BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **18**, No 1
Sofia • 2018 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2018-0003

Packet-Level Link Capacity Evaluation for IP Networks

Seferin T. Mirtchev

Technical University of Sofia, 1756 Sofia, Bulgaria
E-mail: stm@tu-sofia.bg

Abstract: *In recent times, with many applications, the IP networks have become the most powerful tool for sharing information. Best-effort IP interconnected networks deliver data according to the available resources, without any assurance of throughput, delay bounds, or reliability requirements. As a result, their performance is highly variable and cannot be guaranteed. In IP networks, ensuring proper link capacity at the packet level is a challenging problem. In this article, a method to evaluate the link capacity of IP networks at the packet level based on a single server delay system with state-dependent arrival and departure processes is suggested. The dependence of the traffic being carried on the queue length and on the defined waiting time is shown. Presented graphic dependencies allow for defined quality of service, namely the probability of packet loss and admissible delays, to determine the carried traffic of the links.*

Keywords: *Link capacity, packet level, generalised single server delay queue, state-dependent arrival process, throughput, peaked flow, queue length, carried traffic, network congestion, best-effort IP network, overload regime, congestion, packet loss, packet delay, demand-capacity-performance relation.*

1. Introduction

The best-effort IP interconnected networks cannot guarantee quality of service. The planning and building of IP networks is a complex process, which is based on teletraffic engineering, but also uses many practical rules. This leads not only to inefficiencies due to inappropriate sizing but also to misconceptions regarding the effectiveness of traffic controls. Many studies that have been performed have considered Internet traffic in terms of stochastic processes of packet, flow, and session arrivals and the main problem is a lack of awareness of known results and their implications as pointed out by Bonald and Roberts [3]. For assuring Quality of Service (QoS), understanding the demand-capacity-performance relation of IP networks is required.

Operators adjust contention ratios and oversubscription periodically in order to attain the quality of experience and revenue targets while keeping an eye on the expenses. An overprovisioned best-effort IP network can meet most user requirements and has the advantages of relatively low capital and operational costs as pointed out by Bonald and Roberts [3]. This network cannot ensure high throughput and low latency for real-time services.

The packet switching in the Internet and the huge number of different services and applications lead to the specific characteristics of traffic flows. Poisson processes describe the random processes in telephone networks with packet switching. In IP networks, the arrival and service processes have specific properties as long-range dependence, self-similar, or fractal-like behavior as pointed out by Paxson and Floyd [22].

In recent years, increasing interest in the development of models and methods for classical queueing systems (especially the Erlang C formula) to study Internet networks has led to the development of many extensions of previously existing results. For these new models and methods, the phrase "Internet Erlang Formula" is used.

In Iversen [10], a robust and efficient algorithm for evaluating multi-service multi-rate queueing systems, including finite buffer systems and loss systems, based on Erlang formulae is presented. In Erik, Misuth and Weber [6], the possibility of Erlang B and Erlang C formula utilization in the next generation networks is discussed.

A model based on queueing theory for service performance in cloud computing is presented in Vilaplana, Solsona and Teixido [27]. In Salah, Elbadawi and Boutaba [24], an analytical model based on Markov chains to predict the number of cloud instances or virtual machines needed to satisfy a given service level objective performance requirement such as response time, throughput, or request loss probability is presented.

In this article, a method to evaluate the relation between demand, capacity, and performance in IP networks through the generalization of the classical single server delay system with state-dependent arrival and departure rates is suggested. It is shown that a queueing system with state-dependent arrival and departure processes has specific application for analyzing of IP networks. Determining the packet-level link capacity allows for efficient mechanisms operation of the congestion management in the modern telecommunications networks with packet switching.

2. State-of-the-art of the problem in literature

The teletraffic engineering provides useful tools for the stochastic processes modelling in telecommunication networks. Usually the queueing models are widely used in the network planning and evaluation of the quality of service as pointed out by Giambene [8]. There is no single traffic model that can efficiently capture the traffic characteristics of all types of networks. Traffic modeling is a basic requirement for accurate capacity planning as pointed out by Mirtchev at al. [18]. The queueing systems are used to evaluate the parameters of the quality of service as probability of packet loss, average packet delay and throughput.

A new analytical model of IEEE 802.11 network with distributed coordination function to channels access based on single finite queue M/M/1/ k is presented in Kosek-Szott [13]. The proposed model allows evaluating the throughput, the delays and the frame blocking probability.

An issue of congestion in IEEE 802.11 networks caused by the time-variant channel capacity and the contention among neighbouring nodes is presented in Jain at al. [11]. The effect of congestion which degrades the throughput and increases delay and packet drops is carried out by OPNET simulations.

The performance of a network of femto cells with a final capacity using single channel delay system M/M/1/k is evaluated in terms of quality of service parameters such as packet blocking probability, average packet delay and utilization for different buffer sizes as pointed out by Kumar, Aaamir and Qadeer [15].

The nodes of a wireless mesh networks are modelled as a combination of two single channel systems M/M/1/k to distinguish transit and local traffic generated in Chen and Reisslein [5]. The developed analytical model is used to evaluate the throughput and delays of a clustered FiWi network.

An algorithm for routing in Mobile Ad hoc NETworks (MANET) to make more effective use of the network resources by adjusting the traffic probabilistically is propose in Joardar, Bhattacherjee and Giri [12]. The algorithm (Swarm **i**nspired Congestion **a**ware probabilistically Load – SiCaL) identifies the congestion areas among nodes to the destination to avoid the congestion in the intermediate links and also minimise the packet loss in the network.

In order to support multiple applications and to effectively utilize the resources by dynamically scheduling the users, radio resource management procedure is one of the key design roles for improving the LTE system performance. The main scheduling algorithms used in LTE network are analysed in Sulthana and Nakkeeran [25] by means of many quality of service parameters like goodput, delay, packet loss ratio and fairness index using an open source simulator LTEsim.

The proposed traffic scheduling algorithm in wireless mesh network is evaluated by link model using the M/D/1 queue as pointed out by Naeini [21].

The dependence of the improvement in throughput of the service networks when it is added additional capacity on the traffic load is investigated in Ziya [28] by considered the M/M/1/m, M/G/1/m-PS, and M/G/c/c queues.

The hidden Markov models (D-BMAP/D/1/k) is presented in Moltchanov, Koucheryavy and Harju [20] with an arrival process of frames and service process in the wireless channel.

The optimal buffer capacity k for the M/M/1/k queue under some standard cost and reward structures by comparing various Markov reward processes is studied in Hautphenne and Haviv [9].

A complicated single-server queueing system M(n)/G/1/k with state-dependent arrivals and general service distribution is studied in Chao and Rahman [4]. The server is turned on when more than defined number of customers are present, and off only when the system is empty. After the server is turned off, it does not operate until at least a defined number of customers is present in the system.

The feedback information on the buffer state provides the basis for the Transmission Control Protocol (TCP) to regulate carefully the transmission rate of Internet flows. To evaluate this behaviour, a G/G/1-type queue with workload-dependent arrival rate and service speed is considered in Bekker and Boxma [2].

In the cited publications, which are based on the single server delay systems, the task to evaluate the link throughput in the modern telecommunication networks is not directly placed.

3. IP link capacity evaluation

Internet traffic is due to the interaction among millions of users, hundreds of heterogeneous applications, and dozens of sophisticated protocols as pointed out by Swift and Dagli [26]. The technical components of the Internet are complex in themselves and they are augmented by a general unpredictability and diversity of human components as pointed out by Fomenkov at al. [7]. Internet traffic is variable, with individual connections ranging from extremely short to extremely long and from extremely low-rate to extremely high-rate as pointed out by Bonald and Roberts [3].

There are three traffic regimes that help in understanding the performance requirements in IP networks – transparent, elastic and overload are presented in Bonald and Roberts [3].

The transparent regime occurs when all flows have a relatively low peak rate. The probability that the sum of the rates is less than the link capacity is very high. The congestion probability is negligible and the delays are very small in this regime. The contention between arrival packets can be solved by a simple first-in, first-out buffer.

In the elastic regime, some flows have a high peak rate and try to occupy the whole link capacity. The buffers overflow and the networks flows will have significant losses and delays. To improve the quality of service, rates of some flows must be reduced. The transmission control protocol normally realizes the necessary adjustment. With max-min fairness, only the high peak rate flows are constrained for rate reduction. The others maintain their rate and suffer negligible loss, as experienced in the transparent regime as pointed out by Bonald and Roberts [3].

The overload regime occurs when the demand exceeds the link capacity. The congestion probability and the delays are very high for all flows and the performance is very poor in this regime. The bursty traffic flows lead to frequently and lengthy overloads. This regime should be avoided through appropriate traffic engineering.

The stochastic processes that describe the traffic flow in packet switching networks, can be smooth, regular, or peaked as pointed out by Mirtchev [17]. In these three cases, the variance in the number of arrival packets is smaller than, equal to, or greater than the mean value, respectively. The typical transmission of packets in networks is in the form of bursts (a large amount of data sent in a short time). The burstiness is caused by the nature of the data being communicated and it leads to peaked traffic flow in the networks. Hence, there are many studies that describe complicated queueing systems with specific behaviour such as long-range dependence, peakedness, self-similar, and heavy tail distributions. This behaviour for IP network at packet level can be analysed by queueing systems with state-dependent arrival and service rates.

Queueing systems with arrival and/or service rates depending on the system's state arise in various application areas as pointed out by Abouee-Mehrizi and Baron [1] and by Kumar and Dharsana [14]. The state-dependent features make it possible to describe burstiness in IP networks.

4. Generalised single server delay system

Simple models like the classical single server queues can often be used to obtain comprehensive results, e.g., to predict global traffic behaviour as pointed out by Mittal and Jain [19].

In Mirtchev [16], a single server queueing system with adaptable arrival and service speeds based on the amount of work required right after customer arrivals or departures is investigated. The classic M/M/1/k single server delay system is generalised to nonlinear state-dependent arrival and service rate. Two peakedness factors, i.e., *p* for arrival and *q* for departure processes, are defined for the single server queue M(g)/M(g)/1/k with a state-dependent generalised Poisson arrival process *M*(*g*), state-dependent exponentially distributed service time *M*(*g*), and limited waiting rooms *k* (Fig. 1).

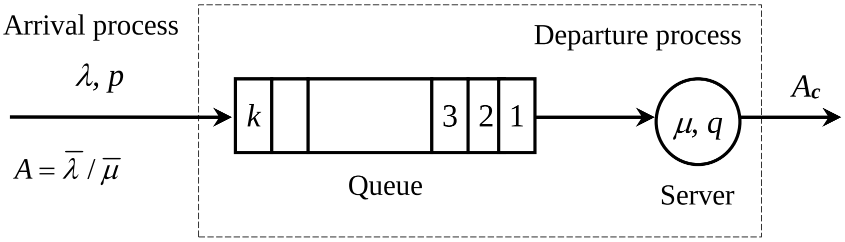

Fig. 1. Generalised single server delay system – M(g)/M(g)/1/k queue

Generalised Poisson arrival processes and generalised Bernoulli departure processes are said to be peaked, regular, or smooth according to whether the corresponding peakedness factors are greater than, equal to, or smaller than one, respectively.

The finite state-transition diagram of the investigated single server queue with state-dependent arrival and service rates is shown in Fig. 2.

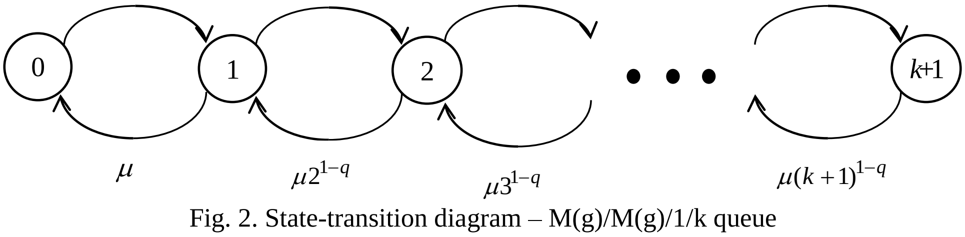

Fig. 2. State-transition diagram – M(g)/M(g)/1/k queue

By applying state-dependent arrival and service intensity to the general solution of the birth and death processes and using traffic intensity when the system is empty $a = \lambda/\mu$, steady state probabilities can be obtained as pointed out by Mirtchev [16]:

(1) $$P_j = \frac{a^j / (j!)^{1/p-q}}{\sum_{i=0}^{k+1} a^i / (i!)^{1/p-q}} \quad \text{for} \quad j = 0, 1, 2, ..., k+1.$$

The average value of the state-dependent arrival intensity is

(2) $$\bar{\lambda} = \sum_{j=0}^{k+1} \lambda_j P_j = \lambda \sum_{j=0}^{k+1} (j+1)^{1-1/p} P_j.$$

The average value of the state-dependent departure intensity is

(3) $$\bar{\mu} = \sum_{j=1}^{k+1} \mu_j \frac{P_j}{1-P_0} = \frac{\mu}{1-P_0} \sum_{j=1}^{k+1} j^{1-q} P_j.$$

The offered traffic is calculated by means of the average arrival and departure intensity

(4) $$A = \bar{\lambda} / \bar{\mu}.$$

The average departure intensity is reciprocal of the mean service time

(5) $$\bar{\mu} = 1 / \bar{\tau}.$$

The carried traffic is equivalent to the probability that the system is busy:

(6) $$A_c = 1 - P_0.$$

The time congestion probability B describes the fraction of time for which all waiting rooms are busy:

(7) $$B = P_{k+1}.$$

The waiting probability is denoted by $P(t_w>0)$ which means that the waiting time probability is greater than 0

(8) $$P(t_w > 0) = \sum_{j=1}^{k} P_j = 1 - P_0 - P_{k+1}.$$

Let us assume the First-In-First-Out (FIFO) discipline. The waiting time distribution function $P(t_w>t')$ is defined as the probability of waiting time exceeding defined time interval t'. From the probability theory it is given by

(9) $$P(t_w > t') = \sum_{i=1}^{k} P_i Q_i(> t').$$

An arbitrary packet enters service when i packets are in the system P_i. Since the service time is exponentially distributed the conditional probability $Q_i(>t')$ that j ($j \le i$) packets terminate in the time interval (0, t'] becomes a generalized Poisson distribution. The probability $Q_l(t')$ that l packets terminate in the time interval (0, t'] *is* a generalized Poisson distribution with a mean value $\bar{\mu}_l$,

(10) $$Q_l(t') = \frac{(\bar{\mu}_l t')^l}{l!} e^{-\bar{\mu}_l t'}.$$

The conditional probability $Q_i(>t')$ that the arbitrary packet has to wait longer than t', given i ($k \ge i \ge 1$) packets in the system, is expressed by

(11) $$Q_i(>t') = \sum_{r=0}^{i-1} \frac{(\bar{\mu}_i t')^r}{r!} e^{-\bar{\mu}_i t'}.$$

The mean departure intensity of the generalized Poisson distribution is

(12) $$\bar{\mu}_i = \left(\sum_{r=1}^{i} \mu_r P_r \right) \Bigg/ \left(\sum_{r=1}^{r} P_r \right).$$

5. Numerical results

In this section, we provide numerical results obtained by a computer program. The maximum carried traffic by the link with a given mean value and variance of the arrival and departure processes at a defined congestion probability due to buffer overflow and the dependence of the carried traffic from the defined time interval, normalized to the average service time, at a certain probability to wait more than a defined time interval and at a determined queue length is calculated by the bisection iteration method in numerical analysis. The accuracy of the applied analytical modeling is very good because we used recursion formulas for state probabilities calculation and robust bisection iteration method that in our case is relatively fast.

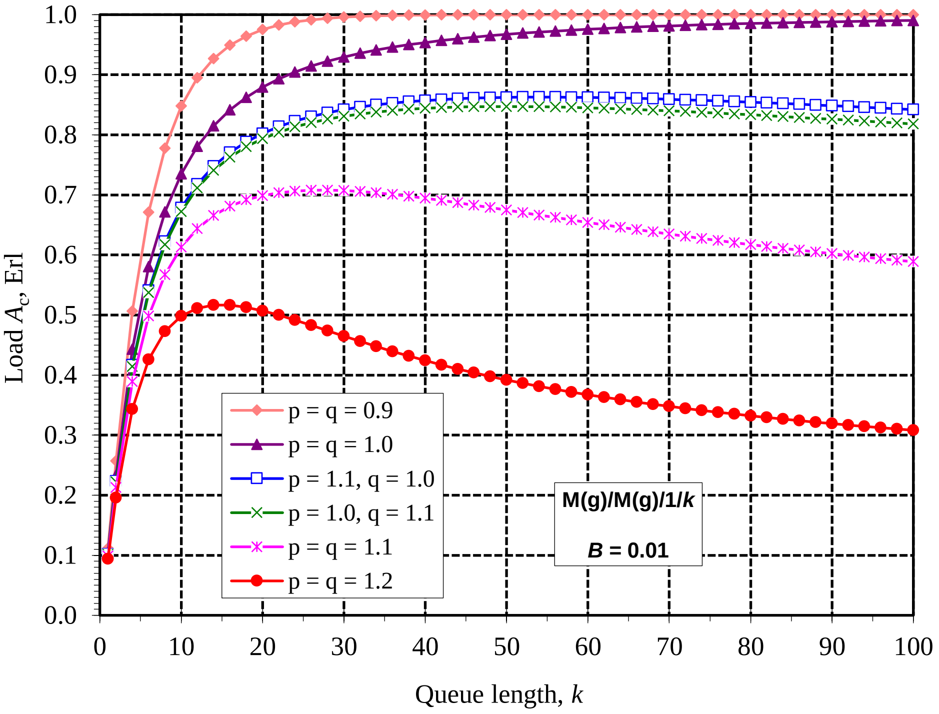

Fig. 3. Admissible load A_c as a function of queue length k for different values of arrival p and service q peakedness factors at defined congestion probability B=0.01 in a single server queue M(g)/M(g)/1/k

Fig. 3 shows the admissible load A_c in a single server queue M(g)/M(g)/1/k as a function of queue length k for different values of arrival p and service q peakedness factors and defined small packet congestion probability B=0.01. We can see high throughput when the arrival and departure processes are smooth and also the behaviour of the classical single server queue M/M/1/k when the peakedness factors p and q are equal to 1. The burstiness of the arrival and departure processes can decrease the carried traffic to 0.3 Erl. High burstiness leads to decrease in the carried traffic when the queue length increases. When the arrival and service processes are peaked it is better to have small queue 15-25 packets for better utilisation and smaller delays.

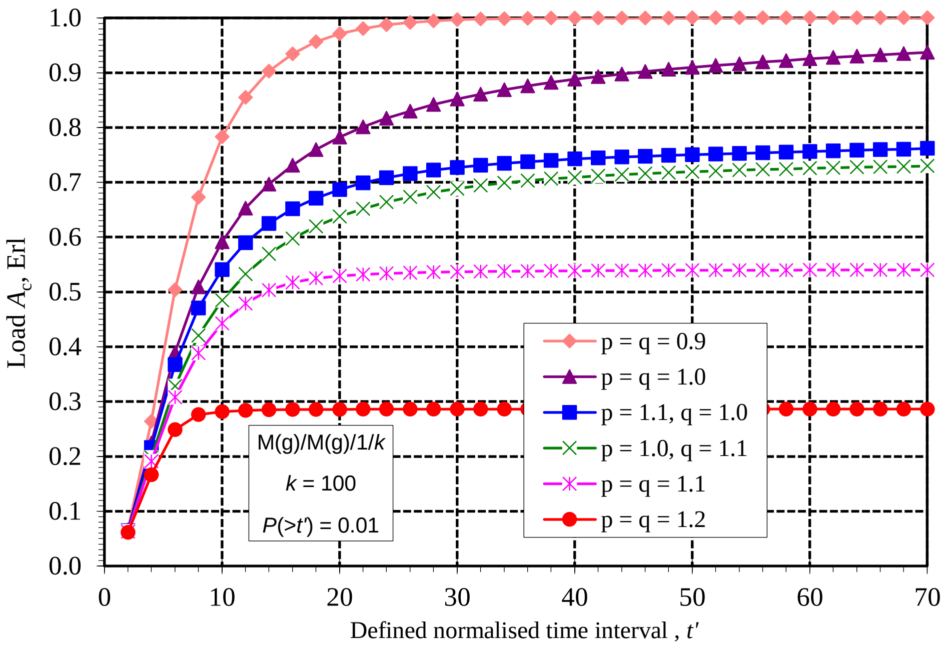

Fig. 4. Admissible load A_c as a function of defined normalized time interval t' for different values of arrival p and service q peakedness factors at determined probability to wait more than a defined time interval $P(>t')$ = 0.01 and queue length k = 100 in a single server queue M(g)/M(g)/1/k

Fig. 4 presents the admissible load A_c in a single server queue M(g)/M(g)/1/k as a function of defined normalised time interval t' for different values of arrival p and service q peakedness factors at small value of the probability to wait more than a defined time interval $P(>t')$ = 0.01 and queue length k = 100 packets. We can see high throughput when the arrival and departure processes are smooth and also the behaviour of the classical single server queue M/M/1/k when the peakedness factors p and q are equal to 1. The burstiness of the arrival and departure processes can decrease the carried traffic more than 0.3 Erl. High burstiness leads to negligible carried traffic changes when the defined time interval increases. When the arrival and departure processes are peaked it is better to have small value of the defined time interval equal to from $10\bar{\tau}$ to $20\bar{\tau}$ for better utilisation and smaller delays.

The presented results show that the influence of the peakedness factors over the performance measures is significant.

In modern IP networks it is usually transmitted packages with big length in the order of 500, 1000 and 1500 bytes. When we know the transmission speed of the link, we can easily calculate the mean service time. By admissible delays for different services the normalised time intervals will be determined. For the calculated minimum time interval and selected small value of the probability to wait more than this time interval the carried traffic will be determined.

6. Conclusion

In this article a method for the evaluation of the link capacity in IP networks at the packet level based on generalized Poisson arrival and Bernoulli departure processes with state-dependent arrival and service rates is proposed. The presented method for determining the link carried traffic at defined admissible congestion probability and at selected small value of the probability to wait more than a defined time interval based on the generalized single server queue M(g)/M(g)/1/k allows for accurate sizing of telecommunication networks and to improve quality of service.

The graphic dependence of the carried traffic on the queue size at defined packet congestion probability, and the dependence of the carried traffic on the defined time interval at a determined probability to wait more than a defined time interval and at a determined queue length gives a possibility to determine the links utilization at defined quality of service. Determining the bandwidth of the links allows for efficient operation of the management of congestion in modern telecommunications networks with packet switching.

The importance of the suggested method is due to its ability to describe peaked, regular, and smooth behavior of the teletraffic flows. The proposed method provides a unified framework to describe behavior that can be found in up-to-date networks. This is the case in general teletraffic systems and it is an important feature in designing telecommunications networks.

In conclusion, we believe that the presented method for link capacity evaluation in IP networks at the packet level will be practically useful when planning telecommunication networks. The determination of the link capacity for IP networks is important to provide quality of services and to avoid congestions and bottlenecks in the network.

References

1. Abouee-Mehrizi, H., O. Baron. State-Dependent M/G/1 Queueing Systems. – Queueing Sys., Vol. **82**, 2016, No 1, pp. 121-148.
2. Bekker, R., O. Boxma. An M/G/1 Queue with Adaptable Service Speed, SPOR-Report (Reports in Statistics, Probability and Operations Research). Eindhoven University of Technology, 2005.
3. Bonald, T., J. Roberts. Internet and the Erlang Formula. – ACM SIGCOMM Comput. Commun. Review, Vol. **42**, 2012, No 1, pp. 23-30.
4. Chao, X., A. Rahman. Analysis and Computational Algorithm for Queues with State-Dependent Vacations II: M(n)/G/1/K. – Jrl Syst Sci & Complexity, 2006, No 19, pp. 191-210.

5. Chen, P., M. Reisslein. A Simple Analytical Throughput-Delay Model for Clustered FiWi Networks. – Photonic Network Communications, Vol. **29**, 2015, No 1, pp 78-95.
6. Erik, C., T. Misuth, A. Weber. Application of Erlang Formulae in Next Generation Networks. – Int. J. Comput. Network Inf. Security, Vol. **4**, 2012, No 1, pp. 59-66.
7. Fomenkov, M., K. Keys, D. Moore, K. Claffy. Longitudinal Study of Internet Traffic from 1998-2003. – In: Proc. of Winter International Symposium on Information and Communication Technologies (WISICT), 2004, pp. 1-6.
8. Giambene, G. Queuing Theory and Telecommunications: Networks and Applications. Springer, 2005.
9. Hautphenne, S., M. Haviv. The Bias Optimal k in the M/M/1/k Queue: An Application of the Deviation Matrix. – Probability in the Engineering and Informational Sciences, Vol. **30**, 2016, pp. 61-78.
10. Iversen, V. The Internet Erlang Formula, Proceedings: Internet of Things, Smart Spaces, and Next Generation Networking. – In: Proc. of 12th International Conference, NEW2AN and 5th Conference, ruSMART 2012, Springer, 2012, pp. 328-337 (Lecture Notes in Computer Science, Vol. **7469**).
11. Jain, A., S. Sharma, R. Jha, K. Manoj. Effect of Congestion on the Performance of IEEE 802.11 Network. – Int. J. of Information and Communication Technology, Vol. **1**, 2008, No 3/4, pp. 318-328.
12. Joardar, S., V. Bhattacherjee, D. Giri. SiCaL: A Swarm Inspired Congestion Aware Probabilistically Load Balance Routing in MANETs. – Int. J. of Information and Communication Technology, Vol. **7**, 2015, No 6, pp. 585-606.
13. Kosek-Szott, K. Throughput, Delay, and Frame Loss Probability Analysis of IEEE 802.11 DCF with M/M/1/K Queues. – In: Proc. of IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), UK, 2013, pp. 2234-2238.
14. Kumar, E., S. Dharsana. Analysis of M/M/1 Queueing System with State Dependent Arrival and Detainment of Retracted Customers. – Malaya Journal of Mathematic, Special Issue, 2015, No 1, pp. 89-98.
15. Kumar, W., S. Aamir, S. Qadeer. Performance Analysis of a Finite Capacity Femtocell Network. – Mehran University Research Journal of Engineering & Technology, Vol. **33**, 2014, No 1, pp. 129-136.
16. Mirtchev, S. Single Server Queueing Model with State Dependent Arrival and Departure Rates. – Electrotechnica & Electronica, Vol. **50**, 2015, No 11-12, pp. 42-48.
17. Mirtchev, S. Palm’s Machine-Repair Model with a Generalised Poisson Input Stream and Constant Service Time. – In: International Conference on Software in Telecommunications and Computer Networks, 2006, Split, Croatia, pp. 81-85.
18. Mirtchev, S., R. Goleva, G. Balabanov, V. Alexiev. Multiserver Loss Queueing System Polya/G/n/0 with Peaked Input Flow. – Int. J. Reasoning-Based Intelligent Systems, Vol. **5**, 2013, No 3, pp. 169-176.
19. Mittal, R., M. Jain. Maximum Entropy Analysis of MX/G/1 Retrial Queue with k-Phases of Heterogeneous Service and Impatient Calls Under Different Vacation Policies. – American Journal of Mathematical and Management Sciences, Vol. **34**, 2015, No 2, pp. 117-146.
20. Moltchanov, D., Y. Koucheryavy, J. Harju. Non-Preemptive iD-BMAPi/D/1/K Queuing System Modelling the Frame Transmission Process over Wireless Channels. – In: Proc. of International Teletraffic Congress (ITC’19), Beijing, China, Vol. **6a**, 2005, pp. 1335-1344.
21. Naeini, V. S. Performance Analysis of WiMAX-Based Wireless Mesh Networks Using an M/D/1 Queuing Model. – International Journal of Wireless and Mobile Computing, Vol. **7**, 2014, No 1, pp. 35-47.
22. Paxson, V., S. Floyd. Wide-Area Traffic: The Failure of Poisson Modeling. – IEEE/ACM Trans. Netw., Vol. **3**, 1995, No 3, pp. 226-244.
23. Roberts, J. Internet Traffic, QoS, and Pricing. – Proceedings of the IEEE, Vol. **92**, 2004, No 9, pp. 1389-1399.
24. Salah, K., K. Elbadawi, R. Boutaba. An Analytical Model for Estimating Cloud Resources of Elastic Services. – J. Network Sys. Managem., Vol. **24**, 2016, No 2, pp. 285-308.

25. Sulthana, S., R. Nakkeeran. Study of Downlink Scheduling Algorithms in LTE Networks. – Journal of Networks, Vol. **9**, 2014, No 12, pp. 3381-3391.
26. Swift, D., C. Dagli. A Study on the Network Traffic of Connexion by Boeing: Modeling with Artificial Neural Networks. – Engineering Applications of Artificial Intelligence, Vol. **21**, 2008, No 8, pp. 1113-1129.
27. Vilaplana, J., S. Solsona, I. Teixido. A Performance Model for Scalable Cloud Computing. – In: Proc. of 13th Australasian Symposium on Parallel and Distributed Computing (AusPDC), 2015, pp. 51-60.
28. Ziya, S. On the Relationships among Traffic Load, Capacity, and Throughput for the M/M/1/m, M/G/1/m-PS, and M/G/c/c Queues. – IEEE Transactions on Automatic Control, Vol. **53**, 2008, No 11, pp. 2696-2701.

Received 17.11.2017; *Accepted 10.12.2017*