IEICE TRANS. COMMUN., VOL.E102–B, NO.5 MAY 2019

INVITED PAPER Special Section on European ICT R&D Project Activities on Broadband Access Technologies in Conjunction with Main Topics of 2016/2017 IEICE ICT Forum

# Overflows in Multiservice Systems

**Mariusz GŁĄBOWSKI**[†a)], ***Senior Member***, **Damian KMIECIK**[†], ***Nonmember***, ***and*** **Maciej STASIAK**[†], ***Member***

**SUMMARY** This article presents a universal and versatile model of multiservice overflow systems based on Hayward's concept. The model can be used to analyze modern telecommunications and computer networks, mobile networks in particular. The advantage of the proposed approach lies in its ability to analyze overflow systems with elastic and adaptive traffic, systems with distributed resources and systems with non-full-availability in primary and secondary resources.

***key words:*** *multiservice traffic, overflow system, elastic traffic, adaptive traffic, distributed systems*

## 1. Introduction

Traffic overflow is one of the best known and at the same time one of the oldest traffic distribution control mechanisms in networks. This mechanism is activated when the resources designated to service a given call stream (the so-called primary resources–PR) are fully occupied and admission of a new call is blocked. When this is the case, this call can be redirected to other, alternative, resources (the so-called secondary resources–SR) that are capable of servicing the call.

The application of the traffic overflow mechanism in telecommunications networks became possible after the introduction of crossbar switching networks [1] to hierarchical telephone networks in the 1950s. The traffic overflow mechanism has also been widely used in packet networks [2]–[4] to optimize traffic distribution by way of equalizing the load of individual connection paths [5]. The traffic overflow mechanism has also found application in IT systems, initially to help organize files [6], and then, in recent years, to increase the efficiency of data centers and cloud solutions [7]. With the case of data centers, the two basic premises for the application of traffic overflow are: equalizing loads [8], [9] and energy consumption limitations [10], [11]. In optical networks, the traffic overflow mechanism is one of the key mechanisms, considered as early as the dimensioning stage for these networks (e.g. in Optical Burst Switching Networks) [12], [13]. Traffic overflow is also widely used in access networks, primarily in mobile wireless networks, [14]. The main reason behind its introduction is optimization of the use of resources offered by different technologies (e.g. 2G (GSM), 3G (UMTS) and 4G (LTE)) that operate within the same area and belong to one operator [14]–[16]. Currently, the overflow mechanism is also being considered as one of the methods for load equalization in self-organizing, self-optimizing and self-configuring networks [17]–[19]. The overflow mechanism has also found its application in smaller structures, such as switching networks, because an introduction of overflow links in the switches of the first stage leads to a significant decrease in the level of the internal blocking in a switching network [20]–[23].

The first mathematical overflow models for single-service networks are proposed in such classic, from today's perspective, works as for example, [1], [24]–[31]. The biggest practical importance for traffic engineering of network systems have the following works: [1], [25], [30]. In [1], [25], the Equivalent Random Traffic (ERT) method is developed, while in [30] a method, called Fredricks-Hayward's method (FH) is proposed that significantly simplifies the approach adopted in the ERT. Both methods are based on the two first moments of overflow traffic, the average value of traffic intensity and its variance (the so-called Riordan's formulas), determined on the basis of a two-dimensional Markov process in a system in which SR have infinite capacity [32]. In the ERT method, overflow traffic moments provide the basis for a determination and definition of equivalent resources (ER), i.e. single, fictitious PR such that traffic that overflows from them is identical with traffic that overflows from a number of real PR. On the basis of the ER it is then possible to determine the blocking probability for calls in an overflow system, and in SR in particular. In the method [30], in turn, Riordan's formulas provide a basis for a determination of the peakedness factor of the total traffic offered to SR, and in consequence the blocking probability in SR. The blocking probability can be determined on the basis of a modification to Erlang's formula (the so-called Hayward's formula), in which overflow traffic and the capacity of SR are divided by the peakedness factor.

From among a number of other single-service models of overflow systems, the following works are notable [26], [33], [34], in which an approximation of an overflow call stream by a Pascal stream is proposed. Then, in [3], [35], [36] overflow systems with queues are analyzed. In [3], [37], overflow traffic is described on the basis of the two and the three first moments of the IPP process (Interrupted Poisson Process) [38]. The problem of a development of analytical models of systems with traffic overflow to which Engset traffic streams (from a finite number of traffic sources) are offered, are discussed in [28], [29], [39], among others.

Works on multiservice traffic overflow were initiated

Manuscript received August 16, 2018.
Manuscript publicized November 22, 2018.

[†]The authors are with the Poznan University of Technology, Poland.
a) E-mail: mariusz.glabowski@put.poznan.pl

DOI: 10.1587/transcom.2018EUI0002

in the 1990s by a commercial implementation of the first multiservice telecommunications network – Integrated Services Digital Network (ISDN) [40]. These works, assisted by ever-growing new network technologies and standards, have been carried out ever since until now. [41], [42] propose a methodology to model multiservice overflow systems on the basis of the approach proposed in [30] for single-service systems. In the case of multiservice systems, PRs are modelled on the basis of a multiservice model of Full Availability Group (FAG), in which the occupancy distribution can be determined on the basis of a simple recurrent formula [43], [44] for Erlang traffic streams, and on the basis of [45], [46] for Engset, Erlang and Pascal traffic streams. Secondary resources (SR), in turn, are modelled on the basis of a modification of the FAG in which offered overflow traffic is divided by appropriate (corresponding) peakedness factors.

The model [41], [42] has provided a basis for a construction of a large number of advanced models of overflow systems. For example, in [47], [48], limited availability of secondary resources is taken into consideration, while in [49], systems with multiservice Engset traffic streams, generated by a finite number of traffic sources, are considered. A possibility to service elastic traffic is taken into consideration in [47], [50], while in [51] a model of an overflow system with adaptive traffic is developed.

Other notable multiservice models of overflow systems include, for example, [52], [53] which propose concepts for modelling traffic offered to SR on the basis of processes of the type: Markov-Modulated Poisson Process (MMPP). Then, in [54], [55], to describe such traffic the Batched Poisson Process (BPP) is used. [56], [57] propose SR models with offered overflow traffic with a required value of the peakedness factor. In [13], [58], [59], overflow systems are considered in which PR and SR have different service parameters (service time, bitrate). In [60], to describe overflow systems, a two-dimensional convolution algorithm is used, while in [61], an overflow system is approximated by an EIG model (Erlang Ideal Grading) [62]. Multiservice switching networks with overflow links are analyzed in [22], [23], [63].

The present article shows a number of different possibilities of modelling multiservice overflow systems on the basis on the methodology proposed in [41], [42]. Multiservice PR and SR models are presented as well as a method for a determination of the parameters of overflew traffic. The article presents different ways for modelling systems with distributed resources, elastic traffic and adaptive traffic. The article also shows a possibility of modelling multiservice PR and SR with non-full-availability. Until now, such systems have not been analyzed in the literature.

The article has been structured as follows. Section 2 presents briefly the ERT and the FH methods elaborated for modelling single-service overflow systems. Section 3 provides a description of multiservice traffic. In Sect. 4, the most general multiservice PR model, a method for a determination of the parameters of overflow traffic and the most general SR model are presented. Section 5 presents a number of the most representative multiservice models with traffic overflow. Section 6 compares the results of the analytical modelling with the results of the simulations for a number of selected structures of overflow systems. Section 7 concludes the article.

## 2. Traffic Overflow in Single-Service Systems

### 2.1 Parameters of Traffic Overflow

If Erlang traffic with the intensity $A$ Erl. is offered to PR with the capacity of $V$ Allocation Units (AUs), then traffic that overflows from such resources (overflow traffic) will be described by the so-called Riordan's formulas [1]:

$$R = A\,[E_V(A)] = A\left[\frac{A^V}{V!}\Big/\sum_{i=0}^{V}\frac{A^i}{i!}\right] \tag{1}$$

$$\sigma^2 = R\left(\frac{A}{V+R+1-A}+1-R\right) \tag{2}$$

where $R$ is the average value of overflow traffic, whereas $\sigma^2$ is the variance of overflow traffic. The symbol $E_V(A)$ defines the blocking probability in the Erlang model. In classic single-service systems, the notion of AU corresponds to and defines those resources that can service one single call, i.e. a link or a channel.

### 2.2 Equivalent Resources – ERT Method

If SR with the capacity $V$ AUs is offered non-Erlang traffic, characterized by the parameters $R$, $\sigma^2$, then a certain amount of equivalent resources are defined (ER). ER have the capacity $\Delta V^*$, while traffic offered to these resources is Erlang traffic with the intensity $A^*$ Erl. The parameters $A^*$, $\Delta V^*$ are chosen in such a way as to describe traffic that overflows from the ER by the parameters $R$, $\sigma^2$. Figure 1 shows, in a schematic way, the concept of the ER. The parameters $A^*$, $\Delta V^*$ can be determined on the basis of Equations (1) and (2) for known values of the parameters $R$, $\sigma^2$ [1], [64]. Having determined the values of the parameters $A^*$, $\Delta V^*$, it is possible to determine the blocking probability $E$ in SR on the basis of the Erlang model for a group with the capacity $(V+\Delta V^*)$ to which traffic with the intensity $A^*$ Erl is offered:

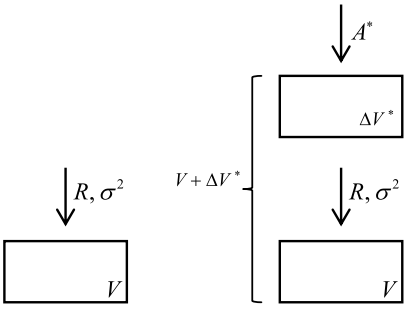


**Fig. 1** The concept of equivalent resources.

$$E = E_{V+\Delta V^*}(A^*) = \frac{(A^*)^{V+\Delta V^*}}{(V+\Delta V^*)!} \Big/ \sum_{i=0}^{V+\Delta V^*} \frac{(A^*)^i}{i!} \tag{3}$$

### 2.3 Hayward's Concept

If SR with the capacity $V$ AUs is offered non-Erlang traffic, characterized by the parameters $R$, $\sigma^2$, then the blocking probability $E$ can be determined on the basis of a modified Erlang's formula (called Hayward's formula) [30]:

$$E = E_{V/Z}(R/Z) \tag{4}$$

where $Z$ is the peakedness factor of offered traffic:

$$Z = \sigma^2/R \tag{5}$$

It should be stressed that both the ER model and the Hayward's concept are approximate models, though their accuracy is comparable with each other and still very high [33].

## 3. Traffic Overflow in Multiservice Systems

### 3.1 Multiservice Traffic

Modern networks are packet networks that service a number of call classes. In traffic theory, the assumption is that in multiservice systems a call is meant to be a packet stream related to a given service, often termed as the flow [65]–[67]. Following numerous research studies, it has been verified that call streams for a large number of services can be approximated by Poisson streams [67], [68]. A call of class $i$ is assigned a certain constant bitrate $c_i$, expressed in kbps and determined on the basis of the maximum bitrate or other methods known from the literature, such as [69]–[72]. The assumption in the article is that call streams are Poisson streams that can be described as follows:

- $M$ – the number of traffic classes offered to the system,
- $\lambda_i$ – intensity of a call stream of class $i$ ($0 < i \leq M$),
- $\mu_i$ – the average service intensity of a call of class $i$,
- $c_i$ – constant bitrate of a call of class $i$,
- $A_i$ – traffic intensity of traffic of class $i$ offered to the system:

$$A_i = \lambda_i/\mu_i \tag{6}$$

In multiservice systems it is possible to determine the value of the allocation unit (AU) on the basis of the bitrates $c_i$. The maximum value of AU can be determined as the GCD (Greatest Common Divisor) of all the bitrates of individual classes [73]:

$$c_{\mathrm{AU}} = \mathrm{GCD}(c_1, c_2, \ldots, c_M) \tag{7}$$

A choice of the allocation unit (AU) allows the so-called discretization of a multiservice system to be made, i.e. the expression of the demands of individual call classes as well as the capacity of the system in AUs:

$$t_i = \lceil c_i/c_{\mathrm{AU}} \rceil, \; V = \lfloor C/c_{\mathrm{AU}} \rfloor \tag{8}$$

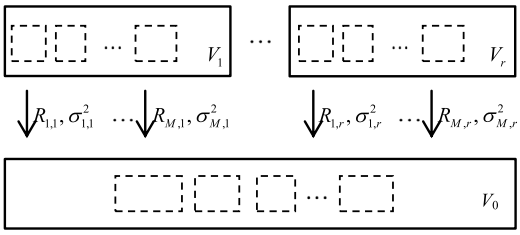


**Fig. 2** General diagram of multiservice overlow.

where $C$ is the bitrate of the resources of a system under consideration, and is expressed in kbps. Observe that after a discretization of a system, demands of particular call classes are dimensionless and are the multiples of AU. A further assumption is that the values of demands for individual classes are known.

### 3.2 General Scheme of Multiservice Traffic Overflow

Figure 2 shows a scheme of multiservice traffic overflow. The system of primary resources is composed of $r$ PRs. Each PR is offered a mixture of $M$ Erlang traffic classes. Traffic that cannot be serviced by PR overflows to SR. The notation in Fig. 2 is the following:

- $M$ – the number of traffic classes offered to each PR,
- $A_{i,j}$ – intensity of traffic of class $i$ offered to PR $j$ ($0 < j \leq r$),
- $t_i$ – the number of AUs demanded by a call of class $i$,
- $V_j$ – capacity of PR $j$,
- $R_{i,j}$ – intensity of traffic of class $i$ that overflows from PR $j$,
- $\sigma_{i,j}$ – variance of traffic of class $i$ that overflows from PR $j$,
- $V_0$ – capacity of SR.

In Fig. 2, all the possibilities of potential distribution of PR/SR are marked with dotted line. The problem of resource distribution will be addressed more thoroughly in Sect. 4.

## 4. Assumptions for Overflow System Modeling

### 4.1 Recurrence Resource Models

The assumption in the article is that PR will be modelled by multiservice FAG models [43]–[46], [54]–[56], [66], [74]–[80] in which the occupancy distribution in a non-state-dependent system is determined in a recurrent way. In the case of modelling of distributed PR, the LAG (Limited Availability Group) model is used [79], [81], [82]. This model also uses a recurrence form of a state-dependent multiservice system.

Another assumption in the article is that SR will be modelled, in line with Hayward's concept, on the basis of modified FAG/LAG models [41] in which the intensities $R_{i,j}$ of overflow traffic in individual classes are divided by the peakedness factors that correspond to these streams $Z_{i,j}$:

$$Z_{i,j} = \sigma^2_{i,j} / R_{i,j} \tag{9}$$

whereas the capacity $V_0$ for SR is divided by the so-called aggregate peakedness factor $Z_0$ [41]:

$$Z_0 = \sum_{j=1}^{r} \sum_{i=1}^{M} Z_{i,j} \frac{R_{i,j} t_{i,j}}{\sum_{k=1}^{r} \sum_{l=1}^{m} R_{k,l} t_{k,l}} \tag{10}$$

The aggregate peakedness factor is then determined on the basis of the weighted average of the peakedness factors of corresponding appropriate traffic classes that overflow from individual PR. The assumption is that the weight of each factor $Z_{i,j}$ is directly proportional to the number of AUs that is necessary for the complete service of overflow traffic $R_{i,j}$.

In its general form, the occupancy distribution in a multiservice system can be written in the following way:

$$\begin{cases} [P(n)]_v = \mathrm{RF_X}\{[P(n-t_i)]_v\} & \text{for } 0 \le n \le v \\ [P(n)]_v = 0 & \text{for } v < n < 0 \\ \sum_{n=0}^{v} [P(n)]_v = 1 \end{cases} \tag{11}$$

where $[P(n)]_v$ is the occupancy probability of $n$ AUs in the resources with the capacity $v$ AUs. The symbol $\mathrm{RF_X}(n)$ (Recurrent Function) depends on the structure of given resources and the implemented traffic management mechanisms. The parameter X = {PR, SR} defines the type of resources under consideration.

Once the distribution $[P(n)]_v$ has been determined it is possible to determine the blocking probability for calls of individual classes in particular PRs/SRs:

$$\left[E_{i,j}\right]_v = \sum_{n=v-t_i+1}^{v} [P(n)]_v \tag{12}$$

where $v = V_j$ for PR and $v = V_0/Z_0$ for SR.

An appropriate explicit form of the function $\mathrm{RF_X}(n)$ will be given in the section devoted to the dicussion on appropriate types of overflow systems.

### 4.2 Modelling of the Parameters of Overflow Traffic

The distribution (11) allows us to determine the blocking probability (12), and in consequence the average value of the intensity of traffic of class $i$ that overflows from $j$-th PR:

$$R_{i,j} = A_{i,j} \left[E_{i,j}\right]_{V_j} \tag{13}$$

Until now, no accurate method for a calculation of the variance $\sigma^2_{i,j}$ of overflow traffic of class $i$ that overflows from the $j$-th PR has been developed. In this article we present two approximate methods for a determination of the parameter $\sigma^2_{i,j}$. Both methods involve a decomposition of the multiservice PR that services $M$ traffic classes into $M$ fictitious single-service PRs with the capacity $V^*_{i,j}$. The knowledge of the parameters $A_{i,j}$ and $V^*_{i,j}$ makes it possible to use Formula (2) to determine the variance:

$$\sigma^2_{i,j} = R_{i,j} \left\{ \frac{A_{i,j}}{V^*_{i,j}/t_{i,j} + R_{i,j} + 1 - A_{i,j}} + 1 - R_{i,j} \right\} \tag{14}$$

where the quotient $V^*_{i,j}/t_{i,j}$ normalizes the system to a single-service model.

In the method [42], the capacities of fictitious resources $V^*_{i,j}$ are determined in relation to the criterion of blocking probability matching, i.e. on the basis of the Erlang model, with the additional assumption that the blocking probability in PR and in fictitious resources is identical:

$$\left[E_{i,j}\right]_{V_j} = E_{V^*_{i,j}}(A_{i,j}) \tag{15}$$

In the method [41], the capacity $V^*_{i,j}$ of fictitious resources is determined with respect to the criterion of carried traffic matching and is defined as the average part of the capacity $V_j$ that is not occupied by calls of those classes that are different from a given class $i$:

$$V^*_{i,j} = \frac{1}{t_i} \left[ V_j - \sum_{k=1, k\neq i}^{M} Y_{k,j} t_k \right] \tag{16}$$

where $Y_{k,j}$ is the average value of the intensity of traffic of class $k$ carried in the $j$-th PR:

$$Y_{k,j} = A_{k,j} \left(1 - \left[E_{k,j}\right]_{V_j}\right) \tag{17}$$

The accuracy of the above methods depends on a given system under scrutiny and on the load area. However, on the basis of the authors' own experience gained during numerous relevant studies, Authors are in position to claim that in general both methods offer the same accuracy.

### 4.3 Sequence of Calculations

Successive steps in modeling a multiservice overflow system of which the general scheme in presented in Fig. 2 can be summarized as follows:

1. Determination, on the basis of (11) and (12), of the occupancy distribution and the blocking probability in PR.
2. Determination, on the basis of (13), (14), (9) and (10), of the parameters of traffic that overflows from PR, i.e. the parameters $R_{i,j}$, $\sigma^2_{i,j}$, $Z_{i,j}$, $Z_0$.
3. Determination, on the basis of (11) and (12), of the occupancy distribution and the blocking probability in SR.

## 5. Overflow System Models

### 5.1 Model of a System with Streaming Traffic

In streaming traffic, the bitrate of a call and its service time are independent of the current load of a network. In such a case, the occupancy distribution of PR comes down to the distribution [43], [44], which means that it will be determined by the distribution (11), in which $v = V_j$, whereas the function $\mathrm{RF_{PR}}_j(n)$ takes on the following form:

$$\mathrm{RF_{PR}}_j(n) = \frac{1}{n} \sum_{i=1}^{M} A_{i,j} t_i \, [P(n-t_i)]_{V_j} \tag{18}$$

The occupancy distribution in SR comes down to the distribution [41] and can be expressed by Formula (11), in which $v = V_0/Z_0$, whereas the function $\mathrm{RF}_{\mathrm{SR}}(n)$ takes on the following form:

$$\mathrm{RF}_{\mathrm{SR}}(n) = \frac{1}{n}\sum_{i=1}^{M}\sum_{j=1}^{r}\frac{R_{i,j}}{Z_{i,j}}t_i\,[P(n-t_i)]_{V_0/Z_0} \tag{19}$$

### 5.2 Model of a System with Elastic Traffic

Modern multiservice networks with differentiated QoS parameters, including the Internet, are packet networks in which calls (packet streams) are influenced by different traffic shaping mechanisms. Elastic traffic [83] is related to the mechanism of thresholdless compression that is related primarily to non-real time services in those networks that employ Transmission Control Protocol (TCP protocol). This mechanism works in the following way. In states with high resource loads, the absence of free AUs to service a new call leads to a compression of all calls that are currently being serviced. This compression is based on an increase in service time and a concurrent decrease in the number of admitted AUs to a value that allows this new call to be serviced (with a decreased number of AUs and extended service time).

Modelling of systems with elastic traffic is based on an introduction of certain fictitious virtual resources $V^{\mathrm{vir}}$ to the system. This means that the capacity of the system in the model is: $V + V^{\mathrm{vir}}$, where $V$ is the real capacity. In states $n \le V$, calls do not undergo the compression mechanism, while the call service in the fictitious service area ($V < n \le V^{\mathrm{vir}}$) corresponds to the compression in a system with the real capacity. In the fictitious service area, the ratio $n/V$ is the measure of the "compression depth" and determines how many times the bitrates of serviced calls in the system are decreased in the occupancy state $n$.

The occupancy distribution in PR with elastic traffic can be reduced to the approximating distribution derived in [66]. This distribution can be written in the form (11), where $v = V + V^{\mathrm{vir}}$, and the function $\mathrm{RF}_{\mathrm{PR}j}(n)$ is equal to:

$$\mathrm{RF}_{\mathrm{PR}j}(n) = \frac{1}{\min(n, V_j)}\sum_{i=1}^{M}A_{i,j}t_i\,[P(n-t_i)]_{V_j+V_j^{\mathrm{vir}}} \tag{20}$$

The occupancy distribution in SR can be approximated by the distribution proposed in [47] which, in the adopted notation for the distribution (11) for $v = (V_0+V_0^{\mathrm{vir}})/Z_0$, can be written as follows:

$$\mathrm{RF}_{\mathrm{SR}}(n) = \frac{1}{\min(n, V_0/Z_0)}\sum_{i=1}^{M}\sum_{j=1}^{r}\frac{R_{i,j}}{Z_{i,j}}t_i[P(n-t_i)]_{\frac{V_0+V_0^{\mathrm{vir}}}{Z_0}} \tag{21}$$

Note that if the compression mechanism is applied only in PR (SR), then to model the resources of SR (PR) the model (19) (model (18)) is used.

### 5.3 Model of a System with Adaptive Traffic

Adaptive traffic is related to the threshold compression mechanism. It is used in the case of the execution of real-time services with Real-Time Transport Protocol (RTP) [84] and Real-Time Control Protocol RTCP [85]. The operation of this mechanism is the following. The resources of a system are divided into load areas (LAs), limited by appropriate occupancy thresholds. When a given system exceeds a predefined occupancy threshold, a new call will be admitted with a decreased, pre-defined earlier, number of AUs that remains unchanged throughout the service execution. Service time does not depend on the load of the system and in each load area is identical. The assumption is that threshold compression mechanisms in PR and SR operate independently and, as a result, calls from PR to SR overflow in their uncompressed form.

Note that the threshold compression mechanism reduces bitrates at the stage of new call admission which remain the same and unchanged during service. The thresholdless compression mechanism, discussed earlier in the previous subsection, influences in turn the calls that are already being serviced.

Modelling of systems with adaptive traffic is based on the introduction of a set of occupancy thresholds for each call classes to the resources. Therefore, for class $i$ the set $q_i$ of thresholds: $\left\{T_{i,1}, T_{i,2}, \ldots, T_{i,q_i}\right\}$ is introduced, where the thresholds are arranged non-decreasingly $0 < T_{i,1} < T_{i,2} < T_{i,q_i} < V$ and $V$ is the capacity of the resources under consideration. In the model, LAs limited by the neighboring thresholds are considered, in which calls are allocated appropriate numbers of AUs from the set $\left\{t_{i,0}, t_{i,1}, \ldots, t_{i,q_i}\right\}$ for service, where $t_{i,0} > t_{i,1} > t_{i,q_i}$. Therefore, in the LA, such that $0 \le n \le T_{i,1}$ calls of class $i$ obtain $t_{i,0}$ AUs. Respectively for the LA: $T_{i,k} < n \le T_{i,k+1}$ it is $t_{i,k}$ AUs, whereas for the LA: $T_{i,q_i} < n \le V$ it is $t_{i,q_i}$ AUs. The occupancy distribution in PR with adaptive traffic, compressed in a threshold way, will be then reduced in its essence to the distribution (11), in which $v = V_j$, whereas Function $\mathrm{RF}_{\mathrm{PR}j}(n)$ takes on the following form [86]–[88]:

$$\mathrm{RF}_{\mathrm{PR}j}(n) = \frac{1}{n}\sum_{i=1}^{M}\sum_{k=0}^{q_i}A_{i,j}t_{i,k}\delta_{i,k}(n-t_{i,k})[P(n-t_{i,k})]_{V_j} \tag{22}$$

where $\delta_{i,k}$ is the conditional transition probability, in this particular case equal to:

$$\delta_{i,k}(n) = \begin{cases} 1 & \text{for } T_{i,k} < n \le T_{i,k+1} \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

For the correspondence with the load intervals pre-defined earlier in (23), the assumption is that $T_{i,0} = 0$ and $T_{i,q_i+1} = V_j$. It should be noted that the parameter $\delta_{i,k} = 1$ only in this load area in which the allocated number of AUs to service

a call is $t_{i,k}$. Another assumption is that the thresholds in PR are arranged in such a way [89] that the blocking phenomenon occurs exclusively in the oldest LA: $T_{i,k} < n \leq V_j$. Then, (12) can be written as follows:

$$\left[E_{i,j}\right]_{V_j} = \sum_{n=V_j-t_{i,q_i}+1}^{V_j} [P(n)]_{V_j} \tag{24}$$

The occupancy distribution in SR comes down in its essence to the distribution [51] and can be expressed by Formula (11), in which $v = V_0/Z_0$, whereas Function $\mathrm{RF}_{\mathrm{SR}}(n)$ takes on the following form:

$$\mathrm{RF}_{\mathrm{SR}}(n) = \frac{1}{n}\sum_{i=1}^{M}\sum_{k=0}^{q_i}\sum_{j=1}^{r}\frac{R_{i,j}}{Z_{i,j}}t_{i,k}\delta_{i,k}(n-t_{i,k})\left[P(n-t_{i,k})\right]_{\frac{V_0}{Z_0}} \tag{25}$$

where:

$$\delta_{i,k}(n) = \begin{cases} 1 & \text{for } \frac{T_{i,k}}{Z_0} < n \leq \frac{T_{i,k+1}}{Z_0} \\ 0 & \text{otherwise} \end{cases} \tag{26}$$

### 5.4 Model of a System with Distributed Resources

The basis for modelling of systems with distributed resources is provided by Limited Availability Group (LAG) [79], [81], [82]. LAG is a set of separated component resources to which multiservice traffic is offered. The notion of "resource separation" is related to the way calls are serviced in the LAG. A call of class $i$ can be serviced, when the LAG has at least $t_i$ free (unoccupied) AUs in one component resource. This means that a call cannot be "divided" between AUs of a number of component resources. The assumption is that the LAG services streaming traffic and that it is composed of $s$ identical component resources, each with the capacity $f$ AUs. This group can model a group of, for example, $s$ cells of a mobile network [48] or an output group of a switching network [22], [23].

The occupancy distribution in the $j$-th PR with distributed resources can be approximated by the distribution (11) dla $v = V_j = s_j f_j$, while Function $\mathrm{RF}_{\mathrm{PR}j}(n)$ takes on the following form [81]:

$$\mathrm{RF}_{\mathrm{PR}j}(n) = \frac{1}{n}\sum_{i=1}^{M} A_{i,j}t_i\xi_{i,j}(n-t_i)\left[P(n-t_i)\right]_{V_j} \tag{27}$$

where $\xi_{i,j}$ is the conditional transition probability for a stream of class $i$ that can be determined combinatorially:

$$\xi_{i,j}(n) = \frac{F(V_j-n, s_j, f_j) - F(V_j-n, s_j, t_i-1)}{F(V_j-n, s_j, f_j)} \tag{28}$$

Function $F(x, s, f)$ w (28) is the number of the arrangements of $x$ free AUs in $s$ separated component resources, each of which with the capacity equal to $f$:

$$F(x,s,f) = \sum_{k=0}^{\lfloor \frac{x}{f+1} \rfloor} (-1)^k \binom{s}{k}\binom{x+s-k(f+1)-1}{s-1} \tag{29}$$

The parameter $\xi_{i,j}$ expresses the number of favorable arrangements (the numerator in Formula (28) to all possible arrangements (the denominator in Formula (28)) of free AUs in the $j$-th PR. The notion of favorable arrangements is understood to mean such arrangements of free AUs in $s_j$ separated component resources that at least one of these resorces can service a call of class $i$ that requires $t_i$ AUs to set up a connection. The blocking probability in distributed PR is determined by Formula:

$$\left[E_{i,j}\right]_{V_j} = \sum_{n=0}^{V_j}\left[1-\xi_{i,j}(n)\right][P(n)]_{V_j} \tag{30}$$

The occupancy distribution in SR with distributed resources comes down to the distribution [41] and can be expressed by Formula (11), in which $v = V_0/Z_0$, while Function $\mathrm{RF}_{\mathrm{SR}}(n)$ takes on in the considered case the following form [47], [48]:

$$\mathrm{RF}_{\mathrm{SR}}(n) = \frac{1}{n}\sum_{i=1}^{M}\sum_{j=1}^{r}\frac{R_{i,j}}{Z_{i,j}}t_i\xi_{i,0}(n-t_i)\left[P(n-t_i)\right]_{\frac{V_0}{Z_0}} \tag{31}$$

where:

$$\xi_{i,0}(n) = \frac{F(V_0/Z_0-n, s_0, f_0/Z_0) - F(V_0/Z_0-n, s_0, t_i-1)}{F(V_0/Z_0-n, s_0, f_0/Z_0)} \tag{32}$$

The PR and SR models presented above refer to streaming traffic. These models can be generalized to include elastic and adaptive traffic, e.g. [47] describes a model with elastic traffic in which SRs are distributed.

### 5.5 Model of a Non-Full-Availability System

A large number of network systems are non-full-availability systems in which new calls that arrive at a given input have access only to a certain part $d$ of resources, lower than the total capacity of the resources $V$. If a new call of class $i$ arrives at the input to the system and $d$ resources, available to this input, cannot service a given call (i.e. do not have $t_i$ free AUs), then this call is lost despite the fact that the system has $t_i$ AUs that could service this call. The basis for modelling multiservice non-full-availability systems is provided by the EIG (Erlangs Ideal Grading) [90]. The first multiservice model of the EIG with identical value of the availability parameter for all classes of calls is proposed in [91]. In [62], [92] an EIG model with variable value of the availability parameter is proposed.

The inputs of the EIG that have access to the same resources are called a load group. A multiservice EIG has the following properties:

1. the number of load groups is equal to the number of possible choices of $d$ AUs from among all $V$ AUs (two load groups differ from each other by at least one AUs),
2. each load group has access to the same number of $d$

AUs of the group,
3. traffic offered to each load group is identical,
4. the occupancy distribution in each of the load groups is identical.

The occupancy distribution in the $j$-th PR with non-full-availability resources approximated by the EIG model, is determined by the distribution (11) for $v = V_j$ and the availability $d_{i,j}$ for calls of class $i$. Function $\mathrm{RF}_{\mathrm{PR}j}(n)$ in such a case takes on the following form [62]:

$$\mathrm{RF}_{\mathrm{PR}j}(n) = \frac{1}{n}\sum_{i=1}^{M} A_{i,j}t_i\left[1-\gamma_{i,j}(n-t_i)\right][P(n-t_i)]_{V_j} \tag{33}$$

where $\gamma_{i,j}$ is the conditional blocking probability for calls of class $i$ in the occupancy state $n$ AUs in the EIG:

$$\gamma_{i,j}(n) = \sum_{x=d_{i,j}-t_i+1}^{\Psi}\binom{d_{i,j}}{x}\binom{V_{i,j}-d_{i,j}}{n-x}\Big/\binom{V_{i,j}}{n} \tag{34}$$

The upper limit of the summation in (34) depends on $n$:

$$\Psi = \begin{cases} n & \text{if } (d_{i,j}-t_i+1) \le n \le d_{i,j} \\ d_{i,j} & \text{if } n > d_{i,j} \end{cases} \tag{35}$$

Having the conditional probabilities $\gamma_{i,j}$ determined, we can determine the blocking probability for calls of class $i$ in PR $j$:

$$\left[E_{i,j}\right]_{V_j} = \sum\nolimits_{n=d_{i,j}-t_i+1}^{V_j}\gamma_{i,j}(n)\,[P(n)]_{V_j} \tag{36}$$

The occupancy distribution in SR with non-full-availability resources, approximated by the EIG model, can be expressed by Formula (11), in which $v = V_0/Z_0$, whereas Function $\mathrm{RF}_{\mathrm{SR}}(n)$ takes on in the considered case the following form:

$$\mathrm{RF}_{\mathrm{SR}}(n) = \frac{1}{n}\sum_{i=1}^{M}\sum_{j=1}^{r}\frac{R_{i,j}}{Z_{i,j}}t_i\left[1-\gamma_{i,0}(n-t_i)\right][P(n-t_i)]_{\frac{V_j}{Z_0}} \tag{37}$$

where:

$$\gamma_{i,0}(n) = \sum_{x=\frac{d_{i,0}}{Z_0}-t_i+1}^{\Psi}\binom{\frac{d_{i,j}}{Z_0}}{x}\binom{\frac{V_{i,0}}{Z_0}-\frac{d_{i,0}}{Z_0}}{n-x}\Big/\binom{\frac{V_{i,j}}{Z_0}}{n} \tag{38}$$

The non-full availability PR and SR models presented above refer to streaming traffic. These models can be easily generalized to include elastic and adaptive traffic.

### 5.6 Model of an Overflow System with BPP Traffic

Multiservice traffic of the type BPP (Bernoulli, Pascal, Poisson) is a mixture of Engset (Bernoulli call stream), Erlang (Poisson call stream) and Pascal traffic (Pascal call stream). Note that BPP traffic includes all "Markovian" dependencies in traffic intensity on a given state of resources. In the case of Erlang traffic, offered traffic $A_{i,\mathrm{Er}}$ is independent of the occupancy state $n$:

$$A_{i,\mathrm{Er}}(n) = A_{i,\mathrm{Er}} \text{ for } i \in M_{\mathrm{Er}} \tag{39}$$

In the case of Engset traffic, traffic intensity $A_{i,\mathrm{En}}$ decreases, while for Pascal traffic $A_{i,\mathrm{Pa}}$ increases along with the increase in $n$:

$$A_{i,\mathrm{En}}(n) = \alpha_{i,\mathrm{En}}\left[S_{i,\mathrm{En}} - N_{i,\mathrm{En}}\right] \text{ for } i \in M_{\mathrm{En}} \tag{40}$$

$$A_{i,\mathrm{Pa}}(n) = \alpha_{i,\mathrm{Pa}}\left[S_{i,\mathrm{Pa}} + N_{i,\mathrm{Pa}}\right] \text{ for } i \in M_{\mathrm{Pa}} \tag{41}$$

where $S_{i,X}$ ($X \in \{\mathrm{En}, \mathrm{Pa}\}$) is the number of traffic sources of class $i$ type $X$, $\alpha_{i,X}$ is the average traffic intensity from one free source of class $i$ type $X$, $N_{i,X}$ is the number of calls of class $i$, type $X$, that are serviced in state $n$, whereas $M_X$ is the set of all traffic classes of type $X$.

The occupancy distribution in PR with BPP traffic can come down then to the distribution [45], [62], i.e. will be determined by the distribution (11), in which $v = V_j$, whereas Function $\mathrm{RF}_{\mathrm{PR}j}(n)$ will take on the following form:

$$\mathrm{RF}_{\mathrm{PR}j}(n) = \frac{1}{n}\left\{\sum_{i\in M_{i,\mathrm{Er}}} A_{i,j,\mathrm{Er}}t_{i,\mathrm{Er}}[P(n-t_{i,\mathrm{Er}})]_{V_j} + \sum_{i\in M_{i,\mathrm{En}}}\alpha_{i,j,\mathrm{En}}\left[s_{i,j,\mathrm{En}} - y_{i,j,\mathrm{En}}(n)\right]t_{i,\mathrm{En}}[P(n-t_{i,\mathrm{En}})]_{V_j} + \sum_{i\in M_{i,\mathrm{Pa}}}\alpha_{i,j,\mathrm{Pa}}\left[s_{i,j,\mathrm{Pa}} + y_{i,j,\mathrm{Pa}}(n)\right]t_{i,\mathrm{Pa}}[P(n-t_{i,\mathrm{Pa}})]_{V_j}\right\} \tag{42}$$

where $y_{i,X}(n)$ is the average value of the number of serviced calls of class $i$, type $X$:

$$y_{i,j,\mathrm{X}} = \frac{\alpha_{i,j,X}\left[S_{i,j,X} \pm y_{i,j,X}(n-t_{i,X})\right][P(n-t_{i,X})]_{V_j}}{[P(n)]_{V_j}} \tag{43}$$

The occupancy distribution in SR with BPP traffic can come down to the distribution [62], i.e. will be determined by the distribution (11), in which $v = V_0/Z_0$, whereas Function $\mathrm{RF}_{\mathrm{PR}j}(n)$ will take on the following form:

$$\mathrm{RF}_{\mathrm{SR}}(n) = \frac{1}{n}\left\{\sum_{i\in M_{i,\mathrm{Er}}}\sum_{j=1}^{r}\frac{R_{i,j,\mathrm{Er}}}{Z_0}t_{i,\mathrm{Er}}[P(n-t_{i,\mathrm{Er}})]_{\frac{V_0}{Z_0}} + \sum_{i\in M_{i,\mathrm{En}}}\sum_{j=1}^{r}\frac{R_{i,j,\mathrm{En}}}{Z_0}t_{i,\mathrm{En}}[P(n-t_{i,\mathrm{En}})]_{\frac{V_0}{Z_0}} + \sum_{i\in M_{i,\mathrm{Pa}}}\sum_{j=1}^{r}\frac{R_{i,j,\mathrm{Pa}}}{Z_0}t_{i,\mathrm{Pa}}[P(n-t_{i,\mathrm{Pa}})]_{\frac{V_0}{Z_0}}\right\} \tag{44}$$

In the case of Engset and Pascal traffic, a decomposition

of the PR, e.g. by the method [42], requires an appropriate formula to determine fictitious capacities $V^*_{i,j,X}$ to be applied. For this purpose, Formula (15) is used for Erlang traffic, whereas for Engset and Pascal traffic, Engset and Pascal formula, respectively [33]:

$$\left[E_{i,j,\mathrm{En}}\right]_{V_j} = \frac{\binom{S_{i,j,\mathrm{En}}}{V^*_{i,j,\mathrm{En}}}\left(\alpha_{i,j,\mathrm{En}}\right)^{V^*_{i,j,\mathrm{En}}}}{\sum_{n=0}^{V^*_{i,j,\mathrm{En}}}\binom{S_{i,j,\mathrm{En}}}{n}\left(\alpha_{i,j,\mathrm{En}}\right)^{n}} \tag{45}$$

$$\left[E_{i,j,\mathrm{Pa}}\right]_{V_j} = \frac{\binom{S_{i,j,\mathrm{Pa}}+V^*_{i,j,\mathrm{Pa}}-1}{V^*_{i,j,\mathrm{Pa}}}\left(\alpha_{i,j,\mathrm{Pa}}\right)^{V^*_{i,j,\mathrm{Pa}}}}{\sum_{n=0}^{V^*_{i,j,\mathrm{Pa}}}\binom{S_{i,j,\mathrm{Pa}}+n-1}{n}\left(\alpha_{i,j,\mathrm{Pa}}\right)^{n}} \tag{46}$$

Since Riordan's formulas (13) and (14) allow us to determine the variance of overflow traffic with the assumption that offered traffic is Erlang traffic, then it is necessary to convert Engset and Pascal traffic $A_{i,j,X}$ into equivalent Erlang traffic $(A_{i,j,X})^*_{\mathrm{Er}}$ that overflows from the equivalent resources with the capacity $\Delta V^*_{i,j,x}$. To achieve that, the concept of equivalent resources (Sect. 2.2) is used. The average value and the variance for Engset and Pascal traffic are equal [33]:

$$A_{i,j,\mathrm{En}} = S_{i,j,\mathrm{En}}\frac{\alpha_{i,j,\mathrm{En}}}{1+\alpha_{i,j,\mathrm{En}}} \tag{47}$$

$$\sigma^2(A_{i,j,\mathrm{En}}) = S_{i,j,\mathrm{En}}\frac{\alpha_{i,j,\mathrm{En}}}{\left(1+\alpha_{i,j,\mathrm{En}}\right)^2} \tag{48}$$

$$A_{i,j,\mathrm{Pa}} = S_{i,j,\mathrm{Pa}}\frac{\alpha_{i,j,\mathrm{Pa}}}{1-\alpha_{i,j,\mathrm{Pa}}}, \tag{49}$$

$$\sigma^2(A_{i,j,\mathrm{Pa}}) = S_{i,j,\mathrm{Pa}}\frac{\alpha_{i,j,\mathrm{Pa}}}{\left(1-\alpha_{i,j,\mathrm{Pa}}\right)^2} \tag{50}$$

For the parameters $A_{i,j,X}$, $\sigma^2(A_{i,j,X})$, it is now possible to determine, on the basis of Eqs. (1) and (2), the parameters of the equivalent resources $(A_{i,j,X})^*_{\mathrm{Er}}$, $\Delta V^*_{i,j,X}$. Now, traffic with the intensity $(A_{i,j,X})^*_{\mathrm{Er}}$ is offered to PR with the capacity $V^*_{i,j,X}+\Delta V^*_{i,j,X}$. Thus defined PR allows us to determine the parameters for trafffic that overflows to SR, i.e., the average value $R_{i,j,X}$ and the variance $\sigma^2_{i,j,X}$. It should be noted that for Engset trafffic with the capacity $\Delta V^*_{i,j,X}$ it takes on negative values. Therefore, to determine offered traffic $(A_{i,j,X})^*_{\mathrm{Er}}$ in (1) the integral form of Erlang's formula [28] is used. This problem is addressed in a more detailed way in [51].

### 5.7 Commentary

This article aims at a presentation of the possibilities offered by Hayward's concept [30] to model multiservice overflow systems in present-day network systems. The authors of the article believe that Hayward's concept is universal, versatile and promising for modelling modern multiservice overflow systems. In Sect. 5.1, the most general model with streaming traffic [41], [42] is discussed, where the parameters of calls, both in PR and SR are fixed (do not change). In Sect. 5.2, an overflow model with elastic traffic compressed thresholdlessly [50] is discussed, whereas in Sect. 5.3 a model with adaptive traffic with threshold compression is presented [51]. In Sects. 5.4 and 5.5, models of multiservice overflow systems are presented in which the occupancy distributions are state-dependent, i.e. a model with distributed resources [47], [48] and a model in which PR and SR can be of non-full-availability nature. In Sect. 5.6, a model of an overflow system with BPP traffic [51] is described. The propositions of the model with non-full-availability resources is published for the first time and has not been verified in simulations yet.

The proposed models show that the Hayward's concept can be used to model overflow in all possible traffic configurations and all traffic shaping mechanisms in PR and SR used in modern network systems. In the near future, the authors intend to solve, on the basis of the Hayward's approach, the problem of overflow in systems with queues in PR and SR as well as the problem of concurrent overflow of a mixture streaming, elastic and adaptive traffic. Since the proposed models are approximate models, the question arises as to the accuracy the models can offer. In the following section the results of a comparison of analytical modelling with simulation modelling for a number of selected overflow systems will be presented. These results show high accuracy of the proposed models that is comparable with the accuracy offered by other multiservice models [89]. Therefore, these models can be successfully used in practice to model and optimize network systems.

## 6. Numerical Results

This section presents results provided by a comparison of the analytical calculations for a number of selected overflow systems with those obtained in the simulation experiments. The results, presented in Figs. 3, 4 and 5 are limited to the systems presented in Sects. 5.1, 5.2 and 5.4, respectively. The parameters of the systems under investigation are listed in Tables 1, 2 and 3, respectively. The aim of this section is to provide a general presentation of the accuracy of methods based of Fredericks-Hayward's methodology, elaborated for modeling multiservice systems with traffic overflows. The detailed evaluation of the accuracy of the presented methods, also for other systems, can be found in the authors' publications cited in particular sections of this article. All simulation experiments were carried out for 5 series, 10000000 calls each. The results of the simulation are marked with appropriate symbols with the 95% confidence intervals.

The results of the study clearly show that the accuracy of the method does not depend on the capacity of both PR and SR. In the same way, any differentiation in demands of individual classes (the $t$ parameter) does not influence the accuracy of the proposed method. The results presented in this article as well as numerous other results published in the literature provide ample evidence as to state that the use of Fredericks-Hayward's concept provides high accuracy in calculations for systems with overflow traffic.

The biggest influence on the accuracy of the models

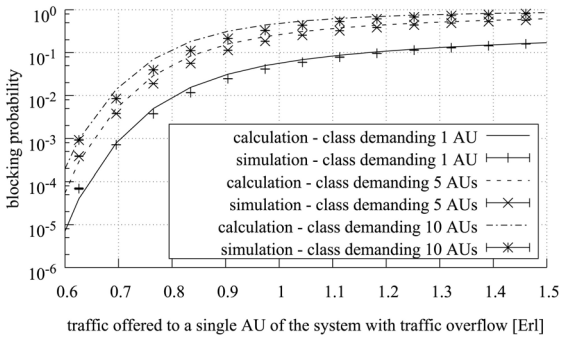


**Fig. 3** Blocking probability in the alternative resources of the system presented in Table 1.

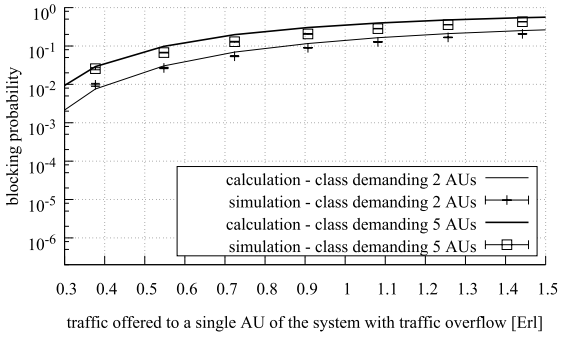


**Fig. 4** Blocking probability in the alternative resources of the system presented in Table 2.

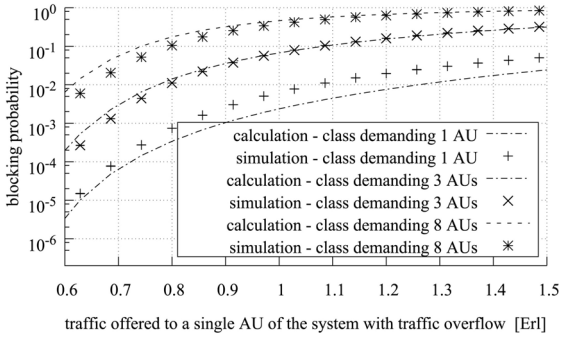


**Fig. 5** Blocking probability in the alternative resources of the system presented in Table 3.

**Table 1** The parameters of the overflow system from Sect. 5.1.

| $j$ | $V_j$ | $t_1$ | $t_2$ | $t_3$ | $V_0$ |
|---|---|---|---|---|---|
| 1 | 60 | 2 | 3 | 6 | 140 |
| 2 | 60 | 1 | 4 | 6 | |
| 3 | 100 | 1 | 5 | 10 | |
| 4 | 100 | 1 | 5 | 10 | |

has the expression $Z_0/V_0$ that usually takes on non-integer values. When this is the case, Authors take advantage of the simplest linear interpolation that provides sufficient accuracy in engineering applications and practice. A method for a determnation of the capacity of fictitious resources (with regard to the criterion of the adjustment of the blocking probability [42] or to the criterion of the adjustment (matching) of carried traffic [41]) has only slight and negligible influence on the accuracy of final results.

**Table 2** The parameters of the overflow system from Sect. 5.2.

| $j$ | $V_j$ | $V_j^{\mathrm{vir}}$ | $t_1$ | $t_2$ | $t_3$ | $t_4$ | $V_0$ |
|---|---|---|---|---|---|---|---|
| 1 | 40 | 40 | – | 3 | – | 5 | 50 |
| 2 | 50 | 50 | 2 | 3 | 4 | – | 50 |

**Table 3** The parameters of the overflow system from Sect. 5.4.

| primary | | | | secondary | | offered traffic | | | |
|---|---|---|---|---|---|---|---|---|---|
| $j$ | $V_j$ | $s_j$ | $f_j$ | $s_0$ | $f_0$ | $M$ | $t_1$ | $t_2$ | $t_3$ |
| 1 | 40 | 2 | 20 | | | 3 | 1 | 3 | 8 |
| 2 | 60 | 2 | 30 | 4 | 30 | 3 | 1 | 3 | 8 |
| 3 | 60 | 3 | 20 | | | 3 | 1 | 3 | 8 |

## 7. Conclusions

This article presents a number of different approximate models of multiservice overflow systems for which a common theoretical base is provided by the Hayward's concept. These models are, in Authors opinion, most versatile and universal and can be used in practice to analyze different scenarios for the operation of overflow systems, and then to solve various design and optimization issues in modern telecommunications and computer networks, such as TCP/IP networks or 4G and 5G mobile networks. The present article discusses, or proposes, a number of multiservice overflow models with streaming, elastic and adaptive traffic, models with distributed and non-full-availability resources and models with BPP traffic. Such a wide range of modelling opportunities proves the large and universal potential that is held by Hayward's concept. It is Authors opinion that any approach that is based on this particular concept will make it possible to solve virtually any practical problem related to traffic overflow in present-day network systems.

In the near future Authors intend to develop further models of overflow systems that will be capable of servicing concurrently streaming, elastic and adaptive traffic with the possibility of queueing this traffic in PR and SR. This will be the most universal model of the multiservice overflow system.

### References

[1] R.I. Wilkinson, "Theories of toll traffic engineering in the USA," Bell Syst. Tech. J., vol.40, pp.421–514, 1956.
[2] M. Głąbowski, S. Hanczewski, and M. Stasiak, "Erlang's ideal grading in DiffServ modelling," Proc. IEEE Africon 2011, pp.1–6, Livingstone, Zambia, IEEE, Sept. 2011.
[3] J. Matsumoto and Y. Watanabe, "Theoretical method for the analysis of queueing system with overflow traffic," Electron. Comm. Jpn. Pt/I, vol.64, no.6, pp.74–83, 1981.
[4] E.W. Wöng, J. Guö, B. Möran, and M. Zukerman, "Information exchange surrogates for approximation of blocking probabilities in overflow loss systems," Proc. 25th International Teletraffic Congress, pp.1–9, 2013.
[5] A.A. Kist and R.J. Harris, "Scheme for alternative packet overflow routing (SAPOR)," Workshop on High Performance Switching and

Routing, 2003, HPSR., pp.269–274, June 2003.
[6] P. Clapson, "Improving the access time for random access files," Commun. ACM, vol.20, no.3, pp.127–135, March 1977.
[7] P. Kühn and M.E. Mashaly, "Multi-server, finite capacity queuing system with mutual overflow," Proc. 2nd European Teletraffic Seminar, M. Fiedler, ed., Karlskrona, Sept. 2013.
[8] M. Mashaly and P.J. Kühn, "Load balancing in cloud-based content delivery networks using adaptive server activation/deactivation," Proc. 24th International Teletraffic Congress, ITC'12, pp.21:1–21:3, 2012.
[9] G. Soni and M. Kalra, "A novel approach for load balancing in cloud data center," Advance Computing Conference (IACC), 2014 IEEE International, pp.807–812, Feb. 2014.
[10] P.J. Kühn, "Systematic classification of self-adapting algorithms for power-saving operation modes of ICT systems," Proc. 2nd International Conference on Energy-Efficient Computing and Networking, e-Energy'11, pp.51–54, New York, NY, USA, ACM, 2011.
[11] M. Yoshino, N. Nishibe, M. Oba, and N. Komoda, "Classification of energy-saving operations from the perspective of system management," 8th IEEE International Conference on Industrial Informatics (INDIN), pp.651–656, July 2010.
[12] C. Gauger, P. Kühn, E. Breusegem, M. Pickavet, and P. Demeester, "Hybrid optical network architectures: Bringing packets and circuits together," IEEE Commun. Mag., vol.44, no.8, pp.36–42, 2006.
[13] M. Wang, S. Li, E. Wong, and M. Zukerman, "Performance analysis of circuit switched multi-service multi-rate networks with alternative routing," J. Lightwave Technol., vol.32, no.2, pp.179–200, Jan. 2014.
[14] S. Fernandes and A. Karmouch, "Vertical mobility management architectures in wireless networks: A comprehensive survey and future directions," IEEE Commun. Surveys Tuts., vol.14, no.1, pp.45–63, First 2012.
[15] X. Wu, B. Murherjee, and D. Ghosal, "Hierarchical architectures in the third-generation cellular network," IEEE Wireless Commun., vol.11, no.3, pp.62–71, June 2004.
[16] M. Głąbowski, S. Hanczewski, and M. Stasiak, "Modelling of cellular networks with traffic overflow," Math. Probl. Eng., vol.2015, p.15, 2015. Article ID 286490.
[17] 3GPP, "Self-configuring and self-optimizing network (SON) use cases and solutions (Release 9)," Technical Report, TR 36.902, 3GPP, Sep. 2010.
[18] N. Zia and A. Mitschele-Thiel, "Self-organized neighborhood mobility load balancing for LTE networks," 2013 IFIP Wireless Days (WD), pp.1–6, Nov. 2013.
[19] M. Głąbowski, S. Hanczewski, and M. Stasiak, "Modelling load balancing mechanisms in self-optimising 4G mobile networks with elastic and adaptive traffic," IEICE Trans. Commun., vol.E99-B, no.8, pp.1718–1726, Aug. 2016.
[20] R. Fortet, Systeme Pentaconta Calcul d'orange, LMT, Paris, 1961.
[21] H. Inose, T. Saito, and M. Kato, "Three-stage time-division switching junctor as alternate route," Electron. Lett., vol.2, no.5, pp.78–84, 1966.
[22] M. Głąbowski and M.D. Stasiak, "Modelling of multiservice switching networks with overflow links for any traffic class," IET Circuits, Devices & Systems, vol.8, no.5, pp.358–366, 2014.
[23] M. Głąbowski and M.D. Stasiak, "Multiservice switching networks with overflow links and resource reservation," Math. Probl. Eng., vol.2016, Article ID 4090656, 17 pages, 2016.
[24] E.W.M. Wong, A. Zalesky, Z. Rosberg, and M. Zukerman, "A new method for approximating blocking probability in overflow loss networks," Computer Networks, vol.51, no.11, pp.2958–2975, 2007.
[25] G. Bretschneider, "Die Berechnung von Leitungsgruppen für berfließenden Verkehr in Fernsprechwählanlagen," Nachrichtentechnische Zeitung (NTZ), no.11, pp.533–540, 1956.
[26] B. Wallström, "A distribution model for telefone traffic with varying call intensity, including overflow traffic," Ericsson Technics, vol.20, no.2, pp.183–202, 1964.
[27] U. Herzog and A. Lotze, "Das RDA-Verfahren, ein Streuwertverfahren für unvollkommene Bündel," NTZ - Nachrichtentechnische Zeitschrift, pp.640–646, 1966.
[28] G. Bretschneider, "Extension of the equivalent random method to smooth traffics," Proc. 7th International Teletraffic Congress, Stockholm, 1973.
[29] R. Schehrer, "On the calculation of overflow systems with a finite number of sources and full availiable groups," IEEE Trans. Commun., vol.26, no.1, pp.75–82, Jan. 1978.
[30] A. Fredericks, "Congestion in blocking systems – A simple approximation technique," Bell Syst. Tech. J., vol.59, no.6, pp.805–827, July–Aug. 1980.
[31] J.F. Shortle, "An equivalent random method with hyper-exponential service," J. Perform. Evaluation, vol.57, no.3, pp.409–422, 2004.
[32] L. Kösten, "Behaviour of overflow traffic and the probabilities of blocking in simple gradings," Proc. 8th International Teletraffic Congress, pp.425/1–425/5, 1976.
[33] V. Iversen, "Teletraffic engineering handbook," Technical Report, Technical University of Denmark, 2010.
[34] C.G. Park and D.H. Han, "Comparisons of loss formulas for a circuit group with overflow traffic," J. Appl. Math. Informatics, vol.30, no.1–2, pp.135–145, 2012.
[35] J.A. Morrison, "Analysis of some overflow problems with queuing," Bell Syst. Tech. J., vol.59, no.8, pp.1427–1462, 1980.
[36] Y. Zhao and E. Gambe, "Analysis on partial overflow queueing systems with two kinds of calls," IEEE Trans. Commun., vol.35, no.9, pp.942–949, Sept. 1987.
[37] J. Matsumoto and Y. Watanabe, "Individual traffic characteristics queueing systems with multiple poisson and overflow inputs," IEEE Trans. Commun., vol.33, no.1, pp.1–9, Jan. 1985.
[38] A. Kuczura, "The interrupted Poisson process as an overflow process," Bell Syst. Tech. J., vol.52, no.3, pp.437–448, 1973.
[39] A. Bachle, "On the calculation of full available groups with offered smooth traffic," Proc. 7th International Teletraffic Congress, pp.223/1–223/6, Stockholm, Sweden, 1973.
[40] H. Akimaru and K. Kawashima, Teletraffic: Theory and Application, Springer, Berlin-Heidelberg-New York, 1999.
[41] M. Głąbowski, K. Kubasik, and M. Stasiak, "Modeling of systems with overflow multi-rate traffic," Telecommun. Syst., vol.37, no.1–3, pp.85–96, March 2008.
[42] Q. Huang, K.T. Ko, and V.B. Iversen, "Approximation of loss calculation for hierarchical networks with multiservice overflows," IEEE Trans. Commun., vol.56, no.3, pp.466–473, March 2008.
[43] J. Kaufman, "Blocking in a shared resource environment," IEEE Trans. Commun., vol.29, no.10, pp.1474–1481, 1981.
[44] J. Roberts, "A service system with heterogeneous user requirements — Application to multi-service telecommunications systems," Proc. Performance of Data Communications Systems and their Applications, G. Pujolle, ed., pp.423–431, Amsterdam, North Holland, 1981.
[45] M. Głąbowski, "Modelling of state-dependent multi-rate systems carrying BPP traffic," Annals of Telecommunications, vol.63, no.7-8, pp.393–407, Aug. 2008.
[46] M. Głąbowski, M. Stasiak, and J. Weissenberg, "Properties of recurrent equations for the full-availability group with BPP traffic," Math. Probl. Eng., vol.2012, p.17, 2012. Article ID 547909.
[47] M. Głąbowski, A. Kaliszan, and M. Stasiak, "Modelling overflow systems with distributed secondary resources," Computer Networks, vol.108, pp.171–183, 2016.
[48] M. Głąbowski, A. Kaliszan, and M. Stasiak, "Analytical modelling of multi-tier cellular networks with traffic overflow," Computer Networks: 24th International Conference, CN 2017, Lądek Zdrój, Poland, June 20–23, 2017, Proceedings, P. Gaj, A. Kwiecień, and M. Sawicki, ed., pp.256–268, Springer International Publishing, Cham, 2017.
[49] M. Głąbowski, K. Kubasik, and M. Stasiak, "Modelling of systems with overflow multi-rate traffic and finite number of traffic sources," Proc. 6th International Symposium on Communication Systems, Net-

works and Digital Signal Processing 2008, pp.196–199, Graz, July 2008.
[50] M. Głąbowski, D. Kmiecik, and M. Stasiak, "Overflow of elastic traffic," 2016 International Conference on Broadband Communications for Next Generation Networks and Multimedia Applications (CoBCom), 2016.
[51] M. Głąbowski, D. Kmiecik, and M. Stasiak, "Modelling of multiservice networks with separated resources and overflow of adaptive traffic," Wireless Communications and Mobile Computing, vol.2018, Article ID 7870164, 17 pages, 2018.
[52] S.P. Chung and J.C. Lee, "Performance analysis and overflowed traffic characterization in multiservice hierarchical wireless networks," IEEE Trans. Wireless Commun., vol.4, no.3, pp.904–918, May 2005.
[53] L.R. Hu and S.S. Rappaport, "Personal communication systems using multiple hierarchical cellular overlays," IEEE J. Sel. Areas Commun., vol.13, no.2, pp.406–415, 1995.
[54] J.S. Kaufman and K.M. Rege, "Blocking in a shared resource environment with batched Poisson arrival processes," J. Perform. Evaluation, vol.24, no.4, pp.249–263, 1996.
[55] I.D. Moscholios, J.S. Vardakas, M.D. Logothetis, and A.C. Boucouvalas, "Congestion probabilities in a batched Poisson multirate loss model supporting elastic and adaptive traffic," Annals of Telecommunications, vol.68, no.5, pp.327–344, June 2013.
[56] L. Delbrouck, "On the steady-state distribution in a service facility carrying mixtures of traffic with different peakedness factors and capacity requirements," IEEE Trans. Commun., vol.31, no.11, pp.1209–1211, 1983.
[57] E.A. van Doorn and F.J.M. Panken, "Blocking probabilities in a loss system with arrivals in geometrically distributed batches and heterogeneous service requirements," IEEE/ACM Trans. Netw., vol.1, no.6, pp.664–677, 1993.
[58] Q. Huang, Y.C. Huang, K.T. Ko, and V. Iversen, "Loss performance modeling for hierarchical heterogeneous wireless networks with speed-sensitive call admission control," IEEE Trans. Veh. Technol., vol.60, no.5, pp.2209–2223, 2011.
[59] B.M. Bakmaz and M.R. Bakmaz, "Solving some overflow traffic models with changed serving intensities," AEU - International Journal of Electronics and Communications, vol.66, no.1, pp.80–85, 2012.
[60] M. Głąbowski, A. Kaliszan, and M. Stasiak, "Two-dimensional convolution algorithm for modelling multiservice networks with overflow traffic," Math. Probl. Eng., vol.2013, p.18, 2013. Article ID 852082.
[61] M. Głąbowski, S. Hanczewski, and M. Stasiak, "Modelling load balancing mechanisms in self-optimising 4G mobile networks," 21st Asia-Pacific Conference on Communications (APCC), pp.1–5, Kyoto, Japan, Oct. 2015.
[62] M. Głąbowski, S. Hanczewski, M. Stasiak, and J. Weissenberg, "Modeling Erlang's Ideal Grading with multi-rate BPP traffic," Math. Probl. Eng., vol.2012, p.35, 2012. Article ID 456910.
[63] S. Hanczewski, M. Sobieraj, and M.D. Stasiak, "The direct method of effective availability for switching networks with multi-service traffic," IEICE Trans. Commun., vol.E99-B, no.6, pp.1291–1301, June 2016.
[64] Y. Rapp, "Planning of junction network in a multi-exchange area," Proc. 4th International Teletraffic Congress, p.4, London, 1964.
[65] D. Bertsekas and R. Gallager, Data Networks, 2nd ed., Prentice-Hall, Upper Saddle River, NJ, USA, 1992.
[66] G. Stamatelos and V. Koukoulidis, "Reservation-based bandwidth allocation in a radio ATM network," IEEE/ACM Trans. Netw., vol.5, no.3, pp.420–428, June 1997.
[67] T. Bonald and J.W. Roberts, "Internet and the Erlang formula," SIGCOMM Comput. Commun. Rev., vol.42, no.1, pp.23–30, Jan. 2012.
[68] V. Paxon and S. Floyd, "Wide-area traffic: The failure of traffic modeling," Proc. 1994 SIGCOMM Conference, pp.257–268, Aug. 1994.
[69] F. Kelly, "Notes on effective bandwidth," Technical Report, University of Cambridge, 1996.
[70] R. Gurein, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its aplication to bandwidth allocation in high-speed networks," J. Sel. Areas Commun., vol.9, no.7, pp.968–981, Sept. 1991.
[71] I. Norros, "On the use on fractional brownian motion in the theory of connectionless networks," J. Sel. Areas Commun., vol.13, no.6, pp.953–962, Aug. 1995.
[72] A. Pras, L. Nieuwenhuis, R. van de Meent, and M. Mandjes, "Dimensioning network links: A new look at equivalent bandwidth," IEEE Netw., vol.23, no.2, pp.5–10, March 2009.
[73] J. Roberts, V. Mocci, and I. Virtamo, eds., Broadband Network Teletraffic, Final Report of Action COST 242, Commission of the European Communities, Springer, Berlin, 1996.
[74] J. Roberts, "Teletraffic models for the Telcom 1 integrated services network," Proc. 10th International Teletraffic Congress, p.1.1.2, Montreal, 1983.
[75] S. Rácz, B.P. Gerő, and G. Fodor, "Flow level performance analysis of a multi-service system supporting elastic and adaptive services," Perform. Evaluation, vol.49, no.1–4, pp.451–469, 2002.
[76] T. Bonald and J. Virtamo, "A recursive formula for multirate systems with elastic traffic," IEEE Commun. Lett., vol.9, no.8, pp.753–755, Aug. 2005.
[77] G. Fodor and M. Telek, "Bounding the blocking probabilities in multirate CDMA networks supporting elastic services," IEEE/ACM Trans. Netw., vol.15, no.4, pp.944–956, Aug. 2007.
[78] V.G. Vassilakis, I.D. Moscholios, and M.D. Logothetis, "The extended connection-dependent threshold model for call-level performance analysis of multi-rate loss systems under the bandwidth reservation policy," Int. J. Commun. Syst., vol.25, no.7, pp.849–873, 2012.
[79] M. Głąbowski, A. Kaliszan, and M. Stasiak, "Modeling product-form state-dependent systems with BPP traffic," Perform. Evaluation, vol.67, no.3, pp.174–197, March 2010.
[80] M. Stasiak, "Queuing systems for the Internet," IEICE Trans. Commun., vol.E99-B, no.6, pp.1224–1242, June 2016.
[81] M. Stasiak, "Blocking probability in a limited-availability group carrying mixture of different multichannel traffic streams," Annales des Télécommunications, vol.48, no.1-2, pp.71–76, 1993.
[82] M. Głąbowski and M. Stasiak, "Multi-rate model of the group of separated transmission links of various capacities," Telecommunications and Networking - ICT 2004, Lecture Notes in Computer Science, vol.3124, pp.1101–1106, Springer Berlin Heidelberg, 2004.
[83] J. Postel, "Transmission control protocol," RFC 793 (INTERNET STANDARD), Sept. 1981. Updated by RFCs 1122, 3168, 6093, 6528.
[84] H. Kaplan, K. Hedayat, N. Venna, P. Jones, and N. Stratton, "An extension to the session description protocol (SDP) and real-time transport protocol (RTP) for media loopback," RFC 6849 (Proposed Standard), Feb. 2013.
[85] J. Ott and C. Perkins, "Guidelines for extending the RTP control protocol (RTCP)," RFC 5968 (Informational), Sept. 2010.
[86] J.S. Kaufman, "Blocking in a completely shared resource environment with state dependent resource and residency requirements," Proc. Eleventh Annual Joint Conference of the IEEE Computer and Communications Societies on One World Through Communications (vol.3), IEEE INFOCOM'92, pp.2224–2232, Los Alamitos, CA, USA, IEEE Computer Society Press, 1992.
[87] J. Kaufman, "Blocking with retrials in a completely shared resource environment," J. Perform. Evaluation, vol.15, no.2, pp.99–116, 1992.
[88] I. Moscholios, M. Logothetis, and G. Kokkinakis, "Connection-dependent threshold model: A generalization of the Erlang multiple rate loss model," J. Perform. Evaluation, vol.48, no.1-4, pp.177–200, May 2002.
[89] M. Stasiak, M. Głąbowski, A. Wiśniewski, and P. Zwierzykowski, Modeling and Dimensioning of Mobile Networks, Wiley, 2011.
[90] E. Brockmeyer, H. Halstrom, and A. Jensen, "The life and works of A.K. Erlang," Acta Polytechnika Scandinavia, vol.6, no.287, 1960.

[91] M. Stasiak, "An approximate model of a switching network carrying mixture of different multichannel traffic streams," IEEE Trans. Commun., vol.41, no.6, pp.836–840, 1993.
[92] S. Hanczewski and M. Stasiak, "Modeling of systems with reservation by Erlang's ideal grading," Proc. 5th Polish-German Teletraffic Symposium, A. Feldmann, P.J. Kuhn, M. Pióro, and A. Wolisz, eds., pp.39–47, Berlin, Oct. 2008.

**Mariusz Głąbowski** received the M.Sc., Ph.D. and D.Sc. (Habilitation) degrees in telecommunication from the Poznan University of Technology, Poland, in 1997, 2001, and 2010, respectively. Since 1997 he has been working in the Department of Electronics and Telecommunications, Poznan University of Technology. He is engaged in research and teaching in the area of performance analysis and modeling of multiservice networks and switching systems. Prof. Mariusz Głąbowski is the author/co-author of 4 books, 12 book chapters and of over 150 papers which have been published in communication journals and presented at national and international conferences.

**Damian Kmiecik** is a Ph.D. student in telecommunications at Poznan University of Technology, Poland. He is the author, and co-author, of 10 scientific papers mostly related to analytical modelling of telecommunications systems.

**Maciej Stasiak** received M.Sc. and Ph.D. degrees in electrical engineering from the Institute of Communications Engineering, Moscow, Russia, in 1979 and 1984, respectively. In 1996 he received D.Sc. degree from Poznan University of Technology in electrical engineering. In 2006 he was nominated as full professor. Between 1983–1992 he worked in Polish industry as a designer of electronic and microprocessor systems. In 1992, he joined Poznan University of Technology, where he is currently Head of the Chair of Communications and Computer Networks at the Faculty of Electronics and Telecommunications. He is the author, and co-author, of more than 250 scientific papers and five books. He is engaged in research and teaching in the area of performance analysis and modelling of queuing systems, multiservice networks and switching systems. Since 2004 he has been actively carrying out research on modelling and dimensioning cellular 4G/5G networks.