed as follows:

$$R_k(s) = \lim_{z\to 0} \frac{1}{k!} \frac{\partial^k Q(s,z)}{\partial z}, \tag{16}$$

where $Q(s,z) \stackrel{def}{=} \frac{z}{A(s,z)-z}$ and

$$A(s,z) \stackrel{def}{=} \sum_{k=0}^{\infty} z^k a_k(s), \quad s > 0, |z| < 1. \tag{17}$$

As it turns out (see [20]) the sequence $\big(R_k(s)\big)$ can also be defined recursively as

$$R_0(s) = 0, \quad R_1(s) = \frac{1}{a_0(s)}, \quad R_{k+1}(s) = R_1(s)\Big(R_k(s) - \sum_{i=0}^{k} a_{i+1}(s)R_{k-i}(s)\Big), \tag{18}$$

where $k \geq 1$.

Treating the last equation (8) as a specific-type boundary condition (see [16]), we can find the representation for $C(s,m)$ explicitly and, utilizing (15), we obtain the following formula for the LT of the conditional transient queue-size distribution in the considered model:

$$\begin{aligned} &\int_0^{\infty} e^{-st}\mathbf{P}\{X(t) = m \,|\, X(0) = n\}dt = \Phi_{K-n}(s,m) \\ &+ \frac{\sum_{i=1}^{K-1} b_{K-i}(s)\Phi_i(s,m) + d(s,m) + \delta_{m,0}(s+\lambda)^{-1} - \Phi_K(s,m)}{\Theta_K(s) - \sum_{i=1}^{K-1} b_{K-i}(s)\Theta_i(s) - \sum_{i=K}^{\infty} b_i(s)}\Theta_{K-n}(s), \end{aligned} \tag{19}$$

where $0 \leq m,n \leq K$ and

$$\Theta_n(s) \stackrel{def}{=} a_0(s)R_{n+1}(s) + \sum_{k=0}^{n} R_{n-k}(s)\Big(a_{k+1}(s) - f^{-1}(s)\sum_{i=k+1}^{\infty} a_i(s)\Big), \tag{20}$$

$$\Phi_n(s,m) \stackrel{def}{=} \sum_{k=0}^{n} R_{n-k}(s)\Big(h_K(s,m)f^{-1}(s)\sum_{i=k+1}^{\infty} a_i(s) - h_{K-k}(s,m)\Big). \tag{21}$$

Transient results for main stochastic characteristics of queueing models with single/multiple vacation policies can also be found in [11], [12], [13] and [17].

4 Threshold-type N-policy

In this section we will present main analytical results for some exemplary queueing models with the mechanism of N-policy. Yadin and Naor in 1963 [27] were the first who studied the $M/G/1$-type queue with this type of threshold control policy.

Let us consider, firstly, the $M/G/1$-type infinite-buffer model. We will apply the mean-value approach to find the representation for the mean waiting time $\mathbf{E}(W)$ in the stationary state of the system. Assume that jobs arrive according to a Poisson process with intensity λ and that the service time of individual job is generally distributed with finite mean $\mathbf{E}(B)$ and the second moment $\mathbf{E}(B^2)$. Moreover, the first service after reaching the threshold level N is preceded by a generally distributed setup time with finite two first moments $\mathbf{E}(S)$ and $\mathbf{E}(S^2)$.

Let us note that we have (see e.g. Adan and Resing, 2001 [1])

$$\begin{aligned}\mathbf{E}(W) =& \mathbf{E}(X_Q)\mathbf{E}(B) + \rho\mathbf{E}(B_R) + \sum_{i=1}^{N} \mathbf{P}\{\text{Number of arriving job is i}\}\Big[\frac{N-i}{\lambda} + \mathbf{E}(S)\Big] \\ &+ \mathbf{P}\{\text{Server is during setup time on arrival}\}\mathbf{E}(S_R), \end{aligned} \tag{22}$$

where X_Q, B_R and S_R stand for the number of jobs waiting in the queue (buffer), residual service time and residual setup time, respectively. Obviously $\rho = \lambda\mathbf{E}(B) < 1$.

Observe that the probability that a job enters the system during the accumulation period (when the processing is suspended) equals to $1-\rho$. Besides, the accumulation period consists of N successive interarrival times plus the following setup time. In consequence, the probability that the arriving job is the ith one during the accumulation period is equal to the product of $1-\rho$ and the proportion of means: of one interarrival time and the duration of the whole accumulation period. So, we have

$$\mathbf{P}\{\text{Number of arriving job is i}\} = (1-\rho)\frac{\lambda^{-1}}{N\lambda^{-1} + \mathbf{E}(S)}, \tag{23}$$

where $i \in \{1, ..., N\}$.

Similarly, we have

$$\mathbf{P}\{\text{Server is during setup time on arrival}\} = (1-\rho)\frac{\mathbf{E}(S)}{N\lambda^{-1} + \mathbf{E}(S)}. \tag{24}$$

Substituting these two relationships into (22), after simplification, we obtain

$$\mathbf{E}(W) = \frac{\rho\mathbf{E}(B_R)}{1-\rho} + \frac{N\lambda^{-1}}{N\lambda^{-1} + \mathbf{E}(S)}\Big[\frac{N-1}{2\lambda} + \mathbf{E}(S)\Big] + \frac{\mathbf{E}(S)}{N\lambda^{-1} + \mathbf{E}(S)}\mathbf{E}(S_R). \tag{25}$$

Now, let us deal with the finite-buffer $M/G/1/K$-type queueing model operating under the N-policy, with Poisson arrival stream with rate λ and generally distributed service times with CDF $F(\cdot)$ with LST $f(\cdot)$ in that, as previously, the maximum system state is K. As usually, the service process is organized according to the FIFO discipline. We are interested in the transient analysis of the queue-size distribution (see Kempa and Kurzyk [18] for more detailed analysis). Observe that the operation of the system can be observed on successive buffer loading periods BL_1, BL_2, ... followed by busy periods BP_1, BP_2, ..., during which the system becomes idle. From the memoryless property of interarrival times follows that initial and completion epochs of successive busy periods are Markov moments. Hence (BL_k) and (BP_k), $k \geq 1$, are sequences of independent random variables with the same CDFs in each sequence separately. For simplicity we identify BL_k, BP_k, $k \geq 1$, with their durations.

Let $X(t)$ be the number of jobs present in the system at time t, including the one being served at this moment (if any). Investigate the queue-size distribution during the first buffer loading period BL_1 starting at time $t = 0$. We get

$$\mathbf{P}\{\big(X(t)=m\big)\cap\big(t\in BL_1\big)\} = I\{0\le m\le N-1\}\frac{(\lambda t)^m}{m!}e^{-\lambda t}, \tag{26}$$

where $t \ge 0$. Introducing the following notation:

$$\widetilde{q}^{BL}(s,m) \stackrel{def}{=} \int_0^\infty e^{-st}\mathbf{P}\{\big(X(t)=m\big)\cap\big(t\in BL_1\big)\}dt, \tag{27}$$

where $s > 0$, we have from (26)

$$\widetilde{q}^{BL}(s,m) = I\{0\le m\le N-1\}\int_0^\infty e^{-(s+\lambda)t}\frac{(\lambda t)^m}{m!}dt = I\{0\le m\le N-1\}\frac{\lambda^m}{(\lambda+s)^{m+1}}. \tag{28}$$

Obviously, each buffer loading period duration has the N-Erlang distribution with parameter λ, so we obtain

$$\widetilde{g}^{BL}(s) = \int_0^\infty e^{-st}dG^{BL}(t) \stackrel{def}{=} \int_0^\infty e^{-st}d\mathbf{P}\{BL_k < t\} = \int_0^\infty e^{-(s+\lambda)t}\frac{\lambda^N}{(N-1)!}t^{N-1}dt = \Big(\frac{\lambda}{\lambda+s}\Big)^N. \tag{29}$$

Let us consider now the evolution of the system during a busy period. Assume temporarily that the service process can be started with arbitrary possible level n of buffer state, where $1 \le n \le K$ (not only at $n = N$). Let $Q_n^{BP}(t,m)$ be transient conditional queue-size distribution at time $t \in BP_1$, namely

$$Q_n^{BP}(t,m) \stackrel{def}{=} \mathbf{P}\{\big(X(t)=m\big)\cap\big(t\in BP_1\big)|X(0)=n\}, \tag{30}$$

where $t > 0$ and $1 \le m,n \le K$. For simplicity let us assume that BP_1 begins at time $t = 0$. Since successive departure epochs are Markov moments then, applying the total probability law with respect to the first departure moment after $t = 0$, we obtain the following system of integral equations:

$$\begin{aligned} Q_1^{BP}(t,m) &= \sum_{i=1}^{K-2}\int_0^t \frac{(\lambda x)^i}{i!}e^{-\lambda x}Q_i^{BP}(t-x,m)dF(x) + \sum_{i=K-1}^{\infty}\int_0^t \frac{(\lambda x)^i}{i!}e^{-\lambda x}Q_{K-1}^{BP}(t-x,m)dF(x) \\ &+ \overline{F}(t)e^{-\lambda t}\Big[I\{1\le m\le K-1\}\frac{(\lambda t)^{m-1}}{(m-1)!} + I\{m=K\}\sum_{i=K-1}^{\infty}\frac{(\lambda t)^i}{i!}\Big], \end{aligned} \tag{31}$$

and, for $2 \le n \le K$,

$$\begin{aligned} Q_n^{BP}(t,m) &= \sum_{i=0}^{K-n-1}\int_0^t \frac{(\lambda x)^i}{i!}e^{-\lambda x}Q_{n+i-1}^{BP}(t-x,m)dF(x) + \sum_{i=K-n}^{\infty}\int_0^t \frac{(\lambda x)^i}{i!}e^{-\lambda x}Q_{K-1}^{BP}(t-x,m)dF(x) \\ &+ \overline{F}(t)e^{-\lambda t}\Big[I\{n\le m\le K-1\}\frac{(\lambda t)^{m-n}}{(m-n)!} + I\{m=K\}\sum_{i=K-n}^{\infty}\frac{(\lambda t)^i}{i!}\Big], \end{aligned} \tag{32}$$

where $\overline{F}(t) \stackrel{def}{=} 1 - F(t)$.

Introducing the following nomenclature:

$$\widetilde{q}_n^{BP}(s,m) \stackrel{def}{=} \int_0^\infty e^{-st} Q_{N-n}^{BP}(t,m)dt, \tag{33}$$

$$a_n(s) \stackrel{def}{=} \int_0^\infty e^{-(s+\lambda)t} \frac{(\lambda t)^n}{n!} dF(t), \tag{34}$$

$$\theta_n(s,m) \stackrel{def}{=} \int_0^\infty e^{-(\lambda+s)t} \overline{F}(t) \Big[I\{n \le m \le K-1\} \frac{(\lambda t)^{m-n}}{(m-n)!} + I\{m=K\} \sum_{i=K-n}^{\infty} \frac{(\lambda t)^i}{i!} \Big] dt, \tag{35}$$

where $s > 0$, we can transform the equations of the system (31)–(32) to the following ones:

$$\widetilde{q}_{K-1}^{BP}(s,m) = \sum_{i=1}^{K-2} a_i(s) \widetilde{q}_{K-i}^{BP}(s,m) + \widetilde{q}_1^{BP}(s,m) \sum_{i=K-1}^{\infty} a_i(s) + \theta_1(s,m), \tag{36}$$

$$\sum_{k=-1}^{n} a_{k+1}(s) \widetilde{q}_{n-k}^{BP}(s,m) - \widetilde{q}_n^{BP}(s,m) = \phi_n(s,m), \tag{37}$$

where $0 \le n \le K-2$ and

$$\phi_n(s,m) \stackrel{def}{=} a_{n+1}(s) \widetilde{q}_0^{BP}(s,m) - \widetilde{q}_1^{BP}(s,m) \sum_{i=n+1}^{\infty} a_i(s) - \theta_{K-n}(s,m). \tag{38}$$

Using the similar method as described in Section 3, we get the representation for $\widetilde{q}_0^{BP}$ (we need only this one) in the form

$$\widetilde{q}_0^{BP}(s,m) = \Pi_1(s,m) \Pi_2^{-1}(s), \tag{39}$$

where

$$\Pi_1(s,m) \stackrel{def}{=} \sum_{i=1}^{K-2} a_i(s) \eta_{K-i}(s,m) - \frac{\theta_K(s,m)}{f(s)} \sum_{i=K-1}^{\infty} a_i(s) + \theta_1(s,m) - \eta_{K-1}(s,m), \tag{40}$$

$$\Pi_2(s) \stackrel{def}{=} \gamma_{K-1}(s) - \sum_{i=1}^{K-2} a_i(s) \gamma_{K-i}(s) - \frac{1}{f(s)} \sum_{i=K-1}^{\infty} a_i(s), \tag{41}$$

$$\gamma_n(s) \stackrel{def}{=} a_0(s) R_{n+1}(s) + \sum_{i=0}^{n} R_{n-i}(s) \Big[a_{i+1}(s) - \frac{1}{f(s)} \sum_{j=i+1}^{\infty} a_j(s) \Big] \tag{42}$$

and

$$\eta_n(s,m) \stackrel{def}{=} \sum_{i=0}^{n} R_{n-i}(s) \Big[\frac{\theta_K(s,m)}{f(s)} \sum_{j=i+1}^{\infty} a_j(s) - \theta_{K-i}(s,m) \Big] \tag{43}$$

and the functional sequence $(R_k(s))$ is defined in (18).

Denoting now by $g_n^{BP}(\cdot)$ the LST of CDF of busy period duration in the system that starts working with $1 \le n \le K$ packets present in the buffer queue, we can formulate a system of equations for $g_1^{BP}(s), ..., g_K^{BP}(s)$ similar to (36)–(37) and find, in particular, the following representation:

$$\widetilde{g}^{BP}(s) \stackrel{def}{=} \widetilde{g}_N^{BP}(s) = \gamma_{K-N}(s) \widetilde{\Pi}_1(s) \Pi_2^{-1}(s) + \widetilde{\eta}_{K-N}(s), \quad n \ge 0, \tag{44}$$

where

$$\widetilde{\eta}_n(s) \stackrel{def}{=} \sum_{i=0}^{n} R_{n-i}(s)\Big[a_0^{-1}(s)\big(1+a_0^{-1}(s)\big)^{-1}\widetilde{\theta}_K(s) - \widetilde{\theta}_{K-i}(s)\Big], \tag{45}$$

$$\widetilde{\theta}_n(s) \stackrel{def}{=} \begin{cases} f(\lambda+s), & n=1, \\ 0, & n \geq 2, \end{cases} \tag{46}$$

and

$$\widetilde{\Pi}_1(s) \stackrel{def}{=} \sum_{i=1}^{K-2} a_i(s)\widetilde{\eta}_{K-i}(s) + \widetilde{\eta}_1(s)\sum_{i=K-1}^{\infty} a_i(s) + \widetilde{\theta}_1(s) - \widetilde{\eta}_{K-1}(s) \tag{47}$$

and $\Pi_2(s)$ and $\gamma_n(s)$ were defined in (41) and (42), respectively.

Furthermore, the law of total probability gives

$$\mathbf{P}\{X(t)=m\} = \sum_{i=1}^{\infty}\Big(\mathbf{P}\{\big(X(t)=m\big)\cap\big(t\in BL_i\big)\} + \mathbf{P}\{\big(X(t)=m\big)\cap\big(t\in BP_i\big)\}\Big). \tag{48}$$

Since BL_i and BP_i, $i \geq 1$, are independent and have identical distributions (in each sequence separately), we get

(49)

$$\mathbf{P}\{\big(X(t)=m\big)\cap\big(t\in BL_i\big)\} = \int_0^t \mathbf{P}\{\big(X(t-y)=m\big)\cap\big(t-y\in BL_1\big)\}d\big(G^{BL}*G^{BP}\big)^{(i-1)*}(y),$$

(50)

$$\mathbf{P}\{\big(X(t)=m\big)\cap\big(t\in BP_i\big)\} = \int_0^t \mathbf{P}\{\big(X(t-y)=m\big)\cap\big(t-y\in BP_1\big)\}d\big[\big(G^{BL}\big)^{i*}*\big(G^{BP}\big)^{(i-1)*}\big](y).$$

Taking LTs of (49)–(50), referring to (48), we obtain now, as main result, the formula for the LT of the queue-size distribution of in the $M/G/1/K$-type system with threshold-type N-policy:

$$\int_0^{\infty} e^{-st}\mathbf{P}\{X(t)=m\}dt = \frac{\widetilde{q}^{BL}(s,m) + \widetilde{g}^{BL}(s)\widetilde{q}_0^{BP}(s,m)}{1-\widetilde{g}^{BP}(s)\widetilde{g}^{BL}(s)}. \tag{51}$$

In [15] a similar model with infinite buffer and batch arrivals is studied in transient state, where additionally setup times are implemented.

5 Active Queue Management - preventional "filtering" of arriving customers

A special-type group of solutions which may result in limitting the access to the service station is associated with the so called Active Queue Management (AQM). The main idea of AQM is in the intervention in the process of qualification for service the arriving customers. In other words, an AQM-type approach allows for dropping the arriving job even when there is a place in the accumulating buffer, differently than in the classical Tail Drop discipline (the accumulation of the buffer is undisturbed until the buffer becomes saturated; then all the arriving customers are being lost). Obviously, such a mechanism may also be called "balking" and may be considered as a type of customer "unpatience", which is known in queueing theory. However, AQM is originally motivated by networking studies and assumes that the "roles" are reversed here. It is not the customer who decides whether to join the system, but the system "manager" that have much

more knowledge about the system operation now and in the past. AQM algorithms dedicated for IP routers have been investigated widely since the paper [9] by Floyd and Jacobson was published in 1993. Usually an arriving packet is rejected (dropped) with probability depending on the actual queue state (as in the classical "balking" discipline: from the point of view of the arriving customer, typically, the only knowledge about the system is the actual queue size at the arriving moment). However, the probability of dropping may also depend on other, even very complex, stochastic characteristics, like the "history" of customer losses or the frequency of empty buffer and full buffer occurrences. In the Internet routers AQM-type packet dropping is used mainly to prevent buffer queues from growing too long, keeping high link utilizations simultaneously. Other goals are stable and predictable buffer queues with low variances of queue sizes and, moreover, desynchronization of TCP sources (see e.g. Chydziński and Chróst, 2011 [6]). Indeed, in the classical Tail Drop approach, in the case of buffer overflow, according to the TCP protocol, all the sources (hosts sending packets directed ot the node) may reduce the intensity of sending the packets (synchronization).

Let us assume the $M/M/1/K$-type finite-buffer queueing system in which the entering job which finds i jobs present in the system is dropped by the "manager" with probability d_i, $i = 0,1,...,K$, $d_K = 1$ (the so called "dropping function"). If λ is the Poisson arrival rate and μ^{-1} denotes mean service time, the system of equilibrium equations for the steady-state queue-size probabilities $p_k \stackrel{def}{=} \mathbf{P}\{X = k\}$, where X is the number of packets present in the system in the stationary state, has the following form (see Kempa, 2011 [14]):

$$\left\{\begin{array}{rcl} \lambda(1-d_0)p_0 & = & p_1, \\ \big[\lambda(1-d_k)+\mu\big]p_k & = & \lambda(1-d_{k-1})p_{k-1}+\mu p_{k+1}, \quad 1 \le k \le K-1, \\ \mu p_K & = & \lambda(1-d_{K-1})p_{K-1}. \end{array}\right. \tag{52}$$

Hence we get

$$p_k = \frac{\lambda^k}{\mu^k}\prod_{i=0}^{k-1}(1-d_i)p_0, \;\; k = 1,2,...K. \tag{53}$$

From the normalization condition

$$\sum_{i=0}^{K} p_i = 1$$

we find

$$p_0 = \Big(1+\sum_{k=1}^{K}\rho^k\prod_{i=0}^{k-1}(1-d_i)\Big)^{-1}, \tag{54}$$

where $\rho = \frac{\lambda}{\mu}$ is the occupation rate of the system, and hence, finally,

$$p_k = \frac{\rho^k\prod_{i=0}^{k-1}(1-d_i)}{1+\sum_{j=1}^{K}\rho^j\prod_{i=0}^{j-1}(1-d_i)}, \qquad k = 0,1,...,K. \tag{55}$$

Obviously, if $d_i \equiv 0$, we get the well-known solution for the "classical" $M/M/1/K$-type queue.

Let us note that after the implementation of the dropping function we get not a usual "thinning" of the Poisson arrival process: starting with n jobs present, the first arriving one is dropped

with probability d_n but the second one either with probability d_{n+1} (if the previous one was qualified for service), or with probability d_n (otherwise).

Now let us consider a more general queueing system in which the service time of individual customer is generally distributed with a CDF $F(\cdot)$ ($M/G/1/K$-type model). Denote by X_n the number of jobs present in the system after departing the nth job, hence $0 \leq X_n \leq K-1$. We are interested in the stationary distribution of X_n, namely $\widehat{p}_k = \lim_{n\to\infty}\{X_n = k\}$, $k = 0,...,K-1$. Obviously, (X_n) is an ergodic Markov chain, so probabilities $\widehat{p}_k$ satisfy the following system of equations (see [6]):

$$\widehat{p}_j = \sum_{i=0}^{K-1} \widehat{p}_i p_{i,j}, \quad 0 \leq j \leq K-1, \tag{56}$$

and $\sum_{j=0}^{K-1} \widehat{p}_j = 1$, where

$$p_{i,j} = \mathbf{P}\{X_{n+1} = j \,|\, X_n = i\}, \;\; 0 \leq i,j \leq K-1, \tag{57}$$

denote one-step transition probabilities of the Markov chain (X_n).

The following representation is true:

$$p_{i,j} = \begin{cases} q_{1,j} & \text{if} \quad i = 0, 0 \leq j \leq K-1, \\ q_{i,j-i+1} & \text{if} \quad 1 \leq i \leq K-1, \\ 0 & \text{otherwise}, \end{cases} \tag{58}$$

where

$$q_{i,j} = \int_0^\infty A_{i,j}(x)dF(x) \tag{59}$$

and $A_{i,j}(x)$ denotes here the probability that up to the fixed time x exactly j jobs are "qualified" for service on condition that initially (at time $t = 0$) we have exactly i jobs present. It is shown in [6] that

$$\int_0^\infty e^{-sx} A_{i,j}(x)dx = \frac{\lambda^j \prod_{k=0}^{j-1}(1-d_{i+k})}{\prod_{k=0}^{j}[s+\lambda(1-d_{i+k})]}, \quad s > 0, \tag{60}$$

where $i, j \geq 0$ and we accept the agreement that $\prod_{k=0}^{-1} = 1$.

In practice, in the next step, $A_{i,j}(x)$ (and hence $q_{i,j}$) are usually found inverting the right side of (60) which is not difficult.

An interesting study on AQM-type model with continuous job volumes and processor sharing can be found in [26].

REFERENCES:

[1] Adan, I., Resing, J., Queueing theory. Eindhoven University of Technology (2001).

[2] Bocharov, P.P., D'Apice, C., Pechinkin, A.V., Salerno, S., Queueing theory. VSP (2004).

[3] Borovkov, A.A., Probabilistic processes in the queueing theory. Nauka (1972).

[4] Choudhury, G., An $M^X/G/1$ queueing system with a setup period and a vacation period. Queueing Syst. **36** (2000), 23–38.

[5] Choudhury, G., A batch arrival queue with a vacation time under single vacation policy. Comp. Oper. Res. **29** (14) (2002), 1941–1955.

[6] Chydziński, A., Chróst, Ł., Analysis of AQM queues with queue size based packet dropping. Int. J. Appl. Math. Comput. Sci. **21** (3) (2011), 567–577.

[7] Cohen, J.W., The single server queue. North-Holland (1982).

[8] Doshi, B.T., Queueing systems with vacations - a survey, Queueing Syst. **1** (1) (1986), 19–66.

[9] Floyd, S., Jacobson, V., Random early detection gateways for congestion avoidance. IEEE/ACM Trans. Netw. **1** (4) (1993), 397–413.

[10] Fuhrmann, S.W., Cooper, R.B., Stochastic decompositions in the $M/G/1$ queue with generalized vacations. Oper. Res. **33** (5) (1985), 1117–1129.

[11] Kempa, W.M., $GI/G/1/\infty$ batch arrival queueing system with a single exponential vacation. Math. Methods Oper. Res. **69** (1) (2009), 81–97.

[12] Kempa, W.M., Some new results for departure process in the $M^X/G/1$ queueing system with a single vacation and exhaustive service. Stoch. Anal. Appl. **28** (1) (2010), 26–43.

[13] Kempa, W.M., Analysis of departure process in batch arrival queue with multiple vacations and exhaustive service. Commun. Stat. - Theory Methods **40** (16) (2011), 2856–2865.

[14] Kempa, W.M., On main characteristics of the $M/M/1/N$ queue with single and batch arrivals and the queue size controlled by AQM algorithms. Kybernetika **47** (6) (2011), 930–943.

[15] Kempa, W.M., On transient queue-size distribution in the batch arrival system with the N-policy and setup times. Math. Commun. **17** (1) (2012), 285–302.

[16] Kempa, W.M., Queue-size distribution in energy-saving model based on multiple vacation policy. In: Current Trends in Analysis and Applications. Proceeding of the 9th ISAAC Congress 2015, 733–740.

[17] Kempa, W.M., Transient workload distribution in the $M/G/1$ finite-buffer queue with single and multiple vacations. Ann. Oper. Res. **239** (2) (2016), 381–400.

[18] Kempa, W.M., Kurzyk, D., Queue-size distribution in a WSN node with power saving algorithm based on N-Policy (submitted).

[19] Kleinrock, L., Queueing systems. Volumes 1 and 2. John Wiley & Sons (1975).

[20] Korolyuk, V.S., Boundary-value problems for compound Poisson processes. Naukova Dumka (1975).

[21] Levy, Y., Yechiali, U., Utilization of idle time in an $M/G/1$ queueing. Management Sci. **22** (2) (1975), 202–211.

[22] Mancuso, V., Alouf, S., Analysis of power saving with continuous connectivity. Comput. Netw. **56** (2012), 2481–2493.

[23] Miller, L.W., Alternating priorities in multi-class queue. Ph.D. dissertation, Cornell University, Ithaca, N.Y. (1964).

[24] Takacs, L., Introduction to the theory of queues. Oxford University Press (1962).

[25] Takagi, H., Queueing analysis. Volume 1: Vacation and priority systems. Volume 2: Finite systems. North-Holland (1991–Vol. 1, 1993–Vol. 2).

[26] Tikhonenko, O., Kempa, W.M., Queueing system with processor sharing and limited memory under control of the AQM mechanism. Autom. Remote Control **76** (10) (2015), 1784–1796.

[27] Yadin, M., Naor, P., Queueing systems with a removable service station. Oper. Res. Quarterly **14** (1963), 393–405.

Wojciech Kempa
Silesian University of Technology, Faculty of Applied Mathematics, Institute of Mathematics,
Gliwice, Poland
E-mail:wojciech.kempa@polsl.pl