



## TD(07) 007

### Meeting details

**(8<sup>th</sup> MCM COST 290 Action, Malaga, Spain, Feb 15-16 2007)**

### **Study of Queueing Systems with State Dependent Mean Service Time**

Seferin Mirtchev  
Technical University of Sofia  
8 Kliment Ohridski St.  
1000 Sofia, Bulgaria  
stm@tu-sofia.bg

#### **Abstract**

This paper deals with a full accessibility loss system and a single server delay system with a Poisson arrival process and state dependent exponentially distributed service time. We use the generalized service flow with nonlinear state dependence mean service time. The idea is based on the analytical continuation of the Binomial distribution and the classic M/M/n/0 and M/M/1/k system. We apply techniques based on birth and death process and state-dependent service rates.

We consider the system M/M(g)/n/0 and M/M(g)/1/k (by Kendal notation) with a generalized departure process – Mg. The output intensity depends nonlinear on the system state with a defined parameter: “*peaked factor - p*”. We obtain the state probabilities of the system using the general solution of the birth and death processes.

The influence of the peaked factor on the state probability distribution, the congestion probability and the mean system time are studied. It is shown that the state-dependent service rates changes significantly the characteristics of the queueing systems. The advantages of simplicity and uniformity in representing both peaked and smooth behaviour make this queue attractive in network analysis and synthesis.

#### **Keywords**

Loss and delay system, Queueing analyses, State dependent service rate, Birth and death process, Peaked and smooth traffic

#### **Working Group N 1**

## 1. INTRODUCTION

Simple models like the classical full accessibility and single-server queues can often be used to obtain comprehensive results, e.g., to predict the global traffic behaviour. When modeling network traffic, packet and connection arrivals are often assumed to be Poisson processes because such processes have attractive theoretical properties.

Many studies on traffic measurements from a variety of packet switching networks, like Ethernet, Internet, ATM and etc., have shown considerable difference between actual network traffic and assumptions in traditional theoretical traffic models. The basic characteristic of traffic in modern telecommunications networks is burstness. That is why there are many studies that generalize the queueing systems by state-dependent arrival and service rates.

In [5] the burstness of the total arrival process is characterized in packet network performance models by the dependence among successive interarrival times, dependence among successive service times and between service and interarrival times. These dependence effects are demonstrated analytically by considering a multiclass single-server queue with batch-Poisson arrival processes.

In [7] the author has modified the generalized Erlang blocking model to permit blocked requests to retry, with reduced resource requirements and arbitrary mean residency requirements. The presented approach modifies a one-dimensional recursion developed for the generalized Erlang model in an intuitively satisfying manner, and results in an approximation scheme that is both efficient and quite accurate. This study arose in the context of high-speed networks in which high bandwidth but non-real-time messages may, upon being blocked, request service with smaller bandwidth and larger residency time.

[9] is focused on the call blocking probabilities calculation in single link loss models where calls of each service-class come from finite sources and compete for the available link bandwidth under the complete sharing policy. There is reviewed the Engset multirate loss model and the single-retry loss model for finite sources in which blocked calls of a service-class may immediately retry once, in order to be connected in the system, with reduced bandwidth and increased service time requirements.

In [11] is considered two generalizations of the Engset model: Permitting the distributions of the holding time and interarrival time to differ from source to source; Permitting the distribution of the time until a source generates a new burst or packet differs according to whether or not the previous burst or packet was successful. The call and time congestions are approximated for the generalization. In this paper, they are improved the accuracy of the approximation, provided an efficient algorithm for its numerical computation and proved its convergence.

In [4] is developed an algorithm for computing exact steady-state blocking probabilities for each class in product-form loss networks to cover general state-dependent arrival and service rates. This generalization allows considering a wide variety of buffered and unbuffered resource-sharing models with non-Poisson traffic as may arise with overflows in the context of alternative routing.

In [10] is developed a numerically exact method for evaluating the time-dependent mean, variance, and higher order moments of the number of entities in a  $Ph_i/Ph_i/\infty$  queueing system, where  $Ph_i$  denotes a time-dependent generalization of a phase-type renewal process.

In [1] a continuous-time M/M/1 queueing system is analyzed in which the server can serve at two different speeds. The actual speed of the server depends on the state (empty or nonempty) of a fluid buffer. Fluid flows continuously into the fluid buffer at a constant rate, but is released from the buffer only during busy periods of the server. Hence, the contents of the fluid buffer are in turn determined by the queueing system. The queueing model serves as a mathematical model for a two-

level *traffic shaper* at the edge of an ATM network. The stationary joint distribution of the number of customers in the system and the contents of the fluid buffer is investigated. From this distribution, various performance measures such as the steady-state sojourn time distribution of a customer is obtained.

In [8] is introduced and evaluated a generalized Poisson arrival process by state-dependent arrival rates. The proposed single server delay system provides a unified framework to model peaked and smooth traffic and makes it attractive in network analysis.

In [3] is presented a queueing system where feedback information about the level of congestion is given right after arrival instants. If the amount of work right after an arrival is smaller or larger than a finite number then the server starts to work at two different service speeds. In addition, they have considered the generalization to the  $N$ -step service speed function.

In [2] a TCP-like linear-increase multiplicative-decrease flow control mechanism is presented. They consider congestion signals that arrive in batches according to a Poisson process. The service times in the queueing model depend on the workload in the system and the transmission rate cannot exceed a certain maximum value.

The Bernoulli-Poisson-Pascal (BPP) method is used to approximate the main congestion functions associated with peaked and smooth traffic in lost-call-cleared systems. The BPP model represents peaked and smooth traffic by two separate models, and cannot represent arbitrary smooth traffic. The BPP traffic models are insensitive to the holding time distribution [4]. The state probabilities for these loss systems only depend on the holding time through the mean value which is included in the offered traffic.

The queueing literature contains many studies about queues with workload-dependent service speeds. In these studies it is usually assumed that the speed of the server is continuously adapted over time based on the buffer content. In many practical situations service speed adaptations are only made at particular points in time, like arrival epochs. For example, feedback information about the buffer state may only be available at such epochs.

In this paper, we consider queueing systems with adaptable service speed based on the amount of work right after customer arrivals or departure. Between these events, the service speed is held fixed and may not be changed until the next customer arrival or depart. We generalize the classical loss and delay queueing systems to nonlinear state-dependent service rate. We use the generalized service flow with nonlinear state dependence mean service time. The idea is based on the analytic continuation of the Binomial distribution and the classic  $M/M/n/0$  and  $M/M/1/k$  system. We apply techniques based on birth and death process and state-dependent service rates.

These generalized models can be used to analyze multiplexing, message storage, traffic regulator and communication network performance.

## 2. GENERALIZED ERLANG DISTRIBUTION

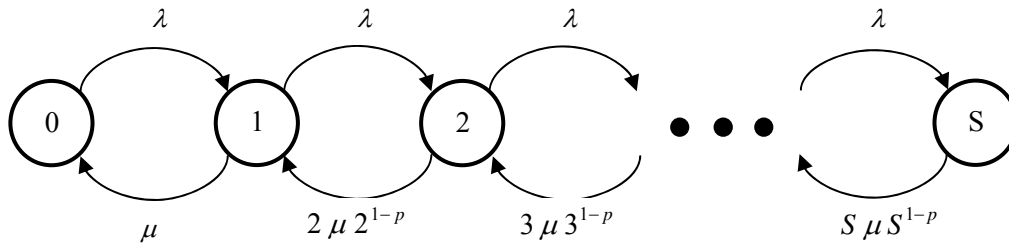
Let us consider a full availability loss system  $M/M(g)/S/0/S$  with a Poisson input stream  $M$ , state dependent exponentially distributed service time  $M(g)$ , number of servers  $S$ , waiting room 0 and number of sources  $S$ . This is a birth and death process and we can use the general solution for the stationary probability of having  $j$  customers in the system [4]

$$P_j = \frac{\prod_{i=0}^{j-1} \lambda_i / \mu_{i+1}}{1 + \sum_{v=1}^S \prod_{i=0}^{v-1} \lambda_i / \mu_{i+1}} \quad j = 1, 2, 3, \dots, S. \quad (1)$$

This generalized queueing system may be described by selecting the birth and death coefficient as follows

$$\lambda_j = \lambda, \quad \mu_j = j \mu j^{1-p} \quad j = 0, 1, 2, \dots, S \quad . \quad (2)$$

The service rate is state-dependent and depends on the peakedness factor  $p$ . This system is always ergodic. The finite state-transition diagram is shown in Fig.1.



**Fig.1.** A state-transition diagram - M/M(g)/S/0/S system

As the number of servers is equal to the number of sources the system has not any losses and delay, the whole offered traffic is carried and it is called the intended traffic load.

The stationary probabilities of having  $j$  customers in the system has generalized Erlang distribution when the service time is state dependent

$$P_j = \frac{a^j / (j!)^{2-p}}{\sum_{i=0}^S a^i / (i!)^{2-p}} \quad j = 0, 1, 2, \dots, S, \quad (3)$$

where  $a = \lambda/\mu$  is traffic intensity.

The intended traffic is the equilibrium number of busy servers

$$A_i = \sum_{j=1}^S j P_j . \quad (4)$$

The variance of the intended traffic is

$$V(A_i) = \sum_{j=0}^S (j - A_i)^2 P_j . \quad (5)$$

The peakedness of the intended traffic is the variance to mean ratio

$$z_i = V(A_i) / A_i . \quad (6)$$

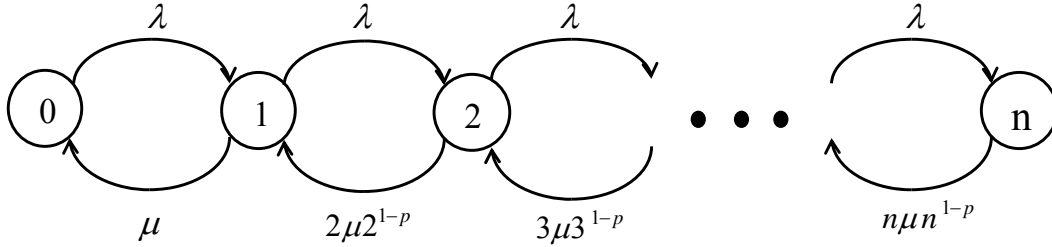
### 3. MODEL DESCRIPTION

#### GENERALIZED FULL ACCESSIBILITY LOSS SYSTEM

Let us consider a multi server system M/M(g)/n/0/S with a Poisson input stream M, state dependent exponentially distributed service time M(g), number of servers n, waiting room 0 and number of sources S ( $S > n$ ). This generalized loss system may be described by selecting the birth-death coefficient as follows

$$\lambda_j = \lambda \quad \mu_j = j \mu j^{1-p} \quad j = 0, 1, 2, \dots, n. \quad (7)$$

The finite state-transition diagram is shown in Fig.2.



**Fig.2.** A state-transition diagram - M/M(g)/n/0/S system

Applying these coefficients to the general solution of the birth and death process and using traffic intensity  $a = \lambda/\mu$  we obtain the steady state probabilities

$$P'_j = \frac{a^j / (j!)^{2-p}}{\sum_{i=0}^n a^i / (i!)^{2-p}} \quad j = 0, 1, 2, \dots, n. \quad (8)$$

The offered traffic is calculated by means of the arrival rate and the mean holding time

$$A = \lambda \bar{\tau} = a \sum_{i=1}^{n+1} \frac{1}{i^{1-p}} P'_{i-1}, \quad \text{erl.} \quad (9)$$

The carried traffic is equivalent to the average number of busy servers

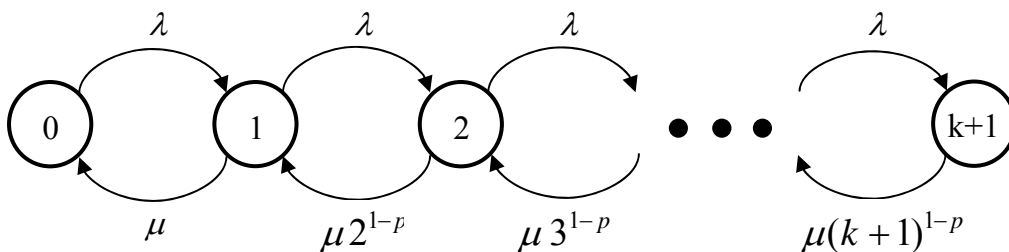
$$A_o = \sum_{i=1}^n i P'_i, \quad \text{erl.} \quad (10)$$

#### GENERALIZED SINGLE SERVER DELAY SYSTEM

Let us consider a single server queue M/M(g)/1/k with a Poisson input stream M, state dependent exponentially distributed service time M(g) and limited waiting rooms k. This generalized queueing system has the birth and death coefficient as follows

$$\lambda_j = \lambda \quad \mu_j = \mu j^{1-p} \quad j = 0, 1, 2, \dots, k+1. \quad (11)$$

The finite state-transition diagram is shown in Fig.3.



**Fig.3.** A state-transition diagram - M/M(g)/1/k/S queue

Applying these coefficients to the general solution of the birth and death process and using traffic intensity  $a = \lambda/\mu$  we obtain the steady state probabilities

$$P_j'' = \frac{a^j / (j!)^{1-p}}{\sum_{i=0}^{k+1} a^i / (i!)^{1-p}} \quad j = 0, 1, 2, \dots, k+1. \quad (12)$$

The offered traffic is calculated by means of the average arrival rate and the mean holding time

$$A = \lambda \bar{\tau} = a \sum_{j=1}^{k+2} \frac{1}{j^{1-p}} P_{j-1}'' . \quad (13)$$

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

$$A_o = 1 - P_0'' = A(1 - P_{k+1}'') . \quad (14)$$

#### 4. PERFORMANCE MEASURES

##### *GENERALIZED FULL ACCESSIBILITY LOSS SYSTEM*

TIME CONGESTION PROBABILITY  $B_t$  describes the fraction of time that all  $n$  servers are busy

$$B_t = P_n' . \quad (15)$$

CALL CONGESTION PROBABILITY  $B_c$  is ratio of the lost traffic (offered minus carried traffic) to the offered traffic

$$B_c = \frac{A - A_o}{A} . \quad (16)$$

TRAFFIC CONGESTION PROBABILITY  $B_a$  is the ratio of the difference between the intended and carried traffic to the intended traffic

$$B_a = \frac{A_i - A_o}{A_i} . \quad (17)$$

##### *GENERALIZED SINGLE SERVER DELAY SYSTEM*

BLOCKING PROBABILITY. The time congestion probability  $B$  describes the fraction of time that all waiting rooms are busy

$$B = P_{k+1}'' . \quad (18)$$

MEAN NUMBER OF CALLS. The mean number of calls present in the system in steady state by definition is

$$L = \sum_{j=1}^{k+1} j P_j'' . \quad (19)$$

MEAN SYSTEM TIME. From the Little formula, we have the mean system time

$$T = L/\lambda . \quad (20)$$

## 5. CALCULATION OF THE STATE PROBABILITY

The traffic intensity  $a$  is not equal to the intended traffic in a case of a generalized Erlang process when the service time is state dependent because we calculate the power of the Erlang unsymmetrical distribution. That is why we have to calculate the intended traffic  $A_i$  and the peakedness  $z_i$  when defining the traffic intensity  $a$  and peakedness factor  $p$ .

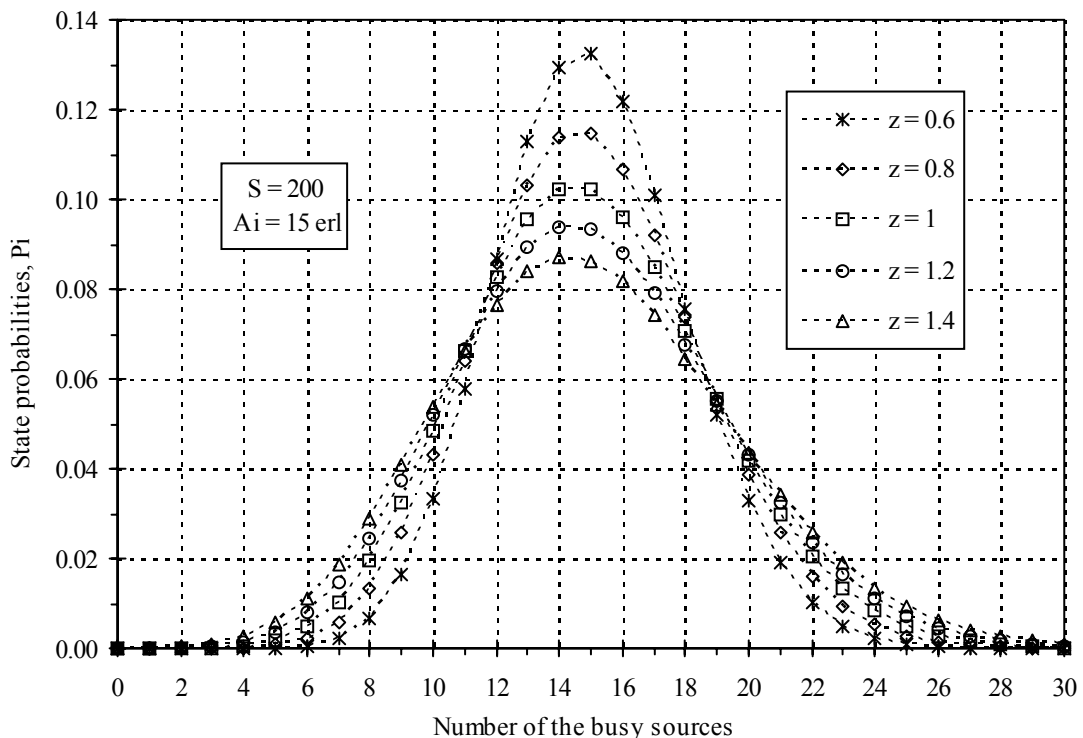
From the practical point of view we first define the intended traffic  $A_i$  and the peakedness  $z_i$  and after that calculate the traffic intensity  $a$  and peakedness factor  $p$ .

A fundamental question about the system defined by equations (3), (5) and (6) is whether there exist solutions  $a, p$  for an arbitrary  $A_i, z_i$ . Although no formal proof seems to exist, this seems to be the case and the solution appears to be unique. We can find solutions of the above system with the iterating method of consecutive replacements.

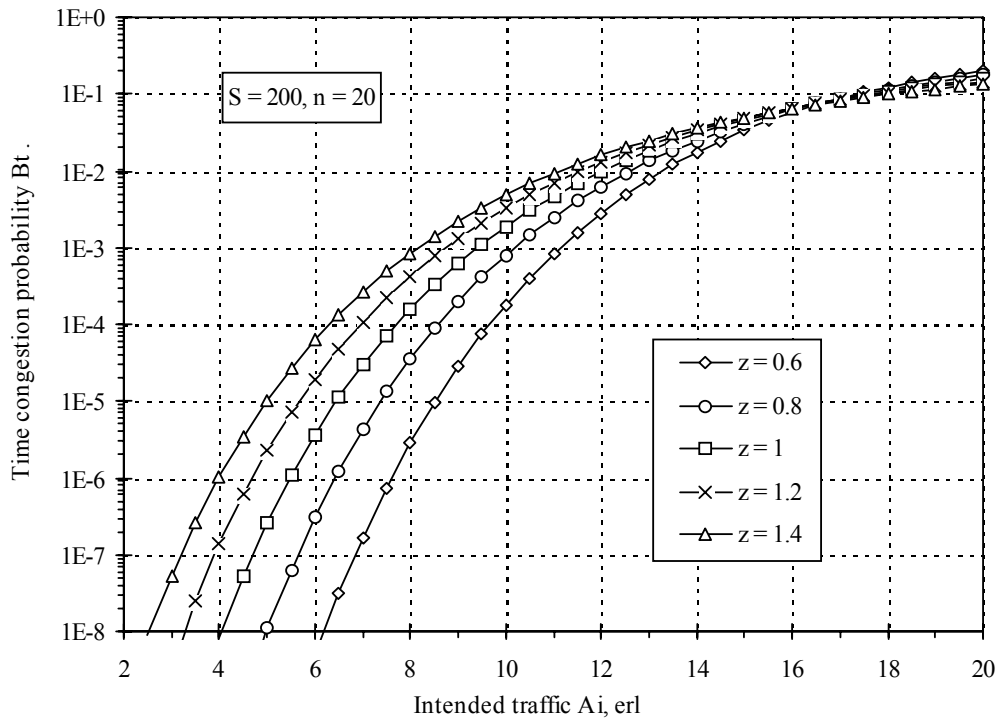
## 6. NUMERICAL RESULTS

In this section we give numerical results obtained by a Pascal program on a personal computer. The described methods were tested on a computer over a wide range of arguments.

Figure 4 shows the generalized Erlang distribution where the intended traffic is  $A_i = 15 \text{ erl}$ , the number of the sources is  $S = 200$  and the peakedness  $z_i$  is change from  $0.6$  to  $1.4$ . It will be seen that when the peakedness  $z_i$  increases the probability distribution becomes broad about the mean.



**Fig.4.** Generalized Erlang distribution when the intended traffic is  $A_i = 15 \text{ erl}$ , the number of the sources is  $S = 200$  and different peakedness  $z_i$



**Fig.5.** Time congestion probability in a full availability loss system with 20 servers, 200 sources and different peakedness  $z_i$

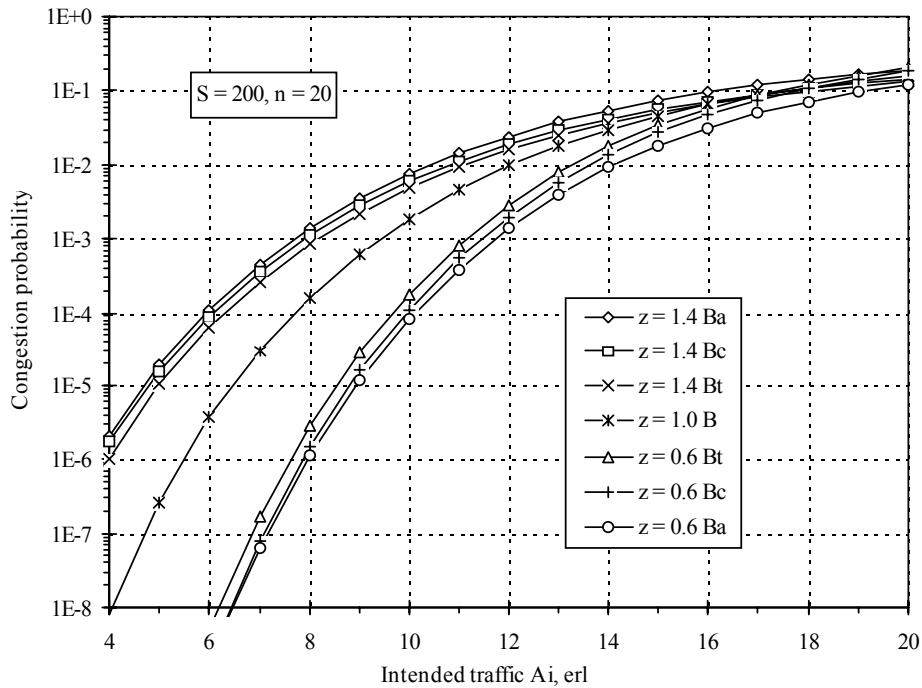
Figure 5 presents the time congestion probability in a full availability loss system with 20 servers, 200 sources and different peakedness  $z_i$  as function of the intended traffic  $A_i$ . When the intended traffic per server is big (0.7 - 1 erl) the influence of the peakedness to the time congestion probability is negligible.

Figure 6 compares together the time, call and traffic congestion probabilities in a full availability loss system with 20 servers, 200 sources and different peakedness of the intended traffic  $z_i$  as function of the intended traffic  $A_i$ .

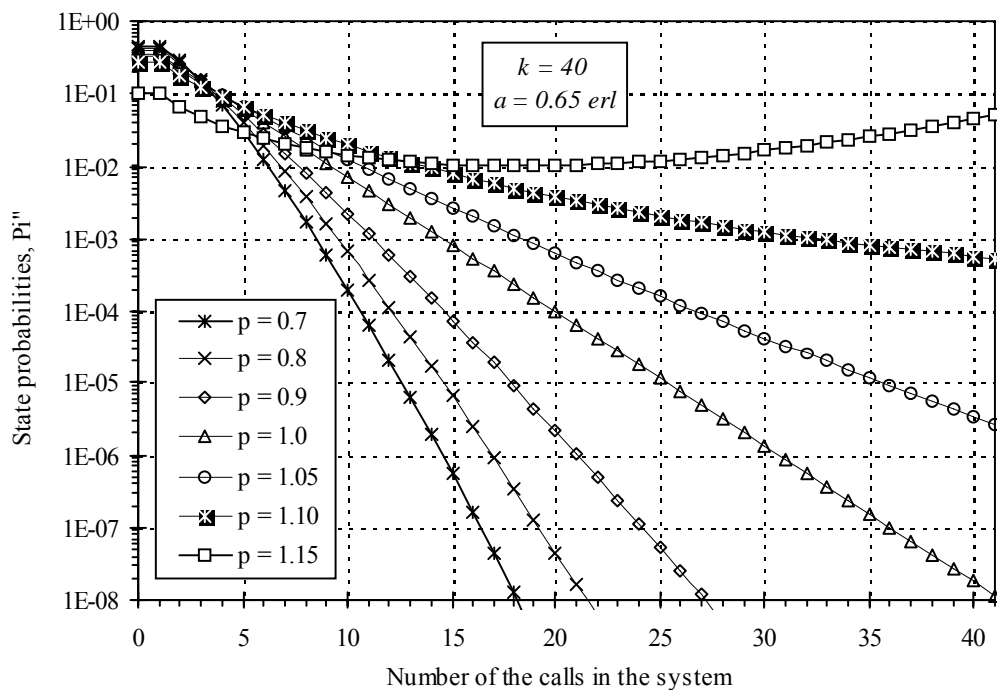
Figure 7 shows the stationary probability distribution in a single server queue M/M(g)/1/k with a state dependent mean service time, 40 waiting positions, 0.65 erl traffic intensity and different peakedness factor  $p$ . We can see that when the peakedness factor is bigger than one the probabilities can increase when the number of the calls in the system increases.

Figure 8 illustrates the dependence of the mean service time from the number of calls in the system and different peakedness factor  $p$  from 0.7 to 1.15. We can see that when the peakedness factor is smaller or bigger than one the mean service time decrease or increase respectively when the number of the calls in the system increases.

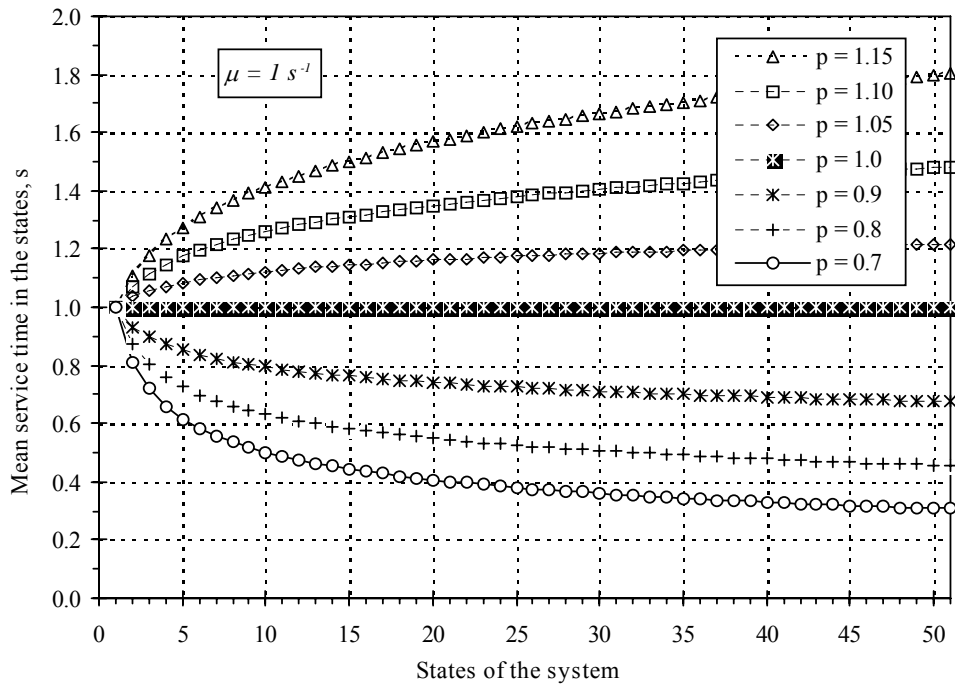
Figure 9 shows the time congestion probability in a single delay system with 0.65 erl traffic intensity and different peakedness factor as function of the buffer size. When the peakedness factor is bigger than 1 the influence of the buffer size on the time congestion probability is negligible. In some cases the time congestion probability can increase when the buffer size increase.



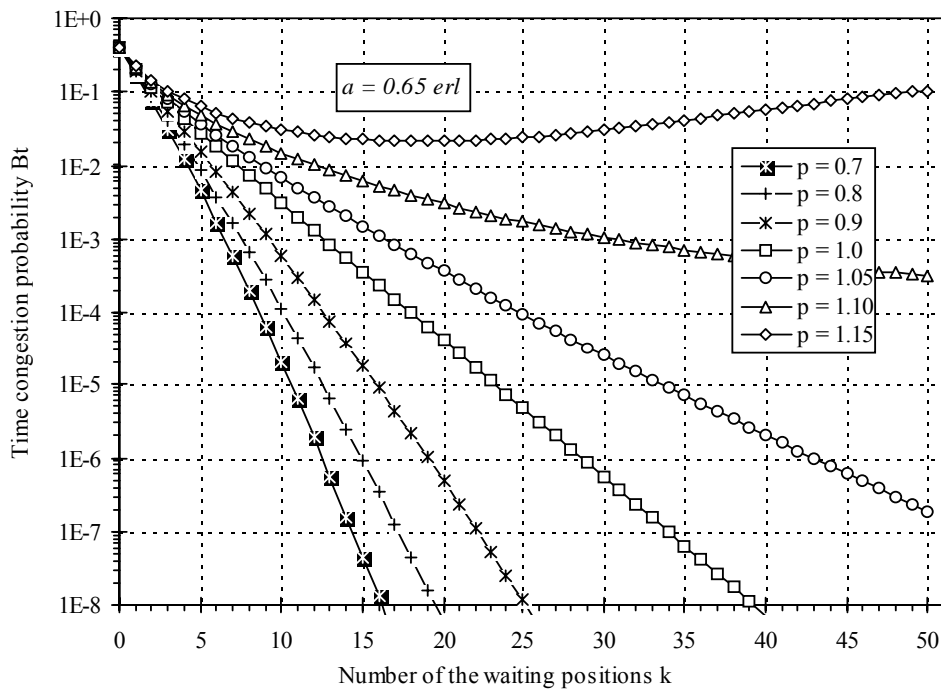
**Fig.6.** Time, call and traffic congestion probabilities in a full availability loss system with 20 servers, 200 sources and different peakedness  $z_i$



**Fig.7.** Stationary probability distribution in a single server queue with state dependent mean service time and different peakedness factor

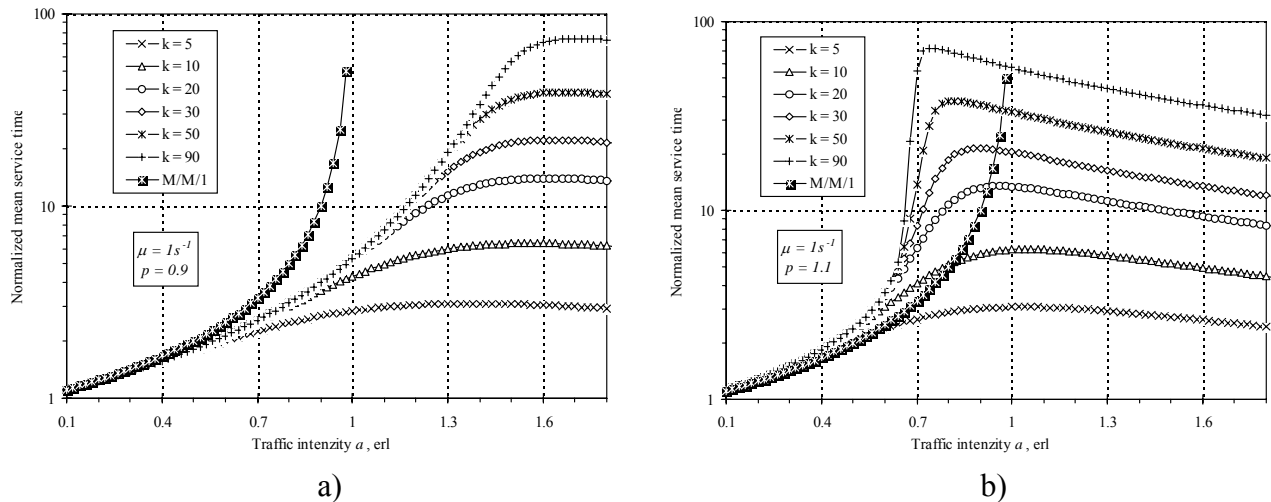


**Fig.8.** Dependence of the mean service time from the number of the calls in the system and different peakedness factor



**Fig.9.** Time congestion probability in a single delay system with state dependent mean service time and different peakedness factor

Figure 10 (a, b) presents the normalized mean system time ( $T' = T/\tau$ ) as function of the traffic intensity when the peakedness factor is 0.9 and 1.1 respectively and different waiting room.



**Fig.10.** Normalized mean system time when the mean service time decrease– a) and increase – b) when the number of the calls in the system increases

It is shown that the influence of the peakedness over the performance measures is significantly.

## 8. CONCLUSION

In this paper a generalized Erlang distribution as a result of state dependent mean service time is introduced and evaluated. A basic model for a loss system  $M/M(g)/n/0/S$  and delay queue  $M/M(g)/1/k$  is examined in detail.

The proposed method provides a unified framework to model peaked, regular and smooth behaviour of the teletraffic systems. Numerical results and subsequent experience have shown that this method is accurate and useful in analysis of queueing systems.

The classic teletraffic system – the Full accessibility loss system is independent of the service time distribution. In this paper is shown that the influence of state dependent service rate over the main parameters of the full availability loss system is significant. The main parameters of this system – state probabilities and call, time and traffic congestion probabilities are defined and presented graphically.

The single server delay system with state dependent service rate can be used as a means for controlling and smoothing the data flow into the telecommunications networks. This system can be used to explain the behaviour of real traffic regulator as “leaky bucket” and “congestion window”.

The importance of the teletraffic systems in a case of state dependent mean service time comes from its ability to describe behaviour that is to be found in up-to-day networks. It is the case in a general teletraffic system, which is an important feature in designing telecommunications networks.

In conclusion, we believe that the presented generalized Erlang distribution and queueing system will be useful in practice. As part of future work, we plan to analyze a regulator in the network.

## REFERENCES

1. Adan I. J. B. F., E. A. van Doorn, J. A. C. Resing, and W. R.W. Scheinhardt. Analysis of a single-server queue interacting with a fluid reservoir. *Queueing Systems*, 29:313–336, 1998.
2. Altman E., K. Avratchenkov, C. Barakat, and R. Nunez-Queija, State dependent M/G/1 type queueing analysis for congestion control in data networks, IEEE INFOCOM, Anchorage, Alaska, April, 2001.
3. Bekker R. and 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.
4. Choudhury G.L., K.K. Leung, W. Whitt, "An inversion algorithm for loss networks with state-dependent rates," *infocom*, p. 513, Fourteenth Annual Joint Conference of the IEEE Computer and Communication Societies (Vol. 2)-Volume, 1995.
5. Fendick K.W., V. Saksena and W. Whitt. Dependence in packet queues, IEEE Transactions on Communications, Vol. 37, Issue 11, Nov. 1989 Page(s):1173 – 1183.
6. Hayes J.F. and T.V.J.Ganesh Babu. Modeling and Analysis of Telecommunications Networks. John Wiley&Sons, 2004.
7. Kaufman J.S. Blocking in a completely shared resource environment with state dependent resource and residency requirements, INFOCOM '92. Eleventh Annual Joint Conference of the IEEE Computer and Communications Societies, May 1992, Page(s):2224 - 2232 vol.3
8. Mirtchev S. and I. Stanev. Evaluation of a Single Server Delay System with a Generalized Poisson Input Stream. ITC19, Beijing, China, Vol.6a, 2005, pp.553-542.
9. Moscholios I. D., M. D. Logothetis and P. I. Nikolaropoulos, "Engset Multi-Rate State-Dependent Loss Models", *Performance Evaluation*, Vol. 59, issue 2-3, pp. 247-277, February, 2005.
10. Nelson B. L., M. R. Taaffe. The *Pht/Pht/∞* Queueing System: Part I—The Single Node *INFORMS Journal on Computing*, Vol. 16, No. 3, 2004, pp. 266–274.
11. Wong M., A. Zalesky and M. Zukerman. On Generalizations of the Engset Model, submitted to IEEE Communications Letters, 2006.