IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 3, 2020, accepted December 24, 2020, date of publication January 5, 2021, date of current version January 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3049437

A Model of a System With Stream and Elastic Traffic

SŁAWOMIR HANCZEWSKI, (Member, IEEE), MACIEJ STASIAK, (Member, IEEE), AND JOANNA WEISSENBERG

Faculty of Computing and Telecommunications, Poznan University of Technology, Poznan 60-965, Poland

Corresponding author: Sławomir Hanczewski (slawomir.hanczewski@put.poznan.pl)

This work was supported by the research project ''Modeling and Service Quality Evaluation of Internet-Based Services'' funded by the Polish National Science Centre under Grant 2016/23/B/ST7/03925.

ABSTRACT This paper presents a multi-service model of a communications system with stream and elastic traffic. The idea of the model is based on an analysis of a two-dimensional Markov process, which approximates the real process in the considered system. Such an approach allows us to describe the use of system resources by stream and elastic traffic with high accuracy. An appropriate simulation model has been developed and implemented in order to verify the proposed model. The presented results have confirmed the validity of all the theoretical assumptions adopted in the study.

INDEX TERMS Stream and elastic traffic, compression, two-dimensioned Markov process.

I. INTRODUCTION

Data traffic in modern networks is based on the IP protocol. Recently, attempts have been made to send various data streams (with different QoS (Quality of Service) requirements) using the IP protocol, e.g., voice (Voice over IP). The problems with ensuring appropriate QoS parameters and adequate levels of security have been solved, and today there are basically no problems with sending any data stream over IP networks. Furthermore, the development of access networks, especially cellular networks, has led to the point where users have access to the network from anywhere on Earth. Reports, published by the largest manufacturers of network devices, unanimously state that the data volume is growing bigger every year [1], [2]. Furthermore, network operators have noticed an enormous increase in traffic load in areas affected by the Coronavirus pandemic. According to [3], [4], the increase in data traffic amounted to 30–50% already in the first few days of the pandemic. Such a dynamic growth in traffic load is a challenge for network operators and shows that network resources are not limitless.

Network operators possess a variety of traffic management mechanisms that allow them to shape the transmitted data streams in such a way that the available network resources are used in an optimal way. The proper adoption of management mechanisms is of particular importance in networks with limited resources, e.g., in mobile networks. Depending on the properties of the transmitted data stream, it is possible to use several traffic management mechanisms, e.g., threshold and thresholdless compression [5]–[8], resource reservation [9], [10], overflow traffic [11], redirections, priorities [12], [13], or queueing [14], [15].

The shaping of data streams means that their parameters are changing, e.g., the transmission speed or delay. Traffic streams are most commonly divided into traffic with a constant bit rate (CBR) and traffic with a variable bit rate (VBR). VBR data streams are usually subjected to thresholdless compression, which is inscribed into the properties of the TCP protocol. This allows the speed of data transmission to be increased or decreased depending on the load of the links. In the case that all the resources of the system are occupied, the transmission speeds of currently serviced streams are reduced in order to service new data streams that arrive at the input of the system. Two types of compressed data streams have been widely address in the literature: so-called adaptive and elastic traffic. If a decrease in bit rate is accompanied by extended service time necessary for transferring all the data, then the generated traffic is elastic traffic (e.g., an e-mail with an attachment). If, however, a decrease in bit rate of a given packet stream is not followed by any changes in the service time, then we consider the traffic to be adaptive traffic. An example of this type of stream is the data stream associated with ''live'' transmission. In this case, when the links are overloaded, the quality of the transmitted video stream

The associate editor coordinating the review of this manuscript and approving it for publication was Francesco Musumeci.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

decreases to maintain data transmission, which reduces the required transmission speed.

Multi-service network systems are often described at the application level in the literature. A call is defined as a stream of CBR packets. The terms stream and flow are also used in the literature, e.g., [16]–[19]. This approach assumes that for the analytical modeling of multi-service network systems, every real stream of VBR packets can be replaced with a CBR stream of an accurate value, known in the literature as equivalent bandwidth (EB) [20]–[22]. Nowadays, especially in Internet modeling, it is assumed that the EB value is equal to the maximum bit rate of real packet streams from VBR. In most proposed models of resource systems, it is assumed that traffic at the call level (i.e., streams in which packet streams are seen as individual request) can be described by Poisson streams [18], [22], [23]. Consequently, it is possible to analyze multi-service systems on the basis of multi-dimensional Markov processes in which the number of dimensions is equal to the number of offered traffic streams. In [24], [25], a basic, recurrent model of a multi-service, full-availability system was proposed. In [26], this model was generalized to include BPP traffic streams, which are a mixture of Bernoulli, Pascal, and Engset traffic streams. In [27], a model of fully available resources with BPP traffic is proposed, which is based on a convolution of occupancy distributions of particular traffic classes.

In [9], the model of [24], [25] is expanded to include elastic traffic resources. It was assumed that this type of traffic can be compressed to certain limit values. In [28], this model is generalized to include an instance of limitless compression of elastic traffic streams. In the model, the bit rate of serviced calls may tend towards zero and, as a consequence, calls will always be serviced, i.e., the phenomenon of lost calls will never occur. In the models of [9] and [28], it was assumed that the multi-dimensional Markov service process is a reversible process. The consequence of this assumption is a state-dependent call service process of all classes and, as a result, an appropriate distribution of the resources between the serviced calls of individual classes. Such a resource distribution can be executed on the basis of the so-called balanced fairness algorithm [17], [29]. In [5], the model of [9] is expanded to include the possibility of simultaneous service of elastic and adaptive traffic. Reference [30] proposed the interpretation of an occupancy distribution, determined in [5], as an occupancy distribution of a multi-service queueing system with stream traffic. In [14] and [15], a model of a queueing system with elastic and adaptive traffic is proposed.

An approximate model of full-availability resources that services a mixture of stream and adaptive traffic is proposed in [31], [32]. In this model, it is assumed that the system services only stream traffic, i.e., adaptive traffic is approximated with such stream traffic that it corresponds to adaptive traffic with the maximum compression coefficient. In [33], a two-dimensional queueing model is proposed. In the model, it is assumed that the occupancy distribution may be represented as the product of two occupancy distributions associated with the relevant types of traffic. Such an approach has made it possible to develop a model of a full-availability resources with a mixture of stream and elastic traffic. In [33], it is assumed that the Markov service process is reversible and, as a result, that the serviced resource division is in agreement with the balanced fairness algorithm. Moreover, the assumption that the service process is reversible has enabled the construction of a two-dimensional occupancy distribution in which one dimension corresponds to the occupancy distribution of stream traffic and the other corresponds to the occupancy distribution of elastic traffic.

In [16], a model of a system that services a mixture of stream (non-adaptive traffic), adaptive and elastic traffic is proposed. In the model the partial overlap link bandwidth policy is adopted according to which the capacity of the system is divided into two parts, a common part (for stream traffic only) and a part reserved for elastic and adaptive traffic streams only. In order to determine the capacity of each part the Authors of [16] have proposed the iterative algorithm of the system allocation procedure with the respect to QoS and GoS requirements. In the algorithm it is assumed that the minimum capacity of the common part is calculated on the basis of Erlang-B formulae, while the capacity of second part is defined as a sum of maximum number of adaptive and elastic traffic streams simultaneously present in the system. The number of adaptive traffic stream is determined on the basis on two dimensional Markov chain (on the assumption that the number of adaptive traffic stream is independent of other traffic streams), and is a basis for calculation of the number of elastic traffic streams. Finally, the blocking probabilities for each class of calls are obtained from the steady state distribution of the Continuous Time Markov Chain.

This article proposes an approximate analytical model of a communications system with stream and elastic traffic. The idea of the model is based on an analysis of a two-dimensional Markov process, which approximates the real process in the considered system. To validate and verify the model proposed here, an original and purpose-made simulator was used. The results of the simulation corroborate the validity and accuracy of the theoretical assumptions adopted in the model.

The remaining part of the article is structured as follows. In Section II, an analytical model of the system with stream and elastic traffic is proposed. Section III presents several sample results of the analytical modeling of various selected systems which are then compared with the relevant results of the digital simulation. Finally, Section IV summarizes the article.

II. A MODEL OF A MULTI-SERVICE SYSTEM WITH STREAM AND ELASTIC TRAFFIC

Let us assume that M_s classes of stream traffic and M_e classes of elastic traffic are offered to the system. If we assume that the calls of elastic traffic do not undergo compression, the offered traffic can be described by the following parameters:

- M - the total number of call classes offered to the system: $M = M_s + M_e$
- $t_{i,u}$ - the demanded bit rate of a call of class i type u, where $u = s$ for stream traffic, and $u = e$ for elastic traffic
- t_{AU} - the bit rate of allocation unit (AU) adopted in the considered system [32]. The value of t_{AU} can be calculated according to the following formula:

$$t_{AU} = \text{GCD}\left\{c_{1,s}, \dots, c_{M_s,s}, c_{1,e}, \dots, c_{M_e,e}\right\}, \qquad (1)$$

  where GCD is the biggest common divisor of the bit rates of all the offered call classes
- $c_{i,u}$ - the demanded number of AUs necessary to service a call of class i type $u = s$:

$$c_{i,u} = \left\lceil \frac{t_{i,u}}{t_{AU}} \right\rceil \qquad (2)$$

- $\lambda_{i,u}$ - the call stream intensity of class i type u
- $\mu_{i,u}$ - the call stream intensity of class i type u
- $a_{i,u}$ - the average traffic intensity of class i type u, defined in relation to calls:

$$a_{i,u} = \frac{\lambda_{i,u}}{\mu_{i,u}} \qquad (3)$$

- $A_{i,u}$ - the average traffic intensity of class i type u, defined in relation to AUs:

$$A_{i,u} = a_{i,u} \cdot c_{i,u} \qquad (4)$$

- V - the total bit rate of the system
- C - the total capacity of the system expressed in AUs:

$$C = \left\lfloor \frac{V}{t_{AU}} \right\rfloor \qquad (5)$$

Let us now consider a model with the real capacity C_r. Introducing the notion of the so-called virtual capacity C_v ($C_v > C_r$) significantly simplifies the analysis of the resources with stream and elastic traffic. The virtual capacity of the system is the capacity of the system from an elastic traffic point of view, which does not affect the real capacity. However, the greater number of calls of elastic traffic can be serviced because of the mechanism of compression (decrease of the demanded number of AUs necessary for a connection of calls of elastic traffic).

A diagram of the multi-service resources with the virtual capacity is shown in Fig. 1. The introduction of the virtual capacity allows the system with traffic compression to be regarded as a system with stream traffic only. In the thus defined real capacity resources, the bit rate of the calls of elastic traffic can be reduced as long as the sum of the AUs occupied in the virtual capacity resources (which is defined as the sum of uncompressed demands of calls of all classes) is less than the virtual capacity. This means that the compression of the calls of elastic traffic in the real resources corresponds to the servicing of the uncompressed calls of elastic traffic in the virtual resources. Therefore, in the analysis of the service process of the considered system, the traffic resources can be divided into two areas: the uncompressed service area (here referred to as the real area) and the compressed area (here referred to as the virtual capacity), where the calls of elastic traffic undergo compression. Let n be the total number of serviced AUs in the multi-service resources. In the real area ($0 \leq n \leq C_r$), the calls of stream and elastic traffic are serviced, whereas in the virtual area ($C_r < n \leq C_v$) the calls of elastic traffic are always serviced and the calls of stream traffic only if the sum of AUs occupied by this type of calls is less than the real capacity.

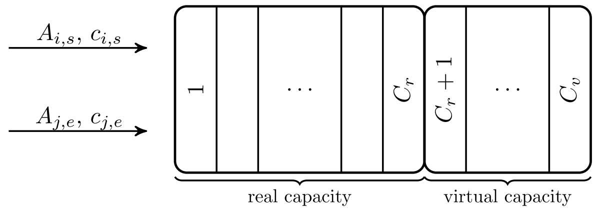

FIGURE 1. Multi-service system with stream and elastic traffic.

The service process in the multi-service system can be considered at microstates and macrostates call levels. The macrostate is defined as a set of natural numbers $\mathbb{N}_+$ determining the number of calls of class i serviced in the system and is denoted by

$$(X_s, X_e) = \left\{x_{1,s}, x_{2,s}, \dots, x_{M_s,s}, x_{1,e}, x_{2,e}, \dots, x_{M_e,e}\right\}, \qquad (6)$$

where $x_{i,u}$ is the number of calls of class i type u ($u = s$ for stream traffic, $u = e$ for elastic traffic) serviced in the system. The total number of $n\left(X_s, X_e\right)$ AUs serviced in the microstate (X_s, X_e) is therefore equal to

$$n\left(X_s, X_e\right) = \sum_{i=1}^{M_s} x_{i,s}c_{i,s} + \sum_{j=1}^{M_e} x_{j,e}c_{j,e}. \qquad (7)$$

The macrostate, in turn, is defined on the basis of the total number of AUs occupied by the calls of stream and elastic traffic in the system, without the division of AUs between individual call classes for both types of traffic. Finally, on the basis of the definition of the microstate and the informal definition of the macrostate, we are in the position to formally define the macrostate (n_s, n_e) as the state in which n_s AUs are occupied by the calls of stream traffic and n_e AUs are occupied by the calls of elastic traffic. We can thus write

$$\Omega\left(n_s, n_e\right) = \\ = \left\{(X_s, X_e) \in \Omega : \ n_s = \sum_{i=1}^{M_s} x_{i,s}c_{i,s} \wedge \right. \\ \left. n_e = \sum_{j=1}^{M_e} x_{j,e}c_{j,e} \right\}, \qquad (8)$$

where Ω is the set of all microstates such that

$$\Omega = \\ = \left\{(X_s, X_e) : \ 0 \leq n_{X_s,X_e} \leq C_v \wedge \sum_{i=1}^{M_s} x_{i,s}c_{i,s} = C_r \right.$$

$$\left.\Rightarrow \sum_{j=1}^{M_e} x_{j,e} c_{j,e} = 0\right\}. \tag{9}$$

The compression of calls of elastic traffic causes a decrease in their bit rates and may simultaneously be associated with an extension of their service times. The average traffic intensity of calls of class i type u in microstate (X_s, X_e), defined in relation to AUs, can be written as follows:

$$A_{i,u}\left(X_s, X_e\right) = a_{i,u}\left(X_s, X_e\right) c_{i,u}\left(X_s, X_e\right), \tag{10}$$

where $a_{i,u}\left(X_s, X_e\right)$ is the average traffic intensity of calls of class i type u in microstate (X_s, X_e) defined in relation to calls. On the basis of (3), we can write

$$a_{i,u}\left(X_s, X_e\right) = \frac{\lambda_{i,u}}{\mu_{i,u}\left(X_s, X_e\right)}. \tag{11}$$

The parameter $c_{i,u}\left(X_s, X_e\right)$ defines the number of resources demanded by calls of class i type u in microstate (X_s, X_e). The parameter $\mu_{i,u}\left(X_s, X_e\right)$, in turn, defines the service intensity of calls of class i type u in microstate (X_s, X_e). In the case of stream traffic, both the demanded number of AUs and the service time of calls do not experience any alternations regardless of the microstate in which the service process is located. We can thus write

$$\forall_{(X_s,X_e)\in\Omega} \quad c_{i,s}\left(X_s, X_e\right) = c_{i,s} \tag{12}$$

and

$$\forall_{(X_s,X_e)\in\Omega} \quad \mu_{i,s}\left(X_s, X_e\right) = \mu_{i,s}. \tag{13}$$

In the virtual area, the demanded number of AUs by calls of elastic traffic decreases, while the service time of calls of these classes increases. Therefore,

$$\forall_{(X_s,X_e)\in\Omega}\, c_{j,e}\left(X_s, X_e\right) = \begin{cases} c_{j,e}, \\ \quad 0 \le n\left(X_s, X_e\right) \le C_r, \\ c_{j,e}\dfrac{C_r}{n\left(X_s, X_e\right)}, \\ \quad C_r < n\left(X_s, X_e\right) \le C_v, \end{cases} \tag{14}$$

and

$$\forall_{(X_s,X_e)\in\Omega}\, \mu_{j,e}\left(X_s, X_e\right) = \begin{cases} \mu_{j,e}, \\ \quad 0 \le n\left(X_s, X_e\right) \le C_r, \\ \mu_{j,e}\dfrac{C_r}{n\left(X_s, X_e\right)}, \\ \quad C_r < n\left(X_s, X_e\right) \le C_v. \end{cases} \tag{15}$$

Notice that both the demanded number of AUs by calls of elastic traffic and the service intensity of these calls do not change for each microstate (X_s, X_e) that

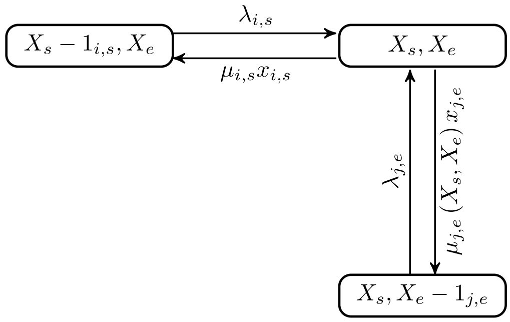

FIGURE 2. A fragment of the Markov process for three adjacent states.

belongs to macrostate (n_s, n_e) defined by (8). As a result, we get

$$\forall_{(X_s,X_e)\in\Omega(n_s,n_e)}\, c_{j,e}\left(X_s, X_e\right) = c_{j,e}\left(n_s, n_e\right) = \begin{cases} c_{j,e}, \\ \quad 0 \le n_s + n_e \le C_r, \\ c_{j,e}\dfrac{C_r}{n_s + n_e}, \\ \quad C_r < n_s + n_e \le C_v, \end{cases} \tag{16}$$

and

$$\forall_{(X_s,X_e)\in\Omega(n_s,n_e)}\, \mu_{j,e}\left(X_s, X_e\right) = \mu_{j,e}\left(n_s, n_e\right) = \begin{cases} \mu_{j,e}, \\ \quad 0 \le n_s + n_e \le C_r, \\ \mu_{j,e}\dfrac{C_r}{n_s + n_e}, \\ \quad C_r < n_s + n_e \le C_v. \end{cases} \tag{17}$$

On the basis of (10), (12), (13), (16), and (17), it is possible to determine the average traffic intensities of calls of stream and elastic traffic in macrostate (n_s, n_e) by the following formulas, respectively:

$$A_{i,s}\left(n_s, n_e\right) = a_{i,s} \cdot c_{i,s} = A_{i,s}, \tag{18}$$

$$A_{j,e}\left(n_s, n_e\right) = a_{j,e}\left(n_s, n_e\right) c_{j,e}\left(n_s, n_e\right) = A_{j,e}. \tag{19}$$

In order to determine the occupancy distribution in the thus defined system, let us consider the service process in three adjacent microstates presented in Fig. 2. Let us now make the assumption that the considered service process is a reversible process and that it satisfies the local balance equations. The local balance equations for the three neighboring microstates from Fig. 2 will take on the following form:

$$\mu_{i,s} x_{i,s} P\left(X_s, X_e\right) = \lambda_{i,s} P\left(X_s - 1_{i,s}, X_e\right), \tag{20}$$

$$\mu_{j,e}\left(X_s, X_e\right) x_{j,e} P\left(X_s, X_e\right) = \lambda_{j,e} P\left(X_s, X_e - 1_{j,e}\right), \tag{21}$$

where parameter $P\left(X_s, X_e\right)$ is the probability of microstate (X_s, X_e), $1_{i,s}$ denotes exactly one call of class i of stream traffic, and parameter $1_{j,e}$ denotes exactly one call of class

j of elastic traffic. Equations (20) and (21), after dividing both sides by $\mu_{i,s}$ and $\mu_{j,e}(X_s, X_e)$, and multiplying by $c_{i,u}$ and $c_{j,e}(X_s, X_e)$, respectively, can be written as follows:

$$x_{i,s}c_{i,s}P(X_s, X_e) = \frac{\lambda_{i,s}c_{i,s}}{\mu_{i,s}}P\left(X_s - 1_{i,s}, X_e\right), \tag{22}$$

$$x_{j,e}c_{j,e}(X_s, X_e)P(X_s, X_e) = \\ = \frac{\lambda_{j,e}c_{j,e}(X_s, X_e)}{\mu_{j,e}(X_s, X_e)}P\left(X_s, X_e - 1_{j,e}\right). \tag{23}$$

In turn, on the basis of (10) and (11), we get:

$$x_{i,s}c_{i,s}P(X_s, X_e) = A_{i,s}P\left(X_s - 1_{i,s}, X_e\right), \tag{24}$$

$$x_{j,e}c_{j,e}(X_s, X_e)P(X_s, X_e) = \\ = A_{j,e}(X_s, X_e)P\left(X_s, X_e - 1_{j,e}\right). \tag{25}$$

In microstates in which the total number of AUs exceeds the real capacity of the system, the service process does not satisfy the local balance equations; therefore, the product $x_{i,u}(X_s, X_e)\, c_{i,u}(X_s, X_e)$ will be substituted with such a value of the service stream $y_{i,u}(X_s, X_e)$ that it will guarantee the reversibility property of the Markov process in the system under consideration:

$$y_{i,s}(X_s, X_e)P(X_s, X_e) = A_{i,s}P\left(X_s - 1_{i,s}, X_e\right), \tag{26}$$

$$y_{j,e}(X_s, X_e)P(X_s, X_e) = \\ = A_{j,e}(X_s, X_e)P\left(X_s, X_e - 1_{j,e}\right). \tag{27}$$

Because of the independence of the traffic stream offered to the system, we can add up, all M_s equations of type (26) and all M_e equations of type (25). Hence,

$$P(X_s, X_e)\left[\sum_{i=1}^{M_s} y_{i,s}(X_s, X_e) + \sum_{j=1}^{M_e} y_{j,e}(X_s, X_e)\right] = \\ = \sum_{i=1}^{M_s} A_{i,s}P\left(X_s - 1_{i,s}, X_e\right) + \\ + \sum_{j=1}^{M_e} A_{j,e}(X_s, X_e)P\left(X_s, X_e - 1_{j,e}\right). \tag{28}$$

The sum in the square brackets determines the total service stream (in terms of AUs) of all the calls in microstate (X_s, X_e). This stream can take on a value equal to the total number of AUs occupied in the system (in the real area) or to the real capacity (in the virtual area). Hence, the occupancy distribution at the microstate level can finally be written in the following form:

$$P(X_s, X_e) = \\ = \begin{cases} \dfrac{1}{\min\{n(X_s, X_e), C_r\}} \\ [3pt]\left[\sum_{i=1}^{M_s} A_{i,s}P\left(X_s - 1_{i,s}, X_e\right) + \right. \\ \left. + \sum_{j=1}^{M_e} A_{j,e}(X_s, X_e)P\left(X_s, X_e - 1_{j,e}\right)\right], \\ \quad \text{for } 0 \le n(X_s, X_e) \le C_v, \\ 0, \quad \text{for all other cases,} \end{cases} \tag{29}$$

where the value of the probability $P(0\ldots, 0, 0\ldots, 0)$ results from the normative condition $\sum_{(X_s,X_e)\in\Omega} P(X_s.X_e) = 1$.

The probability of macrostate (n_s, n_e) is defined as the sum of the probabilities of the microstate that belongs to the set $\Omega(n_s, n_e)$. As a result, we get

$$P(n_s, n_e) = \sum_{(X_s,X_e)\in\Omega(n_s,n_e)} P(X_s, X_e). \tag{30}$$

Now, since the service process in the considered system is by definition a reversible process, in relation to the macrostate, we can add up both sides of (29) over the set $\Omega(n_s, n_e)$. Taking into account (30) and the definition of the macrostate in (8) we can write the probability of macrostate (n_s, n_e) in the following way:

$$P(n_s, n_e) = \\ = \begin{cases} \dfrac{1}{\min\{n_s + n_e, C_r\}} \\ \left[\sum_{i=1}^{M_s} A_{i,s}P\left(n_s - c_{i,s}, n_e\right) + \right. \\ \left. + \sum_{j=1}^{M_e} A_{j,e}(n_s, n_e)P\left(n_s, n_e - c_{j,e}\right)\right], \\ \quad \text{for } 0 \le n_s + n_e \le C_v, \\ 0, \quad \text{for all other cases.} \end{cases} \tag{31}$$

The value of the probability $P(0, 0)$ at the macrostate level results from the normative condition $\sum_{(n_s,n_e)} P(n_s, n_e) = 1$.

The blocking phenomenon occurs if there are not enough AUs available to service a new call. Thus, in the case of calls of class i of stream traffic, it occurs if one of the following three mutually exclusive cases occurs:

1) if there are no calls of elastic traffic serviced in the system and the sum of the demands of serviced calls and a new call of class i exceeds the real capacity C_r
2) if there are calls of elastic traffic serviced in the system and the sum of the demands of serviced calls of stream traffic and a new call of class i is equal to or exceeds the real capacity C_r
3) if there are calls of elastic traffic serviced in the system and the sum of the demands of serviced calls and a new call of class i exceeds the virtual capacity C_v

Conversely, in the case of calls of class i of elastic traffic, it occurs in the following two situations:

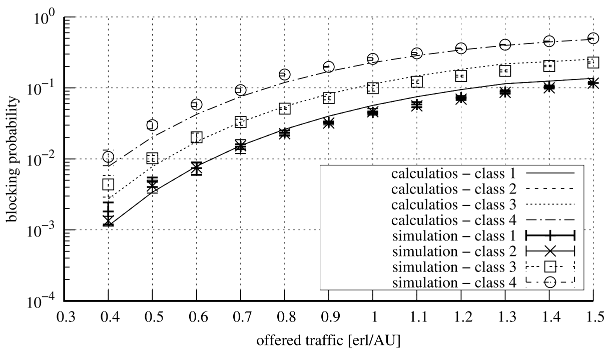

FIGURE 3. Blocking probability in System I.

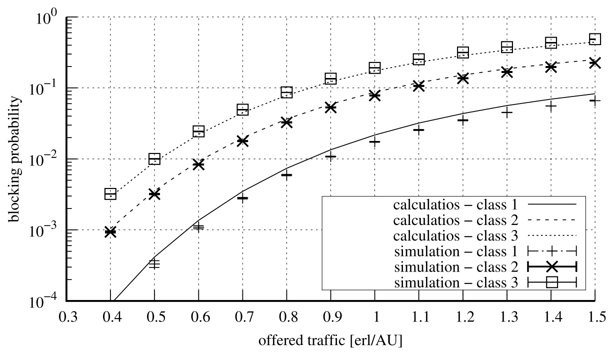

FIGURE 4. Blocking probability in System II.

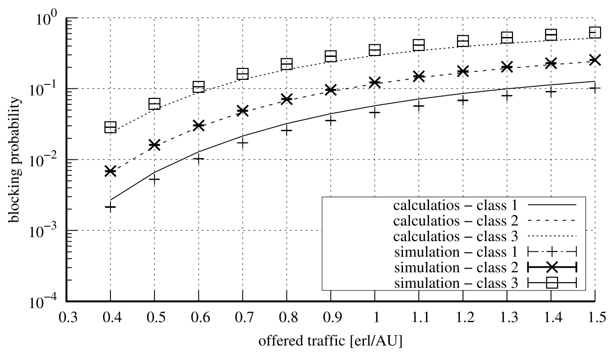

FIGURE 5. Blocking probability in System III.

1) if there are no calls of elastic traffic serviced and the sum of all serviced calls of stream traffic is equal to the real capacity C_r
2) if there are calls of elastic traffic serviced in the system and the sum of the demands of serviced calls and a new call of class i exceeds the virtual capacity C_v

Thus, taking into account the states in which blocking for individual call classes occurs, the blocking probability is equal to:

$$E_{i,u} = \sum_{(n_s,n_e)\in\Omega_{i,u}} [P(n_s, n_e)]_{C_v}, \tag{32}$$

where

$$\begin{aligned}\Omega_{i,s} &\\ = \{&(n_s, n_e) \in \Omega\\ &: \left(C_r - c_{i,s} + 1 \le n_s \le C_r \ \wedge\ n_e = 0\right) \ \vee\ (n_e \neq 0) \wedge\\ &\wedge \left[\left(n_s = C_r - c_{i,s}, \ldots, C_r - 1 \ \wedge 0 < n_s + n_e \le C_v\right) \wedge\right.\\ &\vee \left.\left(n_s < C_r - c_{i,s} \ \wedge\ C_v - c_{i,s} + 1 \le n_s + n_e \le C_v\right)\right]\}\end{aligned} \tag{33}$$

is the set of macrostates in which the blocking of calls of class i of stream traffic occurs, and

$$\begin{aligned}\Omega_{j,e} = \{&(n_s, n_e) \in \Omega : \ (n_s = C_r \ \wedge\ n_e = 0) \ \vee\\ &\vee (n_s = 0, \ldots, C_r - 1 \wedge\\ &\wedge\ C_v - c_{j,e} + 1 \le n_s + n_e \le C_v)\},\end{aligned} \tag{34}$$

is the set of macrostates in which the blocking of calls of class j of elastic traffic occurs.

III. RESULTS

In order to verify the accuracy of the proposed model of a system with stream and elastic traffic, the model's results were compared with the results of simulation experiments. For this purpose, an appropriate simulator of the considered system was designed and implemented (C++).

Figs. 3–5 present the results of the blocking probability in the considered systems. Figure 3 shows the results in System I with the parameters $C_r = 10$ AUs and $C_v = 20$ AUs. The system was offered four call classes: one stream class ($c_{1,s} = 1$ AU) and three elastic classes ($c_{1,e} = 1$ AU, $c_{2,e} = 2$ AUs, $c_{3,e} = 4$ AUs). Fig. 4 presents the results of the blocking probability in System II with the parameters $C_r = 10$ AUs and $C_v = 40$ AUs. The system was offered three call classes: $c_{1,s} = 1$AU, $c_{1,e} = 3$ AUs, $c_{2,e} = 6$ AUs. In Fig. 5, the results for System III with the parameters: $C_r = 20$ AUs, $C_v = 25$ AUs, $c_{1,s} = 1$AU, $c_{1,e} = 2$ AUs, and $c_{2,e} = 5$ AUs are presented.

The results of the simulation are presented in Figs. 3–5 in the form of dots with a confidence interval of 95%, calculated with Student's t-distribution for five series, with 1,000,000 calls each (of the class that generated the lowest number of calls) in each series. For each point of the simulation, the confidence interval is at least two orders of magnitude lower than the results of the simulation. The obtained results are presented as a function of traffic offered per one AU of the system:

$$a = \frac{\sum_{i=1}^{M} A_{i,u}}{C_r}. \tag{35}$$

The presented results confirm the satisfactory accuracy of the proposed model. The accuracy does not depend on the ratio of the real capacity C_r to the total capacity C_v, which takes into account the virtual part of the capacity of the system. The blocking probabilities of calls of stream traffic depend on the real capacity of the system and on the mixture of offered traffic as well. In the case of elastic traffic, the blocking probabilities depend also on the virtual capacity. The greater virtual capacity of the system the smaller value of blocking probability of calls of elastic traffic. It is also worth to point that greater virtual capacity of the system means greater compression of calls of elastic traffic as well as proportional increase of theirs service time. The model is easily programmable and can be a useful tool in engineering practice when assessing the QoS of IP networks. In the next stage of our research, the model will be expanded to a model for queueing systems with stream and elastic traffic.

IV. SUMMARY

The presented paper proposes an analytical model of a communications system with a mixture of stream and elastic traffic. Such a mixture is typical for modern networks. The results obtained by the model are compared with the results of a digital simulation, which confirms the validity of all the theoretical assumptions of the model. The model can be used for the analysis of modern network systems with constant and variable bit rate streams.

REFERENCES

[1] Cisco, "Cisco annual Internet report (2018–2023)," Cisco, San Jose, CA, USA, Tech. Rep. C11-741490-01, 2020.

[2] *Ericsson Mobility Report*, Ericsson, Stockholm, Sweden, 2019.

[3] E. Koeze and N. Popper, "The virus changed the way we Internet," New York Times, New York, NY, USA, Apr. 2020. [Online]. Available: https://www.nytimes.com/interactive/2020/04/07/technology/coronavirus-internet-use.html

[4] C. Labovitz, "Network traffic insights in the time of covid-19: April 9 update," Nokia, Espoo, Finland, Tech. Rep. 2, 2020.

[5] 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.

[6] M. Sobieraj, M. Stasiak, J. Weissenberg, and P. Zwierzykowski, "Analytical model of the single threshold mechanism with hysteresis for multi-service networks," *IEICE Trans. Commun.*, vol. E95-B, no. 1, pp. 120–132, 2012.

[7] I. D. Moscholios, M. D. Logothetis, and A. C. Boucouvalas, "Blocking probabilities of elastic and adaptive calls in the Erlang multirate loss model under the threshold policy," *Telecommun. Syst.*, vol. 62, no. 1, pp. 245–262, May 2016.

[8] I. D. Moscholios, V. G. Vassilakis, and P. G. Sarigiannidis, "Performance modelling of a multirate loss system with batched Poisson arrivals under a probabilistic threshold policy," *IET Netw.*, vol. 7, no. 4, pp. 242–247, Jul. 2018.

[9] 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.

[10] M. Głąbowski, A. Kaliszan, and M. Stasiak, "Asymmetric convolution algorithm for blocking probability calculation in full-availability group with bandwidth reservation," *IET Circuits, Devices Syst.*, vol. 2, no. 1, pp. 87–94, Feb. 2008.

[11] M. Głąbowski, A. Kaliszan, and M. Stasiak, "Modelling overflow systems with distributed secondary resources," *Comput. Netw.*, vol. 108, pp. 171–183, Oct. 2016.

[12] K. Subramaniam and A. A. Nilsson, "Tier-based analytical model for adaptive call admission control scheme in a UMTS-WCDMA system," in *Proc. IEEE 61st Veh. Technol. Conf.*, vol. 4, May 2005, pp. 2181–2185.

[13] S. Hanczewski, M. Stasiak, and P. Zwierzykowski, "Modelling of the access part of a multi-service mobile network with service priorities," *EURASIP J. Wireless Commun. Netw.*, vol. 2015, no. 1, pp. 1–14, Dec. 2015.

[14] M. Stasiak, "Queuing systems for the Internet," *IEICE Trans. Commun.*, vol. E99-B, no. 6, pp. 1224–1242, Jun. 2016.

[15] S. Hanczewski, M. Stasiak, and J. Weissenberg, "Queueing model of a multi-service system with elastic and adaptive traffic," *Comput. Netw.*, vol. 147, pp. 146–161, Dec. 2018.

[16] G. Fodor, S. Rácz, and M. Telek, "On providing blocking probability and throughput guarantees in a multi-service environment," *Int. J. Commun. Syst.*, vol. 15, no. 4, pp. 257–285, May 2002.

[17] T. Bonald and A. Proutiere, "Insensitive bandwidth sharing in data networks," *Queueing Syst.*, vol. 44, no. 1, pp. 69–100, May 2003.

[18] T. Bonald and J. W. Roberts, "Internet and the Erlang formula," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 42, no. 1, pp. 23–30, Jan. 2012.

[19] I. D. Moscholios and M. D. Logothetis, *Efficient Multirate Teletraffic Loss Models Beyond Erlang*. Hoboken, NJ, USA: Wiley, 2019.

[20] J. Y. Hui, "Resource allocation for broadband networks," *J. Sel. Areas Commun.*, vol. 6, no. 9, pp. 1598–1608, Dec. 1988.

[21] F. Kelly, "Notes on effective bandwidth," in *Stochastic Networks*, F. P. Kelly, S. Zachary, and I. B. Ziedins, Eds. London, U.K.: Oxford Univ. Press, 1996, pp. 141–168.

[22] J. Roberts, V. Mocci, and I. Virtamo, Eds., *Broadband Network Teletraffic, Final Report of Action COST 242*. Berlin, Germany: Commission of the European Communities, Springer, 1996.

[23] V. Paxson and S. Floyd, "Wide area traffic: The failure of Poisson modeling," *IEEE/ACM Trans. Netw.*, vol. 3, no. 3, pp. 226–244, Jun. 1995.

[24] J. Kaufman, "Blocking in a shared resource environment," *IEEE Trans. Commun.*, vol. COM-29, no. 10, pp. 1474–1481, Oct. 1981.

[25] J. Roberts, "A service system with heterogeneous user requirements: Application to multi-services telecommunications systems," in *Proc. Perform. Data Commun. Syst. Appl.*, G. Pujolle, Ed. Amsterdam, The Netherlands: North Holland, 1981, pp. 423–431.

[26] 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.

[27] V. Iversen, "The exact evaluation of multi-service loss systems with access control," in *Proc. 7th Nordic Teletraffic Seminar*, Lund, Sweden, Aug. 1987, pp. 56–61.

[28] 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.

[29] T. Bonald, L. Massoulié, A. Proutière, and J. Virtamo, "A queueing analysis of max-min fairness, proportional fairness and balanced fairness," *Queueing Syst.*, vol. 53, nos. 1–2, pp. 65–84, Jun. 2006.

[30] 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), vol. 370, A. Kwiecień, P. Gaj, and P. Stera, Eds. Berlin, Germany: Springer, 2013, pp. 436–445.

[31] M. Stasiak, J. Wiewiora, P. Zwierzykowski, and D. Parniewicz, "Analytical model of traffic compression in the umts network," in *Proc. 6th Eur. Perform. Eng. Workshop, Comput. Perform. Eng.*, in Lecture Notes in Computer Science, vol. 5652, J. Bradley, Ed., London, U.K. Berlin, Germany: Springer-Verlag, Jul. 2009, pp. 79–93.

[32] M. Stasiak, M. Głąbowski, A. Wiśniewski, and P. Zwierzykowski, *Modeling and Dimensioning of Mobile Networks*. Hoboken, NJ, USA: Wiley, 2011.

[33] S. Hanczewski, M. Stasiak, and J. Weissenberg, "Model of a multiservice server with stream and elastic traffic," in *Image Processing and Communications*, M. Choraś and R. S. Choraś, Eds. Cham, Switzerland: Springer, 2020, pp. 287–294.

SŁAWOMIR HANCZEWSKI (Member, IEEE) received the M.Sc. and Ph.D. degrees in telecommunications from the Poznan University of Technology, Poland, in 2001 and 2006, respectively, and the D.Sc. degree from the Poznan University of Technology, in 2020. Since 2001, he has been working with the Poznan University of Technology, where he is currently with the Faculty of Computing and Telecommunications. He is also an Assistant Professor with the Institute of Communications and Computer Networks. He has authored or coauthored more than 60 scientific articles, related mostly to the analytical modeling of communications systems.

MACIEJ STASIAK (Member, IEEE) received the M.Sc. and Ph.D. degrees in electrical engineering from the Institute of Communications Engineering, Moscow, Russia, in 1979 and 1984, respectively, and the D.Sc. degree in electrical engineering from the Poznan University of Technology. From 1983 to 1992, he worked with the Polish Industry Sector as a Designer of Electronic and Microprocessor Systems. In 1992, he joined the Poznan University of Technology. In 2006, he was nominated as a Full Professor. He is currently the Director of the Faculty of Computing and Telecommunications, Institute of Communications and Computer Networks, Poznan University of Technology. He has authored or coauthored more than 300 scientific articles 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 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 the Poznan University of Technology, Poland, in 2015. The main area of her professional activity is Markovian analysis of multiservice networks. Her research interest includes application of stochastic processes theory in telecommunication systems.