Generalized Phase Type Distribution and Its Applications to Analysis of Telecommunication Networks with Heterogeneous Customers
Обобщенное распределение фазового типа и его применение для анализа телекоммуникационных сетей с разнотипными запросами

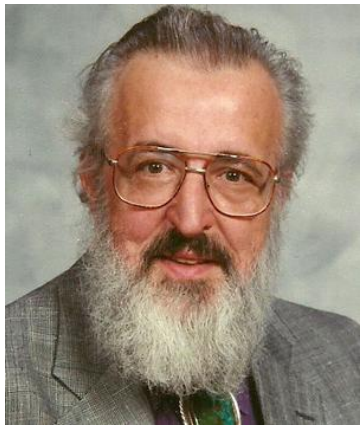
Alexander Dudin

RUDN University, 6, Miklukho-Maklaya st., 117198, Moscow, Russia

Master-class of the Center of Applied Probabilistic Analysis of the Peoples' Friendship University of Russia

Phase-type (PH) Service Time Distribution

PH distribution is a probability distribution constructed by a mixture of exponential distributions.



Marcel Neuts

M. F. Neuts. Matrix-Geometric Solutions in Stochastic Models: an Algorithmic Approach, Chapter 2: Probability Distributions of Phase Type; Dover Publications Inc., 1981.

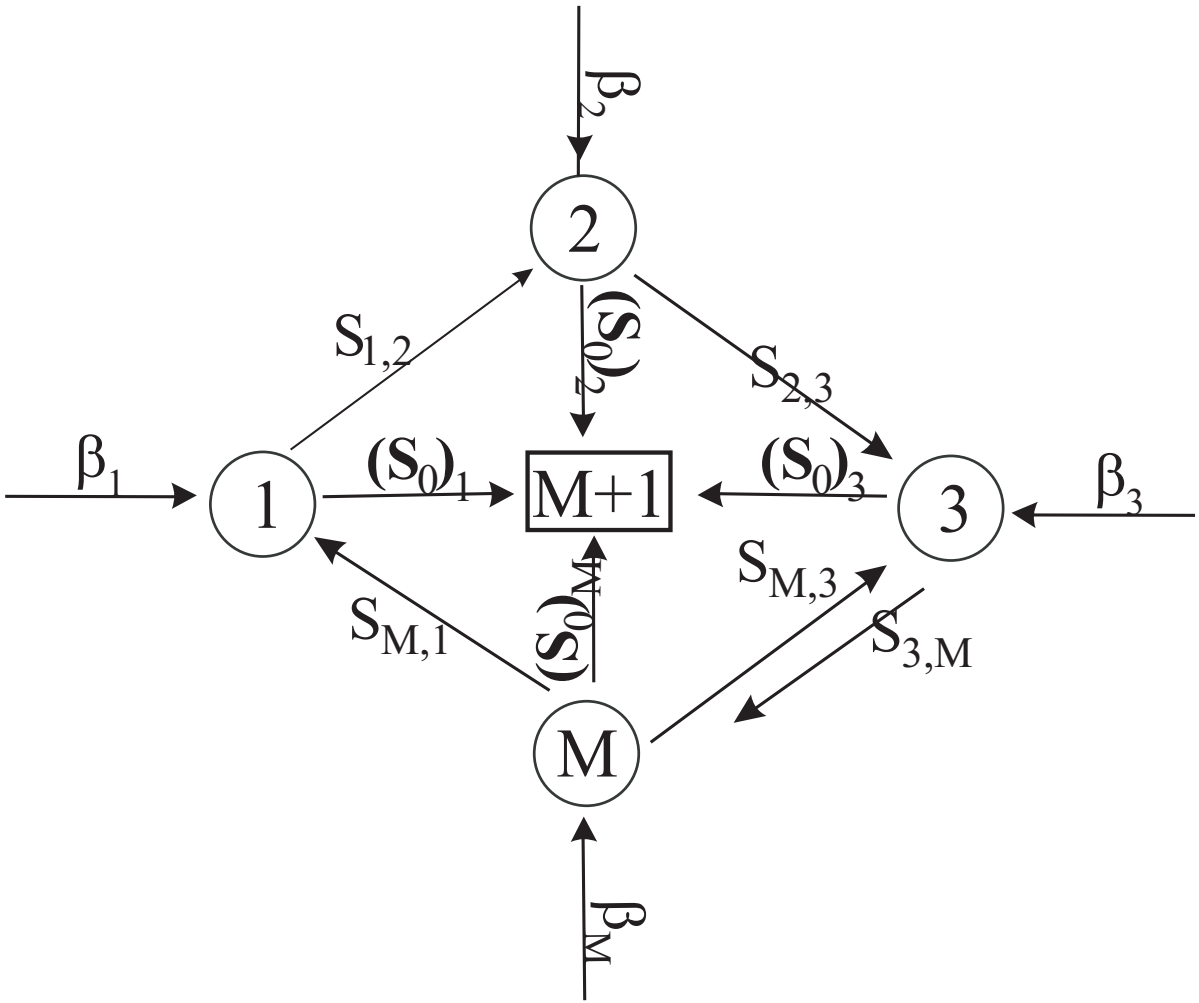
Phase-type distribution

Let us consider Markov process $\eta_t, t \geq 0$, with a finite state space $\{1, \ldots, M, M+1\}$. The states $1, \ldots, M$ are transient and state $M+1$ is absorbing state.

The service time can be interpreted as a time until the Markov process $\eta_t, t \geq 0$, reaches the absorbing state $M+1$ condition on the fact that the initial state of this process is selected among the transient states according to the probabilistic row vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)$.

Transition rates of the process η_t within the set $\{1, \ldots, M\}$ are defined by the sub-generator S and transition rates into the absorbing state are given by the entries of the column vector $\boldsymbol{S}_0 = -S\mathbf{e}$.

Phase-type distribution - graphical interpretation



Phase-type distribution – partial cases

Exponential distribution: $\boldsymbol{\beta} = (1)$, $S = (-\mu)$;

Hyper-exponential distribution:
$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)$, $S = \mathrm{diag}\{-\mu_1, \ldots, -\mu_M\}$;

Erlang distribution:

$$\boldsymbol{\beta} = (1, 0, \ldots, 0),\ S = \begin{pmatrix} -\mu & \mu & 0 & 0 & \ldots 0 & \\ 0 & -\mu & \mu & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & -\mu \end{pmatrix};$$

Coxian distribution:

$$\boldsymbol{\beta} = (1, 0, \ldots, 0),\ S = \begin{pmatrix} -\mu_1 & p_1\mu_1 & 0 & 0 & \ldots 0 & \\ 0 & -\mu_2 & p_2\mu_2 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & -\mu_M \end{pmatrix}.$$

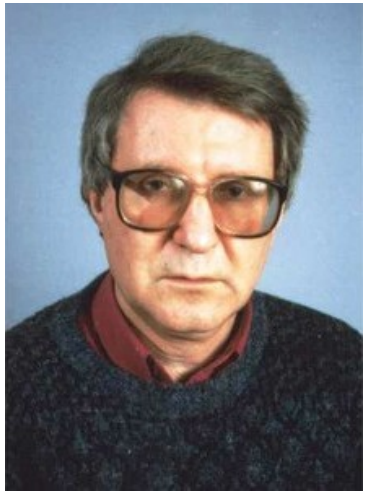
Phase-type distribution – main contributors

RUDN

G.P. Basharin

Phase-type distribution – main contributors

P.P. Bocharov, A.V. Pechinkin, Queueing Theory. Moscow: RUDN, 1995.



P.P. Bocharov

Phase-type distribution – main advantages

The class of *PH* (phase-type) distributions is dense (in the sense of weak convergence) in the set of all probability distributions of non-negative variables. Therefore, the *PH* distribution can be used to approximate of an arbitrary distribution, see

S. Asmussen, Applied Probability and Queues, Springer, 2003.

For methods to estimate the parameters of the *PH* distribution of service time based on real data, see, e.g.,

Asmussen, Nerman, Olsso. Fitting Phase-Type Distributions via the EM Algorithm. Scandinavian Journal of Statistics. (1996) 23 (4): 419-441.

H. Okamura, T. Dohi, Mapfit: An R-Based Tool for PH/MAP Parameter Estimation. Lecture Notes in Computer Science, (2015) 9259: 105-112.

Phase-type distribution – main advantages in case of single-server queues

Method of embedded Markov chains

$$Y_l = \int_0^\infty P(l,t)dB(t)$$

– complicated two-dimension recursion for general $B(t)$
– simple recursion for PH.

Phase-type distribution – main advantages in case of single-server queues

$$P(n,t) = \sum_{j=0}^{\infty} e^{-\delta t} \frac{(\delta t)^j}{j!} K_n^{(j)}, \ n \geq 1,$$

where

$$K_0^{(j)} = (I + \delta^{-1} D_0)^j,$$

the matrices $K_n^{(j)}, \ n \geq 1, j \geq 0$, are found from the double recursion

$$K_0^{(0)} = I, \ K_n^{(0)} = O, \ n \geq 1,$$

$$K_0^{(j+1)} = K_0^{(j)}(I + \delta^{-1} D_0),$$

$$K_n^{(j+1)} = \delta^{-1} \sum_{i=0}^{n-1} K_i^{(j)} D_{n-i} + K_n^{(j)}(I + \delta^{-1} D_0), \ n \geq 1, \ j \geq 0.$$

Phase-type distribution – main advantages in case of single-server queues

$$Y_l = Z_l(I_{\bar{W}} \otimes \mathbf{S}_0),\ l \geq 0,$$

where the matrices $Z_l,\ l \geq 0$, are found from recursion

$$Z_0 = -(I_{\bar{W}} \otimes \boldsymbol{\beta})(D_0 \oplus S)^{-1},$$

$$Z_l = -\sum_{i=0}^{l-1} Z_i(D_{l-i} \otimes I_M)(D_0 \oplus S)^{-1},\ l \geq 1.$$

Phase-type distribution – main advantages in case of multi-server queues

N-server queues with arbitrary distribution of service time. Method of embedded Markov chains is not applicable.

$$\{\chi_t^{(1)}, \chi_t^{(2)}, \ldots, \chi_t^{(N)}\},\ t \geq 0,$$

where $\chi_t^{(n)} \in 1, \ldots, M$ - the phase of service in the n-th busy server, $n = \overline{1, N}$.

There are different ways for busy servers enumeration. E.g., number 1 has the server providing the longest service till this moment. Servers may be renumerated or not in case of service completion and immediate start of the next service.

The dimension of the state space $\bar{K} = M^N$.
Disadvantage − −dimension may be high : $N = 10$, $M = 4$ the dimension $\bar{K} = 4^{10} = 1048576$.
Example:
Breuer L., Dudin A.N., Klimenok V.I. A retrial $BMAP/PN/N$ system // Queueing Systems. 2002. V. 40. P. 433-457.

Breuer L., Klimenok V., Birukov A., Dudin A., Krieger U. Modeling the access to a wireless network at hot spots // European Transactions on Telecommunications. 2005. V. 16. P.309-316
$N = 7$.

Method Ramaswami-Lucantoni:

$$\{\eta_t^{(1)}, \eta_t^{(2)}, \ldots, \eta_t^{(M)}\},\ t \geq 0,$$

where $\eta_t^{(m)} \in 1, \ldots, N$ - the number of servers on phase m, $m = \overline{1, M}$.
The dimension of the state space $\tilde{K} = C_{N+M-1}^{M-1}$.
$N = 10$, $M = 4$ the dimension $\bar{K} = C_{13}^{3} = 286$.

$$286 - - - 1048576$$

V. Ramaswami, Independent Markov processes in parallel. Communications in Statistics – Stochastic Models, 1985, 1: 419-432.

V. Ramaswami, D.M. Lucantoni, Algorithms for the multi-server queue with phase-type service. Communications in Statistics – Stochastic Models, 1985, 1: 393-417.

Matrices $P_n(\vec{\beta})$, $A_n(N,S)$, $L_{N-n}(N,\tilde{S})$ describe intensities of transitions of the process

$$\{\eta_t^{(1)}, \eta_t^{(2)}, \ldots, \eta_t^{(M)}\},\ t \geq 0,$$

at the moment when new service starts, the transitions without service completion take place, service is completed, correspondingly.

Example:
Kim C.S., Mushko V.V., Dudin A. Computation of the steady state distribution for multi-server retrial queues with phase type service process // Annals of Operations Research. 2012. V. 201. No. 1. P. 307-323.

$N > 20$.
Kim C.S., Dudin S., Taramin O., Baek J. Queueing system $MMAP/PH/N/N+R$ with impatient heterogeneous customers as a model of call center // Applied Mathematical Modelling. 2013. V. 37. No 3. P. 958-976.

Multi-server queues with heterogeneous customers:

Let us consider a system with K-types of customers. The service time distribution of type-k customer has PH distribution with irreducible representation $(\beta^{(k)}, S^{(k)})$, $k = \overline{1, K}$.

Markov chain:

$$\tau_t = \{n_t, n_t^{(1)}, n_t^{(2)}, \dots, n_t^{(K)}, \eta_t^{(1,1)}, \dots, \eta_t^{(1,M_1)},$$

$$\eta_t^{(2,1)}, \dots, \eta_t^{(2,M_2)}, \dots, \eta_t^{(K,1)}, \dots, \eta_t^{(K,M_K)}\}, t \geq 0,$$

where, during the moment t,

n_t is the number of busy servers,

$n_t^{(k)}$ is the number of type-k customers on service,

$n_t^{(k)} = \overline{0, K}$, $k = \overline{1, K}$, $\sum_{k=1}^{K} n_t^{(k)} = n_t$

$\eta_t^{(k,m)}$ is the number of servers at phase m of PH service process of type-k customers, $m = \overline{1, M_k}$,

$\eta_t^{(k,m)} = \overline{0, n_t^{(k)}}$, $\sum_{m=1}^{M_k} \eta_t^{(k,m)} = n_t^{(k)}$, $k = \overline{1, K}$,

In order to simplify the investigation of the considered model, instead of separate considering the service times of K types of customers, that have phase-type distribution with irreducible representation $(\beta^{(k)}, S^{(k)})$ and finite state space $\{1, \ldots, M_k, M_k + 1\}$, we propose to use generalized *PH* distribution, with an irreducible representation $(\beta_1, \ldots, \beta_K, S)$ where

$$\beta_k = \left(\mathbf{0}_{\sum_{m=2}^{k-1} M_m}, \beta^{(k)}, \mathbf{0}_{\sum_{m=k+1}^{K} M_m} \right), \; k = \overline{1, K},$$

and the matrix S of the form

$$S = \begin{pmatrix} S^{(1)} & O & \ldots & O \\ O & S^{(2)} & \ldots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \ldots & S^{(K)} \end{pmatrix}.$$

The time having such a distribution can be interpreted as the time until the underlying Markov process $\eta_t, t \geq 0$, with the finite state space $\{1, \ldots, \mathcal{M}, \mathcal{M}+1\}$, where $\mathcal{M} = \sum_{k=1}^{K} M_k$, reaches the single absorbing state $\mathcal{M}+1$. The initial state of this process is selected among the states $\{1, \ldots, \mathcal{M}\}$ depending on the type of a customer who is chosen for service. If an arbitrary type-k customer is chosen for service, the initial state of this process is selected according to the probabilistic row vector $\boldsymbol{\beta}_k$.

Dudin, A., Kim, C., Dudina, O., Dudin, S. Multi-server queueing system with a generalized phase-type service time distribution as a model of call center with a call-back option. Annals of Operations Research, (2016) 239(2): 401-428.

Instead of the components

$$\tau_t = \{n_t, n_t^{(1)}, n_t^{(2)}, \ldots, n_t^{(K)}, \eta_t^{(1,1)}, \ldots, \eta_t^{(1,M_1)}, \eta_t^{(2,1)}, \ldots, \eta_t^{(2,M_2)}, \ldots, \eta_t^{(K,1)}, \ldots, \eta_t^{(K,M_K)}\}, t \geq 0,$$

of the Markov chain we consider the the components

$$\zeta_t = \{n_t, \eta_t^{(1)}, \ldots, \eta_t^{(\mathcal{M})}\}, \quad t \geq 0,$$

where

n_t, $n_t = \overline{0, K}$, is the number of customers on service,

$\eta_t^{(m)}$ is the number of servers at phase m of generalized service process of customers, $\eta_t^{(m)} = \overline{0, n_t}$, $m = \overline{1, \mathcal{M}}$, $\sum_{m=1}^{\mathcal{M}} \eta_t^{(m)} = n_t$, during the moment t, $t \geq 0$.

Is it reasonable to use the *GPH* distribution? Is the dimension of the state space now equal to $(M_1 + \cdots + M_K)^N$?

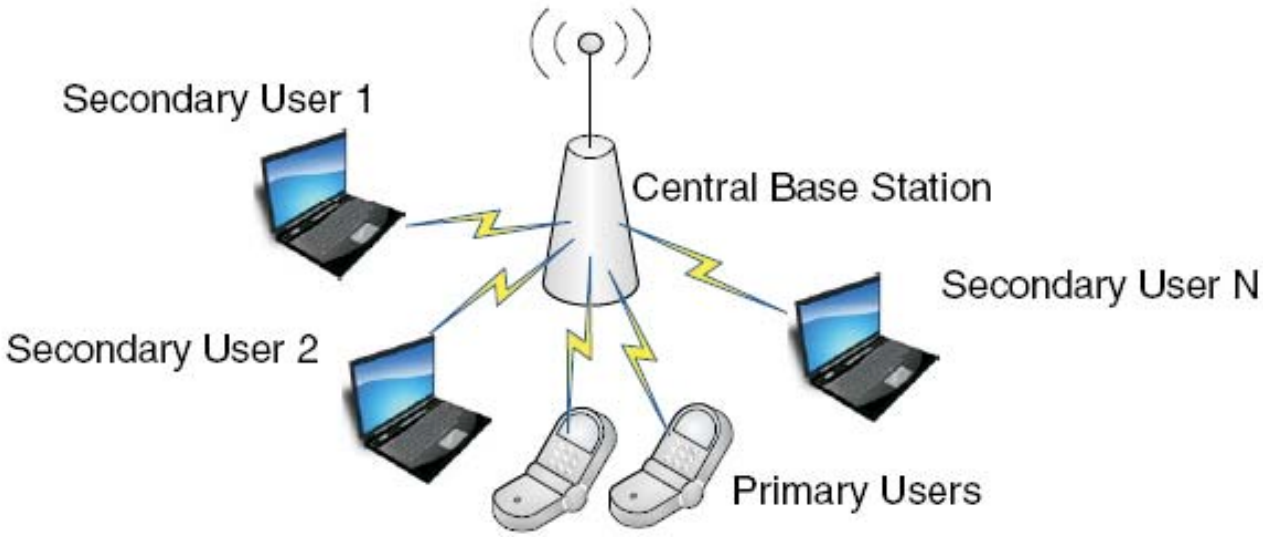
NO, if Ramaswami-Lucantoni approach is used!

Lemma. Use of the generalized phase-type with state space $(1, \ldots, M_1 + \cdots + M_K + 1)$ instead of consideration of K phase-type service processes with dimensions $M_k + 1$, $k = \overline{1, K}$, of the state space does not change (increase or decrease) the dimension of the stochastic process that describes the behavior of system (using Ramaswami-Lucantoni approach).

Cognitive Radio

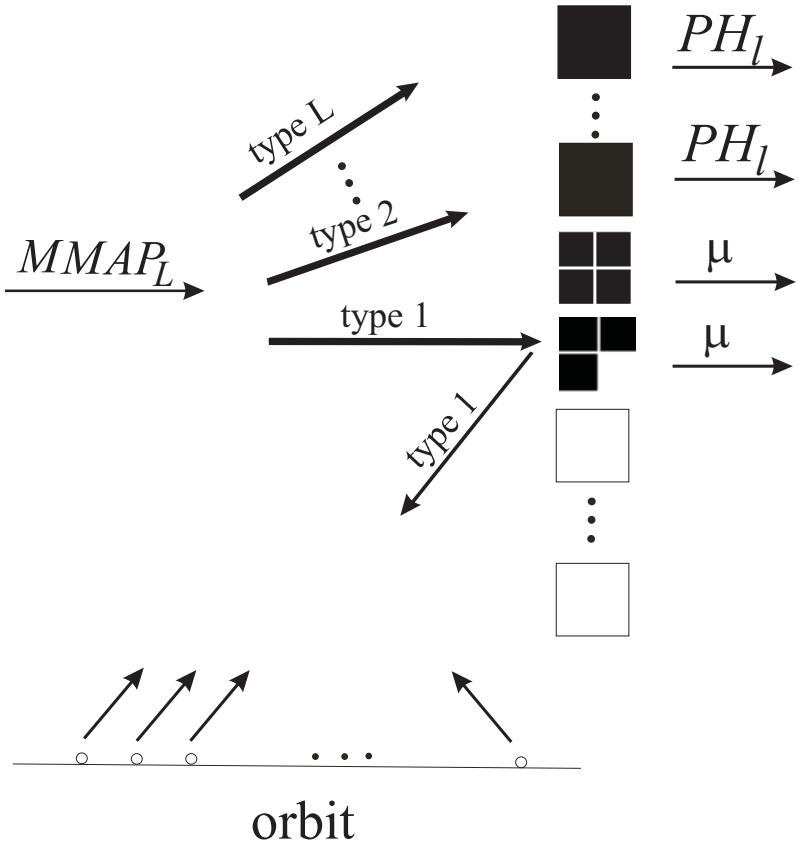
Recently, the technology of cognitive radio has attracted considerable attention of many researchers as a promising technology for optimization of the utilization of scare radio frequency spectrum. Dynamic spectrum access allows effectively use radio frequency and prevent its under-utilization in many real world networks. It enables unlicensed users to temporarily "borrow"unused spectrum while ensuring that the rights of the incumbent license holders are respected. Problems of optimization of joint access of the primary and secondary users can be effectively solved by means of queueing theory. So, although the term Cognitive Radio was first introduced only recently (in 1999), literature devoted to application of queueing theory to cognitive radio is already extensive. A comprehensive survey
I.F. Akyildiz, W.Y. Lee, M.C. Vuran, S. Mohanty, Next generation dynamic spectrum access cognitive radio wireless networks: A survey, Computer Networks, 50(13), pp. 2127-2159, 2006.
devoted to cognitive radio was published in 2006.

Cognitive Radio

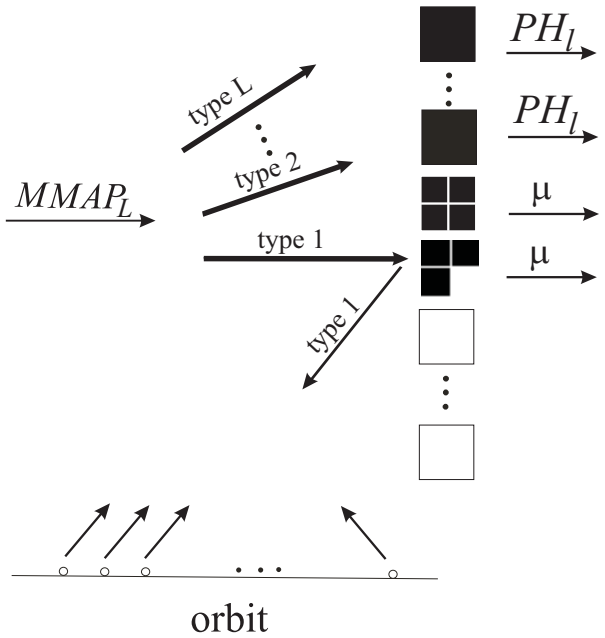


Cognitive Radio System

Priority retrial queue with many types of customers and servers reservation as a model of cognitive radio system



Structure of the queue



The system has K identical servers without buffer. Each server consists of k identical sub-servers. Customers of L types arrive to the system, $2 \leq L < \infty$.

Arrival of two types of customers is defined by the *MMAP* – Marked Markovian Arrival Process. Arrivals in the *MMAP* are directed by an irreducible continuous-time Markov chain ν_t, $t \geq 0$, with the finite state space $\{0, ..., W\}$.

The sojourn time of the chain ν_t in the state ν is exponentially distributed with the parameter λ_ν. After this time expires, with probability $p^{(0)}_{\nu,\nu'}$ the chain ν_t jumps to the state ν' without generation of customers, $\nu, \nu' = \overline{0, W},\ \nu \neq \nu'$, or with probability $p^{(l)}_{\nu,\nu'}$ it jumps to the state ν' with generation of type-l customer, $l = \overline{1, L}$, $\nu, \nu' = \overline{0, W}$. Here notation $\nu = \overline{0, W}$ means that ν takes the values in the set $\{0, 1, \ldots, W\}$.

The *MMAP* is completely characterized by the square matrices D_0, D_l, $l = \overline{1, L}$, defined as follows: $(D_l)_{\nu,\nu'} = \lambda_\nu p^{(l)}_{\nu,\nu'}$, $\nu, \nu' = \overline{0, W}$, $l = \overline{1, L}$, и $(D_0)_{\nu,\nu} = -\lambda_\nu, \nu = \overline{0, W},\ (D_0)_{\nu,\nu'} = \lambda_\nu p^{(0)}_{\nu,\nu'},\ \nu, \nu' = \overline{0, W}, \nu \neq \nu'$.

The matrix $D(1) = D_0 + \sum_{l=1}^{L} D_l$ is the generator of the Markov chain $\nu_t, t \geq 0$.

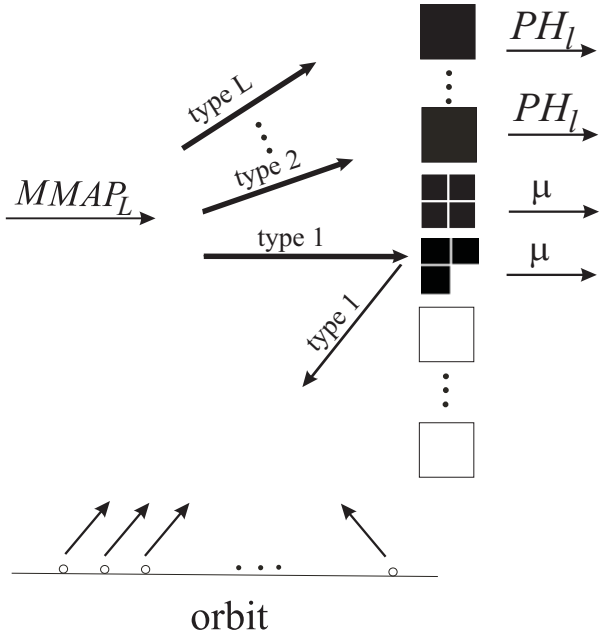
The stationary distribution vector $\boldsymbol{\theta}$ of this process satisfies the system of equations

$$\boldsymbol{\theta} D(1) = \mathbf{0},\ \boldsymbol{\theta}\mathbf{e} = 1.$$

The average intensity of type-l customers arrival λ_l is defined by the formula $\lambda_l = \boldsymbol{\theta} D_l \mathbf{e}, \quad l = \overline{1, L}$.

The squared coefficient of variation c_{var} of intervals between successive arrivals is defined by $c_{var} = 2\lambda\boldsymbol{\theta}(-D_0)^{-1}\mathbf{e} - 1$.

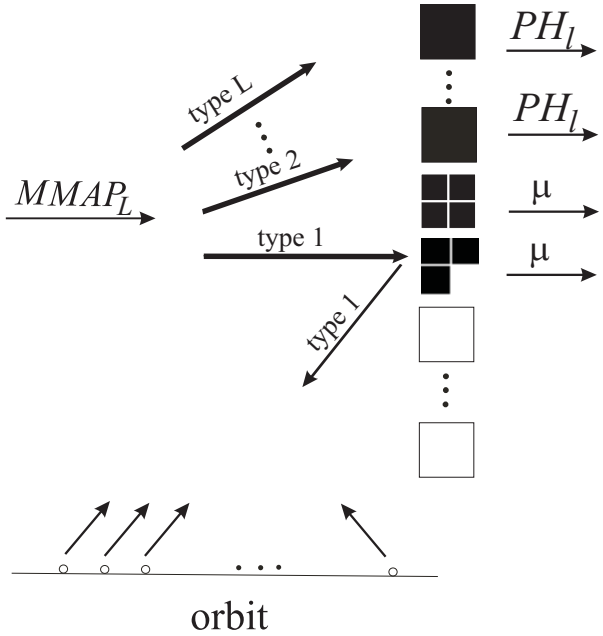
The coefficient of correlation c_{cor} of two successive intervals between arrivals is defined by $c_{cor} = (\lambda\boldsymbol{\theta}(-D_0)^{-1}(D(1) - D_0)(-D_0)^{-1}\mathbf{e} - 1)/c_{var}$.



Type-1 customers are low-priority customers and all other types of customers have preemptive priority over type-1 customers. Priority customers of any type do not have any privilege over other types of priority customers.

Service to a priority customer is provided by a whole server. Service to type-1 customer is provided by a sub-server.

An arriving priority customer is always admitted to the system except the situation when, during its arrival moment, all servers are occupied by priority customers.



If during a priority customer arrival epoch all servers are busy, but at least one of them provides service to type-1 customers, service in one of such servers is terminated. Type-1 customers, service of which was provided by the terminated server, move into orbit and retry later.

Admission to the system of type-1 customers is restricted via the threshold mechanism defined as follows. Some preassigned threshold M is fixed, $0 < M \leq K$. The incoming type-1 customer is accepted for service in the system if the number of totally busy servers is less than M. Otherwise, this customer goes to orbit.

A customer in orbit repeats the attempts to get access after a time interval having an exponential distribution with the parameter $\alpha, \alpha > 0$.
The customers staying in orbit may be impatient and leave the system, independently of each other, after a random amount of time having an exponential distribution with the parameter $\gamma,\ \gamma \geq 0$.
The service time of type-1 customers is exponentially distributed with the rate μ.
The service time of type-l customer has a phase-type distribution PH_l with an irreducible representation $(\boldsymbol{\beta}^{(l)}, S^{(l)})$, $l = \overline{2, L}$.

We propose to use generalized PH distribution with an irreducible representation $(\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_L, S)$ where

$\boldsymbol{\beta}_l = \left(\mathbf{0}_{\sum_{m=2}^{l-1} M_m}, \boldsymbol{\beta}^{(l)}, \mathbf{0}_{\sum_{m=l+1}^{L} M_m}\right)$, $l = \overline{2, L}$, and the matrix S of the form

$S = \operatorname{diag}\{S^{(2)}, \ldots, S^{(L)}\}$.

Cognitive Radio. Particular case

$L = 2$, $k = 1$, exponential service time distributions

Sun B., Lee M.H., Dudin S.A., Dudin A.N. Analysis of multiserver queueing system with opportunistic occupation and reservation of servers. Mathematical Problems in Engineering. 2014. ID 178108. P. 1-13.

Consider the process

$$\xi_t = \{i_t, n_t, s_t, \nu_t, \eta_t^{(1)}, \ldots, \eta_t^{(\mathcal{M})}\}, \quad t \geq 0,$$

where

i_t, $i_t \geq 0$, is the number of type-1 customers in orbit,

n_t, $n_t = \overline{0, K}$, is the number of priority customers on service,

s_t, $s_t = \overline{0, \min\{K - n_t, M\}k}$, is the number of type-1 customers on service,

ν_t, $\nu_t = \overline{0, W}$, is the state of the underlying process of the *MMAP*,

$\eta_t^{(m)}$ is the number of servers at phase m of generalized service process of priority customers, $\eta_t^{(m)} = \overline{0, n_t}$, $m = \overline{1, \mathcal{M}}$, $\sum_{m=1}^{\mathcal{M}} \eta_t^{(m)} = n_t$, during the moment t, $t \geq 0$.

The process ξ_t, $t \geq 0$, is an irreducible regular continuous-time Markov chain.

Lemma 2. *The generator Q of the Markov chain ξ_t has the following block structure:*

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & \dots & Q_{0,k} & O & O & \dots \\ Q_{1,0} & Q_{1,1} & \dots & Q_{1,k} & Q_{1,k+1} & O & \dots \\ O & Q_{2,1} & \dots & Q_{2,k} & Q_{2,k+1} & Q_{2,k+2} & \dots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Remark 1. It can be verified that the following limits exist:

$$Y_0 = \lim_{i\to\infty} R_i^{-1} Q_{i,i-1},\ Y_1 = \lim_{i\to\infty} R_i^{-1} Q_{i,i} + I,$$

$$Y_l = \lim_{i\to\infty} R_i^{-1} Q_{i,i+l-1},\ l = \overline{2, k+1},$$

where the matrix R_i is a diagonal matrix with the diagonal entries defined as the moduli of the corresponding diagonal entries of the matrix $Q_{i,i},\ i \geq 0$.

Ergodicity condition and stationary probabilities of the system states

Existence of the limits $Y_l,\ l = \overline{0, k+1}$, implies that the Markov chain $\xi_t,\ t \geq 0$, belongs to the class of continuous-time asymptotically quasi-Toeplitz Markov chains ($AQTMC$).

The sufficient condition for the ergodicity of the $AQTMC$ $\xi_t,\ t \geq 0$, is the fulfillment of the inequality

$$\mathbf{y} Y_0 \mathbf{e} > \mathbf{y} \sum_{l=2}^{k+1} (l-1) Y_l \mathbf{e} \tag{1}$$

where the row vector $\mathbf{y}$ is the unique solution to the system of linear algebraic equations

$$\mathbf{y} \sum_{l=0}^{k+1} Y_l = \mathbf{y},\ \mathbf{y}\mathbf{e} = 1. \tag{2}$$

Lemma 3. If the intensity of impatience $\gamma > 0$ the ergodicity condition is fulfilled for any other system parameters.

If condition (2) is fulfilled then the stationary probabilities exist:

$$\pi(i,n,s,\nu,\eta^{(1)},\ldots,\eta^{(\mathcal{M})}) = \lim_{t\to\infty} P\{i_t = i, n_t = n, s_t = s, \nu_t = \nu,$$

$$\eta_t^{(1)} = \eta^{(1)},\ldots,\eta_t^{(\mathcal{M})} = \eta^{(\mathcal{M})}\}, i \geq 0,\ n = \overline{0,K}.$$

Let us form the row vectors of the stationary probabilities $\boldsymbol{\pi}_i$ as follows:

$$\boldsymbol{\pi}(i,n,s) = (\boldsymbol{\pi}(i,n,s,0),\boldsymbol{\pi}(i,n,s,1),\ldots,\boldsymbol{\pi}(i,n,s,W)),$$

$$\boldsymbol{\pi}(i,n) = (\boldsymbol{\pi}(i,n,0),\boldsymbol{\pi}(i,n,1),\ldots,\boldsymbol{\pi}(i,n,\tilde{N}_n)),\ \tilde{N}_n = k\min\{K-n,M\},$$

$$\boldsymbol{\pi}_i = (\boldsymbol{\pi}(i,0),\boldsymbol{\pi}(i,1),\ldots,\boldsymbol{\pi}(i,K)),\ i \geq 0.$$

The probability vectors $\boldsymbol{\pi}_i$, $i \geq 0$, satisfy the following system of linear algebraic equations:

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \dots) Q = \mathbf{0}, \quad (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \dots)\mathbf{e} = 1.$$

V.I. Klimenok, A.N. Dudin, "Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory *Queueing Systems*, vol. 54, pp. 245-259, 2006.

Performance measures

- The average number of priority customers on service is
$N_{prior} = \sum_{i=0}^{\infty} \sum_{n=1}^{K} n\pi(i,n)\mathbf{e}.$
- The average number of type-1 customers on service is
$N_{type-1} = \sum_{i=0}^{\infty} \sum_{n=0}^{K} \sum_{s=1}^{\tilde{N}_n} s\pi(i,n,s)\mathbf{e}.$
- The average number of customers in orbit is $L_{orbit} = \sum_{i=1}^{\infty} i\pi_i\mathbf{e}.$
- The intensity of output of type-1 customers is
$\lambda_{out}^{(non-priority)} = \mu N_{type-1}.$
- The intensity of output flow of priority customers is

$$\lambda_{out}^{(prior)} = \sum_{i=1}^{\infty} \sum_{n=1}^{K} \pi(i,n)(I_{(\tilde{N}_n+1)\bar{W}} \otimes L_{K-n}(K,\tilde{S}))\mathbf{e}.$$

- The intensity of output flow of type-l customers $\lambda_{out}^{(l)}$, $l = \overline{2, L}$, who get service in the system is calculated as

$$\lambda_{out}^{(l)} = \sum_{i=1}^{\infty} \sum_{n=1}^{K} \pi(i, n)(I_{(\tilde{N}_n+1)\bar{W}} \otimes L_{K-n}(K, \tilde{S}^{(l)}))\mathbf{e}.$$

- The loss probability of type-1 customers is

$$P_1^{(loss)} = 1 - \frac{\lambda_{out}^{(1)}}{\lambda_1} = \frac{\gamma L_{orbit}}{\lambda_1}.$$

- The loss probability of type-l, $l = \overline{2, L}$, customers is computed as

$$P_l^{(loss)} = \lambda_l^{-1} \sum_{i=0}^{\infty} \pi(i, K)(I_{\tilde{N}_K} \otimes D_l \otimes I_{T_K})\mathbf{e}.$$

- The loss probability of a priority customer is

$$P_{prior}^{(loss)} = 1 - \lambda_{out}^{(prior)} / \lambda_{prior}$$

where $\lambda_{prior} = \sum_{l}^{L} \lambda_l$.

- The blocking (loss) probability of an arbitrary customer is

$$P^{(loss)} = 1 - \frac{\lambda_{out}^{(non-priority)} + \lambda_{out}^{(priority)}}{\lambda}.$$

The probability that arriving type-1 customer is not granted immediate access to the system (goes to orbit) is

$$P^{(ent-orbit)} = \lambda_1^{-1} \sum_{i=0}^{\infty} \sum_{n=0}^{K} \sum_{s=\max\{0,(M-n)k\}}^{\tilde{N}_n} \boldsymbol{\pi}(i,n,s)(D_1 \otimes I_{T_n})\mathbf{e}.$$

- The intensity of type-1 customers whose service was terminated is calculated as

$$\gamma^{(termination-orbit)} = \sum_{i=0}^{\infty}\sum_{n=0}^{K-1}\sum_{s=(K-n-1)k+1}^{\tilde{N}_n} ((1-\delta_{s-\lfloor\frac{s}{k}\rfloor k,0})(s-\lfloor\frac{s}{k}\rfloor k)+$$

$$\delta_{s-\lfloor\frac{s}{k}\rfloor k,0}k) \times \pi(i,n,s)((\sum_{l=2}^{L} D_l)\otimes I_{T_n})\mathbf{e}.$$

- The probability that arrival of a priority customer will lead to forced termination of service of non-priority customers is

$$P^{(termination)} = \lambda_{prior}^{-1}\sum_{i=0}^{\infty}\sum_{n=0}^{K-1}\sum_{s=(K-n-1)k+1}^{\tilde{N}_n} \pi(i,n,s)((\sum_{l=2}^{L} D_l)\otimes I_{T_n})\mathbf{e}.$$

Experiment 1.

Dependence of the main system performance measures on the number k of sub-channels in a channel.
We consider the system with $K = 5$ servers and two types of customers: type-1 customers are non-priority and type-2 customers are priority.
The average intensity of type-1 customers arrival is $\lambda_1 = 2.666$, the average intensity of type-2 customers arrival is $\lambda_2 = 1.333$, the coefficients of correlation $c_{cor} = 0.2$ and the coefficient of variation $c_{var} = 12.35$.
The service time of type-2 customers is defined by $S^{(2)} = (-1.5)$ and $\beta^{(2)} = (1)$.
If the type-1 customer is served by the whole server ($k = 1$, subchannel is the same as the channel), the service rate is $\mu = 3$. If the channel is divided into k subchannels, then the service intensity decreases proportionally, i.e., $\mu = \frac{3}{k}$.
The intensity of impatience of each customer from orbit is $\gamma = 0.1$, the retrial intensity of each customer from orbit is $\alpha = 0.15$.

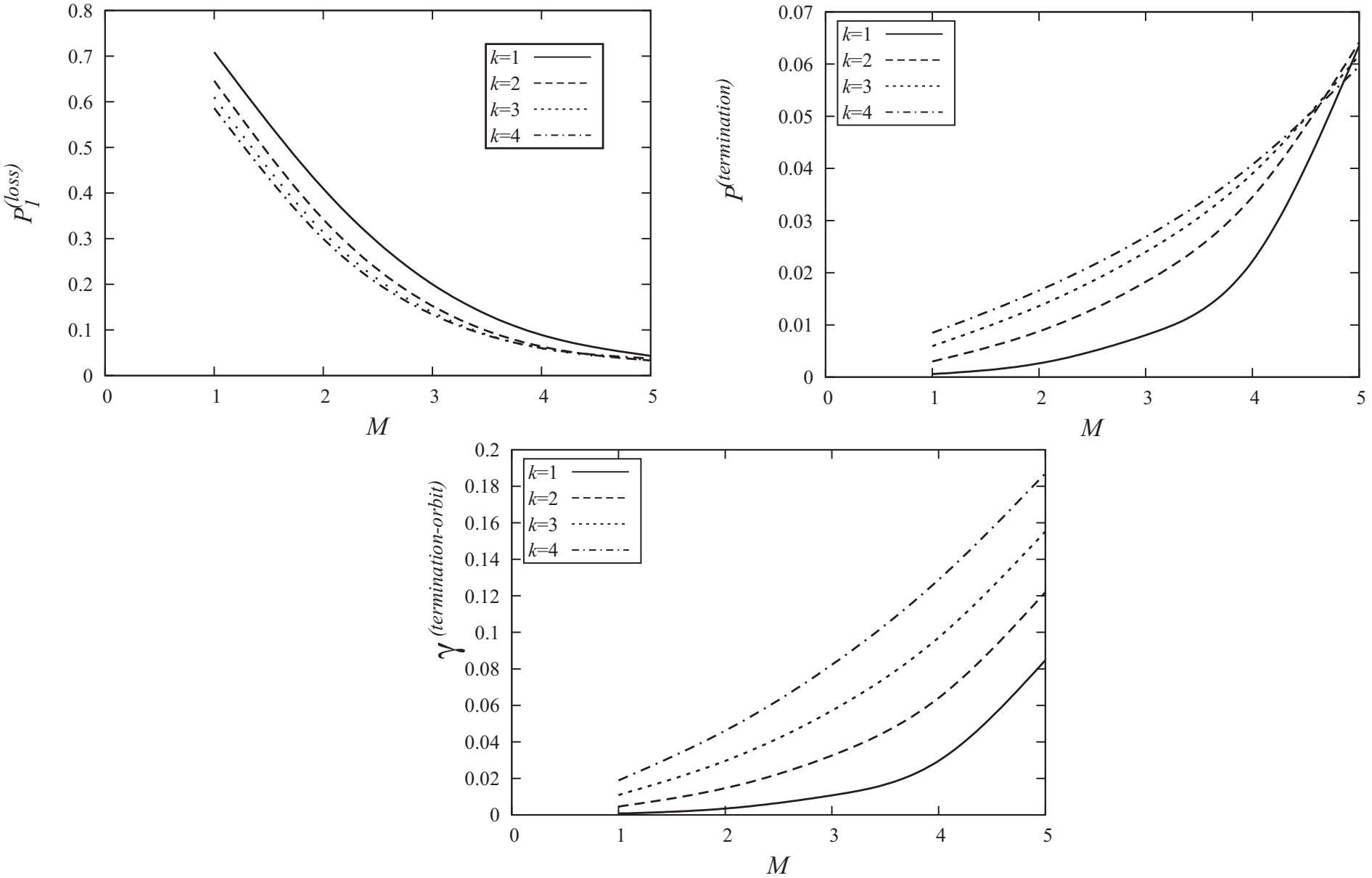


Figure 2: Dependence of some of performance measures on the threshold M for different values of the number of subchannels $k = \overline{1,4}$

Experiment 2.

Importance of account the impact of variance of the service time.
We consider one type of priority customers with three service time distributions with the same mean service rate $\tilde{\mu} = 1.5$, but different variance .
The first service process, coded as $var = 1$, has the coefficient of variation equal to 1 and is defined by the matrix $S^{(2)} = (-1.5)$ and the vector $\beta^{(2)} = (1)$.
The second service process, coded as $var = 0.5$, has the coefficient of variation equal to 0.5 and is defined by the matrix $S^{(2)} = \begin{pmatrix} -3 & 3 \\ 0 & -3 \end{pmatrix}$, and the vector $\beta^{(2)} = (1, 0)$.
The third service process, coded as $var = 25$, has the coefficient of variation equal to 25 and is defined by the matrix $S^{(2)} = \begin{pmatrix} -0.09312 & 0 \\ 0 & -7.323 \end{pmatrix}$ and the vector $\beta^{(2)} = (0.05, 0.95)$.
Let us assume that the rest of system parameters is the same as in the previous example and $k = 2$.

Table 1. Performance measures of priority customers

	$P_{prior}^{(loss)}$	$\lambda_{out}^{(prior)}$	N_{prior}
$var = 0.5$	0.00592	1.32516	0.88344
$var = 1$	0.00584	1.32526	0.88351
$var = 25$	0.00424	1.32739	0.88494

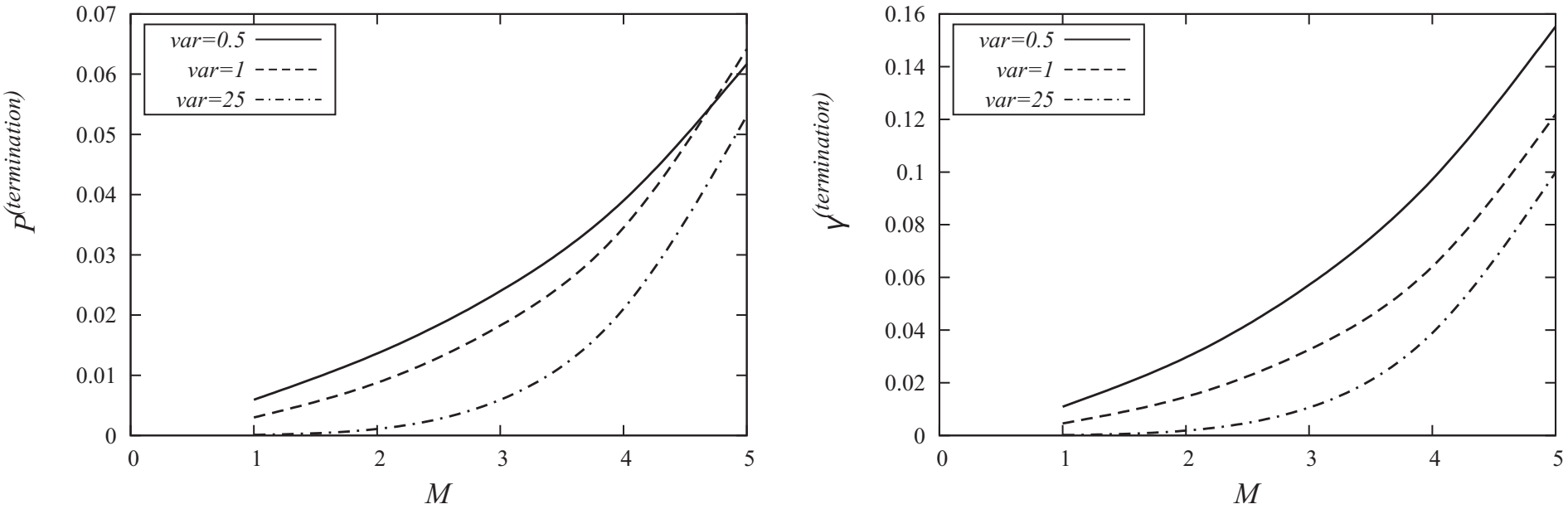


Figure 3: Dependence of the probability that arrival of a priority customer leads to termination of service of non-priority customers and the intensity of type-1 customers whose service was terminated on the parameter M for different coefficients of variation in the service process

Experiment 3.

In this experiment, we investigate the importance of account of types of primary customers.
System with $K = 5$ servers (each server consists of $k = 2$ subchannels) and three types of customers arrive to the system: type-1 customers are non-priority customers, type-2 and type-3 customers are priority customers.
The arrival flow is defined by the matrices

$$D_0 = \begin{pmatrix} -5 & 0 \\ 0 & -1 \end{pmatrix}, D_1 = \begin{pmatrix} 0.5 & 0.5 \\ 0.01 & 0.5 \end{pmatrix},$$

$$D_2 = \begin{pmatrix} 3.6 & 0.1 \\ 0 & 0 \end{pmatrix}, D_3 = \begin{pmatrix} 0 & 0.3 \\ 0.02 & 0.47 \end{pmatrix}.$$

This arrival flow has $\lambda_1 = 0.5258$, $\lambda_2 = 0.11935$, $\lambda_3 = 0.48387$, $c_{cor} = 0.06577$ and $c_{var} = 1.2$.
$\mu = 1$, $\gamma = 0.1$, $\alpha = 0.15$, $S^{(2)} = (-\mu_2)$, $\beta^{(2)} = (1)$, $\mu_2 = 0.1141975$, $S^{(3)} = (-\mu_3)$, $\beta^{(3)} = (1)$, $\mu_3 = 3$.

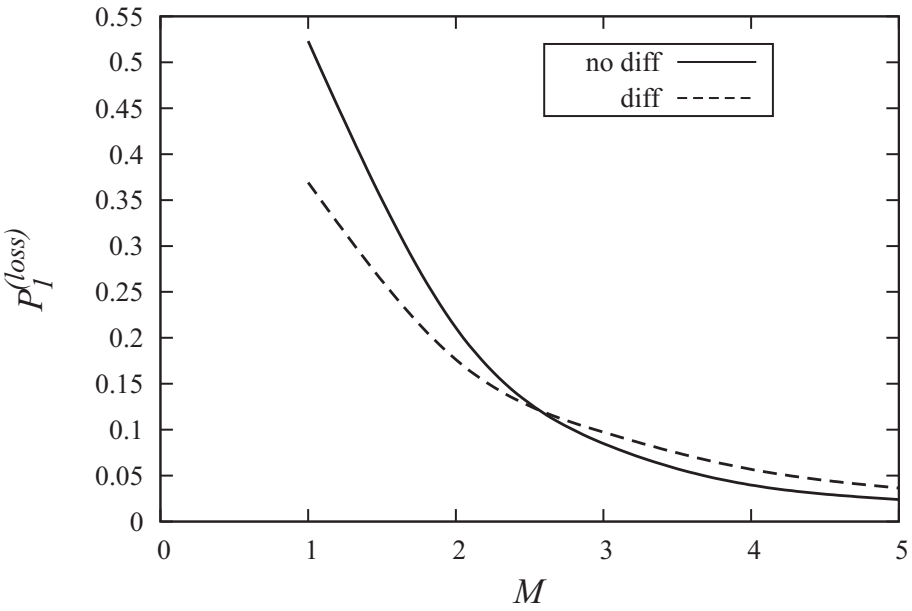
Now let us consider the case when priority customers are not differentiated. The arrival flow is defined by the matrices

$$D_0 = \begin{pmatrix} -5 & 0 \\ 0 & -1.0 \end{pmatrix}, D_1 = \begin{pmatrix} 0.5 & 0.5 \\ 0.01 & 0.5 \end{pmatrix}, D_2 = \begin{pmatrix} 3.6 & 0.4 \\ 0.02 & 0.47 \end{pmatrix},$$

The arrival flow of priority customers has the intensity $\lambda_p = \lambda_2 + \lambda_3 = 0.48387 + 0.11935 = 0.60322$. The coefficients of correlation and variation do not change when the flows are superposed. The mean service time of priority customers can be found as $b_p = \frac{\mu_2^{-1}\lambda_2 + \mu_3^{-1}\lambda_3}{\lambda_p} = 2$.
So, the service time of priority customers is defined by the matrix $S^{(2)} = (-\frac{1}{b_p}) = (-0.5)$ and the vector $\beta^{(2)} = (1)$.

Table 2. Performance measures of priority customers

	$P_2^{(loss)}$	$P_3^{(loss)}$	$P_{prior}^{(loss)}$	$\lambda_{out}^{(prior)}$	N_{prior}
with differentiation	0.33349	0.02691	0.08757	0.5504	0.85356
without differentiation	-	-	0.05523	0.56991	1.13982



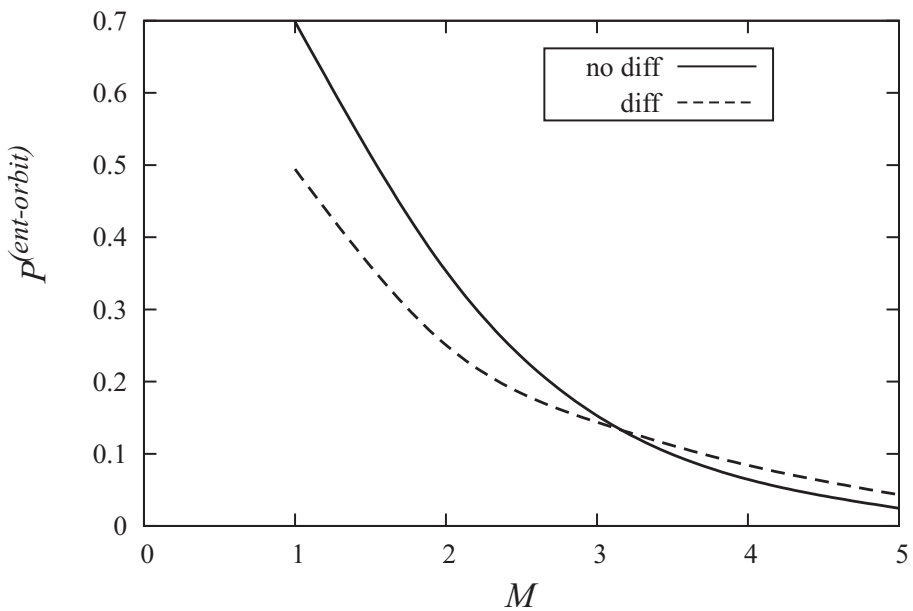


Figure 4: Dependence of the loss probabilities of type-1 customers and the probability of type-1 customer goes to orbit upon arrival on the parameter M

Dudin A.N., Lee M.H., Dudina O.S., Lee S.K. Analysis of priority retrial queueing system with many types of customers and servers reservation as the model of information transmission in cognitive radio system. IEEE Transactions on Communications. 2016. V.64. No 12. DOI: 10.1109/TCOMM.2016.2606379

Conclusion

We illustrated application of *GPH* distribution to analysis of a multi-server priority retrial queue with many types of customers. Different types of priority customers have different distribution of a server occupation time. Non-priority customers may share a server (occupy a sub-channel). Reservation of servers aiming to decrease the forced termination probability of non-priority customers is assumed. The formulated purpose (to analyse the generalized model and clarify whether or not the account of the use of sub-channels, presence of many types of primary customers, more general service time distributions essentially impacts on the key performance measures of the system) is achieved.

Mathematical analysis of the considered queueing model is implemented via investigation of stationary behavior of state inhomogeneous multi-dimensional continuous time Markov chain. Behavior of some performance measures of the system as a function of the reservation threshold is numerically illustrated. It creates an opportunity of further solving various optimization problems for the model under study. Importance of account of correlation and variation in arrival process, variation in service process, the number of sub-channels in the channel, separate consideration of priority customers of different type is numerically illustrated.

New Journal "Queueing Models and Service Management"

It is an international refereed journal devoted to the publication of original research papers specializing in queueing systems, queueing networks, reliability and maintenance, service system optimization, service management, and applications in queueing models or networks. The journal publishes theoretical papers using analytical methods or developments of significant methodologies. QMSM publishes works of originality, quality and significance, with particular emphasis given to practical results. Practical papers, illustrating the applications of queueing and service management problems, are of special interest.

https://qmsm.cs.pu.edu.tw/

Thank you for attention!