Modeling of Multi-Service Queueing Systems With Traffic Overflow

Mariusz Głąbowski, *Senior Member, IEEE*, Sławomir Hanczewski, *Member, IEEE*, Damian Kmiecik, Maciej Stasiak, *Member, IEEE*, and Joanna Weissenberg

***Abstract*—This article proposes an analytical model of a multi-service hierarchical system with multi-service overflow traffic. To model the primary and secondary resources of this system, the state-dependent queue service discipline was used. In order to model the secondary resources with the dedicated queue for overflow traffic, the Hayward's approach was generalized and applied. To evaluate its accuracy, the results of analytical modeling were compared with the data obtained during the simulation experiments carried out in the study. Both the data presented in the article and the results obtained by the present authors in numerous comparative studies clearly indicate that the proposed model makes it possible to evaluate the values of the blocking probability with the accuracy that provides its reliable practical application at the stage of network dimensioning. The overflow mechanism has particular significance in networks with limited resources, such as mobile networks.**

***Index Terms*—Analytical modeling, queuing systems, overflow traffic, multi-service systems.**

I. Introduction

NETWORK operators have been using a wide variety of traffic management mechanisms over many years. These mechanisms allow us to use network resources in the most efficient manner possible, particularly in a situation where these resources are limited. One of the oldest mechanisms used for this purpose is traffic overflow. The mechanism was used for the first time in hierarchical telephone networks [1], [2], [3], [4], [5]. The concept of traffic overflow is very simple: calls that cannot be serviced by a given network resource can be transferred or redirected to another resource, which prevents new calls from being lost. According to the adopted system terminology, the network resources at which a new call arrives first are termed primary resources. Those resources to which calls are redirected are, in turn, called secondary or alternative resources [1].

Despite the passage of time, overflow mechanisms are still widely used in modern telecommunications and computer networks. In addition, the overflow mechanism is of particular significance in networks with limited resources, such as mobile networks. The development of technology used in mobile networks causes radio resources to require cooperation of wireless communication systems of different generations (from 2G to 5G) to guarantee their effective operation and sustainability.

Received 23 December 2025; revised 14 May 2026; accepted 16 May 2026. Date of publication 25 May 2026; date of current version 8 June 2026. This research is funded by the Ministry of Science and Higher Education: Grant 0313/SBAD/1316. The associate editor coordinating the review of this article and approving it for publication was V. Fodor. *(Corresponding author: Sławomir Hanczewski.)*

The authors are with the Faculty of Computing and Telecommunications, Poznan University of Technology, 60-695 Poznań, Poland (e-mail: slawomir.hanczewski@put.poznan.pl).

Digital Object Identifier 10.1109/TNSM.2026.3696894

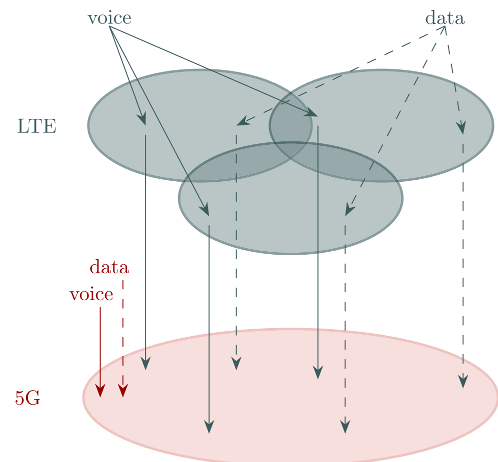

Fig. 1. Overflow mechanism in mobile networks.

The interworking of different access technologies becomes especially important in 5G networks, particularly in the context of resource congestion and fluctuating coverage. One of the mechanisms that ensures continuous access to services is dynamic traffic transfers between different radio interfaces, including from 5G to LTE. Such redirections, implemented in real time at both the access and the core layers, allow users to maintain active sessions even in the event of quality degradation or temporary unavailability of the chosen technology. These mechanisms are supported by functions of Self-Organizing Networks (SON), which allow flexible adjustment of user allocations to current network conditions. Figure 1 presents an exemplary scheme for realizing traffic overflow between cells in LTE and 5G mobile networks (it should be emphasized that each operator defines how the overflow mechanism operates within their specific network). In the case considered, traffic related to voice and data transmission is primarily offered to the LTE cell. If this cell cannot service a new call, the call is transferred (overflows) to the 5G system cell. Only when the 5G cell lacks sufficient resources to service the specific call is it lost. The overflow mechanism operates independently in each LTE cell. It is worth noting that, in the presented scenario, the 5G cell services both calls that cannot be serviced by the LTE cells and those calls that are offered to it directly (in traffic engineering, such calls are referred to as own traffic).

In the following, we first review the main applications and analytical models of overflow traffic, and then present a new model of a multi-service overflow system with queuing in both primary and secondary resources.

A. Related Work

Traffic overflow mechanisms have been extensively applied in modern telecommunications networks. Such an expanded network infrastructure makes it possible for the operator to transfer users' connections between the available resources of the same network (i.e., between cells from overloaded nodes to underutilized ones) as well as between network resources of different types (the so-called vertical connection transfer) [6], [7], [8], [9], [10]. This approach provides operational continuity of access to offered services. The traffic overflow mechanism is also used in self-organizing networks (self-optimizing, self-configuring networks) as an effective method to balance and equalize the load.

Similar overflow mechanisms are also applied beyond mobile networks. It is becoming increasingly common to use the cloud bursting mechanism [11], [12], [13], which allows for the transfer of the redundant traffic from a private cloud to public cloud resources, especially during sudden surges in demand for computing power, disk space, or bandwidth. Using global traffic balancing systems, such as Google Cloud Load Balancing [14], it is possible to dynamically manage traffic between private and public environments, taking into account the current state of congestion, latency and routing policies.

The traffic overflow mechanism is also used in packet networks [15], [16], [17], [18]. The mechanism, similarly as in the case of self-optimizing cellular networks, allows the link load to be equalized. A good example of the application of the traffic overflow mechanism in packet networks is the Scheme for Alternative Packet Overflow Routing (SAPOR) [19]. In this solution, the traffic overflow process is applied to new data streams only, making it possible to employ alternative routes in the routing process. The concept of overflow traffic is also used in data centers. Similarly as in the examples presented above, it makes it possible to equalize the load, which is of great importance in the context of energy saving.

The problem of modeling multi-service overflow systems with queuing was addressed for the first time in [20]. This work proposed a model of a multi-service overflow system with queueing for the primary resources to which Erlang traffic streams were offered. The problem of modeling multi-service systems in which queues with limited capacity are used both for primary and for secondary resources has not been addressed as yet.

Analytical modelling of complex service systems has a long tradition in teletraffic and performance analysis. Important contributions include methods for describing overflow traffic, e.g., [1], [3], [4], [16], [21], [22], [23], [24], [25], [26], [27], [28], [29], and [30] and the study of overflow processes in single-queue systems [3], [31], [32], [33], [34], [35], [36]. These works established methodological foundations for the analytical study of systems with overflow traffic.

A multi-service queuing model with state-dependent queue service discipline, called SD FIFO (State Dependent FIFO), has been proposed and analyzed in [37], [38], [39], and [40]. These works build on models of multi-service systems with elastic traffic [41], [42]. In SD FIFO model, the service process is approximated by a reversible Markov process, which ensure appropriate service stream values in the compression or queuing states. This reversible Markov process enforces an appropriate allocation of the server resources (the number of allocation units, AUs, of the individual classes being serviced) and, consequently, the numbers of AUs of each class waiting in the queue. This means that, in a given state of the process, the number of AUs of class i in service depends on the number of AUs of the remaining classes in service. A system defined in this way may be regarded as M virtual queuing systems, where M is the number of classes, in which the bit rates of the individual servers are variable and depend on the numbers of calls of each class that are being serviced and are waiting in the corresponding queues. Such an allocation of the resources in the server and in the queues, determined in each state on the basis of the Markov process, is consistent with the balanced fairness resource allocation algorithm, which provides state-dependent service of all call classes by admitting calls from the (virtual) queues in a balanced manner and prevents any single class from monopolizing the service.

This means that, in a situation where there are, for example, 10 AUs of class i followed by 3 AUs of class j in the queue, the algorithm will not first send all 10 AUs of class i to service and only then the 3 AUs of class j, but will take allocation units of classes i and j in parallel from the corresponding virtual queues. Such behavior of the algorithm is not related to prioritization, because the algorithm takes AUs of all classes in parallel (although not in equal numbers), without favoring any class. This means that, if in a given state the AUs of a given class are taken in the largest number, then in another state they may be taken in the smallest number, always as determined by the controlling Markov process. This behavior makes the SD FIFO model an idealized model and difficult to implement in practice, since a control device would have to recompute the allocation of the server resources on the basis of the Markov process after each state change. However, simulation studies show that the model provides a good approximation of the behavior of multi-service queueing systems with FIFO discipline.

B. Research Contribution

The present study goes beyond the scope of these works and enables the analysis of multi-service hierarchical telecommunication systems that include queuing for both primary and secondary resources and accommodate Erlang, Engset and Pascal traffic classes with a compression mechanism. The article proposes a model of a multi-service overflow system with queuing for its primary and secondary resources, to which multi-service Erlang, Engset and Pascal traffic streams are offered.

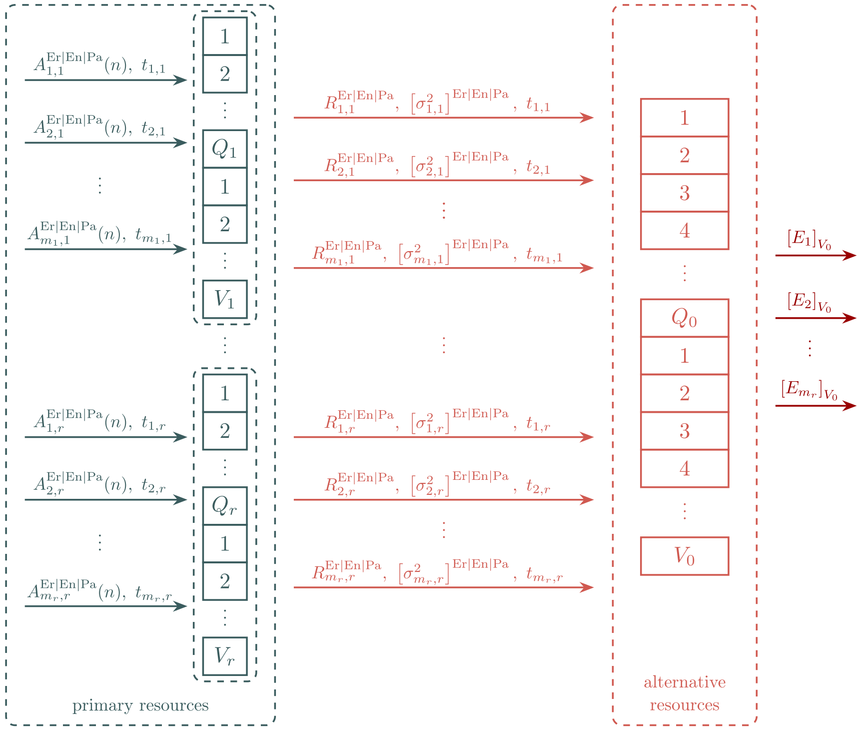

Fig. 2. Scheme of the multi-service overflow system with queues in the primary resources.

Building on the existing SD FIFO multi-service queuing model introduced in [37], [38], [39], and [40], this model is used as the analytical framework for describing the joint service and queuing process in the primary and secondary resources.

- A gap is identified in the analytical modelling of multi-service overflow systems in which limited-capacity queues are employed in both primary and secondary resources, extending earlier work on multi-service overflow systems with queuing only in primary resources [20].
- The existing SD FIFO multi-service queueing model [37], [38], [39], [40] is adapted to the analysis of hierarchical overflow systems with Erlang, Engset and Pascal traffic, providing a unified framework for joint service and queuing.
- Hayward's method, originally developed for single-service systems without additional traffic management mechanisms [4], is generalized to multi-service secondary resources with queues by introducing a mixture of multiple degenerate traffic streams and a collective peakedness factor determined as a weighted average of the per-class peakedness factors.
- The OverFlow-Q-ErEnPa method is developed as an iterative analytical framework for evaluating blocking probabilities and mean queue lengths in multi-service hierarchical overflow networks with Erlang, Engset and Pascal traffic at the dimensioning stage.

II. Traffic Overflow System in Systems With Queueing in Primary and Secondary Resources

The present article considers a multi-service overflow system, in which calls of different traffic classes in the case of the occupancy of the primary resources will first be directed to the queue, where they will be waiting for the primary resources to be released. The assumption is that the queues have limited capacity and it is only when they are fully occupied that calls will be directed to the system of secondary resources. When all secondary resources are occupied, calls will wait in a queue that is related to secondary resources. In the case of the occupancy of a queue related to the secondary resources, calls will be lost. Figure 2 shows the general scheme of the traffic overflow system with dedicated queues for the primary and secondary resources.

The capacities of servers and queues, as well as the other parameters of the analytical description introduced further on, are expressed in AUs. The allocation unit is defined as such a bit rate that the bit rates of all calls, as well as the server capacities, are integer multiples of it [43], [44], and [45].

The primary resources of the system presented in Figure 2 are composed of r component primary resources, where each resource s has the server capacity V_s and queue capacity Q_s, expressed in allocation units (AU). The secondary resource has the server capacity of V_0 AU and the queue capacity equal to Q_0 AU.

The primary resource s $(1 \leq s \leq r)$ is offered m_s traffic classes from the set M_s, including m_s^{Er} classes that belong to the resource M_s^{Er} Erlang traffic classes, m_s^{En} class that belong to the set M_s^{En} Engset traffic classes and m_s^{Pa} classes that belong to the set M_s^{Pa} of Pascal traffic classes, while $m_s^{\mathrm{Er}}+m_s^{\mathrm{En}}+m_s^{\mathrm{Pa}} = m_s$, $M_s^{\mathrm{Er}} \cup M_s^{\mathrm{En}} \cup M_s^{\mathrm{Pa}} = M_s$ and $m = \sum_{s=1}^{r} m_s$.

In Figure 2, the following notation for appropriate traffic intensities is adopted:

$A_{c,s}^{\mathrm{X}}(n)$– the average intensity of traffic of class c $(c \in M_s)$ type X (X = Er|En|Pa) offered to the primary resource s, in the occupancy state n AU,

$R^{\mathrm{X}}_{c,s}$– the average intensity of traffic of class c ($c \in M_s$) that overflows from the primary resource s, with the assumption that traffic of class c offered to the primary resources s is of type X (X = Er|En|Pa),

$[\sigma^2_{c,s}$ $^{\mathrm{X}}$–]variance of traffic of class c of type X that overflows from the resource s of the primary resource system,

$t_{c,s}$– the number of AUs demanded by a call of class c to be serviced in the primary resource s,

n– system occupancy state, defined as the total number of AUs occupied in the system by calls undergoing the service process.

Erlang traffic is independent of the state of the service system, whereas Engset and Pascal traffic is system state dependent, particularly on the number of already serviced calls of Engset and Pascal type [45]:

$$\forall_{1 \le s \le r} \forall_{i \in M_s^{\mathrm{Er}}} \forall_{0 \le n \le V_s} A^{\mathrm{Er}}_{i,s}(n) = A^{\mathrm{Er}}_{i,s} = \frac{\lambda_{i,s}}{\mu_{i,s}}, \quad (1)$$

$$\forall_{1 \le s \le r} \forall_{j \in M_s^{\mathrm{En}}} A^{\mathrm{En}}_{j,s}(n) = \alpha^{\mathrm{En}}_{j,s}\left[S^{\mathrm{En}}_{j,s} - n^{\mathrm{En}}_{j,s}(n)\right], \quad (2)$$

$$\forall_{1 \le s \le r} \forall_{l \in M_s^{\mathrm{Pa}}} A^{\mathrm{Pa}}_{l,s}(n) = \alpha^{\mathrm{Pa}}_{l,s}\left[S^{\mathrm{Pa}}_{l,s} + n^{\mathrm{Pa}}_{l,s}(n)\right], \quad (3)$$

where:

$A^{\mathrm{Er}}_{i,s}$– the average intensity of traffic of class i Erlang type offered to the primary resource s, regardless of the occupancy state n AU,

$\lambda^{\mathrm{Er}}_{i,s}$– the average intensity of calls of class i Erlang type,

$\mu^{\mathrm{Er}}_{i,s}$– the average service intensity for calls of class i of Erlang type in resource s,

$A^{\mathrm{En}}_{j,s}(n)$– the average intensity of traffic of class j Engset type offered to the primary resource s, in the occupancy state n AU,

$A^{\mathrm{Pa}}_{l,s}(n)$– the average intensity of traffic of class l Pascal type offered to the primary resource s, in the occupancy state n AU,

$\alpha^{\mathrm{En}}_{j,s}$– the average intensity of traffic of class j Engset type, generated by one free source in the resource s:

$$\alpha^{\mathrm{En}}_{j,s} = \frac{\gamma^{\mathrm{En}}_{j,s}}{\mu^{\mathrm{En}}_{j,s}}, \quad (4)$$

$\gamma^{\mathrm{En}}_{j,s}$– the average intensity of calls of class j Engset type, generated by one free source,

$\mu^{\mathrm{En}}_{j,s}$– the average service intensity for calls of class j Engset type in the resource s,

$\alpha^{\mathrm{Pa}}_{l,s}$– the average intensity of traffic of class l Pascal type, generated by one free source:

$$\alpha^{\mathrm{Pa}}_{l,s} = \frac{\gamma^{\mathrm{Pa}}_{l,s}}{\mu^{\mathrm{Pa}}_{l,s}}, \quad (5)$$

$\gamma^{\mathrm{Pa}}_{l,s}$– the average intensity of calls of class l Pascal type, generated by one single source,

$\mu^{\mathrm{Pa}}_{l,s}$– the average service intensity for calls of class l Pascal type in the resource s,

$S^{\mathrm{En}}_{j,s}$– the number of traffic sources of class j Engset type, related to the primary resource s,

$S^{\mathrm{Pa}}_{l,s}$– the number of traffic sources of class l Pascal type, related to the primary resource s,

$n^{\mathrm{En}}_{j,s}(n)$– the average values of the number of active sources of class j Engset type in the primary resource s in the occupancy state n AU,

$n^{\mathrm{Pa}}_{l,s}(n)$– the average values of the number of active sources of class l Pascal type in the primary resource s in the occupancy state n AU.

The assumption adopted in this paper that the a mixture of Erlang, Engset and Pascal traffic is offered to the system covers all possible behaviors of the traffic streams offered to the system. In the case of Engset traffic, the call intensity decreases with increasing occupancy state of the system. For Erlang traffic, the call intensity is constant and does not depend on the occupancy state of the system. For Pascal traffic, the call intensity increases with increasing occupancy state of the system.

If a call of class c cannot be serviced by the primary resource s, it is redirected to a queue of a complement resource s with the capacity Q_s AU. When this queue has no space available, i.e. $t_{c,s}$ AU, a call of class c will be redirected to the secondary resource with the capacity V_0 AU. If a call of class c cannot be serviced by the secondary resource with the capacity V_0, it is redirected to a queue in the secondary resource with the capacity Q_0 AU. If the call cannot be stored due to lack of free $t_{c,0}$ AU in the resources of the queue with the capacity Q_0, the call will be lost.

III. Models of Resources in Overflow System

The model of a traffic overflow system with queueing in the primary and secondary resources is composed of a model of the primary resources, a model to determine the average value and variance of individual classes of overflow traffic and a model of the secondary resources. These three models make it possible to determine all important characteristics of the overflow system, in particular the blocking probability for call streams that are offered to the system. To model each of the primary resources, a multi-service SD FIFO model [38] was used, whereas to determine the parameters of traffic that overflows from the primary resources, the method developed in [46] was used. To model the secondary resources with the dedicated queue, a new SD FIFO model was proposed that takes into account handling of traffic streams with the peakedness factor different than one.

A. Model of Component Primary Resources

In this article, to model component primary resources with queueing a recursive occupancy distribution in the SD FIFO queueing system is used [38], that, in line with the adopted notation for this article, can be written for Erlang, Engset and Pascal traffic as follows:

$$[P_n]_{V_s+Q_s} = \frac{1}{\kappa}\left\{\sum\nolimits_{i\in M_s^{\mathrm{Er}}} A_{i,s}^{\mathrm{Er}} t_{i,s}\left[P_{n-t_{i,s}}\right]_{V_s+Q_s} + \sum\nolimits_{j\in M_s^{\mathrm{En}}} A_{j,s}^{\mathrm{En}}(n-t_{j,s}) t_{j,s}\left[P_{n-t_{j,s}}\right]_{V_s+Q_s} + \sum\nolimits_{l\in M_s^{\mathrm{Pa}}} A_{l,s}^{\mathrm{Pa}}(n-t_{l,s}) t_{l,s}\left[P_{n-t_{l,s}}\right]_{V_s+Q_s}\right\}, \quad (6)$$

where $\kappa = \min(n, V_s)$, n $(0 \leq n \leq V_s + Q_s)$ is the number of AU occupied by calls that are serviced and queued in the component resource s, whereas the distribution $[P_n]_{V_s+Q_s}$ determines the occupancy probability n AU in the primary resource s with the server capacity V_s and queue capacity Q_s.

The dependence between the value of offered traffic and the number of serviced calls of a given class of Engset traffic (Formula (2)) and Pascal traffic (Formula (3)) makes it necessary to apply the iteration procedure to determine the occupancy distribution in the considered full-availability primary resources on the basis of Formula (6). Exactly as in the solution presented in [47], it is possible to make the values of intensity of offered Erlang-Engset-Pascal traffic dependent on the occupancy state of the system, and in consequence to determine the occupancy distribution in the system with a finite number of traffic sources, using Formula (6). The method [47] is based on the assumption that the average number of calls of a given class serviced in a given occupancy state in a group with a finite number of traffic sources is similar to the average number of calls serviced in this state for the case of an infinite number of traffic sources. This particular approach assumes that the number $n_{j,s}^{\mathrm{En}}(n)$ and $n_{l,s}^{\mathrm{Pa}}(n)$ of active Engset sources of class j and Pascal sources of class l in the occupancy state n AUs in the resource s can be approximated by the value of the parameters of the service streams $y_{j,s}^{\mathrm{En}}(n)$ and $y_{l,s}^{\mathrm{Pa}}(n)$ that determine the average number of serviced calls of classes j and l in state n of occupied AUs in the resource s:

$$n_{j,s}^{\mathrm{En}}(n) \approx y_{j,s}^{\mathrm{En}}(n), \quad n_{l,s}^{\mathrm{Pa}}(n) \approx y_{l,s}^{\mathrm{Pa}}(n). \quad (7)$$

According to the adopted assumption, the parameters for service streams $y_{c,s}^{\mathrm{X}}(n)$ can be determined on the basis of Formula [47]:

$$y_{c,s}^{\mathrm{X}}(n) = \begin{cases} A_{c,s}^{\mathrm{X}}(n-t_{c,s})\dfrac{\left[P_{n-t_{c,s}}\right]_{V_s+Q_s}}{[P_n]_{V_s+Q_s}} & \text{for } 0 \leq n \leq V_s + Q_s, \\ 0 & \text{in other cases,} \end{cases} \quad (8)$$

in which, the probabilities $[P_n]_{V_s+Q_s}$ will be determined using the recurrence (6), with the initial assumption that the value of offered traffic is independent of the number of active Engset and Pascal traffic sources[1]:

$$\forall_{1\leq s\leq r}\forall_{j\in M_s^{\mathrm{En}}}\, A_{j,s}^{\mathrm{En}}(n) = \alpha_{j,s}^{\mathrm{En}} S_{j,s}^{\mathrm{En}}, \quad (9)$$

$$\forall_{1\leq s\leq r}\forall_{l\in M_s^{\mathrm{Pa}}}\, A_{l,s}^{\mathrm{Pa}}(n) = \alpha_{l,s}^{\mathrm{Pa}} S_{l,s}^{\mathrm{Pa}}. \quad (10)$$

By having the traffic values: A_i^{Er} (Formula (1)) and $A_j^{\mathrm{En}}(n)$ (Formula (2)) and $A_l^{\mathrm{Pa}}(n)$ (Formula (3)), it is possible to determine the distribution of probabilities based on the distribution (6). The state probabilities obtained on the basis of Equation (6) provide input data for the next iteration, in which, according to the assumed adopted, the value of the parameter $y_{c,s}^{\mathrm{X}}(n)$ that determines the number $n_{c,s}^{\mathrm{X}}(n)$ of traffic sources serviced of class c type X in the occupancy state n AU of the primary resource s can be determined. To determine the average number of serviced calls (occupied traffic sources) of individual classes, denote the occupancy distribution determined in the system with Erlang–Engset–Pascal traffic in the ι-iteration by $[P_n]_{V_s+Q_s}^{[\iota]}$ and the average number of serviced calls of class c, type X, determined in Step ι by $\left[y_{c,s}^{\mathrm{X}}(n)\right]^{[\iota]}$. Then,

$$\left[n_{c,s}^{\mathrm{X}}(n)\right]^{[\iota+1]} = \left[y_{c,s}^{\mathrm{X}}(n)\right]^{[\iota+1]} = \begin{cases} \dfrac{\left[A_{c,s}^{\mathrm{X}}(n-t_{c,s})\right]^{[\iota+1]}\left[P_{n-t_c}\right]_{V_s+Q_s}^{[\iota]}}{[P_n]_{V_s+Q_s}^{[\iota]}} & \text{for } 0 \leq n \leq V_s + Q_s, \\ 0 & \text{in other cases.} \end{cases} \quad (11)$$

The iteration process terminates when the required accuracy is obtained.

The method to determine the occupancy distribution and the blocking probability in the primary resource s with the server capacity V_s and queue capacity Q_s with Erlang-Engset-Pascal traffic is presented below:

1) Determination of iteration step $\iota = 0$.
2) Determination of initial values $\left[n_{j,s}^{\mathrm{En}}(n)\right]^{[\iota]}$, $\left[n_{l,s}^{\mathrm{Pa}}(n)\right]^{[\iota]}$:
   $\forall_{1\leq j\leq m_s^{\mathrm{En}}}\forall_{0\leq n\leq V_s}\ \left[n_{j,s}^{\mathrm{En}}(n)\right]^{[\iota]} = 0,$
   $\forall_{1\leq l\leq m_s^{\mathrm{Pa}}}\forall_{0\leq n\leq V_s}\ \left[n_{l,s}^{\mathrm{Pa}}(n)\right]^{[\iota]} = 0.$
3) Increase in the iteration step: $\iota = \iota + 1$.
4) Determination of the intensity value of the Engset traffic of class j and class l Pascal traffic in ι-iterations–Formulas (2) and (3).
5) Determination of state probabilities $[P_n]_{V_s+Q_s}^{[\iota]}$–Formula (6).
6) Determination of the average number of serviced Engset calls $\left[n_{j,s}^{\mathrm{En}}(n)\right]^{[\iota]}$ and Pascal calls $\left[n_{l,s}^{\mathrm{Pa}}(n)\right]^{[\iota]}$ in the occupancy state n AU in the resource s–Formula (11).

[1]This assumption means that in each state n the number of AUs occupied by calls of an Engset stream of class j and Pascal of class l is equal to the number of AUs occupied by corresponding equivalent Erlang streams.

7) Repetition of steps 3–6 until the required accuracy ϵ of the iteration process is achieved:

$$\forall_{1\le s\le r}\forall_{0\le n\le V}\left|\frac{\left[n_{j,s}^{\mathrm{En}}(n)\right]^{[\iota-1]}-\left[n_{j,s}^{\mathrm{En}}(n)\right]^{[\iota]}}{\left[n_{j,s}^{\mathrm{En}}(n)\right]^{[\iota]}}\right|\le\epsilon, \quad (12)$$

$$\forall_{1\le s\le r}\forall_{0\le n\le V}\left|\frac{\left[n_{l,s}^{\mathrm{Pa}}(n)\right]^{[\iota-1]}-\left[n_{l,s}^{\mathrm{Pa}}(n)\right]^{[\iota]}}{\left[n_{l,s}^{\mathrm{Pa}}(n)\right]^{[\iota]}}\right|\le\epsilon. \quad (13)$$

The blocking probability in this system results from the finite capacity of the queue, and for calls of class c can be determined by the following Formula:

$$\left[E_c^{\mathrm{X}}\right]_{V_s+Q_s}=\sum_{n=V_s+Q_s-t_{c,s}+1}^{V_s+Q_s}[P_n]_{V_s+Q_s}\quad \text{for } c\in M_s^{\mathrm{X}}. \quad (14)$$

Formula (14) determines the aggregated sum of blocking states for calls of class c, that is, those states in which the number of free AUs in the queue is lower than the number of AUs required to set up a connection of class c. The service discipline of the SD FIFO queue corresponds to the resource allocation for each class of traffic offered based on the balanced fairness algorithm.

This model approximates well different algorithms for resource division for serviced calls, in particular the proportional algorithm [38].

IV. Queue Parameters

The knowledge of the occupancy distribution in primary resources allows us to determine the queue parameters based on the methodology presented in [38]. According to this methodology, the average number of calls of class c of type X present in the queuing system s ($x_{c,s}^{\mathrm{X}}(n)$) is approximated by the average number of AUs occupied by calls of class c of type X serviced in a fictitious loss system s' (without a queue) with a server capacity V' equal to the sum of the server and a queue capacities of the primary system $V'=V_s+Q_s$. In accordance with the adopted method for calculating queue parameters, they can be determined as follows:

The average queue length for class c of type X in the occupancy state n of the primary resource s is equal to:

$$q_{c,s}^{\mathrm{X}}(n)=[x_{c,s}^{\mathrm{X}}(n)-y_{c,s}^{\mathrm{X}}(n)]t_{i,s}. \quad (15)$$

The average queue length for class c of type X in the primary resource s is equal to:

$$q_{c,s}^{\mathrm{X}}=\sum_{n=V_s+1}^{V_s+Q_s}[x_{c,s}^{\mathrm{X}}(n)-y_{c,s}^{\mathrm{X}}(n)]t_{i,s}[P_n]_{V_s+Q_s}, \quad (16)$$

where:

- $y_{c,s}^{\mathrm{X}}(n)$ defines the average number of calls of class c of type X serviced in the primary resource s with the capacity V_s AU and the queue Q_s AU that can be determined on the basis of Equation (8).
- $x_{c,s}^{\mathrm{X}}(n)$ defines the average number of calls of class c of type X in the resource s' with the capacity V_s' which is composed of real resources with the capacity equal to the sum $V_s'=V_s+Q_s$.

Since in the adopted model, the parameter $x_{c,s}^{\mathrm{X}}(n)$ is approximated by the average number of calls of class c of type X serviced in a fictitious loss system s' (without a queue), with a server capacity V' equal to the sum of the server and queue capacities of the primary system $V'=V_s+Q_s$. Therefore, the value of this parameter can be determined on the basis of Equation (8), which in this case can be written in the following from:

$$x_{c,s}^{\mathrm{X}}(n)=\begin{cases}A_{c,s}^{\mathrm{X}}(n-t_{c,s})\dfrac{\left[P_{n-t_{c,s}}\right]_{V'}}{[P_n]_{V'}} & \text{for } 0\le n\le V',\\ 0 & \text{in other cases.}\end{cases} \quad (17)$$

In this case, the occupancy distribution $\left[P_{n-t_{c,s}}\right]_{V'}$ can be determined on the basis of the recursive dependence (6), with the assumption that the capacity is $V_s'=V_s+Q_s$ and the buffer capacity is equal to zero. For this system we get:

$$\begin{aligned}[P_n]_{V_s'}=\frac{1}{n}\Bigg\{&\sum_{i\in M_s^{\mathrm{Er}}}A_{i,s}^{\mathrm{Er}}t_{i,s}\left[P_{n-t_{i,s}}\right]_{V_s'}\\&+\sum_{j\in M_s^{\mathrm{En}}}A_{j,s}^{\mathrm{En}}(n-t_{j,s})t_{j,s}\left[P_{n-t_{j,s}}\right]_{V_s'}\\&+\sum_{l\in M_s^{\mathrm{Pa}}}A_{l,s}^{\mathrm{Pa}}(n-t_{l,s})t_{l,s}\left[P_{n-t_{l,s}}\right]_{V_s'}\Bigg\}. \end{aligned}\quad (18)$$

By having the average number of calls that are in the queue of individual classes in the resource s, the average total queue length can be determined from the following Formula:

$$q_s=\sum_{c=1}^{m_s}q_{c,s}^{\mathrm{X}}, \quad (19)$$

Based on Little's Law, the average waiting time can be determined using the following formula:

$$T_{c,s}^{\mathrm{X}}=\frac{q_{c,s}^{\mathrm{X}}}{\lambda_{c,s}^{\mathrm{X}}}. \quad (20)$$

The resource whose average number of calls has been determined as $x_{c,s}^{\mathrm{X}}(n)$ in each state n occupies exactly n AUs. In the case of a resource whose average number of calls has been determined as $y_{c,s}^{\mathrm{X}}(n)$, the maximum value of the occupied AUs is V_s, while $(0\le n\le V')$. The inclusion of this particular property, along with the above formulas, allows us to determine the average queue length of a queue that includes calls of all classes using the following formula:

$$q_s=\sum_{n=V_s+1}^{V_s+Q_s}(n-V_s)[P_n]_{V_s+Q_s}. \quad (21)$$

A. Model of Traffic That Overflows From Primary Resources

1) Decomposition of Primary Resources: To determine the traffic parameters of the traffic that overflows from the primary resources, we will apply a modification of the method proposed in [48]. The proposed modification is based on the decomposition of each of the component primary resource s with the server capacity V_s and queue capacity Q_s into m_s fictitious resources without queues, each with the capacity $v_{c,s}^{\mathrm{X}}$,

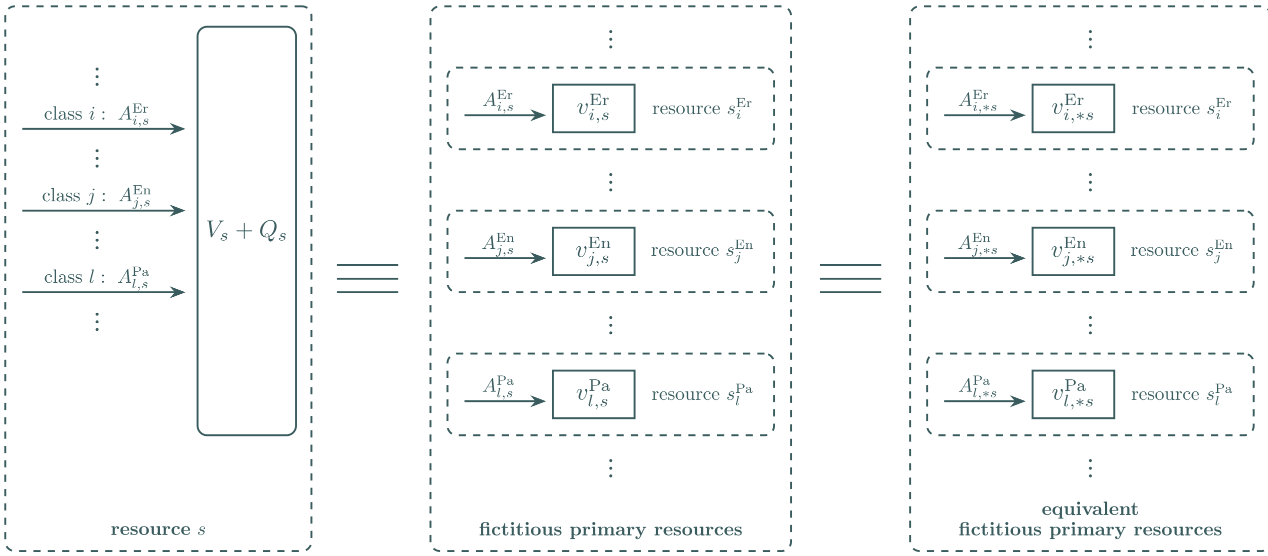

Fig. 3. Decomposition of the primary resource s into m_s fictitious primary resources and m_s equivalent fictitious primary resources.

where X denotes the type of traffic (Erlang, Engset, or Pascal). Then, each fictitious primary resource s_c^{X} is converted into an equivalent fictitious primary resource. Each equivalent fictitious primary resource s_c^{X} (the number of equivalent fictitious primary resources is equal to the number of fictitious primary resources) is characterized by the equivalent capacity $v^{\text{X}}_{c,*s}$ and the equivalent Erlang traffic offered with intensity $A^{\text{X}}_{c,*s}$. These parameters are selected in such a way that traffic that overflows from the equivalent fictitious primary resource s_c^{X} has exactly the same parameters (average value and variance) as traffic that overflows from the fictitious primary resource s_c^{X}.

The need for decomposition results from the fact that each primary resource s transports multi-service traffic streams, and, consequently, the direct determination of variance for traffic of individual classes on the basis of Riordan's formulas [1] is not possible. Each fictitious primary resource s_c^{X} services exclusively calls of one class c, which in turn makes the application of Riordan's formulas [1] possible to determine the variance of traffic of class c that overflows from a fictitious primary resource s_c^{X}. To determine the capacity of a fictitious primary resource s_c^{X}, we can assume that the blocking probability $\left[E_c^{\text{X}}\right]_{v^{\text{X}}_{c,s}}$ for calls of class c type X in the fictitious primary resource s_c^{X} with capacity $v^{\text{X}}_{c,s}$ is exactly the same as the blocking probability $\left[E_c^{\text{X}}\right]_{V_s+Q_s}$ of calls of this class in the primary resource s with the capacity $V_s + Q_s$:

$$\forall_{c\in M_s^{\text{X}}}\ \left[E_c^{\text{X}}\right]_{v^{\text{X}}_{c,s}} = \left[E_c^{\text{X}}\right]_{V_s+Q_s}. \tag{22}$$

The method for the decomposition of the primary resource s into m_s fictitious primary resources and m_s equivalent fictitious primary resources is presented in Figure 3.

Note that, on the basis of (14) it is possible to determine the blocking probabilities for individual traffic classes that are offered to the primary resource s. The result of the adopted assumption–expressed by Formula (22)–that the obtained probabilities are exactly the same as the probabilities in the corresponding fictitious primary resources. Since each fictitious primary resource services only one traffic class, therefore, the capacities $v^{\text{X}}_{c,s}$ of the decomposed fictitious primary resources can be determined on the basis of single-service full-availability models for Erlang, Engset and Pascal traffic. In these models, it is possible to determine the blocking probability as the probability of the occupancy of all AUs:

$$\forall_{c\in M_s^{\text{X}}}\ \left[E_c^{\text{X}}\right]_{v^{\text{X}}_{c,s}} = \left[E_c^{\text{X}}\right]_{V_s+Q_s} = \left[P_{v^{\text{X}}_{c,s}}\right]_{v^{\text{X}}_{c,s}}. \tag{23}$$

Thus, in the case of Erlang traffic, the capacity of fictitious primary resources can be determined on the basis of the second Erlang formula:

$$\left[E_i^{\text{Er}}\right]_{v^{\text{Er}}_{i,s}} = \left[P_{v^{\text{Er}}_{i,s}}\right]_{v^{\text{Er}}_{i,s}} = E_{v^{\text{Er}}_{i,s}}\left(A^{\text{Er}}_{i,s}\right) = \frac{\dfrac{\left(A^{\text{Er}}_{i,s}\right)^{v^{\text{Er}}_{i,s}}}{\left(v^{\text{Er}}_{i,s}\right)!}}{\sum\limits_{n=0}^{v^{\text{Er}}_{i,s}} \dfrac{\left(A^{\text{Er}}_{i,s}\right)^n}{n!}}. \tag{24}$$

In Formula (24), the parameters $\left[E_i^{\text{Er}}\right]_{v^{\text{Er}}_{i,s}}$ and $A^{\text{Er}}_{i,s}$ are known, therefore, on the basis of these parameters it is possible to determine the parameter $v^{\text{Er}}_{i,s}$. In a similar way, the capacities of fictitious resources to which single-service Engset and Pascal traffic is offered, can be determined on the basis of appropriate formulas that determine the blocking probability in the Engset and Pascal model:

$$\left[E_j^{\text{En}}\right]_{v^{\text{En}}_{j,s}} = \left[P_{v^{\text{En}}_{j,s}}\right]_{v^{\text{En}}_{j,s}} = \frac{\binom{S^{\text{En}}_{j,s}}{v^{\text{En}}_{j,s}} \left(\alpha^{\text{En}}_{j,s}\right)^{v^{\text{En}}_{j,s}}}{\sum\limits_{n=0}^{v^{\text{En}}_{j,s}} \binom{S^{\text{En}}_{j,s}}{n} \left(\alpha^{\text{En}}_{j,s}\right)^n}, \tag{25}$$

$$\left[E_l^{\text{Pa}}\right]_{v^{\text{Pa}}_{l,s}} = \left[P_{v^{\text{Pa}}_{l,s}}\right]_{v^{\text{Pa}}_{l,s}} = \frac{\binom{S^{\text{Pa}}_{l,s} + v^{\text{Pa}}_{l,s} - 1}{v^{\text{Pa}}_{l,s}} \left(\alpha^{\text{Pa}}_{l,s}\right)^{v^{\text{Pa}}_{l,s}}}{\sum\limits_{n=0}^{v^{\text{Pa}}_{l,s}} \binom{S^{\text{Pa}}_{l,s} + n - 1}{n} \left(\alpha^{\text{Pa}}_{l,s}\right)^n}. \tag{26}$$

Formulas (24)–(26) apply to single-service systems in which admission of a new call indicates the occupancy of one AU.

The next element of the decomposition of the primary resource s is the exchange (conversion) of the fictitious primary resource s_c^{X} into the equivalent fictitious primary resource s_c^{X}. The need for such an exchange is also related to the possibility of determining the variance of traffic that overflows from the equivalent primary resource s_c^{X} on the basis of Riordan's formulas [1]. Since these formulas make it possible to determine the variance exclusively in the case where the offered traffic is Erlang traffic, then the change of the fictitious primary resource s_c^{X} to the equivalent fictitious primary resource s_c^{X} is related to the change of Engset and Pascal traffic into equivalent Erlang traffic.

Let the variance of traffic of class c type X offered to the resource s be denoted by $[\sigma^2_{c,\Delta s}]^{\mathrm{X}}$. Equivalent Erlang traffic $A^{\mathrm{En}}_{j,*s}$ and $A^{\mathrm{Pa}}_{l,*s}$ is such traffic offered to a number of fictitious additional resources with capacities $\Delta v^{\mathrm{En}}_{j,s}$, $\Delta v^{\mathrm{Pa}}_{l,s}$ such that the traffic that overflows from these resources is equal, in terms of average value and variance, to Engset traffic $\left(A^{\mathrm{En}}_{j,s}, [\sigma^2_{j,\Delta s}]^{\mathrm{En}}\right)$ and Pascal traffic $\left(A^{\mathrm{Pa}}_{l,s}, [\sigma^2_{l,\Delta s}]^{\mathrm{Pa}}\right)$. Note that in the notation of equivalent traffic, e.g. in $A^{\mathrm{En}}_{j,*s}$, the symbol En, which indicates the original nature of this traffic, is retained in the upper index. The "asterisk" introduced in the lower index denotes that it is already equivalent to Erlang traffic. In the case of original Erlang traffic, we have: $A^{\mathrm{Er}}_{i,*s} = A^{\mathrm{Er}}_{i,s}$, $[\sigma^2_{i,\Delta s}]^{\mathrm{Er}} = A^{\mathrm{Er}}_{i,s}$ and $\Delta v^{\mathrm{Er}}_{i,s} = 0$.

The parameters of the equivalent fictitious primary resource s_c^{X} can be determined on the basis of the Equivalent Random Traffic method (ERT) [49]. The average values and variances of Erlang, Engset and Pascal traffic can be determined using the following dependencies [49]:

- Erlang traffic:

$$A^{\mathrm{Er}}_{i,s},\ [\sigma^2_{i,\Delta s}]^{\mathrm{Er}} = A^{\mathrm{Er}}_{i,s}, \tag{27}$$

- Engset traffic:

$$A^{\mathrm{En}}_{j,s} = S^{\mathrm{En}}_{j,s}\frac{\alpha^{\mathrm{En}}_{j,s}}{1+\alpha^{\mathrm{En}}_{j,s}},\ \left[\sigma^2_{j,\Delta s}\right]^{\mathrm{En}} = S^{\mathrm{En}}_{j,s}\frac{\alpha^{\mathrm{En}}_{j,s}}{\left(1+\alpha^{\mathrm{En}}_{j,s}\right)^2}, \tag{28}$$

- Pascal traffic:

$$A^{\mathrm{Pa}}_{l,s} = S^{\mathrm{Pa}}_{l,s}\frac{\alpha^{\mathrm{Pa}}_{l,s}}{1-\alpha^{\mathrm{Pa}}_{l,s}},\ \left[\sigma^2_{l,\Delta s}\right]^{\mathrm{Pa}} = S^{\mathrm{Pa}}_{l,s}\frac{\alpha^{\mathrm{Pa}}_{l,s}}{\left(1-\alpha^{\mathrm{Pa}}_{l,s}\right)^2}. \tag{29}$$

The average value R and variance σ^2 for traffic that overflows from the primary resource with the capacity V, to which Erlang traffic A is offered can be determined, according to the ERT method, on the basis of Riordan's formulas. In the traffic conversion scheme considered, the group with capacity $V = \Delta v^{\mathrm{X}}_{c,s}$ offers equivalent Erlang traffic with intensity $A = A^{\mathrm{X}}_{c,*s}$. From this group, traffic with the average value $R = A^{\mathrm{X}}_{c,s}$ and variance $\left[\sigma^2_{c,\Delta s}\right]^{\mathrm{X}}$ overflows. Therefore, in the notation adopted for the present work, Riordan's formulas can be written as follows:

$$A^{\mathrm{X}}_{c,s} = A^{\mathrm{X}}_{c,*s}E_{\Delta v^{\mathrm{X}}_{c,s}}(A^{\mathrm{X}}_{c,*s}), \tag{30}$$

$$\left[\sigma^2_{c,\Delta s}\right]^{\mathrm{X}} = A^{\mathrm{X}}_{c,s}\left(\frac{A^{\mathrm{X}}_{c,*s}}{\Delta v^{\mathrm{X}}_{c,s}+1-A^{\mathrm{X}}_{c,*s}+A^{\mathrm{X}}_{c,s}}+1-A^{\mathrm{X}}_{c,s}\right). \tag{31}$$

Formulas (30) and (31) allow the pair of parameters ($A^{\mathrm{X}}_{c,*s}$, $\Delta v^{\mathrm{X}}_{c,s}$) to be determined on the basis of known values of the pair of parameters ($A^{\mathrm{X}}_{c,s}$, $\left[\sigma^2_{c,\Delta s}\right]^{\mathrm{X}}$), which depending on the traffic under consideration, is described by Formulas (27)–(29). After determining the values of the parameters ($A^{\mathrm{X}}_{c,*s}$, $\Delta v^{\mathrm{X}}_{c,s}$), the capacity of the equivalent fictitious primary resource s_c^{X} can be evaluated, i.e. the parameter $v^{\mathrm{X}}_{c,*s}$ [1]:

$$v^{\mathrm{X}}_{c,*s} = v^{\mathrm{X}}_{c,s} + \Delta v^{\mathrm{X}}_{c,s}. \tag{32}$$

Note that the pair of parameters ($A^{\mathrm{X}}_{c,*s}$, $v^{\mathrm{X}}_{c,*s}$) for each equivalent fictitious primary resource s_c^{X} determines the Erlang model for the full-availability group [50].

2) Parameters of Overflow Traffic: Overflow traffic of class c that overflows from the primary resource s is equivalent to traffic that overflows from the equivalent fictitious primary resource s_c and will be characterized by two parameters: the average value of the intensity of the traffic $R^{\mathrm{X}}_{c,s}$ and variance $[\sigma^2_{c,s}]^{\mathrm{X}}$. Since the equivalent fictitious primary resource s_c^{X} is determined by the single-service Erlang model, therefore, the parameters $R^{\mathrm{X}}_{c,s}$ and $[\sigma^2_{c,s}]^{\mathrm{X}}$ can be determined on the basis of Riordan's formulas that in the adopted notation will be written as follows:

$$R^{\mathrm{X}}_{c,s} = A^{\mathrm{X}}_{c,*s}E_{v^{\mathrm{X}}_{c,*s}}(A^{\mathrm{X}}_{c,*s}), \tag{33}$$

$$\left[\sigma^2_{c,s}\right]^{\mathrm{X}} = R^{\mathrm{X}}_{c,s}\left(\frac{A^{\mathrm{X}}_{c,*s}}{v^{\mathrm{X}}_{c,*s}+1-A^{\mathrm{X}}_{c,*s}+R^{\mathrm{X}}_{c,s}}+1-R^{\mathrm{X}}_{c,s}\right). \tag{34}$$

The peakedness factor $Z^{\mathrm{X}}_{c,s}$ for traffic of class c that overflows from the equivalent fictitious primary resource s_c^{X} is defined as the ratio between the variance and the average value:

$$Z^{\mathrm{X}}_{c,s} = \frac{\left[\sigma^2_{c,s}\right]^{\mathrm{X}}}{R^{\mathrm{X}}_{c,s}}. \tag{35}$$

B. Model of Secondary Resources

As a result of the decomposition operation and the replacement of primary resources with equivalent fictitious primary resources, the overflow traffic streams, initially generated according to the Erlang, Engset and Pascal distribution, are characterized by the following parameters: the average value of traffic intensity $R^{\mathrm{X}}_{c,s}$, variance $[\sigma^2_{c,s}]^{\mathrm{X}}$ and the number of AUs t_c necessary for a given connection to be set up. These streams are offered to the system of secondary resources with the server capacity V_0 AU to which the queue with the capacity limited to Q_0 AU is related.

To model the occupancy distribution in multi-service secondary resources with queueing, to which a mixture of overflow traffic is offered, Harward's approach, proposed in [46], can be generalized. The generalization is based on the division of the parameters of the secondary resource system (traffic intensity and capacity) by the peakedness factor (coefficient) of the traffic offered. As the result, the occupancy distribution in the system, of secondary resources with queueing, can be written in the adopted notation as follows:

$$[P_n]_{\frac{V_0}{Z}+Q_0} = \frac{1}{\kappa}\Bigg\{\sum_{s=1}^{r}\sum_{i\in M_s^{\mathrm{Er}}}\frac{R_{i,s}^{\mathrm{Er}}}{Z_{i,s}^{\mathrm{Er}}}t_{i,s}\left[P_{n-t_{i,s}}\right]_{\frac{V_0}{Z}+Q_0} + \sum_{s=1}^{r}\sum_{j\in M_s^{\mathrm{En}}}\frac{R_{j,s}^{\mathrm{En}}}{Z_{j,s}^{\mathrm{En}}}(n-t_{j,s})t_{j,s}\left[P_{n-t_{j,s}}\right]_{\frac{V_0}{Z}+Q_0} + \sum_{s=1}^{r}\sum_{l\in M_s^{\mathrm{Pa}}}\frac{R_{l,s}^{\mathrm{Pa}}}{Z_{l,s}^{\mathrm{Pa}}}(n-t_{l,s})t_{l,s}\left[P_{n-t_{l,s}}\right]_{\frac{V_0}{Z}+Q_0}\Bigg\}, \quad (36)$$

where $\kappa = \min(n, V_0)$, n $(0 \le n \le \frac{V_0}{Z} + Q_0)$ is the number of AUs occupied by calls that are in the secondary resource (serviced and queued), whereas the distribution $[P_n]_{\frac{V_0}{Z}+Q_0}$ determines the occupancy probability n AU in the secondary resource with the server capacity V_0 and queue capacity Q_0.

sIn Formula (36), the parameters of traffic that overflows from the primary resources are determined by Formulas (33)–(35). The parameter Z is the aggregated peakedness factor. The introduction of this notion results from the necessity, according to Hayward's concept, to normalize the total capacity of the system of secondary resources $\frac{V_0}{Z} + Q_0$ to which an appropriate mixture of traffic is offered that overflows from each of r primary resources. The aggregated peakedness factor can be approximately determined on the basis of the weighted average of the peakedness factors for individual traffic classes that are offered to the secondary resources system:

$$Z = \sum_{s=1}^{r}\sum_{c=1}^{m_s} Z_{c,s}^{\mathrm{X}}\frac{R_{c,s}^{\mathrm{X}}t_{c,s}}{\sum_{k=1}^{r}\sum_{l=1}^{m}R_{l,k}^{\mathrm{X}}t_{l,k}}. \quad (37)$$

The blocking probability for traffic of class c in the secondary resources with queueing (regardless of from which primary resources this traffic overflows) is equal to:

$$\left[E_c\right]_{V_0+Q_0} = \sum_{n=\frac{V_0}{Z}+Q_0-t_c+1}^{\frac{V_0}{Z}+Q_0}\left[P_n\right]_{\frac{V_0}{Z}+Q_0}. \quad (38)$$

C. Parameters of Queues of Secondary Resources

Similarly as in the case of the primary resources, it is possible to determine the parameters of the queues for the secondary resources. This determination will be based on the occupancy distribution determined by Formula (36). Using the methodology presented to determine the queue parameters for the primary resources, the average queue length for calls of class c type X in the secondary resources can be determined on the basis of the following dependence:

$$q_{c,0}^{\mathrm{X}} = \sum_{n=\frac{V_0}{Z}+1}^{\frac{V_0}{Z}+Q_0} q_{c,0}^{\mathrm{X}}(n)[P_n]_{\frac{V_0}{Z}+Q_0}, \quad (39)$$

where:

- $q_{c,0}^{\mathrm{X}}(n)$ - is the average number of calls of class c, type X in the queue in the occupancy state of the secondary resources n:

$$q_{c,0}^{\mathrm{X}}(n) = \sum_{n=\frac{V_0}{Z}+1}^{\frac{V_0}{Z}+Q_0}[x_{c,0}^{\mathrm{X}}(n) - y_{c,0}^{\mathrm{X}}(n)][P_n]_{\frac{V_0}{Z}+Q_0}, \quad (40)$$

- $y_{c,0}^{\mathrm{X}}(n)$ determines the average number of calls of class c type X serviced in the secondary resource 0 with the capacity V_0 AU, which can be determined on the basis of Equation (8).
- $x_{c,0}^{\mathrm{X}}(n)$ determines the average number of calls of class c type X in resource $0'$ with capacity $V_0' = V_0 + Q_0$

The value of the parameter $x_{c,0}^{X}(n)$ can be determined on the basis of Equation (8) that for this particular case can be written in the following form:

$$x_{c,0}^{\mathrm{X}}(n) = \begin{cases} A_{c,0}^{\mathrm{X}}(n-t_c)\dfrac{\left[P_{n-t_c}\right]_{V_0'}}{[P_n]_{V_0'}} & \text{for } 0 \le n \le V_0', \\ 0 & \text{in other cases.}\end{cases} \quad (41)$$

For this case, the occupancy distribution $\left[P_{n-t_c}\right]_{V_0'}$ can be determined on the basis of the recursive dependence (6), with the accompanying assumption that the resource capacity is $V_s' = V_s + Q_s$ and the buffer capacity is equal to zero. For this system we get that:

$$[P_n]_{V_s'} = \frac{1}{n}\Bigg\{\sum_{i\in M_s^{\mathrm{Er}}}A_{i,s}^{\mathrm{Er}}t_{i,s}\left[P_{n-t_{i,s}}\right]_{V_s'} + \sum_{j\in M_s^{\mathrm{En}}}A_{j,s}^{\mathrm{En}}(n-t_{j,s})t_{j,s}\left[P_{n-t_{j,s}}\right]_{V_s'} + \sum_{l\in M_s^{\mathrm{Pa}}}A_{l,s}^{\mathrm{Pa}}(n-t_{l,s})t_{l,s}\left[P_{n-t_{l,s}}\right]_{V_s'}\Bigg\}. \quad (42)$$

The average queue length (that takes into consideration all call classes) is equal:

$$q_0 = \sum_{n=\frac{V_0}{Z}+1}^{\frac{V_0}{Z}+Q_0}\left(n - \frac{V_0}{Z}\right)[P_n]_{\frac{V_0}{Z}+Q_0}. \quad (43)$$

The average duration time for calls of class c in the queue of the secondary resources is equal to:

$$T_{c,0} = \frac{q_{c,0}^{X}}{\lambda_{c,0}}. \quad (44)$$

1) Sequence of Calculations: The models presented earlier in the article make it possible to develop the Overflow-Q-ErEnPa method (Overflow-Erlang Engset Pascal) to determine the blocking probability and other characteristics in the multi-service overflow system with queueing in the primary and secondary resources to which Erlang–Engset–Pascal traffic is offered. This method is presented below.

The OverFlow-Q-ErEnPa method is an iterative scheme that successively updates the occupancy distributions of the primary and secondary resources on the basis of the recurrent

TABLE I
PARAMETERS OF TRAFFIC OFFERED TO THE EXEMPLARY OVERFLOW NETWORKS

Network no.	Resource s	Class 1			Class 2			Class 3		
		traffic	sources	$t_{1,s}$	traffic	sources	$t_{2,s}$	traffic	source	$t_{3,s}$
1	$V_1 = 30,\ Q_1 = 10$	Erlang	–	1	Engset	30	2	Pascal	20	4
	$V_2 = 30,\ Q_2 = 10$	Erlang	–	1	Engset	30	2	Pascal	20	4
	$V_0 = 20,\ Q_0 = 0$									
2	$V_1 = 120,\ Q_1 = 30$	Erlang	–	8	Engset	50	6	Pascal	40	4
	$V_2 = 50,\ Q_2 = 10$	Erlang	–	3	Pascal	20	5			
	$V_0 = 50,\ Q_0 = 20$									
3	$V_1 = 50,\ Q_1 = 0$	Erlang	–	1	Engset	30	2	Pascal	20	4
	$V_2 = 60,\ Q_2 = 0$	Erlang	–	1	Engset	30	2	Pascal	20	4
	$V_0 = 50,\ Q_0 = 20$									

Overflow-Q-ErEnPa method

1) Determination of the occupancy distribution $[P_n]_{V^s}$ in the primary resource s, where $1 \le s \le r$–Formula (6).
2) Determination of the blocking probability $\left[E^{\mathrm{X}}_{c,s}\right]_{V_s}$ for traffic streams of all classes in the primary resource s, where $1 \le s \le r$, $1 \le c \le m_s$–Formula (14).
3) Determination of the parameters of the queues for the primary resources (s, where $1 \le s \le r$)–Formulas (15)–(21).
4) Determination of the capacity $v^{\mathrm{X}}_{c,s}$ for each fictitious primary resource s^{X}_c–Formulas (24)–(26).
5) Determination of the variance $[\sigma^2_{c,\Delta s}]^{\mathrm{X}}$ of traffic offered to the primary resource–Formulas (27)–(29).
6) Determination of the parameters of the equivalent fictitious primary resource-go s^{X}_c, i.e. the equivalent intensity of Erlang traffic $A^{\mathrm{X}}_{c,*s}$ and the equivalent capacity $v^{\mathrm{X}}_{c,*s}$–Formulas (30)–(32).
7) Determination of overflow traffic parameters: the average value of the intensity of the traffic $R^{\mathrm{X}}_{c,s}$ and variance $\left[\sigma^2_{c,s}\right]^{\mathrm{X}}$–Formulas (33)–(34).
8) Determination of the aggregated peakedness factor Z–Formula (37).
9) Determination of the occupancy distribution $[P_n]_{V_0/Z}$ in the system of secondary resources–Formula (36).
10) Determination of the blocking probability $[E_c]_{V_0}$ for traffic streams of the call classes offered to the system of secondary resources–Formula (38).
11) Determination of the parameters of the queues of the secondary resources for all classes of calls offered to the secondary resources–Formulas (39)–(44).

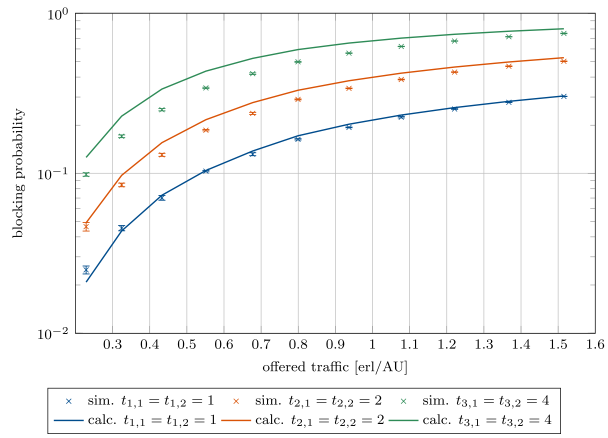

Fig. 4. Blocking probability in Network 1.

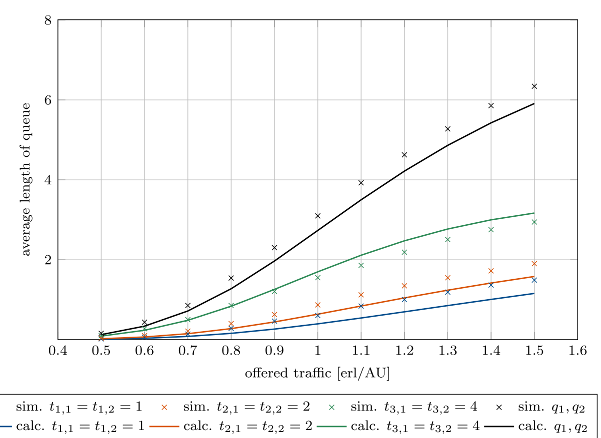

Fig. 5. Average length of queue in Network 1 (primary resources).

equations (6), (18), and (36). The convergence of recurrence-based iterative schemes for full-availability multirate systems with BPP traffic has been rigorously analysed in our previous work [45], where the MIM-BPP algorithm built on the same class of recurrence relations was proved to be convergent. Since the recurrent equations used in the present model satisfy the same assumptions (finite capacities of the system and BPP traffic) and belong to this class, the iterative procedure employed here is convergent as well.

V. THE RESULTS OF MODELING OF SELECTED OVERFLOW SYSTEMS WITH QUEUES IN THE PRIMARY AND SECONDARY RESOURCES AND ERLANG–ENGSET–PASCAL TRAFFIC

The presented OverflowQ-ErEnPa analytical method to model telecommunications systems with Erlang, Engset and Pascal overflow traffic is an approximate method. To assess its accuracy, the analytical results were compared with data obtained from dedicated simulation experiments conducted in this study. The simulation experiments were carried

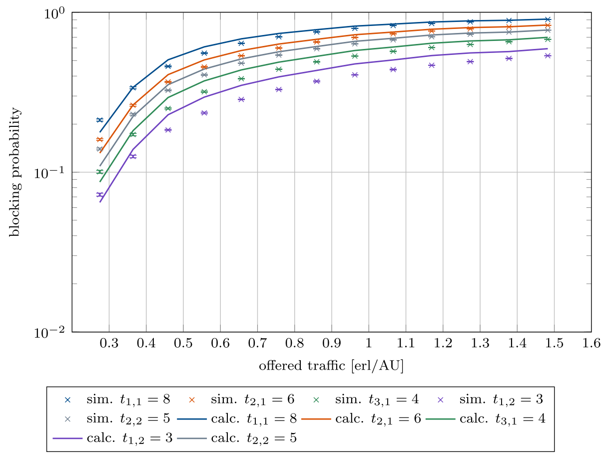

Fig. 6. Blocking probability in Network 2.

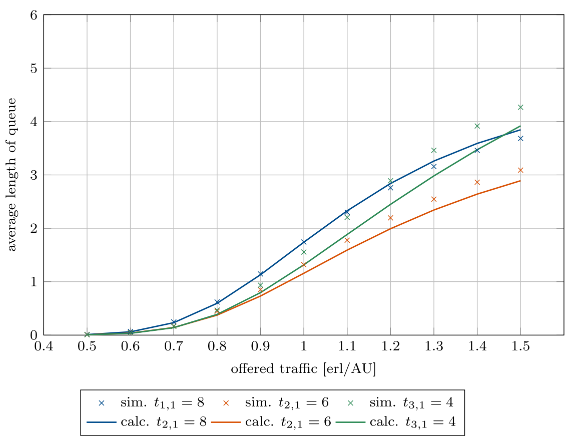

Fig. 7. Average length of queue in Network 2 (primary resources 1).

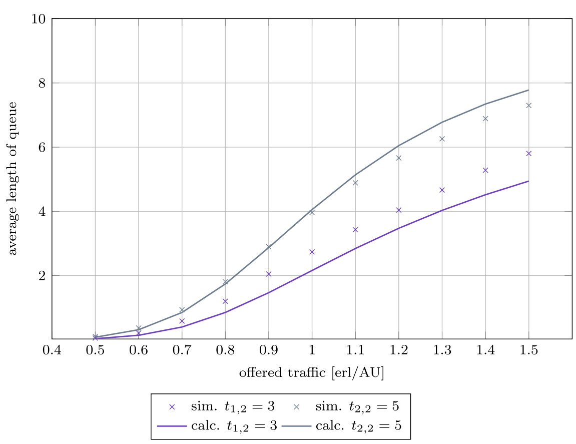

Fig. 8. Average length of queue in Network 2 (primary resources 2).

out using a discrete-event simulator implemented in C++. The traffic sources were configured to emulate the Erlang, Engset and Pascal service classes adopted in the analytical model. Each configuration was simulated for 1000000 call arrivals to achieve statistically reliable estimates. Independent series were executed under distinct random seeds to mitigate stochastic variability. Following a preliminary run,

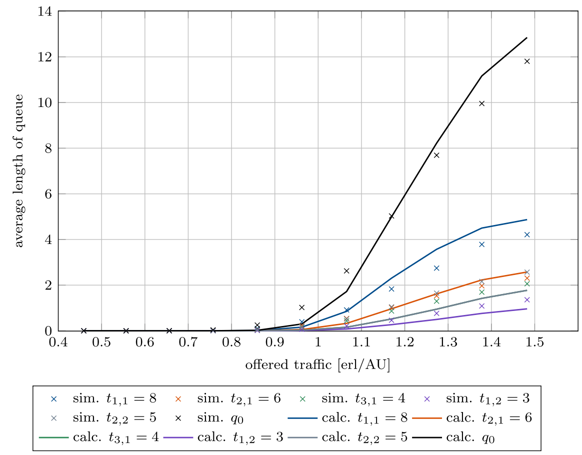

Fig. 9. Average length of queue in Network 2 (secondary resources).

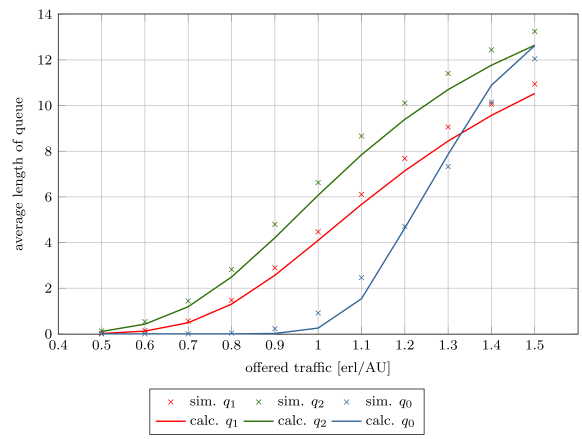

Fig. 10. Total average length of queue in Network 2 (primary and secondary resources).

simulation results were collected under steady-state conditions. Blocking probabilities and mean queue lengths were estimated with 95% confidence intervals computed using the Student's t-distribution based on the results of independent series. This experimental setup ensures reproducibility and enables a consistent comparison between the analytical and simulation results. Figures 4–10 show the study results obtained for two networks with overflow traffic and queues whose parameters are described in Table I.

Figures 4–5 present the empirical results concerning Network No. 1, where queue configurations were exclusively allocated to primary resources. Figure 4 illustrates the blocking probability, while Figure 5 shows the average queue length across specific classes and in aggregate.

Figures 6–10 present the results pertaining to Network No. 2, featuring queueing configurations that encompass both primary and secondary resources. Figure 6 illustrates the blocking probability for all traffic classes offered to Network No. 2. The average queue lengths within Network No. 2 are shown in Figures 7–10. Specifically, Figures 7 and 8 show the average queue lengths within primary resources 1 and 2, respectively. Furthermore, the average queue length within the secondary resources of Network No. 2 is illustrated in

TABLE II
THE ABSOLUTE ERROR OF THE QUEUE LENGTH IN NETWORK NO. 2

a	sim. q_1	calc. q_1	sim. q_2	calc. q_2	sim. q_0	calc. q_0	Δ_{q_1}	Δ_{q_2}	Δ_{q_0}
0.5	0.01920	0.01663	0.12884	0.10466	0.00000	0.00000	0.00256	0.02418	0.00000
0.6	0.14322	0.12026	0.53510	0.42593	0.00010	0.00000	0.02296	0.10917	0.00010
0.7	0.56626	0.48813	1.44179	1.18835	0.00306	0.00000	0.07813	0.25344	0.00306
0.8	1.47430	1.30211	2.82364	2.49164	0.03724	0.00022	0.17219	0.33201	0.03702
0.9	2.89155	2.56649	4.79599	4.20926	0.22648	0.01798	0.32506	0.58673	0.20850
1.0	4.47167	4.09667	6.62772	6.07296	0.91440	0.25398	0.37500	0.55476	0.66042
1.1	6.10920	5.67153	8.66699	7.84436	2.46343	1.53994	0.43767	0.82263	0.92349
1.2	7.68361	7.14295	10.10917	9.39687	4.70888	4.63400	0.54066	0.71229	0.07488
1.3	9.06351	8.44640	11.40654	10.69887	7.32781	7.86255	0.61711	0.70766	0.53475
1.4	10.06118	9.57067	12.44157	11.76918	10.15501	10.89104	0.49051	0.67239	0.73603
1.5	10.94290	10.52967	13.25031	12.64422	12.04797	12.62722	0.41323	0.60609	0.57926

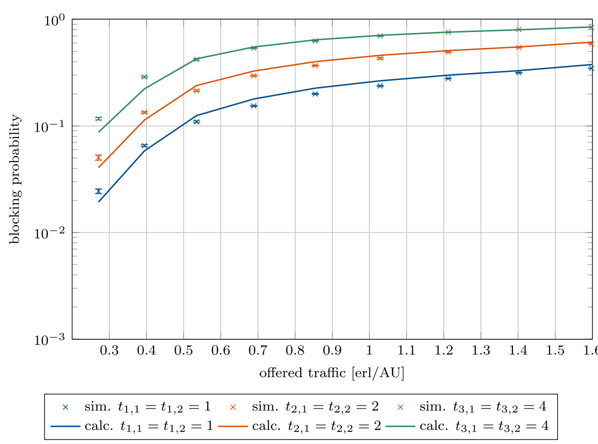

Fig. 11. Blocking probability in Network 3.

Figure 9, with the complete queue length across the network provided in Figure 10.

Finally, Figures 11 and 12 show the blocking probability for all traffic classes offered to Network No. 3 and the corresponding average queue length. In particular, Network No. 3 is characterized by queue configurations solely allocated to secondary resources.

TABLE III
THE ABSOLUTE ERROR OF THE BLOCKING PROBABILITY IN NETWORK NO. 3

a	sim. $E_{1,s}$ $t_{1,1}, t_{1,2}$	calc. $E_{1,s}$ $t_{1,1}, t_{1,2}$	sim. $E_{2,s}$ $t_{2,1}, t_{2,2}$	calc. $E_{2,s}$ $t_{2,1}, t_{2,2}$	sim. $E_{3,s}$ $t_{3,1}, t_{3,2}$	calc. $E_{3,s}$ $t_{3,1}, t_{3,2}$	$\Delta_{E_{1,s}}$ $t_{1,1}, t_{1,2}$	$\Delta_{E_{2,s}}$ $t_{2,1}, t_{2,2}$	$\Delta_{E_{3,s}}$ $t_{3,1}, t_{3,2}$
0.27	0.05049	0.04059	0.02440	0.01932	0.11715	0.08736	0.00990	0.00508	0.02979
0.39	0.13402	0.11364	0.06536	0.05831	0.28866	0.22301	0.02039	0.00704	0.06565
0.53	0.21437	0.23953	0.10955	0.12509	0.42114	0.42720	0.02516	0.01554	0.00606
0.69	0.29636	0.32773	0.15424	0.17920	0.53723	0.55085	0.03137	0.02496	0.01362
0.85	0.36890	0.40158	0.19990	0.22638	0.62717	0.64325	0.03268	0.02649	0.01609
1.03	0.43219	0.46022	0.23747	0.26539	0.69969	0.70942	0.02803	0.02792	0.00974
1.21	0.49275	0.50974	0.27893	0.29972	0.75836	0.75999	0.01699	0.02079	0.00163
1.40	0.54467	0.55173	0.31566	0.33034	0.80473	0.79930	0.00705	0.01468	0.00543
1.60	0.58843	0.61225	0.34778	0.37724	0.83944	0.84979	0.02381	0.02947	0.01035
1.80	0.62891	0.64169	0.38121	0.40139	0.86918	0.87163	0.01278	0.02018	0.00245

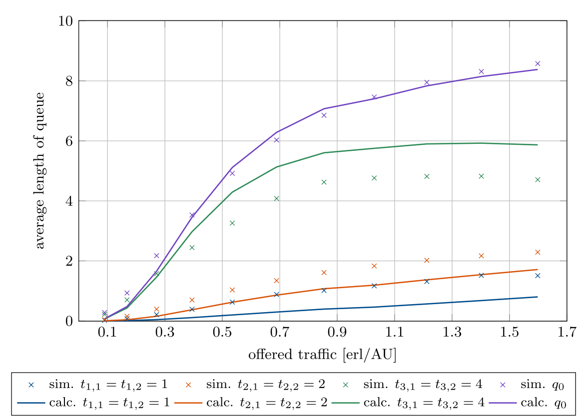

Fig. 12. Average length of queue in Network 3.

TABLE IV
SYSTEM PARAMETERS

Symbol	Description
r	number of primary resources (components) in the system
s	index of the primary resource ($1 \leq s \leq r$)
V_s	server capacity of primary resource s, expressed in allocation units (AU)
Q_s	queue capacity of primary resource s, expressed in allocation units (AU)
V_0	server capacity of the secondary resource, expressed in allocation units (AU)
Q_0	queue capacity of the secondary resource, expressed in allocation units (AU)

TABLE V
TRAFFIC AND MODELING PARAMETERS

Symbol	Description
M_s	set of traffic classes offered to primary resource s
m_s	number of traffic classes offered to primary resource s (cardinality of set M_s)
M_s^{Er}	subset of Erlang traffic classes offered to primary resource s
m_s^{Er}	number of Erlang traffic classes offered to primary resource s (cardinality of set M_s^{Er})
M_s^{En}	subset of Engset traffic classes offered to primary resource s
m_s^{En}	number of Engset traffic classes offered to primary resource s (cardinality of set M_s^{En})
M_s^{Pa}	subset of Pascal traffic classes offered to primary resource s
m_s^{Pa}	number of Pascal traffic classes offered to primary resource s (cardinality of set M_s^{Pa})
m	total number of traffic classes in the system ($m = \sum_{s=1}^{r} m_s$)
c	index of a traffic class ($c \in M_s$)
i	index of an Erlang traffic class ($i \in M_s^{\text{Er}}$)
j	index of an Engset traffic class ($j \in M_s^{\text{En}}$)
l	index of a Pascal traffic class ($l \in M_s^{\text{Pa}}$)
$t_{c,s}$	number of AUs demanded by a call of class c to be serviced in the primary resource s
n	occupancy state, total number of AUs occupied in the considered resource or system
X	type of traffic: Er - Erlang traffic type, En - Engset traffic type, Pa - Pascal traffic type
$S_{j,s}^{\text{En}}$	number of Engset traffic sources of class j in primary resources
$S_{l,s}^{\text{Pa}}$	number of Pascal traffic sources of class l in primary resources
$n_{j,s}^{\text{En}}(n)$	average number of active Engset sources of class j in state n
$n_{l,s}^{\text{Pa}}(n)$	average number of active Pascal sources of class l in state n
$A_{c,s}^{\text{X}}(n)$	intensity of traffic of class c, type X, offered to primary resource s in occupancy state n
$\alpha_{j,s}^{\text{En}}$	average intensity of traffic of class j Engset type, generated by one free source in the resource s
$\gamma_{j,s}^{\text{En}}$	average intensity of calls of class j Engset type, generated by one free source in the resource s
$\mu_{j,s}^{\text{En}}$	average service intensity for calls of class j Engset type in the resource s
$\alpha_{l,s}^{\text{Pa}}$	average intensity of traffic of class l Pascal type, generated by one free source in the resource s
$\gamma_{l,s}^{\text{Pa}}$	average intensity of calls of class l Pascal type, generated by one free source in the resource s
$\mu_{l,s}^{\text{Pa}}$	average service intensity for calls of class l Pascal type in the resource s
$A_{c,s}^{\text{X}}(n)$	intensity of traffic of class c, type X offered to primary resource s,
$\left[E_c^{\text{X}}\right]_{V_s+Q_s}$	blocking probability of calls of class c of type X in primary resource s
$y_{c,s}^{\text{X}}(n)$	average number of serviced calls of classes c type X in state n of occupied AUs in the primary resource s
$x_{c,s}^{\text{X}}(n)$	average number of serviced calls of classes c type X in state n of occupied AUs in the primary resource s' with the capacity $V' = V_s + Q_s$
$v_{c,s}^{\text{X}}$	capacity of factitious resources for class c type X in primary resource s
$v_{c,*s}^{\text{X}}$	capacity of equivalent factitious resources for class c type X in primary resource s
$A_{c,*s}^{\text{X}}(n)$	intensity of equivalent Erlang traffic of class c, type X offered to factitious resource s,
$R_{c,s}^{\text{X}}$	average intensity of traffic of class c ($c \in M_s$) that overflows from the primary resource s, with the assumption that traffic of class c offered to the primary resources s is of type X
$[\sigma_{c,\Delta s}^2]^{\text{X}}$	variance of traffic of class c of type X that overflows from the resource s of the primary resource system
$Z_{c,s}^{\text{X}}$	peakedness factor $Z_{c,s}^{\text{X}}$ for traffic of class c that overflows from the equivalent fictitious primary resource s_c^{X}
Z	aggregated peakedness factor
$y_{c,0}^{\text{X}}(n)$	average number of serviced calls of classes c type X in state n of occupied AUs in the secondary resource 0
$x_{c,0}^{\text{X}}(n)$	average number of serviced calls of classes c type X in state n of occupied AUs in the secondary resource $0'$ with the capacity $V_0' = V_0 + Q_0$

The analytical model proposed in this article is approximate as it assumes the reversibility of the service process. However, the results presented clearly indicate that the model accurately approximates systems with traffic overflow mechanisms. It should be emphasized that regarding the blocking probability, the accuracy of the proposed model depends neither on the queue definitions nor on the traffic offered to the system. However, in the case of the average queue length, the calculation

TABLE VI
PERFORMANCE METRICS

Symbol	Description
$[E_c]_{V_0+Q_0}$	blocking probability for traffic of class c in the secondary resources with queueing
q_0	average queue length in the secondary resources
$q_{c,0}^{\mathrm{X}}$	average queue length for calls of class c type X in the secondary resources
$T_{c,0}$	average duration time for calls of class c in queue of the secondary resources
q_s	average queue length in the primary resources s
$q_{c,s}^{\mathrm{X}}$	average queue length for calls of class c type X in the primary resources s
$T_{c,s}$	average duration time for calls of class c in queue of the primary resources s

results mirror the pattern observed in the results obtained from simulations.

Tables II and III present the absolute errors calculated for the numerical results used to prepare the two selected plots shown in Figures 10 and 11. Specifically, Table II contains the absolute error for the average queue length in the primary and secondary resources of network No. 2, while Table III provides the errors for the blocking probability in network No. 3. The discrepancies between the analytical model results and the digital simulation data stem from the approximate nature of the proposed model. However, the obtained absolute errors confirm the validity of the adopted modeling assumptions.

The values in the tables are presented as a function of the offered traffic per AU of the respective network resources. For the average queue length errors, the traffic is referenced to a single AU of the primary resources; for the blocking probability errors, it is referenced to a single AU of the secondary resources. This distinction arises from the fact that in network No. 2, queues occur in both primary and secondary resources, whereas blocking occurs only in the secondary resources. Referencing the results to the offered traffic per AU of the primary resources allows for a consistent comparison of the obtained results.

VI. CONCLUSION

This article proposes an analytical model of a multi-service hierarchical system with overflow traffic. The assumption is that the system queues will be introduced to the primary and secondary resources of the considered traffic overflow system. This will allow the blocking probability to be reduced. To assess the analytical model developed, the analytical results obtained are compared with the simulation data. Both the data presented in the article and the results obtained by the present authors in numerous comparative studies clearly indicate that the proposed model makes it possible to evaluate the values of the blocking probability with the accuracy that provides its reliable practical application at the stage of network dimensioning. Future research will focus on extending the proposed analytical model to systems that employ compression mechanisms, enabling flexible distribution of resources among multiple service classes. Such systems will include mixed traffic types, including stream, elastic and adaptive calls, which better reflect the characteristics of contemporary telecommunication networks. Moreover, the developed analytical framework can be used to improve network dimensioning and control in hierarchical and cloud-based environments.

APPENDIX

The notation used throughout the paper is summarized Tables IV-VI.

CODE AND DATA AVAILABILITY

The simulator used to generate the numerical results reported in this paper is properiatery research tool that is made freely available for academic use through hosted web interface. Academic users may request access by completing the registration form and accepting the academic end-user licence agreement at https://perfectsoft.it/rnd/network-dimensioning-engine/

REFERENCES

[1] R. I. Wilkinson, "Theories for toll traffic engineering in the U. S. A.," *Bell Syst. Tech. J.*, vol. 35, no. 2, pp. 421–514, Mar. 1956.

[2] Y. Rapp, "Planning of exchange locations and boundaries," *Ericsson Technics*, vol. 18, pp. 91–114, Jun. 1962.

[3] B. Wallström, "Congestion studies in telephone systems with overflow facilities," *Ericsson Technics*, vol. 22, no. 3, pp. 187–351, 1966.

[4] A. A. Fredericks, "Congestion in blocking systems—A simple approximation technique," *Bell Syst. Tech. J.*, vol. 59, no. 6, pp. 805–827, Jul. 1980.

[5] R. Syski, *Introduction to Congestion Theory in Telephone Systems* (Studies in Telecommunication), vol. 4, 2nd ed. Amsterdam, The Netherlands: North-Holland, 1986.

[6] X. Wu, B. Mukherjee, and D. Ghosal, "Hierarchical architectures in the third-generation-cellular network," *IEEE Wireless Commun.*, vol. 11, no. 3, pp. 62–71, Jun. 2004.

[7] 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, 1st Quart., 2012.

[8] Y.-B. Lin, L.-F. Chang, and A. Noerpel, "Modeling hierarchical microcell/macrocell PCS architecture," in *Proc. IEEE Int. Conf. Commun. ICC 95*, vol. 1, Jun. 1995, pp. 405–409.

[9] S. Li, D. Grace, J. Wei, and D. Ma, "Guaranteed handover schemes for a multilayer cellular system," in *Proc. 7th Int. Symp. Wireless Commun. Syst.*, Sep. 2010, pp. 300–304.

[10] N. Zia and A. Mitschele-Thiel, "Self-organized neighborhood mobility load balancing for LTE networks," in *Proc. IFIP Wireless Days (WD)*, Nov. 2013, pp. 1–6.

[11] *What Is Cloud Bursting–Definition — Microsoft Azure*. [Online]. Available: https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-cloud-bursting

[12] M. Femminella, M. Palmucci, G. Reali, and M. Rengo, "Attribute-based management of secure kubernetes cloud bursting," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 1276–1298, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10440131

[13] F. Ullah, S. Dhingra, X. Xia, and M. A. Babar, "Evaluation of distributed data processing frameworks in hybrid clouds," *J. Netw. Comput. Appl.*, vol. 224, Apr. 2024, Art. no. 103837. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1084804524000146

[14] *Google Cloud Networking in Depth: Cloud Load Balancing Deconstructed*. [Online]. Available: https://cloud.google.com/blog/products/networking/google-cloud-networking-in-depth-cloud-load-balancing-deconstructed

[15] E. W. M. Wong, J. Guo, B. Moran, and M. Zukerman, "Information exchange surrogates for approximation of blocking probabilities in overflow loss systems," in *Proc. 25th Int. Teletraffic Congr. (ITC)*, Sep. 2013, pp. 1–9.

[16] J. Matsumoto and Y. Watanabe, "Theoretical method for the analysis of queueing system with overflow traffic," *Electron. Commun. Jpn. (Part I: Communications)*, vol. 64, no. 6, pp. 74–83, Jun. 1981, doi: 10.1002/ecja.4410640610.

[17] I. Maity, S. Misra, and C. Mandal, "DART: Data plane load reduction for traffic flow migration in SDN," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1765–1774, Mar. 2021.

[18] E. Rapaport, I. Poese, P. Zilberman, O. Holschke, and R. Puzis, "Spillover today? Predicting traffic overflows on private peering of major content providers," *IEEE Trans. Netw. Service Manage.*, vol. 18, no. 4, pp. 4169–4182, Dec. 2021.

[19] A. A. Kist and R. J. Harris, "Scheme for alternative packet overflow routing (SAPOR)," in *Proc. Workshop High Perform. Switching Routing, HPSR.*, 2003, pp. 269–274.

[20] M. Glabowski, D. Kmiecik, and M. Stasiak, "Modelling overflow systems with queuing in primary resources," in *Proc. Int. Conf. Heterogeneous Netw. Quality, Rel., Secur. Robustness*, Ho Chi Minh City, Vietnam, Nov. 2019, pp. 148–157, doi: 10.1007/978-3-030-14413-5_12.

[21] G. Bretschneider, "Die berechnung von leitungsgruppen für berfließenden verkehr in fernsprechwählanlagen," *Nachrichtentechnische Zeitun*, vol. 9, no. 11, pp. 533–540, 1956.

[22] A. Baechle, "On the calculation of full available groups with offered smoothed traffic," in *Proc. 7th Int. Teletraffic Congr.*, 1973, pp. 1–6.

[23] A. Kuczura, "The interrupted Poisson process as an overflow process," *Bell Syst. Tech. J.*, vol. 52, no. 3, pp. 437–448, Mar. 1973.

[24] L. Kosten, "Behaviour of overflow traffic and the probabilities of blocking in simple gradings," in *Proc. 8th Int. Teletraffic Congr.*, 1976, pp. 1–5. [Online]. Available: https://gitlab2.informatik.uni-wuerzburg.de/itc-conference/itc-conference-public/-/raw/master/itc08/kosten762.pdf?inline=true

[25] R. Schehrer, "On the calculation of overflow systems with a finite number of sources and full availiable groups," *IEEE Trans. Commun.*, vol. TC-26, no. 1, pp. 75–82, Jan. 1978.

[26] 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. TC-31, no. 11, pp. 1209–1211, Nov. 1983.

[27] W. H. Haemers, B. Sanders, and R. Wilcke, "Simple approximation techniques for congestion functions for smooth and peaked traffic," in *Proc. 10th Int. Teletraffic Congr.*, 1983, pp. 1–7, Paper 4.4b.1.

[28] 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. 664–667, Jun. 1993.

[29] J. S. Kaufman and K. M. Rege, "Blocking in a shared resource environment with batched Poisson arrival processes," *Perform. Eval.*, vol. 24, no. 4, pp. 249–263, Feb. 1996. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0166531694000298

[30] 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, doi: 10.1109/TWC.2005.847031.

[31] H. Heffes, "Analysis of first-come first-served queuing systems with peaked inputs," *Bell Syst. Tech. J.*, vol. 52, no. 7, pp. 1215–1228, Sep. 1973.

[32] J. A. Morrison, "Analysis of some overflow problems with queuing," *Bell Syst. Tech. J.*, vol. 59, no. 8, pp. 1427–1462, Oct. 1980.

[33] G. Brune, "On delay and loss in a switching system for voice and data with internal overflow," in *Proc. 11th Int. teletraffic Congr.*, 1985, pp. 1–7.

[34] J. Matsumoto and Y. Watanabe, "Individual traffic characteristics queueing systems with multiple Poisson and overflow inputs," *IEEE Trans. Commun.*, vol. C-33, no. 1, pp. 1–9, Jan. 1985.

[35] Y. Zhao and E. Gambe, "Analysis on partial overflow queueing systems with two kinds of calls," *IEEE Trans. Commun.*, vol. C-35, no. 9, pp. 942–949, Sep. 1987.

[36] B. Urgaonkar, G. Pacifici, P. Shenoy, M. Spreitzer, and A. Tantawi, "An analytical model for multi-tier internet services and its applications," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 33, no. 1, pp. 291–302, Jun. 2005, doi: 10.1145/1071690.1064252.

[37] S. Hanczewski, M. Stasiak, and J. Weissenberg, "The queueing model of a multiservice system with dynamic resource sharing for each class of calls," in *Computer Networks (Communications in Computer, and Information Science)*, A. Kwiecień, P. Gaj, and P. Stera, Eds., Cham, Switzerland: Springer, 2013, pp. 436–445.

[38] S. Hanczewski, M. Stasiak, and J. Weissenberg, "A queueing model of a multi-service system with state-dependent distribution of resources for each class of calls," *IEICE Trans. Commun.*, vol. 97, no. 8, pp. 1592–1605, Aug. 2014.

[39] M. Stasiak, "Queuing systems for the internet," *IEICE Trans. Commun.*, no. 6, pp. 1234–1242, Jun. 2016.

[40] J. Weissenberg and M. Weissenberg, "Model of a queuing system with BPP elastic and adaptive traffic," *IEEE Access*, vol. 10, pp. 130771–130783, 2022.

[41] G. M. Stamatelos and V. N. Koukoulidis, "Reservation-based bandwidth allocation in a radio ATM network," *IEEE/ACM Trans. Netw.*, vol. 5, no. 3, pp. 420–428, Jun. 1997, doi: 10.1109/90.611106.

[42] S. Rácz, B. P. Gerő, and G. Fodor, "Flow level performance analysis of a multi-service system supporting elastic and adaptive services," *Perform. Eval.*, vol. 49, nos. 1–4, pp. 451–469, Sep. 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0166531602001153

[43] J. Roberts, "Performance evaluation and design of multiservice networks," Commission of the European Communities, Brussels, Belgium, Final Rep. COST 224, 1992.

[44] J. Roberts, V. Mocci, and I. Virtamo, "Broadband network teletraffic," Commission of the European Communities, Springer Verlag, Berlin, Germany, Final Rep. Action COST 242, 1996.

[45] M. Głabowski, M. Stasiak, and J. Weissenberg, "Properties of recurrent equations for the full-availability group with BPP traffic," *Math. Problems Eng.*, vol. 2012, no. 1, p. 17, Jan. 2012.

[46] M. Głąbowski, K. Kubasik, and M. Stasiak, "Modeling of systems with overflow multi-rate traffic," *Telecommun. Syst.*, vol. 37, nos. 1–3, pp. 85–96, Mar. 2008.

[47] M. Głąbowski, A. Kaliszan, and M. Stasiak, "Modeling product-form state-dependent systems with BPP traffic," *Perform. Eval.*, vol. 67, no. 3, pp. 174–197, Mar. 2010.

[48] 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, Mar. 2008.

[49] M. Glabowski, D. Kmiecik, and M. Stasiak, "Modelling of multiservice networks with separated resources and overflow of adaptive traffic," *Wireless Commun. Mobile Comput.*, vol. 2018, no. 1, Aug. 2018, Art. no. 7870164. [Online]. Available: https://www.hindawi.com/journals/wcmc/2018/7870164/

[50] E. Brockmeyer, H. Halstrom, and A. Jensen, "The life and works of A.K. Erlang," *Acta Polytechnika Scandinavia*, vol. 6, no. 287, p. 275, 1960.

Mariusz Głąbowski (Senior Member, IEEE) received the M.Sc., Ph.D., and D.Sc. (Habilitation) degrees in telecommunication from Poznan University of Technology, Poland, in 1997, 2001, and 2010, respectively. In 2020, he was nominated as a Full Professor. Since 1997, he has been with the Institute of Communications and Computer Networks, Faculty of Computing 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. He is the author or co-author of four books, 12 book chapters, and over 150 papers, which have been published in communication journals and presented at national and international conferences.

Sławomir Hanczewski (Member, IEEE) received the M.Sc. and Ph.D. degrees in telecommunications and the D.Sc. degree in ICT from Poznan University of Technology, Poland, in 2001, 2006, and 2020, respectively. Since 2001, he has been with the Faculty of Computing and Telecommunications, Poznan University of Technology. He is currently an Assistant Professor with the Institute of Communications and Computer Networks. He is the author or co-author of more than 60 scientific papers, related mostly to the analytical modeling of communications systems.

Damian Kmiecik received the Ph.D. degree in telecommunications from Poznan University of Technology, Poland, in 2021. He is the author and co-author of more than 20 scientific papers, mostly related to the analytical modeling of telecommunications systems.

Maciej Stasiak (Member, IEEE) received the M.Sc. and Ph.D. degrees in electrical engineering in 1979 and 1984, respectively, and the D.Sc. degree in electrical engineering from Poznan University of Technology in 1996. In 2006, he was nominated as a Full Professor. Between 1983 and 1992, he was with the Polish industry sector as a Designer of electronic and microprocessor systems. In 1992, he joined Poznan University of Technology, where he is currently the Director of the Institute of Communications and Computer Networks, Faculty of Computing and Telecommunications. He is the author or co-author of more than 300 scientific papers and five books. He is engaged in research and teaching in the area of performance analysis and modeling of queuing systems, multi-service networks, and switching systems. Since 2004, he has been actively carrying out research on the modeling and dimensioning of cellular networks.

Joanna Weissenberg received the M.Sc. degree in mathematics from Casimir the Great University, Poland, in 2007, and the Ph.D. degree in telecommunication from Poznan University of Technology, Poland, in 2015. The main area of her professional activity is Markovian analysis of multiservice networks. Her research interests include the application of stochastic processes theory in telecommunication systems and the analysis of queueing delay models in nervous systems.