Modelling Overflow Systems with Queuing in Primary Resources

Mariusz Głąbowski(✉), Damian Kmiecik, and Maciej Stasiak

Faculty of Electronics and Telecommunications, Poznan University of Technology,
Poznań, Poland
mariusz.glabowski@put.poznan.pl

Abstract. This article proposes a new method to determine the characteristics of multiservice overflow systems with queueing systems. A number of methods have been developed that have the advantage of determining the parameters of traffic directed to secondary resources as well as providing a way to model these resources. The accompanying assumption is that queues with limited capacities are used in primary resources. The results of analytical calculations are compared with the results of simulation experiments for a number of selected structures of overflow systems with queueing in primary resources. The results of the study confirm high accuracy of the proposed method and, in consequence, the accuracy of the theoretical assumptions adopted for the proposed method.

Keywords: Overflow system with queueing · Multiservice traffic · Blocking probability

1 Introduction

Traffic overflow is one of the oldest and best known mechanisms for traffic distribution optimization in networks. Traffic overflow is based on the principle that when certain resources, called primary resources, are fully occupied, the traffic overflow mechanism allows calls that are still offered to the resources to be directed (i.e. to overflow) to some alternative resources, called secondary resources [24].

Overflow systems in single-service telecommunications systems with losses have been widely addressed in the literature, e.g. [5,26]. In [26], the analytical modeling of single-service overflow systems is expanded to include the ERT method (Equivalent Random Traffic). The work [5] proposes a simple method for an analysis of overflow systems, i.e. the so-called Hayward's method based on a modification of Erlang B Formula. The author of [13] points out the fact that the call stream offered to secondary resources can be approximated, with acceptable accuracy, by a Pascal call stream. Single-service overflow models with queueing have been addressed is a large number of works [15,18–20]. The queue to which overflow traffic with the peakedness factor higher than unity is offered is analyzed

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2019
Published by Springer Nature Switzerland AG 2019. All Rights Reserved
T. Q. Duong et al. (Eds.): Qshine 2018, LNICST 272, pp. 148–157, 2019.
https://doi.org/10.1007/978-3-030-14413-5_12

in [11]. Works such as [15,20] are devoted to the analysis of the two-dimensional Markov process in an overflow system in which queues are introduced both to primary and secondary resources. An overflow system that services two traffic streams, one of which is queued is described in [3]. In [18] and [19], overflow traffic in systems with queues is described on the basis of the two and three first moments of the IPP process (Interrupted Poisson Process) [16].

The problem of multiservice traffic overflow has been widely discussed in the literature. The article [17] analyzes overflow systems on the basis of Markov-Modulated Poisson Process (MMPP), while in [14] these systems are analysed on the basis of Batched Poisson Process (BPP). The works [4,14] provide analyses of traffic that is characterized by an appropriate peakedness factor that can approximate overflow traffic. A methodology for dimensioning multiservice overflow systems is proposed in [7,12]. In the proposed methods, Hayward's approach is generalized [5] to model secondary resources, which is based on a division of the averaged values offered to secondary resources and the capacities of secondary resources by the peakedness factors of relevant traffic scenarios. The paper [6] considers a model in which secondary resources are of distributed nature, which means that they are composed of a number of separated resources with full availability [8]. In [25], overflow systems in which traffic can change the service parameters, such as the service time and bitrate in secondary resources, are analyzed. Then, in [6], a possibility of elastic traffic service in the overflow system is investigated. As yet, multiservice overflow models with queueing have not been considered in the literature.

This work proposes a model of a multiservice overflow system with queueing for primary resources. To model a queueing system of primary resources, a multiservice queueing model with state-dependent service disciple called SD FIFO (State Dependent FIFO) is used [10,23]. The model is based on a queueing interpretation of the system of resources with losses that supports elastic traffic [21,22]. The SD FIFO discipline corresponds to a resource distribution in a multiservice server according to the balanced fairness algorithm [1,9] that approximates very well other resource distribution algorithms, e.g. the proportional (with regard to offered traffic) algorithm. To model secondary resources, this article takes advantage of the generalized Hayward model [7,12]. The article is structured as follows. Section 2 provides an outline of the subject of research, i.e. a multiservice overflow system in which calls offered to primary resources can be queued. Section 3 includes a model of primary resources, description of a method for a determination of the parameters of traffic that overflows from these resources and a model of secondary resources. In Sect. 4, the results of analytical calculations are compared with the results of simulation experiments for a number of selected structures of overflow systems with queueing for calls in primary (and secondary) resources. Section 5 sums up the article.

2 Traffic Overflow Systems with Queueing for Primary Resources

This study considers multiservice overflow systems in which, when primary resources are fully occupied, calls of different traffic classes will be first directed to queues, where they will stay until enough resources to serve them have been released. Until now, the literature of the subject on multiservice overflow systems has not considered a possibility of queueing of calls before they are directed to secondary resources. The assumption in this article is that queues have a limited capacity and that calls that cannot wait in a queue will be directed to a system of secondary resources. Figure 1 shows a general diagram of the traffic overflow system with queueing in primary resources. The following notation is used:

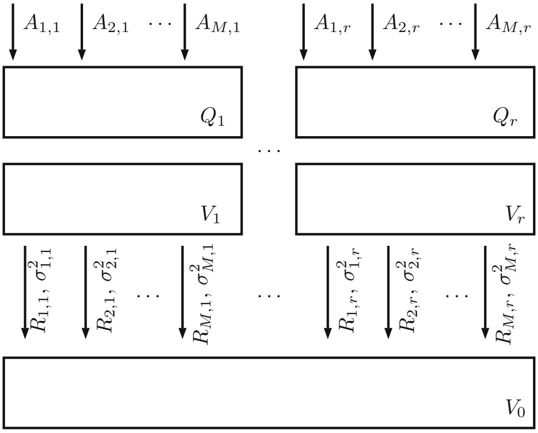
$A_{1,1}$
$A_{2,1}$
$\cdots$
$A_{M,1}$
$A_{1,r}$
$A_{2,r}$
$\cdots$
$A_{M,r}$
Q_1
Q_r
$\cdots$
V_1
V_r
$R_{1,1}, \sigma^2_{1,1}$
$R_{2,1}, \sigma^2_{2,1}$
$\cdots$
$R_{M,1}, \sigma^2_{M,1}$
$\cdots$
$R_{1,r}, \sigma^2_{1,r}$
$R_{2,r}, \sigma^2_{2,r}$
$\cdots$
$R_{M,r}, \sigma^2_{M,r}$
V_0

Fig. 1. Diagram of a multiservice overflow system with queues in primary resources

- r – number of component primary resources in the overflow system,
- Q_k – queue capacity in primary resource k $(1 \leq k \leq r)$,
- V_k – capacity of primary resource k $(1 \leq k \leq r)$,
- V_0 – capacity of secondary resource,
- M – number of traffic classes of traffic offered in overflow system,
- $A_{i,k}$ – traffic intensity of class i $(1 \leq i \leq M)$ offered to resources k $(1 \leq k \leq r)$,
- $R_{i,k}$ – traffic intensity of class i $(1 \leq i \leq M)$ that overflows from resource k $(1 \leq k \leq r)$,
- $\sigma^2_{i,k}$ – variance of traffic intensity of class i $(1 \leq i \leq M)$ that overflows from resource k $(1 \leq k \leq r)$.

2.1 Traffic Offered to Primary Resources

Modern networks, including the Internet, are packet networks. Packets that belong to a service that is being executed create traffic streams that can be analysed in exactly the same way as calls (or flows) are analyzed [1,2,21]. A mathematical analysis of the internal structure of these streams (calls) is very complex and frequently leads to solutions that exclude them from being applied in practice. Hence, multiservice systems are analyzed at the call level. The result of many studies show that calls can be approximated by streams of "Poisson-like" character [1,2]. Such an approach makes it possible to discretize the system and then to construct models that are based on multi-dimensional Markov processes. Discretization is a change in the variable bitrate (VBR) of a packet stream that belongs to a given call in such a way as to convert it into a fixed, constant fictitious bitrate (CBR). The assumption in the most recent works, in particular those that refer to modeling of the TCP/IP network, is that such a fictitious bitrate is equal to the maximum bitrate (flow rate) of a real packet stream that corresponds to calls. The knowledge of fictitious bitrates for individual call classes makes it possible to determine the allocation unit (AUs) of a given overflow system. The maximum value of the AU is described as the Greatest Common Divisor (GCD) of all fictitious bitrates of the calls offered to an overflow system:

$$c_{\mathrm{AU}} = \mathrm{GCD}(c_1, c_2, \ldots, c_M), \tag{1}$$

where c_i is the maximum bitrate of a packet (call) stream of class i, whereas c_{AU} is the bitrate of the allocation unit. Having determined the value of the AU, both the capacity of the system V and the volume of the resources necessary for a connection of class i to be set up is discretized, i.e. expressed in AUs:

$$V = \frac{C}{c_{\mathrm{AU}}}, \tag{2}$$

$$t_i = \frac{c_i}{c_{\mathrm{AU}}}, \tag{3}$$

where C the total bitrate of a given resource of the system. The assumption in the model proposed in this article is that both the capacity of the primary and secondary resources and the demanded volume of resources necessary for calls of individual classes to be set up is expressed in AUs.

2.2 Primary and Secondary Resources

Fig. 1 shows a general system for traffic overflow. The primary resources are composed of r component primary resources, where each resource k has the capacity V_k, expressed in AUs. Each component resource can service M Erlang traffic classes. If a call of class i cannot be serviced by a primary resource k, it is directed to a queue in the component resource k with the capacity Q_k AUs. A lack of free space, i.e. t_i AUs, in this queue, will cause the call of class i to be directed to the secondary resource with the capacity of V_0 AUs. A lack of free t_i AUs in the secondary resources will, in turn, lead to irrevocable loss of the call.

3 Models of Resources in the Overflow System

The model of the traffic overflow system with queueing in primary resources is composed of a model of primary resources, a model for a determination of the average value and variance for individual overflow traffic classes and a model of secondary resources. These three models allow us to determine all important characteristics of the overflow system, in particular the blocking probability for call streams offered to the system. To model each of the primary resources, a multiservice SD FIFO model was used [10], whereas to determine the parameters of traffic that overflows from the primary resources, the method developed in [7] was used.

3.1 Model of Component Primary Resources

In the article, to model the component primary resources with queueing, a recurrent occupancy distribution in a queueing SD FIFO system [10] was used that, according to the notation adopted in the article, can be written in the following way:

$$[P(n)]_{V_k+Q_k} = \begin{cases} \frac{1}{n}\sum_{i=1}^{M} A_{i,k} t_i \left[P(n-t_i)\right]_{V_k+Q_k} & \text{for } 0 \leq n \leq V_k, \\ \frac{1}{V_k}\sum_{i=1}^{M} A_{i,k} t_i \left[P(n-t_i)\right]_{V_k+Q_k} & \text{for } V_k < n \leq V_k + Q_k, \end{cases} \tag{4}$$

where n is the number of AUs occupied by calls that are currently in a component resource k (serviced and queued), whereas the distribution $[P(n)]_{V_k+Q_k}$ defines the occupancy probability n AUs in a system with the capacity of the component resource V_k and the queue capacity Q_k. The blocking probability in such a system results from the finite capacity of the queue and for calls of class i can be determined by the formula:

$$E_{i,k} = \sum_{n=V_k+Q_k-t_i+1}^{V_k+Q_k} [P(n)]_{V_k+Q_k} \, . \tag{5}$$

Formula (5) determines the sum of all blockable states for calls of class i, i.e. those states in which the number of free AUs in a queue is lower than the number of AUs demanded for a call of class i to be set up. Queuing service discipline for a SD FIFO queue corresponds to a resource allocation for each class of offered traffic on the basis of the balanced fairness algorithm.

The model approximates very well different resource distribution algorithms for serviced calls, in particular the proportional algorithm [10].

3.2 Model of Traffic that Overflows Form Primary Resources

Traffic of class i that overflows from the primary resource k can be described by the three following parameters: the average value of traffic intensity $R_{i,k}$, the variance $\sigma^2_{i,k}$ and the number of AUs t_i necessary for a given connection to be executed. The parameter $R_{i,k}$ can be determined on the basis of the loss in traffic in the component resources with queues that are in the blocking state. Therefore, the average value $R_{i,k}$ of the traffic intensity of traffic of class i that overflows from a component queuing system k, can be, on the basis of (5), determined by the following formula:

$$R_{i,k} = A_{i,k} E_{i,k}. \tag{6}$$

To determine the variance of traffic that overflows from the primary resources, a modification to the method proposed in [12] will be used. The proposed modification is based on a decomposition of each of the component primary resources V_k with the queue Q_k into M fictitious resources with the capacities $V^*_{i,k}$, each of them servicing exclusively calls of one class i with the traffic intensity $A_{i,k}$. The assumption is that the blocking probability of class i in a single-service fictitious resource is exactly the same as the blocking probability of class i in the primary component queueing system k. Since the fictitious resource also services Erlang traffic, then its capacity can be determined on the basis of Erlang's function:

$$E_{i,k} = E_{V^*_{i,k}(A_{i,k})} = \frac{(A_{i,k})^{V^*_{i,k}}}{(V^*_{i,k})!} \Bigg/ \sum_{n=0}^{V^*_{i,k}} \frac{(A_{i,k})^n}{n!}. \tag{7}$$

Now, on the basis of Riordan's formula, the variance can be approximated by the variance of traffic that overflows from the fictitious resources [26]:

$$\sigma^2_{i,k} = R_{i,k} \left(\frac{A_{i,k}}{V^*_{i,k} + 1 - A_{i,k} + R_{i,k}} + 1 - R_{i,k} \right). \tag{8}$$

The peakedness factor of traffic that overflows from the primary component resources k is then equal to:

$$Z_{i,k} = \frac{\sigma^2_{i,k}}{R_{i,k}}. \tag{9}$$

3.3 Model of Secondary Resources

To model the occupancy distribution in multiservice secondary resources to which a mixture of overflow traffic is offered, the generalized Hayward distribution, proposed in [7] can be used. According to the notation adopted in the article, this distribution can be written in the following way:

$$n\,[P(n)]_{V_0/Z_0} = \sum_{i=1}^{M} \sum_{k=1}^{r} \frac{R_{i,k}}{Z_{i,k}} t_i\,[P(n-t_i)]_{V_0/Z_0}\,, \tag{10}$$

where the coefficients $Z_{i,k}$ are determined on the basis of (9), whereas the parameter Z_0 is the so-called aggregate peakedness factor. The factor Z_0 can be determined on the basis of the weighted average of the coefficients $Z_{i,k}$, assuming that the weights are determined by the proportion of AUs demanded by calls of individual classes in relation to the demands of all traffic classes. Therefore:

$$Z_0 = \sum_{i=1}^{M} \sum_{k=1}^{r} Z_{i,k} \frac{R_{i,k} t_i}{\sum_{s=1}^{r} \sum_{j=1}^{M} R_{j,s} t_j}. \tag{11}$$

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

$$E_{i,0} = \sum_{n=V_0/Z_0-t_i+1}^{V_0/Z_0} \left[P(n)\right]_{V_0/Z_0}. \tag{12}$$

Formulas (10)–(12) allow the blocking probabilities in a multiservice overflow system with queues in the primary resources to be determined.

4 Results of Modeling of a Selected Overflow System with Queues in Primary Resources

The method for a determination of the blocking probability in systems with overflow traffic and call queuing capabilities in primary resources presented in the article is an approximate method. To evaluate its accuracy and adopted assumptions for the method, the results of the analytical calculations of the blocking probability in the alternative resources were compared with the data obtained in simulation experiments. Studies were carried out for a large number of network structures with traffic overflow. Due to the space constraints, Figs. 2, 3 and 4 present the results for an example network composed of two primary resources and one secondary resource. The parameters of the system under investigation

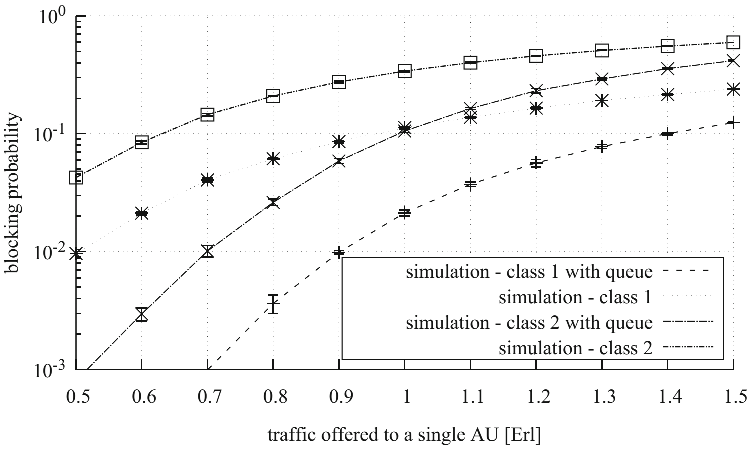

Fig. 2. Blocking probability in the primary resources no. 1

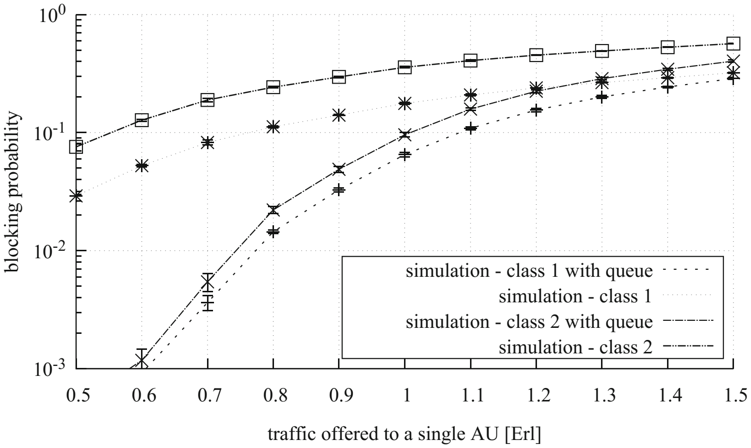

Fig. 3. Blocking probability in the primary resources no. 2

were as follows: $r = 2$, $M = 3$, $V_0 = 15$; $V_1 = 20$, $Q_1 = 10$, $t_1 = 1$, $t_3 = 3$; $V_2 = 10$, $Q_2 = 10$, $t_1 = 1$, $t_2 = 2$.

The impact of using the queue in primary resources of the presented system can be observed in Figs. 2 and 3. The calls blocked in the primary resources are then offered to the alternative resources. The comparison of the simulation and analytical results of the blocking probability in the alternative resources can be further observed in Fig. 4. The data obtained on the basis of the simulation study are presented as points with the confidence intervals calculated after the t-Student distribution (with 95-percent confidence level) for 5 series with 10000 calls each (the classes with the lowest call intensity). The obtained value of the

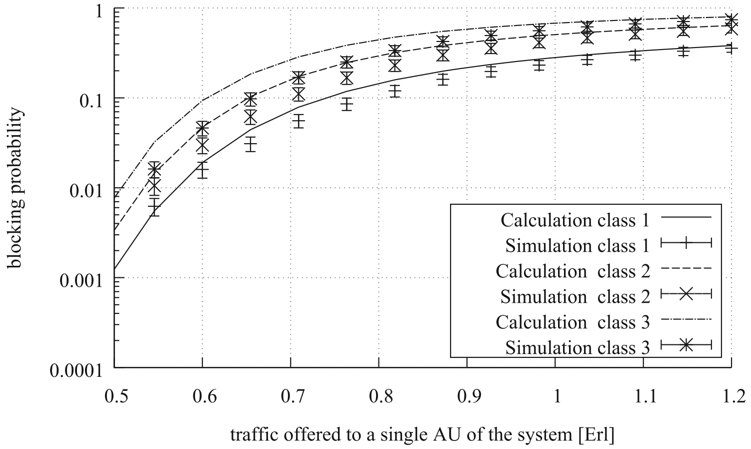

Fig. 4. Blocking probability in the alternative resources

confidence interval for each of the results of the simulation is lower by at least an order of magnitude than the average value obtained in the simulation study. In a large number of instances, the value of the confidence interval is lower that the height of the symbol representing the simulation result.

5 Conclusions

This article proposes an analytical model of a multiservice hierarchical telecommunications system. The assumption in the model is that queues, that allow the blocking probability to be decreased, are introduced in the primary resources of the traffic overflow system under consideration. To validate the developed analytical model, the results of the calculations are compared with the simulation data. Both the data presented in the article and the results gathered by the authors as a result of relevant comparative study indicate that the proposed model provides high accuracy in calculations. Further studies will be expanded to include other types of offered traffic and a possibility of introducing queues also in alternative resources.

Acknowledgements. This paper was developed as a result of the research project 2016/23/B/ST7/03925 entitled “Modelling and service quality evaluation of Internet-based services” funded by the National Science Centre.

References

1. Bonald, T., Massoulié, L., Proutière, A., Virtamo, J.: A queueing analysis of max-min fairness, proportional fairness and balanced fairness. Queueing Syst. **53**(1), 65–84 (2006). https://doi.org/10.1007/s11134-006-7587-7
2. Bonald, T., Roberts, J.W.: Internet and the Erlang formula. ACM SIGCOMM Comput. Commun. Rev. **42**(1), 23–30 (2012). https://doi.org/10.1145/2096149.2096153
3. Brune, G.: On delay and loss in a switching system for voice and data with internal overflow. In: Proceedings of 11th International Teletraffic Congress, pp. 2.1–2.7. North-Holland, Kyoto (1985)
4. Delbrouck, L.: On the steady-state distribution in a service facility carrying mixtures of traffic with different peakedness factors and capacity requirements. IEEE Trans. Commun. **31**(11), 1209–1211 (1983)
5. Fredericks, A.: Congestion in blocking systems - a simple approximation technique. Bell Syst. Tech. J. **59**(6), 805–827 (1980)
6. Glabowski, M., Kaliszan, A., Stasiak, M.: Modelling overflow systems with distributed secondary resources. Comput. Netw. **108**, 171–183 (2016). https://doi.org/10.1016/j.comnet.2016.08.015
7. Glabowski, M., Kubasik, K., Stasiak, M.: Modeling of systems with overflow multi-rate traffic. Telecommun. Syst. **37**(1–3), 85–96 (2008). https://doi.org/10.1007/s11235-008-9070-8
8. Głąbowski, M., Stasiak, M.: Multi-rate model of the group of separated transmission links of various capacities. In: de Souza, J.N., Dini, P., Lorenz, P. (eds.) ICT 2004. LNCS, vol. 3124, pp. 1101–1106. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-27824-5_143

9. Haddad, J.P., Mazumdar, R.R.: Congestion in large balanced multirate networks. Queueing Syst. **74**(2), 333–368 (2013). https://doi.org/10.1007/s11134-012-9322-x
10. Hanczewski, S., Stasiak, M., Weissenberg, J.: A queueing model of a multi-service system with state-dependent distribution of resources for each class of calls. IEICE Trans. Commun. **E97–B**(8), 1592–1605 (2014)
11. Heffes, H.: Analysis of first-come first-served queuing systems with peaked inputs. Bell Syst. Tech. J. **52**(7), 1215–1228 (1973). https://doi.org/10.1002/j.1538-7305.1973.tb02014.x
12. Huang, Q., Ko, K.T., Iversen, V.B.: Approximation of loss calculation for hierarchical networks with multiservice overflows. IEEE Trans. Commun. **56**(3), 466–473 (2008)
13. Iversen, V.: Teletraffic engineering handbook. Technical report, Technical University of Denmark, Lyngby (2010)
14. Kaufman, J.S., Rege, K.M.: Blocking in a shared resource environment with batched Poisson arrival processes. J. Perform. Eval. **24**(4), 249–263 (1996). https://doi.org/10.1016/0166-5316(94)00029-8
15. Kaufman, L., Seery, J.B., Morrison, J.A.: Overflow models for dimension PBX feature packages. Bell Syst. Tech. J. **60**(5), 661–676 (1981). https://doi.org/10.1002/j.1538-7305.1981.tb00255.x
16. Kuczura, A.: The interrupted Poisson process as an overflow process. Bell Syst. Tech. J. **52**(3), 437–448 (1973). https://doi.org/10.1002/j.1538-7305.1973.tb01971.x
17. Lagrange, X., Godlewski, P.: Performance of a hierarchical cellular network with mobility-dependent hand-over strategies. In: Proceedings of Vehicular Technology Conference - VTC, pp. 1868–1872. IEEE, Atlanta, April 1996. https://doi.org/10.1109/VETEC.1996.504082
18. Matsumoto, J., Watanabe, Y.: Individual traffic characteristics queueing systems with multiple poisson and overflow inputs. IEEE Trans. Commun. **33**(1), 1–9 (1985). https://doi.org/10.1109/TCOM.1985.1096202
19. Meier-Hellstern, K.S.: Parcel overflows in queues with multiple inputs. In: Proceedings of 12th International Teletraffic Congress, pp. 3.1–3.8. North-Holland, Torino (1988)
20. Morrison, J.A.: Analysis of some overflow problems with queuing. Bell Syst. Tech. J. **59**(8), 1427–1462 (1980). https://doi.org/10.1002/j.1538-7305.1980.tb03373.x
21. Rácz, S., Gerő, B.P., Fodor, G.: Flow level performance analysis of a multi-service system supporting elastic and adaptive services. Perform. Eval. **49**(1–4), 451–469 (2002). https://doi.org/10.1016/S0166-5316(02)00115-3
22. Stamatelos, G.M., Koukoulidis, V.N.: Reservation-based bandwidth allocation in a radio ATM network. IEEE/ACM Trans. Netw. **5**(3), 420–428 (1997). https://doi.org/10.1109/90.611106
23. Stasiak, M.: Queuing systems for the internet. IEICE Trans. Commun. **E99–B**(6), 1224–1242 (2016)
24. Stasiak, M., Glabowski, M., Wiśniewski, A., Zwierzykowski, P.: Modeling and Dimensioning of Mobile Networks. Wiley, Hoboken (2011)
25. Wang, M., Li, S., Wong, E., Zukerman, M.: Performance analysis of circuit switched multi-service multi-rate networks with alternative routing. J. Lightwave Technol. **32**(2), 179–200 (2014). https://doi.org/10.1109/JLT.2013.2289925
26. Wilkinson, R.I.: Theories of toll traffic engineering in the USA. Bell Syst. Tech. J. **40**, 421–514 (1956)