IEICE TRANS. COMMUN., VOL.E99–B, NO.6 JUNE 2016

INVITED PAPER *Special Section on European ICT R&D Project Activities on Broadband Access Technologies in Conjunction with Main Topics of 2015 IEICE ICT Forum*

Queuing Systems for the Internet

Maciej STASIAK[†a)], ***Member***

SUMMARY This article proposes a versatile model of a multiservice queueing system with elastic traffic. The model can provide a basis for an analysis of telecommunications and computer network systems, internet network systems in particular. The advantage of the proposed approach is a possibility of a determination of delays in network nodes for a number of selected classes of calls offered in modern telecommunications networks.
key words: *multiservice traffic, queueing system, elastic traffic, Internet*

1. Introduction

Any analysis of multiservice networks requires appropriate queueing models for network nodes to be first developed, proven to be robust across a variety of specifications and validated per acceptable guidelines. A number of works discuss queueing models for modern packet networks, in particular models that make dimensioning of the Internet network and mobile networks possible. The bulk of the proposals considered in the literature of the subject lead to the application of a single-service Erlang model [1] that can be used to dimension systems of the Internet network within the context of M/G/R PS (M/G/R Processor Sharing) models [2]–[4]. Such an approach results from the close analogy of the PS mechanism and the mechanisms for providing reliable packet stream service in TCP/IP networks. These mechanisms are designed to distribute resources equally between serviced packet streams through full or partial equalisation of their throughput. This throughput equalisation can be effected by the introduction of a possibility to compress (and decompress) traffic, i.e. to decrease (increase) the flow capacity of streams, with simultaneous prolongation (or shortening) of their service time. In traffic engineering, traffic that is subjected to compression is called elastic traffic (e.g. TCP traffic) [5]. An application of M/G/R PS models to dimension systems with multiservice elastic traffic is then based on an "averaging" of the characteristics of all call streams and, in consequence, leads to a single-service stream serviced by an Erlang queue [1]. Proper dimensioning of such a queueing system is based on a choice of an appropriate "compression depth", i.e. such a guaranteed minimum throughput for packets for which the parameter, called in the M/G/R PS model the delay factor and dependent on the Erlang's Delay Formula, does not exceed the required value [3]. The Erlang's Delay Formula also provides the basis for the approach adopted in [6] in which the max-min fairness algorithm is used to allocate resources to individual packet streams in Internet systems [7]–[9]. In the case of the occupancy of the system, such a resource allocation algorithm, adopted in [6], prompts compression of those packet streams that require the highest throughput. Throughput of the remaining classes are not affected (no compression). The dimensioning of a system with the max-min fairness algorithm is based on finding such a guaranteed minimum bit rate to which streams with the highest bit rates can be compressed, so that the probability of waiting, determined on the basis of the Erlang's Delay Formula, will not exceed the required value. The basic shortcoming of the approaches presented above is the technical transformation of multiservice packet streams into single-service streams. If effected, this will lead to a possibility to estimate general, approximated queueing characteristics of the system, without a possibility of an accurate evaluation of queueing for individual packet streams.

The works of [10] and [11] propose an occupancy distribution for a multiservice full-availability system with losses. In [12], this distribution is generalised within the context of the full-availability system with finite compression, i.e. where elastic traffic streams serviced in the system can be compressed within determined limits. When the compression limit is exceeded, this causes a stream to be lost. In [13], the model [12] has been expanded to include a possibility of unlimited compression, which means that, in the case of the lack of free resources, streams of serviced traffic will always undergo the compression mechanism. As a result, this unlimited compression is equivalent to a lossless traffic service. Models [12] and [13] have been developed on the basis of a multi-dimensional Markov process that introduce appropriate values for service streams in those occupancy states in which compression takes place. The distribution of resources for serviced packet streams, adopted on the basis of the multi-dimensional service process of elastic traffic is consistent with the so-called balanced fairness algorithm [9], [14]. This algorithm ensures state-dependent service for all packet streams within the states in which compression is applicable.

Papers [15] and [16] propose a multiservice queueing model based on a queueing interpretation of the occupancy distribution in a full-availability system with elastic traffic [12]. The discipline for the service of the queue in [16] has been labelled SD FIFO (State Dependent FIFO). It is consistent with the allocation of resources in a multiservice server

Manuscript received January 25, 2016.
[†]The author is with the Faculty of Electronics and Telecommunications, Poznan University of Technology, Poland.
a) E-mail: maciej.stasiak@put.poznan.pl
DOI: 10.1587/transcom.2015EUI0001

for each class of offered traffic on the basis of the balanced fairness algorithm. Thus defined queueing system can be considered as M virtual queueing systems (M is the number of traffic classes offered to the system) with variable capacities of servers dependable on the number of streams of individual traffic classes that are being serviced or waiting in M virtual queues. This model makes it possible to determine the length of individual queues for particular classes of calls. The model does not include a possibility of servicing elastic traffic. The present article proposes a generalisation of the queuing model [16] to include a possibility to estimate queueing characteristics for individual elastic traffic streams. The idea behind this generalisation was originally presented in [17]. The possibility of traffic compression in the queueing system allows the model to be applied to analyse and dimension multiservice network systems, including Internet systems.

The article has been structured in the following way. Section 2 presents and discusses the basic traffic parameters that will be then applied to the analytical models presented in the article. Section 3 covers the basic multiservice model of system with losses as well as its generalisation to include a possibility of elastic traffic service. Section 4 presents a multiservice queueing model with the SD FIFO discipline. The latter section also proposes a generalistaion of the SD FIFO model for elastic traffic. In Sect. 5 the results of the analytical modelling are compared with the results of the simulation experiments for a number of selected traffic management scenarios in network systems.

2. Traffic Parametrization

2.1 Description of Traffic

The assumption in the article is that a call (frequently termed in the literature of the subject as flow, e.g. in [6], [9]) is defined as a stream of packets, or its part, related to a given service. As a result of a variety of studies, e.g. [18], there is a substantial body of evidence that call streams can be treated as Poisson streams. The application of the Poisson distribution to a description of the call stream enables us then to apply "Erlangesque" approach to the analysis of the considered systems. Such an approach greatly simplifies any analysis of a given system by providing a solution to an appropriate Markov process that corresponds to a given service process in the system. This constructed model, however, requires the adoption of the assumption of constant bit rates (CBR) for individual calls. In fact, the majority of packet streams transmitted over networks have variable bit rate (VBR). In traffic engineering this problem is solved by a replacement of variable bit rates of calls of individual classes by constant bit rates. CBR values can be chosen on the basis of the maximum bit rates of given VBR streams, or on the basis of the so-called equivalent bandwidth (EB) determined for offered call streams [19] and [20]. The first approach can be reduced in its essence to an "oversizing" of the system, since the maximum bit rates have only transient nature, whereas in other periods bit rates are generally lower than the maximum. It should be stressed, however, that the application of the maximum bit rate as a measure for a demand of a given call for the resources of the system is consistent with the engineering principle of dimensioning networks for "the worst case scenario". The other approach, far more realistic, proposes an adoption of CBR values for individual call streams based on the equivalent bandwidth that is determined according to the principle that the result of a service of a VBR call should be equal to the result of a service of a CBR call. Most of the EB evaluation methods available in the literature of the subject are based on heuristic algorithms that take into account such parameters as the maximum and average bit rate of a call, variance of the bit rate, admissible delay for a packet, and other parameters characteristic for this type of traffic stream and service system, e.g. [21]–[24].

The further assumption adopted in this article is that call throughputs of all classes are consistently determined by means of an application of one of the presented methods. A choice as to the method for a determination of CBRs for particular classes of calls depends on individual arrangements between the system designer and the network operator and has no direct influence upon the mathematical shape of the constructed model.

The assumption in the models considered in the present article is that all call streams offered to the system are Poisson streams and that they can be described by the following parameters:

- M - the number of call classes offered to the system,
- λ_i - intensity of the call stream of class i ($0 < i \leq M$),
- μ_i - average service intensity for a call of class i,
- c_i - bit rate of class i, (EB or maximum throughput),
- A_i - traffic intensity of traffic of class i:

$$A_i = \lambda_i/\mu_i. \tag{1}$$

2.2 System Discretisation

The knowledge of constant bit rates c_i, assigned to individual call classes makes it possible to carry on with the so-called bandwidth discretisation [21], which is based on a designation of an allocation unit (AU) for the system under investigation. AU determines such a bit rate c_{AU} that bit rates of offered calls are its multiple numbers [21]. The maximum value of an allocation unit can be described by the following formula:

$$c_{\mathrm{AU}} = \mathrm{GCD}(c_1, c_2, ..., c_M), \tag{2}$$

where the acronym GCD stands for Greatest Common Divisor.

A choice of an allocation unit enables us to express demands of individual classes t_i and the capacity of the system V in AUs:

$$t_i = \lceil c_i/c_{\mathrm{AU}} \rceil, \; V = \lfloor C/c_{\mathrm{AU}} \rfloor, \tag{3}$$

where C is the bit rate of the system.

It is convenient in engineering practice to adopt that the allocation unit will be equal to 1 bps, (or 1 kbps, 1 mbps, etc., depending on bit rate units adequate for a considered system) [16], i.e.:

$$c_{\mathrm{AU}} = 1 \text{ bps}. \tag{4}$$

Observe that by selecting the allocation unit on the basis of (4), the capacity of the system and demands of individual calls (expressed in AUs) take on values equal to values of bit rates of the system and individual calls:

$$V = C,\ t_i = c_i \text{ for } 1 \le i \le M. \tag{5}$$

3. Multiservice Models with Losses

3.1 Model of Multiservice Server without Traffic Compression

This system, otherwise called the multiservice full-availability group, will be termed in the article as the multiservice server with the capacity C_r AUs. The system is offered M Erlang traffic classes described by the parameters presented in Sect. 2. Figure 1 shows a schematic diagram of a multiservice full-availability system. The occupancy distribution in the multiservice server can be determined on the basis of recurrence equations [10] and [11]:

$$\begin{cases} [P_n]_{C_r} = \frac{1}{n} \sum\limits_{i=1}^{M} A_i c_i [P_{n-c_i}]_{C_r} & \text{for } 0 \le n \le C_r, \\ [P_n]_{C_r} = 0 & \text{otherwise}, \\ \sum\limits_{n=0}^{C_r} [P_n]_{C_r} = 1, \end{cases} \tag{6}$$

where $[P_n]_{C_r}$ is the occupancy probability for n AUs in the multiservice server with the capacity C_r AUs.

The blocking probability E_i for calls of a given class i $(0 < i \le M)$ in the multiservice server is determined by the lack of sufficient number of c_i AUs required to set up a connection:

$$E_i = \sum_{n=C_r-c_i+1}^{C_r} [P_n]_{C_r}. \tag{7}$$

Equations (6) and (7) form a generalisation of a model of single-service full-availability group [25] onto an instance of a number of traffic classes with differentiated demands.

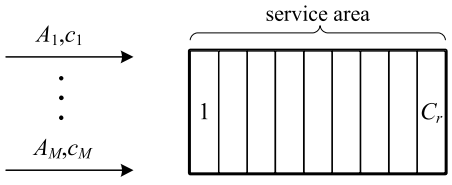

Fig. 1 Schematic diagram of multiservice server.

The formulae can be applied to analyse multiservice systems with losses in which traffic streams do not undergo any traffic formation mechanisms such as, for example, compression. From this perspective, the application of recurrence (6) to analyse TCP/IP systems is limited.

3.2 Model of Multiservice Server with Traffic Compression

This server is also called in the literature the full-availability group with elastic traffic. Elastic traffic service is as follows: a lack of free resources for a new call of a given class is followed by compression of all calls being serviced in the system, i.e. results in a decrease in their bit rates to the value at which a new call can be serviced (also in compressed form). At the same time, the service time for all calls gets prolonged, thus making transmission of all data possible. Figure 2 shows a schematic diagram of a multiservice server with traffic compression. Calls are compressed until the number of busy AUs in the server, determined on the basis of the sum of all uncompressed calls of all classes, exceeds a given virtual capacity of the server C_v, where $C_v > C_r$. If this capacity is exceeded, then a call will be aborted. The occupancy states of the virtual capacity of the server (i.e. such states n that $C_r < n \le C_v$) determine the compression area for elastic traffic. A choice of the value C_v is a "compression depth" indicator that is equal to the ratio of the virtual capacity of the server to the real capacity C_v/C_r and determines how many times the total bit rate of serviced calls in the system can be decreased at the maximum. The occupancy distribution in the multiservice server with traffic compression and the blocking probability E_i for each class of calls i $(0 < i \le M)$ can be determined on the basis of the recurrence equations [12]:

$$\begin{cases} [P_n]_{C_v} = \frac{1}{\min(n,C_r)} \sum\limits_{i=1}^{M} A_i c_i [P_{n-c_i}]_{C_v} & \text{for } 0 \le n \le C_v, \\ [P_n]_{C_v} = 0 & \text{otherwise}, \\ \sum\limits_{n=0}^{C_v} [P_n]_{C_v} = 1, \end{cases} \tag{8}$$

$$E_i = \sum_{n=C_v-c_i+1}^{C_v} [P_n]_{C_v}. \tag{9}$$

The parameter C_r in the expression $\max(n, C_r)$ determines the maximum possible service stream, expressed in the total number of busy AUs. The number of busy AUs will never

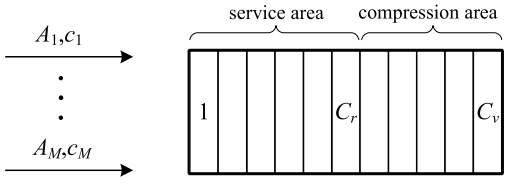

Fig. 2 Schematic diagram of multiservice server with traffic compression.

exceed the real capacity of the multiservice server. The average number of calls $y_i(n)$ of class i serviced in state n is expressed by Formula [16]:

$$y_i(n) = \frac{A_i[P_{n-c_i}]_{C_v}}{[P_n]_{C_v}}. \tag{10}$$

The distribution of the resources of the server between serviced call classes, determined by Formula (10), results from the analysis of a reversible Markov process - that leads to the occupancy distribution (8) - and is consistent with the operation of the balanced fairness algorithm for allocation of resources in a multiservice server.

3.3 Comments

The method for a calculation of the explicit occupancy distribution in a multiservice server has been addressed in a number of works, notably in [26]–[31]. The most effective and simple algorithms can be obtained on the basis of the recurrence distribution (6) [10] and [11]. In turn, the so-called convolution algorithms [29] and [31] allow the occupancy distribution for a multiservice server with Erlang, Engset and Pascal traffic to be determined. [32] proposes occupancy distribution for traffic that is characterised by any peakedness factor. In [33] and [34], a model of a multiservice server with call streams of the batched Poisson type is developed. [35] and [36] propose a recursive generalisation of the distribution (6) for Erlang, Engset and Pascal traffic.

Distribution (8) is proposed in [12] to be applied in an analysis of a multiservice server with limited traffic compression, and is then generalised in [13] to include a case of unlimited compression ($C_v \to \infty$), which means that each call can infinitely decrease its bit rate and increase its service time. Distribution (8) is also used in approximated analysis of multiservice servers with Erlang elastic and adaptive traffic [37], Engset traffic [38], [39] and with a call stream of the batched Poisson type [40]. Adaptive traffic, when free resources are missing, is compressed, i.e. bit rate of serviced calls is decreased, whereas the service time - unlike elastic traffic - does not change.

3.4 Area of Application

In systems with multiservice traffic, Formulae (6) and (7) perform the same function as the Erlang First Formula in systems with single-service traffic. It is worthwhile to add at this point that single-service multidimensional occupancy distributions in the full-availability group (a number of classes with identical demands) were also address by Erlang himself [1].

Distribution (6) can be used to analyse TCP/IP systems operating in low-load areas of a network, where the blocking probability can be, from the engineering point of view, ignored. This distribution can be applied to an approximate analysis of a multiservice server that concurrently services traffic without compression and adaptive and elastic traffic [41]–[43]. It is assumed in these models that traffic is compressed to the maximum, because it is only in such circumstances that losses may be induced. Models can describe certain systems of the TCP/IP network well and that also applies to interfaces of the 3G and 4G mobile networks in which delay for calls can be ignored, e.g. systems that service Real Time (RT) traffic [43].

Distribution (8) can be applied to a description of the above systems with elastic traffic, TCP/IP systems and mobile networks in particular, in which buffering is either non-existent or can be ignored.

4. Multiservice Queueing Models

4.1 Queueing Model without Traffic Compression

The queueing system is composed of a multiservice server with the capacity C_r AUs and a shared multiservice queue with the capacity of U AUs. Figure 3 shows a schematic diagram of a multiservice queueing system without traffic compression. The system operates as follows: a lack of free resources for a new call of a given class causes this call to be directed to the queue. It is proved in [15] and [16] that the distribution (8) has a queueing interpretation in which the compression area $(C_v - C_r)$ corresponds to the capacity of the queue U. Therefore, the occupancy distribution in this queueing system can be written in the following way:

$$\begin{cases} [P_n]_{C_r+U} = & \\ \frac{1}{\min(n,C_r)} \sum_{i=1}^{M} A_i c_i [P_{n-c_i}]_{C_r+U} & \text{for } 0 \le n \le C_r + U, \\ [P_n]_{C_r+U} = 0 & \text{otherwise}, \\ \sum_{n=0}^{C_r+U} [P_n]_{C_r+U} = 1. & \end{cases} \tag{11}$$

The average number of calls of particular classes, serviced in the server in state n AUs of the system (state n determines the number of AUs being serviced in the server and waiting in the queue), is determined by Formula (10) that can be rewritten in the notation of the considered queueing system:

$$y_i(n) = \frac{A_i[P_{n-c_i}]_{C_r+U}}{[P_n]_{C_r+U}}. \tag{12}$$

The average number of calls $x_i(n)$ of class i $(0 < i \le M)$ that are in the system in state n AUs, i.e. the average number

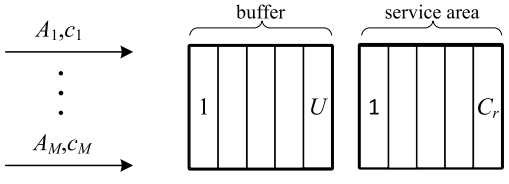

Fig. 3 Schematic diagram of queueing system without compression.

of calls of class i being serviced in the server and waiting in the queue in state n, can be expressed by the following recurrence formula:

$$x_i(n) = \frac{1}{\min(n, C_r)[P_n]_{C_r+U}} \times \times \left\{ \sum_{k=1}^{M} A_k c_k x_i(n - c_k)[P_{n-c_k}]_{C_r+U} + A_i c_i [P_{n-c_i}]_{C_r+U} \right\}, \tag{13}$$

where $x_i(0) = 0$. Formula (13) can be proved in a similar way as in [13] for the average number of serviced calls of class i in a server with infinite compression.

Formulae (12)–(13) make the important parameters of the system possible to be determined, e.g. the average length of the queue Q_i, expressed in the number of waiting calls of class i, and the average length of the queue q_i, expressed in the number of waiting AUs:

$$Q_i = \sum_{n=C_r+1}^{C_r+U} [x_i(n) - y_i(n)][P_n]_{C_r+U}, \tag{14}$$

$$q_i = Q_i c_i = c_i \sum_{n=C_r+1}^{C_r+U} [x_i(n) - y_i(n)][P_n]_{C_r+U}. \tag{15}$$

The total average length of the queue q for calls of all classes, expressed in the number of waiting AUs, can be expressed by formula [16]:

$$q = \sum_{n=C_r+1}^{C_r+U} [n - C_r][P_n]_{C_r+U}. \tag{16}$$

The average waiting times in the queue can be determined on the basis of (14) with the application of Little's formula [16]. Since the considered system has a finite queue, then it is a blocking system. The blocking probability E_i for calls of class i results from the lack of free c_i AUs in the queue:

$$E_i = \sum_{n=C_r+U-c_i+1}^{C_r} [P_n]_{C_r+U}. \tag{17}$$

The system under consideration is defined in [16] as a system with the SD FIFO (State Dependent FIFO) service discipline. Under the provision of this discipline a call stream of each class is always serviced in the server, while the distribution of resources for particular call classes depends on the total number of calls in a given state of the service process and is determined by Formula (12). Since this formula results from the analysis of a reversible Markov process, then the balanced fairness algorithm will be used to allocate resources in the server. Determined by Formulae (11)–(17), this model can be presented in the form of M virtual queueing systems [16] in which bit rates (capacities) for individual servers are changeable (state-dependent) and consistent with (12).

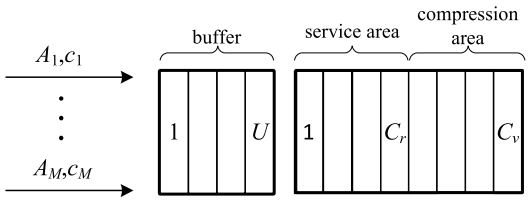

Fig. 4 Schematic diagram of queueing system with compression.

4.2 Queueing Model with Traffic Compression

A queueing system is composed of a multiservice server with the capacity C_v AUs and a shared queue with the capacity U AUs. Figure 4 shows a schematic diagram of a multiservice queueing system with traffic compression. The system services elastic traffic. If the server cannot service a call due to a lack of appropriate number of AUs, then currently serviced calls are compressed, i.e. bit rate is decreased and the service time increases. Calls are compressed until the number of busy AUs in the server, determined on the basis of uncompressed demands of calls of all classes, exceeds the virtual capacity of the server C_v, where $C_v > C_r$. If this capacity is exceeded, then a new call in its uncompressed form will queued.

The queueing model [16] was constructed on the basis of an appropriate interpretation of the model of a system with losses and elastic traffic [12]. Both models have identical occupancy distribution that, depending on the adopted interpretation, describes a system with losses and traffic compression or a queueing system. By giving the traffic compression interpretation to given sets of states and the queuing interpretation to others [17], we can get the occupancy distribution in the queueing system with call compression in the server that can be written in the following form:

$$\begin{cases} [P_n]_{C_v+U} = & \\ \frac{1}{\min(n,C_r)} \sum_{i=1}^{M} A_i c_i [P_{n-c_i}]_{C_v+U} & \text{for } 0 \le n \le C_v + U, \\ [P_n]_{C_v+U} = 0 & \text{otherwise}, \\ \sum_{n=0}^{C_v+U} [P_n]_{C_v+U} = 1, & \end{cases} \tag{18}$$

where the parameter C_r in the expression $\max(n, C_r)$ defines, similarly as in (8), the maximum possible service stream, expressed in the total number of busy AUs in the server.

The states ($0 \le n \le C_r$) determine the operation of the system without compression (server without compression, empty buffer). The states ($C_r < n \le C_v$) determine the operation of the system with compression (server with compression, empty buffer). The states ($C_v < n \le C_v + U$) determine the operation of the system in the queueing mode (server with compression, buffer partly or entirely occupied).

The average number of calls $y_i(n)$ of class i ($0 < i \le$

M) serviced in the server in state n AUs in the system and the average number of calls $x_i(n)$ of class i that are in the system in state n are determined by Formulae (12) and (13) that, in the notation of the considered queueing model with compression, will be written as follows:

$$y_i(n) = \frac{A_i[P_{n-c_i}]_{C_v+U}}{[P_n]_{C_v+U}}, \tag{19}$$

$$x_i(n) = \frac{1}{\min(n, C_r)[P_n]_{C_v+U}} \times \times \left\{ \sum_{k=1}^{M} A_k c_k x_i(n-c_k)[P_{n-c_k}]_{C_v+U} + A_i c_i [P_{n-c_i}]_{C_v+U} \right\}, \tag{20}$$

where $x_i(0) = 0$. The parameters: the average length of the queues Q_i and q_i, expressed in the number of waiting calls and the number of waiting AUs of class i, respectively, the average queue length q for calls of all classes, expressed in the number of waiting AUs, blocking probability E_i for calls of class i, are determined by Formulae (14)–(17) that, in the notation of the considered system, can be rewritten in the following form:

$$Q_i = \sum_{n=C_v+1}^{C_v+U} [x_i(n) - y_i(n)][P_n]_{C_v+U}, \tag{21}$$

$$q_i = Q_i c_i = c_i \sum_{n=C_v+1}^{C_v+U} [x_i(n) - y_i(n)][P_n]_{C_v+U}, \tag{22}$$

$$q = \sum_{n=C_v+1}^{C_v+U} [n - C_r][P_n]_{C_v+U}, \tag{23}$$

$$E_i = \sum_{n=C_v+U-c_i+1}^{C_v+U} [P_n]_{C_v+U}. \tag{24}$$

The queueing model with traffic compression in the server proposed above is characterised, as in the previous model, by the SD FIFO service discipline. This means that the multiservice server services calls of all classes and can be presented in the form of M virtual queueing systems [16] in which the capacities of individual virtual servers are defined by Formula (19). Such a division of resources is consistent with the balanced fairness algorithm that allocates resources to particular classes in multiservice systems.

4.3 Comments

Distribution (11) is proposed in [15] and [16]. In [15], the relationship of the distribution (11) with a model of a server with infinite compression is discussed [13]. The idea of the multiservice queueing system with traffic compression is presented for the first time in [17].

4.4 Areas of Application

The presented models of the multiservice queueing system and the multiservice queueing system with elastic traffic (11) and (18) are exact models for the SD FIFO queue to which the corresponding algorithm is the balanced fairness algorithm for resource allocation in a multiservice server. In these models the server guarantees service to all classes of calls, therefore it can be used to approximate a large number of network systems. In [44], a possibility of the application of the distribution (11) to analyse radio interfaces of mobile LTE network is reported. Distribution (18) makes it possible to take into consideration a possibility of traffic compression in queueing system. Its main advantage is that it can be particularly well-suited for analysis, dimensioning and optimization of TCP/IP systems.

5. Case Study

This section presents the results of a comparison of the analytical calculations, relative to the models described in the article, with the results of the simulation experiments for a number of selected queueing systems.

5.1 System Parameters

The following structural parameters were adopted in the study: $C_r = 50$ AUs, $C_v = 80$ AUs, $U = 30$ AUs, where 1 AU = 10 Kbps. The system was offered three traffic classes with the parameters: $c_1 = 1$ AUs, $c_2 = 3$ AUs and $c_3 = 7$ AUs. Traffic was offered in the proportion $A_1c_1 : A_2c_2 : A_3c_3 = 1 : 1 : 1$. Figures 5–8 show the results of the analytical modelling and the simulation in relation to the average traffic a, offered to one AU in the server:

$$a = \sum_{i=1}^{M} A_i c_i / C_r. \tag{25}$$

All simulation experiments were carried out for 5 series, 1 000 000 calls each. The results of the simulation are indicated by appropriate symbols with the 95% confidence interval.

5.2 Results

Figure 5 shows a comparison of the average queue lengths (21) in the system with call compression (distribution (18)), expressed in the number of calls of individual classes, with the results of the simulation of the FIFO queueing system with the maximum use of the resources of the server. This means that if a server runs low on resources, lower than the number of AUs demanded by the first call in the queue, then it starts its service with the available number of AUs. This mode of operation for the queue service is somewhat similar to the operation of many network systems.

Figure 6 shows a graph presenting the average queue lengths (14) in the system with call compression in which the resources of the server are allocated according to the max-min fairness algorithm [6], i.e. the system induces compression of these packet streams that require the highest throughput. Throughputs of the remaining classes are

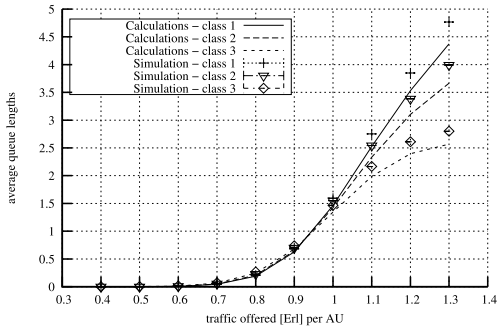

Fig. 5 The average queue lengths (FIFO dyscipline).

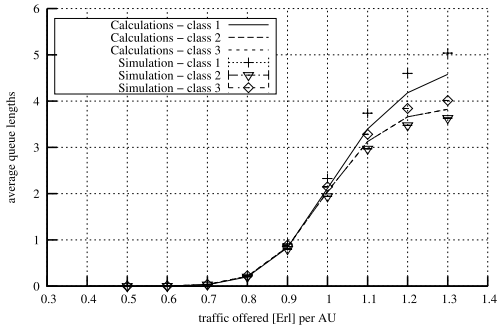

Fig. 6 The average queue lengths (max-min fairness algorithm).

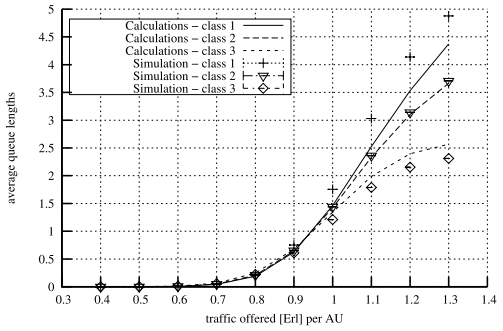

Fig. 7 The average queue lengths (proportional algorithm).

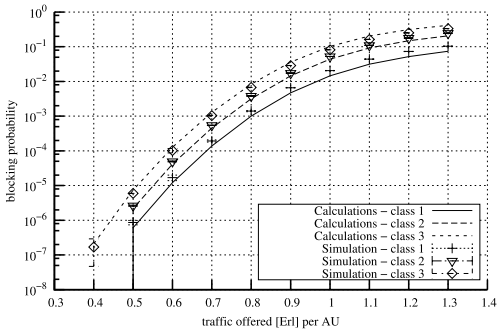

Fig. 8 The blocking probability (proportional algorithm).

not compressed. The assumption in the experiment was that calls of two older classes were compressed to the value $c_2 = c_3 = c_{min} = 2$ AUs. Distribution (11) was used for modelling using the approach [41]–[43] according to which calls are forwarded to the queue only at the maximum compression.

Figure 7 and Fig. 8 present graphs of the average queue lengths and the blocking probabilities (Formulae (21) and (24)) in the queueing system with compression in which a weight algorithm was used to allocate resources in the server. The assumption was that weights of resource allocation w_i for particular classes of calls were proportional to offered traffic:

$$w_i = A_i c_i / \sum_{k=1}^{M} A_k c_k. \tag{26}$$

All the results for the presented models attest to and concur well with the results of the simulation.

6. Conclusions

This article presents two models of multiservice queueing systems. One of the models supports elastic traffic. Both models make a determination of individual queue parameters for particular classes of calls possible. Such an approach may prove to be useful in an analysis of different scenarios for the operation of multiservice networks, in particular for TCP/IP network and the 4G and 5G mobile networks with elastic traffic. The presented SD FIFO queueing models can approximate both systems with variable and constant distribution of resources in a server. The models can thus provide a basis for development of appropriate engineering methods for dimensioning and optimization of multiservice network systems. Additionally, the models proposed in the article can contribute to the development of further studies on multiservice systems with finite and infinite queues and different service disciplines in queues.

The article presents only a fraction of all possible applications for the proposed models. Approximation boundaries, and consequently a choice of appropriate models that would best suit different real scenarios for the operation of network systems, should be investigated within the context of relevant demands on the part of engineers and operators responsible for adequate dimensioning and maintenance of networks.

Acknowledgment

The presented work has been funded by the Polish Ministry of Science and Higher Education within the status activity task "*Structure, analysis and design of modern switching system and communication networks*" in 2016.

References

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

[2] J.W. Cohen, "The multiple phase service network with generalized processor sharing," Acta Informatica, vol.12, no.3, pp.245–284, 1979.
[3] K. Lindberger, "Balancing quality of service, pricing and utilisation in multiservice networks with stream and elastic traffic," Proc. 16th International Teletraffic Congress, Edinburgh, pp.1127–1136, UK, 1999.
[4] N. Dukkipati, M. Kobayashi, R. Zhang-Shen, and N. McKeown, "Processor sharing flows in the Internet," Quality of Service — IWQoS 2005, Lecture Notes in Computer Science, vol.3552, pp.271–285, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.
[5] IETF, "RFC 793: Transmission control protocol," 1981, edited by Jon Postel. [Online]. Available: https://tools.ietf.org/html/rfc793
[6] T. Bonald and J.W. Roberts, "Internet and the Erlang formula," SIGCOMM Comput. Commun. Rev., vol.42, no.1, pp.23–30, Jan. 2012.
[7] J. Bertsekas and R.G. Gallager, Data Networks, 2nd ed., Prentice Hall, 1992.
[8] F. Kelly, "Mathematical modelling of the internet," in Mathematics Unlimited — 2001 and Beyond, B. Engquist and W. Schmid, Eds., pp.685–702, Springer, 2001.
[9] 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, no.1-2, pp.65–84, 2006.
[10] J. Kaufman, "Blocking in a shared resource environment," IEEE Trans. Commun., vol.29, no.10, pp.1474–1481, 1981.
[11] J. Roberts, "A service system with heterogeneous user requirements — Application to multi-service telecommunications systems," Proc. Performance of Data Communications Systems and their Applications, G. Pujolle, Ed., pp.423–431, 1981.
[12] 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, June 1997.
[13] T. Bonald and J. Virtamo, "A recursive formula for multirate systems with elastic traffic," IEEE Commun. Lett., vol.9, no.8, pp.753–755, 2005.
[14] J.-P. Haddad and R.R. Mazumdar, "Congestion in large balanced multirate networks," Queueing Syst., vol.74, no.2-3, pp.333–368, 2013.
[15] S. Hanczewski, M. Stasiak, and J. Weissenberg, "The queueing model of a multiservice system with dynamic resource sharing for each class of calls," Computer Networks, Communications in Computer and Information Science, vol.370, pp.436–445, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.
[16] 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.E97-B, no.8, pp.1592–1605, Aug. 2014.
[17] S. Hanczewski, D. Kmiecik, M. Stasiak, and J. Weissenberg, "Multiservice queuing system with elastic traffic," Proc. IEICE Gen. Conf., 2016, BS-3-18, March 2016.
[18] V. Paxson and S. Floyd, "Wide area traffic: The failure of Poisson modeling," IEEE/ACM Trans. Netw., vol.3, no.3, pp.226–244, June 1995.
[19] J.Y. Hui, "Resource allocation for broadband networks," IEEE J. Sel. Areas. Commun., vol.6, no.9, pp.1598–1608, Dec. 1988.
[20] F. Kelly, "Notes on effective bandwidth," Tech. Rep., University of Cambridge, 1996.
[21] J. Roberts, V. Mocci, and I. Virtamo, Eds., Broadband Network Teletraffic, Final Report of Action COST 242, Commission of the European Communities, Springer, Berlin, 1996.
[22] R. Guerin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," IEEE J. Sel. Areas. Commun., vol.9, no.7, pp.968–981, Sept. 1991.
[23] I. Norros, "On the use of fractional Brownian motion in the theory of connectionless networks," IEEE J. Sel. Areas. Commun., vol.13, no.6, pp.953–962, Aug. 1995.
[24] A. Pras, L. Nieuwenhuis, R. van de Meent, and M. Mandjes, "Dimensioning network links: A new look at equivalent bandwidth," IEEE Netw., vol.23, no.2, pp.5–10, March 2009.
[25] A. Erlang, "Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges," Elektrotechnikeren, vol.13, p.5, 1917.
[26] L. Gimpelson, "Analysis of mixtures of wide- and narrow-band traffic," IEEE Trans. Commun., vol.13, no.3, pp.258–266, Sept. 1965.
[27] J. Aein, "A multi-user-class, blocked-calls-cleared, demand access model," IEEE Trans. Commun., vol.26, no.3, pp.378–385, March 1978.
[28] H. Inose, An Introduction to Digital Integrated Communications Systems, University of Tokyo Press, Tokyo, 1979.
[29] V. Iversen, "The exact evaluation of multi-service loss system with access control," Seventh Nordic Teletraffic Seminar, pp.56–61, Lund, Aug. 1987.
[30] A. Nilson and M. Perry, "Provisioning models for digital loop carriers," Proc. 13th International Teletraffic Congress, vol. Discussion Circle, pp.271–276, Copenhagen, 1991.
[31] K.W. Ross, Multiservice Loss Models for Broadband Telecommunication Networks, Springer London, London, 1995.
[32] L. Delbrouck, "On the steady-state distribution in a service facility carrying mixtures of traffic with different peakedness factors and capacity requirements," IEEE Trans. Commun., vol.31, no.11, pp.1209–1211, 1983.
[33] 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, 1993.
[34] J.S. Kaufman and K.M. Rege, "Blocking in a shared resource environment with batched Poisson arrival processes," Perform. Evaluation, vol.24, no.4, pp.249–263, 1996.
[35] M. Głąbowski, "Modelling of state-dependent multirate systems carrying BPP traffic," Ann. Telecommun., vol.63, no.7-8, pp.393–407, 2008.
[36] M. Głąbowski, M. Stasiak, and J. Weissenberg, "Properties of recurrent equations for the full-availability group with BPP traffic," Math. Probl. Eng., vol.2012, pp.1–17, 2012.
[37] S. Rácz, B.P. Gerö, and G. Fodor, "Flow level performance analysis of a multi-service system supporting elastic and adaptive services," Perform. Evaluation, vol.49, no.1-4, pp.451–469, 2002.
[38] G. Fodor and M. Telek, "Bounding the blocking probabilities in multirate CDMA networks supporting elastic services," IEEE/ACM Trans. Netw., vol.15, no.4, pp.944–956, Aug. 2007.
[39] V.G. Vassilakis, I.D. Moscholios, and M.D. Logothetis, "Call-level performance modelling of elastic and adaptive service-classes with finite population," IEICE Trans. Commun., vol.E91-B, no.1, pp.151–163, Jan. 2008.
[40] I. Moscholios, J. Vardakas, M. Logothetis, and A. Boucouvalas, "Congestion probabilities in a batched Poisson multirate loss model supporting elastic and adaptive traffic," Ann. Telecommun., vol.68, no.5-6, pp.327–344, June 2013.
[41] M. Stasiak, J. Wiewióra, P. Zwierzykowski, and D. Parniewicz, "Analytical model of traffic compression in the UMTS network," Computer Performance Engineering, Lecture Notes in Computer Science, vol.5652, pp.79–93, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.
[42] M. Stasiak and P. Zwierzykowski, "Analytical model of the Iub interface carrying HSDPA traffic in the UMTS network," Management Enabling the Future Internet for Changing Business and New Computing Services, Lecture Notes in Computer Science, vol.5787, pp.536–539, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.
[43] M. Stasiak, M. Głąbowski, A. Wiśniewski, and P. Zwierzykowski, Modeling and Dimensioning of Mobile Networks: From GSM to LTE, John Wiley & Sons, 2011.
[44] S. Hanczewski, M. Stasiak, and P. Zwierzykowski, "Modelling of the LTE radio interface for NRT traffic," 2014 16th International Telecommunications Network Strategy and Planning Symposium

(Networks), pp.1–7, 2014.

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