7

Queues with session arrivals as models for optimizing the traffic control in telecommunication networks

Sergey Dudin [1] and Moon Ho Lee [2]
[1]*Belarusian State University*
Belarus
[2]*Chonbuk National University*
Korea

1. Introduction

Many problems in routing, scheduling, flow control, resources allocation and capacity management in telecommunication, production, and transportation networks can be solved with help of queueing theory. Typically, a user of a network generates not a single item (packet, job, pallet, etc) but a whole bunch of items and service of this user assumes sequential transmission of all these items. This is why the batch arrivals are often assumed in analysis of queueing systems. It is usually assumed that, at a batch arrival epoch, all requests of this batch arrive into the system simultaneously. However, the typical feature of many nowadays networks is that requests arrive in batch, while arrival of requests belonging to a batch is not instantaneous but is distributed in time. We call such batches as sessions. The first request of a session arrives at the session arrival epoch while the rest of requests arrive one by one in random intervals. The session size is random and it may be not known a priori at the session arrival epoch. Such a situation is typical, e.g., in modeling transmission of video and multimedia information. This situation is also typical in IP networks, e.g., in World Wide Web with Hypertext Transfer Protocol (HTTP) where a session can be interpreted as a HTTP connection and a request as a HTTP request. This situation is also discussed in literature with respect to the modeling the Scheme of Alternative Packet Overflow Routing (*SAPOR*) in *IP* networks.

In this scheme, the session is called as *flow* and represents a set of packets that should be sequentially routed in the same channel. When a packet arrives, it is determined (e.g. by means of *IP* address) if the packet is a part of a flow, already tracked. If the packet belongs to an existing flow, the packet is marked for transmission. If the flow is not yet tracked and the channel capacity is still available, the packet is admitted into the system and flow count is increased. Otherwise the flow is routed on the overflow link (or is dropped at all) and the packet is rejected in the considered channel. Tracked flows are cleared after they are finished. Clearing of an inactive flow is done if no more packets belonging to this flow are received within a certain time interval. Tracking and clearing of flows is performed by

means of a token mechanism. The number of tokens, which defines the maximal number of flows that can be admitted into the system simultaneously, is very important control parameter. If this number is small, the channel may be underutilized. If this number is too large, the channel may become congested. Average delivering time and jitter may increase essentially and Grade of Service becomes bad. So, the problem of defining the optimal number of tokens is of practical importance and non-trivial. In (Kist et al., 2005), performance measures of the *SAPOR* scheme of routing in *IP* networks are evaluated by means of computer simulation.

Analogous situation also naturally arises in modeling information retrieving in relational data bases where, besides the CPU and disc memory, some additional "threads" or "connections" should be provided to start the user's application processing. In this interpretation, session means application while requests are queries to be processed within this application and tokens are threads or connections.

In the paper (Lee et al., 2007), the Markovian queueing model with a finite buffer that suits for analytical performance evaluation and capacity planning of the *SAPOR* routing scheme as well as for modelling the other real-world systems with time distributed arrival of requests in a session is considered. To the best of our knowledge, such kind of queueing models was not considered and investigated in literature previously. In (Lee et al., 2007), the problem of the system throughput maximization subject to restriction of the loss probability for requests from accepted sessions is solved. In the paper (Kim et al., 2009), the analysis given in (Lee et al., 2007) is extended in three directions. Instead of the stationary Poisson arrival process of sessions, the *Markov Arrival Process (MAP)* is considered. It allows catching the effect of correlation of flow of sessions. The presented numerical results show that the correlation has profound effect on the system performance measures. The second direction is consideration of the *Phase type (PH)* service process instead of an exponential service time distribution assumed in (Lee et al., 2007). Because *PH* type distributions are suitable for fitting an arbitrary distribution, this allows to take into account the service time distribution and variance of this time in particular, carefully. The third direction of extension is the following one. It is assumed in (Lee et al., 2007), that the loss (due to a buffer overflow) of the request from the accepted session never causes loss of a whole session itself. More realistic assumption in some situations is that the session might be lost (terminates connection ahead of schedule). E.g., it can happen if the percentage of lost voice or video packets (and quality of speech or movie) becomes unacceptable for the user. To take such a possibility into account in some extent, it is assumed in this paper that the loss of a request from the admitted session, with fixed probability, leads to the loss of a session to which this request belongs. Influence of this probability is numerically investigated in the paper (Kim et al., 2009).

In the present paper, the modification of model from (Kim et al., 2009) to the case of an infinite buffer is under study. In contrast to the model with a finite buffer considered in (Kim et al., 2009) where the problem of the throughput maximization was solved under constraint on the probability of the loss of a request from an accepted session, here we do not have such a loss. So, the problem of the throughput maximization is solved under constraint on the average sojourn time of requests from the accepted sessions. In section 2, the mathematical model is described in detail. Stability condition, which is not required in the model (Kim et al., 2009) with a finite state space but is very important in the model with an infinite buffer space, is derived in a simple form. This condition creates an additional

constraint in maximization problem. The steady state joint distribution of the number of sessions and requests in the system is analyzed by means of the matrix analytical technique and expressions for the main performance measures of the system are given in section 3. Section 4 is devoted to consideration of the request and the session sojourn time distribution. Section 5 contains numerical illustrations and their short discussion and section 6 concludes the paper.

2. Mathematical model

We consider a single server queueing system with a buffer of an infinite capacity. The requests arrive to the system in sessions. Sessions arrive according to the Markov Arrival Process. Sessions arrival in the MAP is 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 Markov chain ν_t, $t \geq 0$, in the state ν has an exponential distribution with the parameter $\lambda_\nu, \nu = \overline{0,W}$. After this sojourn time expires, with probability $p_k(\nu,\nu')$, the process ν_t, $t \geq 0$, transits to the state ν', and k sessions, $k = 0,1$, arrive into the system. The intensities of jumps from one state into another, which are accompanied by an arrival of k sessions, are combined into the matrices $D_k, k = 0,1$, of size $(W+1)\times(W+1)$. The matrix generating function of these matrices is $D(z) = D_0 + D_1 z, |z| \leq 1$. The matrix $D(1)$ is the infinitesimal generator of the process $\nu_t, t \geq 0$. The stationary distribution vector $\boldsymbol{\delta}$ of this process satisfies the equations $\boldsymbol{\delta} D(1) = \mathbf{0}, \boldsymbol{\delta}\mathbf{e} = 1$. Here and in the sequel $\mathbf{0}$ is the zero row vector and $\mathbf{e}$ is the column vector of appropriate size consisting of 1's. In case the dimensionality of the vector is not clear from the context, it is indicated as a lower index, e.g. $\mathbf{e}_{\bar{W}}$ denotes the unit column vector of dimensionality $\bar{W} = W + 1$.
The average intensity λ (fundamental rate) of the MAP is defined as

$$\lambda = \boldsymbol{\delta} D_1 \mathbf{e}.$$

The variance v of intervals between session arrivals is calculated as

$$v = 2\lambda^{-1}\boldsymbol{\delta}(-D_0)^{-1}\mathbf{e} - \lambda^{-2},$$

the squared coefficient c_{var} of variation is calculated by

$$c_{var} = 2\lambda\boldsymbol{\delta}(-D_0)^{-1}\mathbf{e} - 1,$$

while the correlation coefficient c_{cor} of intervals between successive group arrivals is given by

$$c_{cor} = (\lambda^{-1}\boldsymbol{\delta}(-D_0)^{-1}D_1(-D_0)^{-1}\mathbf{e} - \lambda^{-2})/v.$$

For more information about the MAP, its special cases and properties and related research see (Fisher & Meier-Hellstern, 1993), (Lucantoni, 1991) and the survey paper by S.

Chakravarthy (Chakravarthy, 2001). Usefulness of the *MAP* in modeling telecommunication systems is mentioned in (Heyman & Lucantoni, 2003), (Klemm et al., 2003). Note, that the problem of constructing the *MAP*, which fits well a real arrival process, is not very simple. However, this problem has practical importance and is intensively solving. For relevant references and the fitting algorithms see, e.g., (Heyman & Lucantoni, 2003), (Klemm et al., 2003), (Asmussen et al., 1996) and (Panchenko & Buchholz, 2007).

Following (Kist et al., 2005) , we assume that admission of sessions (they are called *flows* in (Kist et al., 2005) and called *threads, connections, sessions, exchanges, windows*, etc. in different real-world applications) is restricted by means of *tokens*. The total number of available tokens is assumed to be $K, K \geq 1$. Further we consider the number K as a control parameter and solve the corresponding optimization problem.

If there is no token available at a session arrival epoch the session is rejected. It leaves the system forever. If the number of available tokens at the session arrival epoch is positive this session is admitted into the system and the number of available tokens decreases by one. We assume that one request of a session arrives at the session arrival epoch and if it meets free server, it occupies the server and is processed. If the server is busy, the request moves to the buffer and later it is picked up for the service according to the First Came - First Served discipline.

After admission of the session, the next request of this session can arrive into the system in an exponentially distributed with the parameter γ time. The number of requests in the session has the geometrical distribution with the parameter $\theta, 0 < \theta < 1,$ i.e., probability that the session consists of k requests is equal to $\theta^{k-1}(1-\theta), k \geq 1.$ The average size of the session is equal to $(1-\theta)^{-1}$.

If the exponentially distributed with the parameter γ time since arrival of the previous request of a session expires and new request does not arrive, it means that the arrival of the session is finished. The token, which was obtained by this session upon arrival, is returned to the pool of available tokens. The requests of this session, which stay in the system at the epoch of returning the token, must be completely processed by the system. When the last request is served, the sojourn time of the session in the system is considered finished.

The service time of a request is assumed having *PH* distribution. It means the following. Request's service time is governed by the directing process $\eta_t, t > 0,$ which is the continuous time Markov chain with the state space $\{1,\ldots,M\}.$ The initial state of the process $\eta_t, t \geq 0,$ at the epoch of starting the service is determined by the probabilistic row-vector $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_M)$. The transitions of the process $\eta_t, t \geq 0,$ that do not lead to the service completion, are defined by the irreducible matrix S of size $M \times M$. The intensities of transitions, which lead to the service completion, are defined by the column vector $\mathbf{S}_0 = -S\mathbf{e}$. The service time distribution function has the form $B(x) = 1 - \boldsymbol{\beta} e^{Sx}\mathbf{e}$. Laplace-Stieltjes transform $\int_0^\infty e^{-sx} dB(x)$ of this distribution function is $\boldsymbol{\beta}(sI - S)^{-1}\mathbf{S}_0.$ The average service time is given by $b_1 = \boldsymbol{\beta}(-S)^{-1}\mathbf{e}$. The matrix $S + \mathbf{S}_0\boldsymbol{\beta}$ is assumed to be irreducible. The more detailed description of the *PH* -type distribution and its partial cases can be found, e.g., in the book (Neuts, 1981). Usefulness of *PH* distribution in description of service

process in telecommunication networks is stated, e.g., in (Pattavina & Parini, 2005) and (Riska et al., 2002).

It is intuitively clear that the described mechanism of arrivals restriction by means of tokens is reasonable. At the expense of some sessions rejection, it allows to decrease the sojourn time and jitter for admitted sessions. This is important in modeling real-world systems because the quality of transmission of accepted information units should satisfy imposed requirements of Quality of Service. Quantitative analysis of advantages and shortcomings of this mechanism as well as optimal choice of the number of tokens requires calculation of the main performance measures of the system under the fixed value K of tokens in the system. These measures can be calculated based on the knowledge of stationary distribution of the random process describing dynamics of the system under study.

3. Stationary distribution of the system states

Let us assume that the number $K, K \geq 1,$ of tokens is fixed and let

- i_t be the total number of requests in the system, $i_t \geq 0,$
- k_t be the number of sessions having token for admission to the system, $k_t = \overline{0,K},$
- ν_t and η_t be the states of the directing processes of the MAP arrival process and PH service process, $\nu_t = \overline{0,W}, \eta_t = \overline{1,M},$

at the epoch $t, t \geq 0.$

Note that when $i_t = 0,$ i.e., requests are absent in the system, the value of the component $\eta_t,$ which describes the state of the service directing process, is not defined. To avoid special treatment of this situation, without loss of generality, we assume that if the server becomes idle the state of the component η_t is chosen randomly according to the probabilistic vector $\boldsymbol{\beta}$ and is kept until the next service beginning moment.

It is obvious that the four-dimensional process $\xi_t = \{i_t, k_t, \nu_t, \eta_t\}, t \geq 0,$ is the irreducible regular continuous time Markov chain.

Let us enumerate the states of this Markov chain in lexicographic order and refer to (i,k) as macro-state consisting of $M_1 = (W+1)M$ states $(i,k,\nu,\eta), \nu = \overline{0,W}, \eta = \overline{1,M}.$

For the use in the sequel, introduce the following notation:

- $\gamma^- = \gamma(1-\theta), \gamma^+ = \gamma\theta, \quad \Gamma^- = \gamma^- I_{M_1}, \Gamma^+ = \gamma^+ I_{M_1}, \Gamma = \gamma I_{M_1};$
- $\tilde{C}_K = diag\{0,1,\ldots,K\}$ is the diagonal matrix with the diagonal entries $\{0,1,\ldots,K\}, \quad C_K = \tilde{C}_K \otimes I_{M_1};$
- $R_K = diag\{1,\ldots,K\} \otimes I_{M_1};$ I is an identity matrix, O is a zero matrix;
-

$$A = \begin{pmatrix} O & O & O & \ldots & O & O \\ \Gamma^- & -\Gamma & O & \ldots & O & O \\ O & 2\Gamma^- & -2\Gamma & \ldots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \ldots & K\Gamma^- & -K\Gamma \end{pmatrix}, E^+ = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \\ 0 & 0 & 0 & \ldots & 0 \end{pmatrix};$$

- $$A_1 = \begin{pmatrix} -\Gamma & O & \dots & O & O \\ \Gamma^- & -2\Gamma & \dots & O & O \\ O & 2\Gamma^- & \dots & O & O \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O\dots & (K-1)\Gamma^- & -K\Gamma \end{pmatrix}, \tilde{E} = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix};$$
- $\delta_{i,j}$ is Kronecker delta, $\delta_{i,j}$ is equal to 1, if $i = j$ and equal to 0 otherwise;
- $\otimes$ is the symbol of Kronecker product of matrices;
- $\oplus$ is the symbol of Kronecker sum of matrices;
- $\mathbf{b}^T$ denotes transposed vector $\mathbf{b}$.

Let Q be the generator of the Markov chain $\xi_t, t \geq 0,$ with blocks $Q_{i,j}$ consisting of intensities $(Q_{i,j})_{k,k'}$ of the Markov chain $\xi_t, t \geq 0,$ transitions from the macro-state (i,k) to the macro-state $(j,k'), k,k' = \overline{0,K}.$ The diagonal entries of the matrix $Q_{i,i}$ are negative and the modulus of the diagonal entry of $(Q_{i,i})_{k,k}$ defines the total intensity of leaving the corresponding state (i,k,ν,η) of the Markov chain. The block $Q_{i,j}, i,j \geq 0,$ has dimension $K_1 \times K_1,$ where $K_1 = (K+1)M_1.$

Lemma 1. Generator Q has the three block diagonal structure:

$$Q = \begin{pmatrix} Q_{0,0} & Q_0 & O & O & \dots \\ Q_2 & Q_1 & Q_0 & O & \dots \\ O & Q_2 & Q_1 & Q_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

where non-zero blocks $Q_{i,j}$ are defined by

$$Q_{0,0} = A + I_{K+1} \otimes D_0 \otimes I_M + \tilde{E} \otimes D_1 \otimes I_M,$$
$$Q_1 = A + I_{K+1} \otimes (D_0 \oplus S) + \tilde{E} \otimes D_1 \otimes I_M,$$
$$Q_0 = \gamma^+ C_K + E^+ \otimes D_1 \otimes I_M,$$
$$Q_2 = I_{K+1} \otimes I_{W+1} \otimes \mathbf{S}_0 \boldsymbol{\beta}.$$

Proof of the lemma consists of analysis of the Markov chain $\xi_t, t \geq 0,$ transitions during the infinitesimal interval of time and further assembling the corresponding transition intensities into the matrix blocks. Value γ^- is the intensity of a token releasing due to the finish of the session arrival, γ^+ is the intensity of a new request in the session arrival.

Let us investigate the Markov chain $\xi_t, t \geq 0,$ defined by the generator $Q.$ To this end, at first we should derive conditions under which this Markov chain is ergodic (positive recurrent).

Theorem 1. Markov chain $\xi_t = \{i_t, k_t, \nu_t, \eta_t\}, t \geq 0,$ is ergodic if and only if the following inequality is fulfilled:

$$\mu > \Lambda = \gamma^+ \sum_{k=0}^{K} k\mathbf{x}_k \mathbf{e}_{W+1} + \sum_{k=0}^{K-1} \mathbf{x}_k D_1 \mathbf{e}_{W+1}, \tag{1}$$

where μ is the average service rate defined by

$$\mu^{-1} = b_1 = \boldsymbol{\beta}(-S)^{-1}\mathbf{e}$$

and $\mathbf{x} = (\mathbf{x}_0, \ldots, \mathbf{x}_K)$ is the vector of the stationary distribution of the system $MAP/M/K/0$ with the MAP arrival process, defined by the matrices D_0 and D_1 and the average service rate γ^-.

Proof. It follows from (Neuts, 1981) that the ergodicity condition of the Markov chain $\xi_t = \{i_t, k_t, \nu_t, \eta_t\}, t \geq 0,$ is the fulfillment of inequality

$$\mathbf{y}Q_2\mathbf{e} > \mathbf{y}Q_0\mathbf{e}, \tag{2}$$

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

$$\mathbf{y}(Q_0 + Q_1 + Q_2) = \mathbf{0}, \mathbf{y}\mathbf{e} = 1. \tag{3}$$

It is easy to verify that

$$Q_0 + Q_1 + Q_2 = B \otimes I_M + I_{(K+1)(W+1)} \otimes (S + \mathbf{S}_0\boldsymbol{\beta}),$$

where B is the generator of the Markov chain, which describes behavior of the $MAP|M|K|0$ system with the MAP arrival process defined by matrices D_0 and D_1 and average service rate γ^-:

$$B = \begin{pmatrix} D_0 & D_1 & O & \ldots & O & O \\ \gamma^- I & D_0 - \gamma^- I & D_1 & \ldots & O & O \\ 0 & 2\gamma^- I & D_0 - 2\gamma^- I & \ldots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \ldots & K\gamma^- I & D(1) - K\gamma^- I \end{pmatrix}.$$

According to the definition, vector $\mathbf{x}$ satisfies equations

$$\mathbf{x}B = \mathbf{0}, \mathbf{x}\mathbf{e} = 1. \tag{4}$$

By direct substitution into (3), it can be verified that the vector $\mathbf{y}$, which is solution to the

system (3), can be represented in the form $\mathbf{y} = \mathbf{x} \otimes \boldsymbol{\psi}$, where $\boldsymbol{\psi}$ is the unique solution of the system of linear algebraic equations

$$\boldsymbol{\psi}(S + \mathbf{S}_0 \boldsymbol{\beta}) = \mathbf{0}, \boldsymbol{\psi}\mathbf{e} = 1. \tag{5}$$

By substituting vector $\mathbf{y} = \mathbf{x} \otimes \boldsymbol{\psi}$ into inequality (2), after some transformations we get inequality (1). Theorem 1 is proven.
In what follows we assume that condition (1) is fulfilled. Then the following limits (stationary probabilities) exist:

$$\pi(i,k,\nu,\eta) = \lim_{t\to\infty} P\{i_t = i, k_t = k, \nu_t = \nu, \eta_t = \eta\}, i \geq 0, k = \overline{0,K}, \nu = \overline{0,W}, \eta = \overline{1,M}.$$

Let us combine these probabilities into the row-vectors

$$\boldsymbol{\pi}(i,k,\nu) = (\pi(i,k,\nu,1), \pi(i,k,\nu,2), \ldots, \pi(i,k,\nu,M)),$$
$$\boldsymbol{\pi}(i,k) = (\boldsymbol{\pi}(i,k,0), \boldsymbol{\pi}(i,k,1), \ldots, \boldsymbol{\pi}(i,k,W)),$$
$$\boldsymbol{\pi}_i = (\boldsymbol{\pi}(i,0), \boldsymbol{\pi}(i,1), \ldots, \boldsymbol{\pi}(i,K)), \ i \geq 0.$$

The following statement directly stems from the results in (Neuts, 1981).
Theorem 2. The stationary probability vectors $\boldsymbol{\pi}_i, i \geq 0,$ are calculated by

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_0 R^i, i \geq 0,$$

where the matrix R is the minimal non-negative solution to the equation

$$R^2 Q_2 + R Q_1 + Q_0 = O,$$

and the vector $\boldsymbol{\pi}_0$ is the unique solution to the system of linear algebraic equations

$$\boldsymbol{\pi}_0 (Q_{0,0} + R Q_2) = \mathbf{0}, \boldsymbol{\pi}_0 (I - R)^{-1} \mathbf{e} = 1.$$

Having stationary probability vectors $\boldsymbol{\pi}_i, i \geq 0,$ been computed, we can calculate different performance measures of the system. Some of them are given in the following statements.
Corollary 1. The probability distribution of the number of requests in the system is computed by

$$\lim_{t\to\infty} P\{i_t = i\} = \boldsymbol{\pi}_i \mathbf{e}, i \geq 0.$$

The average number L of requests in the system is computed by

$$L = \sum_{i=0}^{\infty} i \boldsymbol{\pi}_i \mathbf{e} = \boldsymbol{\pi}_0 R (I - R)^{-2} \mathbf{e}.$$

The probability distribution of the number of sessions in the system is computed by

$$\lim_{t\to\infty} P\{k_t = k\} = \sum_{i=0}^{\infty} \boldsymbol{\pi}(i,k)\mathbf{e} = \boldsymbol{\pi}_0 (I - R)^{-1} (\mathbf{e}^{(k)} \otimes \mathbf{e}),\ k = \overline{0,K},$$

where the column vector $\mathbf{e}^{(k)}$ has all zero entries except the k th one, which is equal to 1, $k = \overline{0,K}$.

The average number Z of sessions in the system is computed by

$$Z = \sum_{k=1}^{K} \sum_{i=0}^{\infty} k\boldsymbol{\pi}(i,k)\mathbf{e} = \boldsymbol{\pi}_0 (I - R)^{-1} \sum_{k=1}^{K} k(\mathbf{e}^{(k)} \otimes \mathbf{e}).$$

The distribution function $R(t)$ of a time, during which arrivals from an arbitrary session occur, is computed by

$$R(t) = (1-\theta)\sum_{l=1}^{\infty} \int_0^t \theta^{l-1} \frac{\gamma(\gamma y)^{l-1}}{(l-1)!} e^{-\gamma y} dy = 1 - e^{-\gamma(1-\theta)t}.$$

The average number T of requests processed by the system at unit of time (throughput) is computed by

$$T = \sum_{i=1}^{\infty} \sum_{k=0}^{K} \sum_{\nu=0}^{W} \sum_{\eta=1}^{M} \pi(i,k,\nu,\eta)(\mathbf{S}_0)_\eta = \boldsymbol{\pi}_0 R (I - R)^{-1} (\mathbf{e}_{(K+1)(W+1)} \otimes \mathbf{S}_0).$$

Remark 1. In contrast to the model with a finite buffer, see (Lee et al., 2007) and (Kim et al., 2009), where the arriving session can be rejected not only due to the tokens absence but also due to the buffer overloading, distribution of the number of sessions in the model under study does not depend on the number of requests in the system. It is defined by formula

$$\lim_{t\to\infty} P\{k_t = k\} = \mathbf{x}_k \mathbf{e},\ k = \overline{0,K},$$

where the vectors $\mathbf{x}_k,\ k = \overline{0,K},$ are the entries of the vector $\mathbf{x} = (\mathbf{x}_0, \ldots, \mathbf{x}_K)$ which satisfies the system (5). However, distribution $\boldsymbol{\pi}(i,k), i \geq 0, k = \overline{0,K},$ does not have multiplicative form because the number of requests in the system depends on the number of sessions currently presenting in the system.

Remark 2. It can be verified that the considered model with the infinite buffer has the steady state distribution of the process $\xi_t = \{i_t, k_t, \nu_t, \eta_t\}, t \geq 0,$ coinciding with the steady state distribution of the queueing model of the $MAP/PH/1$ type with the phase service time distribution having irreducible representation $(\boldsymbol{\beta}, S)$ and the MAP arrival process defined by the matrices $\tilde{D}_0$ and $\tilde{D}_1$ having the form

$$\tilde{D}_0 = \begin{pmatrix} D_0 & O & O & \dots & O & O \\ \gamma^- I & D_0 - \gamma I & O & \dots & O & O \\ 0 & 2\gamma^- I & D_0 - 2\gamma I & \dots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \dots & K\gamma^- I & D(1) - K\gamma I \end{pmatrix}, \tilde{D}_1 = \begin{pmatrix} O & D_1 & O & \dots & O & O \\ O & \gamma^+ I & D_1 & \dots & O & O \\ 0 & O & 2\gamma^+ I & \dots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \dots & O & K\gamma^+ I \end{pmatrix}.$$

It is easy to verify that the fundamental rate of this MAP is equal to Λ which is defined in (1). So, stability condition (1) is intuitively clear: the average service rate should exceed the average arrival rate. Note that the first summand in expression $\Lambda = \gamma^+ \sum_{k=0}^{K} k\mathbf{x}_k \mathbf{e}_{W+1} + \sum_{k=0}^{K-1} \mathbf{x}_k D_1 \mathbf{e}_{W+1}$, for the rate Λ represents the rate of requests from already accepted sessions, i.e., the rate of requests who are not the first in a session. The second summand is the rate of the sessions arrival.

Theorem 2. The probability $P_b^{(loss)}$ of an arbitrary session rejection upon arrival is computed by

$$P_b^{(loss)} = \sum_{i=0}^{\infty} \boldsymbol{\pi}(i,K) \frac{(D_1 \otimes I_M)}{\lambda} \mathbf{e} = \mathbf{x}_K \frac{D_1}{\lambda} \mathbf{e}.$$

The probability $P_c^{(loss)}$ of an arbitrary request rejection upon arrival is computed by

$$P_c^{(loss)} = \sum_{i=0}^{\infty} \boldsymbol{\pi}(i,K) \frac{(D_1 \otimes I_M)}{\tilde{\Lambda}} \mathbf{e} = \mathbf{x}_K \frac{D_1}{\tilde{\Lambda}} \mathbf{e}$$

where $\tilde{\Lambda} = \Lambda + \mathbf{x}_K D_1 \mathbf{e}$.

Proof of formula for probability $P_b^{(loss)}$ accounts that the session is rejected upon arrival if and only if the number of sessions in the system at this epoch is equal to K. So

$$P_b^{(loss)} = \frac{\sum_{i=0}^{\infty} \boldsymbol{\pi}(i,K)(D_1 \otimes I_M)\mathbf{e}}{\sum_{i=0}^{\infty} \sum_{k=0}^{K} \boldsymbol{\pi}(i,k)(D_1 \otimes I_M)\mathbf{e}} = \sum_{i=0}^{\infty} \boldsymbol{\pi}(i,K) \frac{(D_1 \otimes I_M)}{\lambda} \mathbf{e}.$$

Rejection of a request can occur only if this request is the first in a session and the number of sessions in the system at this session arrival epoch is equal to K. So

$$P_c^{(loss)} = \frac{\sum_{i=0}^{\infty} \boldsymbol{\pi}(i,K)(D_1 \otimes I_M)\mathbf{e}}{\sum_{i=0}^{\infty} \sum_{k=0}^{K} \boldsymbol{\pi}(i,k)(D_1 + k\gamma^+ I) \otimes I_M \mathbf{e}} = \sum_{i=0}^{\infty} \boldsymbol{\pi}(i,K) \frac{D_1 \otimes I_M}{\tilde{\Lambda}} \mathbf{e}.$$

4. Distribution of the sojourn times

Let $V_b(x)$, $V_c(x)$ and $V_c^{(a)}(x)$ be distribution functions of the sojourn time of an arbitrary session, an arbitrary request, which is the first in a session, and an arbitrary request from the admitted session, which is not the first in this session, in the system under study, and $v_b(s)$, $v_c(s)$ and $v_c^{(a)}(s)$, $Re\, s > 0$, be their Laplace-Stieltjes transforms (*LST*s):

$$v_b(s) = \int_0^\infty e^{-sx} dV_b(x),\ v_c(s) = \int_0^\infty e^{-sx} dV_c(x),\ v_c^{(a)}(s) = \int_0^\infty e^{-sx} dV_c^{(a)}(x).$$

Formulae for calculation of the *LSTs* $v_c(s)$ and $v_c^{(a)}(s)$ are the following:

$$v_c(s) = \frac{1}{\lambda}[\boldsymbol{\pi}_0(D_1 \otimes I_M)\mathbf{e}\boldsymbol{\beta}(sI - S)^{-1}\mathbf{S}_0 + \sum_{i=1}^{\infty}\boldsymbol{\pi}_i(D_1 \otimes I_M)(\mathbf{e}_{(K+1)(W+1)} \otimes ((sI - S)^{-1}\mathbf{S}_0))(\boldsymbol{\beta}(sI - S)^{-1}\mathbf{S}_0)^i] =$$

$$= \frac{1}{\lambda}[\boldsymbol{\pi}_0(D_1 \otimes I_M)\mathbf{e}\boldsymbol{\beta}(sI - S)^{-1}\mathbf{S}_0 + \boldsymbol{\pi}_0 R(\boldsymbol{\beta}(sI - S)^{-1}\mathbf{S}_0)\times$$

$$\times(I - R(\boldsymbol{\beta}(sI - S)^{-1}\mathbf{S}_0))^{-1}(D_1 \otimes I_M)(\mathbf{e}_{(K+1)(W+1)} \otimes ((sI - S)^{-1}\mathbf{S}_0))],$$

$$v_c^{(a)}(s) = \frac{1}{\sum_{k=1}^{K}\sum_{i=0}^{\infty} k\gamma^+\boldsymbol{\pi}(i,k)\mathbf{e}}\sum_{k=1}^{K} k\gamma^+[\boldsymbol{\pi}(0,k)\mathbf{e}\boldsymbol{\beta}(sI - S)^{-1}\mathbf{S}_0 +$$

$$+\sum_{i=1}^{\infty}\boldsymbol{\pi}(i,k)(\mathbf{e}_{W+1} \otimes ((sI - S)^{-1}\mathbf{S}_0))(\boldsymbol{\beta}(sI - S)^{-1}\mathbf{S}_0)^i].$$

Formulae for the average sojourn time V_c of an arbitrary request, which is the first in a session, the average sojourn time V_c^* of an arbitrary non-rejected request, which is the first in a session, and the average sojourn time $V_c^{(a)}$ of an arbitrary request from the admitted session, which is not the first in this session, are as follows:

$$V_c = \frac{1}{\lambda}[\boldsymbol{\pi}_0(D_1 \otimes I_M)\mathbf{e}b_1 + \sum_{i=1}^{\infty}\boldsymbol{\pi}_i(D_1 \otimes I_M)((\mathbf{e}_{(K+1)(W+1)} \otimes (-S)^{-1}\mathbf{e}) + \mathbf{e}ib_1)] =$$

$$= \frac{\boldsymbol{\pi}_0}{\lambda}[(I + R(I - R)^{-2})(D_1 \otimes I_M)\mathbf{e}b_1 + R(I - R)^{-1}(D_1 \otimes I_M)(\mathbf{e}_{(K+1)(W+1)} \otimes (-S)^{-1}\mathbf{e}],$$

$$V_c^* = \frac{V_c}{1 - P_b^{(loss)}},$$

$$V_c^{(a)} = \frac{\sum_{k=1}^{K} k\gamma^+[\boldsymbol{\pi}(0,k)\mathbf{e}b_1 + \sum_{i=1}^{\infty}\boldsymbol{\pi}(i,k)((\mathbf{e}_{W+1} \otimes (-S)^{-1}\mathbf{e}) + \mathbf{e}ib_1)]}{\sum_{k=1}^{K}\sum_{i=0}^{\infty} k\gamma^+\boldsymbol{\pi}(i,k)\mathbf{e}}.$$

If the service time distribution is exponential, expression for the average sojourn time V_c of an arbitrary arriving request, which is the first in a session, becomes simpler:

$$V_c = \frac{\boldsymbol{\pi}_0}{\lambda} b_1 (I - R)^{-2} D_1 \mathbf{e}.$$

Derivation of formula for calculation of the *LST* $v_b(s)$ is more involved. Recall that the sojourn time of an arbitrary session in the system lasts since the epoch of the session arrival into the system until the moment when the arrival of a session is finished and all requests, which belong to this session, leave the system. We will derive expression for the *LST* $v_b(s)$ by means of the method of collective marks (method of additional event, method of catastrophes), for references see, e.g., (Kasten & Runnenburg, 1956) and (Danzig, 1955). To this end, we interpret the variable s as the intensity of some virtual stationary Poisson flow of catastrophes. So, $v_b(s)$ has meaning of probability that no one catastrophe arrives during the sojourn time of an arbitrary session.

We will tag an arbitrary session and will keep track of its staying in the system. Let $v(s,i,l,k,\nu,\eta)$ be the probability that catastrophe will not arrive during the rest of the tagged session sojourn time in the system conditional that, at the given moment, the number of sessions processed in the system is equal to $k, k = \overline{1,K},$ the number of requests is equal to $i, i \geq 0,$ the last (in the order of arrival) request of a tagged session has position number $l, l = \overline{0,i},$ in the system, and the states of the processes $\nu_t, \eta_t, t \geq 0,$ are ν, η . Position number 0 means that currently there is no one request of the tagged session in the system.

It follows from the formula of total probability that if we will have functions $v(s,i,l,k,\nu,\eta)$ been calculated the Laplace-Stieltjes transform $v_b(s)$ can be computed by

$$v_b(s) = P_b^{(loss)} + \frac{1}{\lambda} \sum_{i=0}^{\infty} \sum_{k=0}^{K-1} \sum_{\nu=0}^{W} \sum_{\eta=1}^{M} \sum_{\nu'=0}^{W} \pi(i,k,\nu,\eta) \lambda_\nu p_{\nu,\nu'}^{(1)} \times v(s,i+1,i+1,k+1,\nu',\eta). \tag{6}$$

The system of linear algebraic equations for functions $v(s,i,l,k,\nu,\eta)$ is derived by means of formula of total probability in the following form:

$$\begin{aligned} v(s,i,l,k,\nu,\eta) = [\lambda_\nu \sum_{\nu'=0}^{W} p_{\nu,\nu'}^{(1)} ((1-\delta_{k,K}) v(s,i+1,l,k+1,\nu',\eta) + \\ + \delta_{k,K} v(s,i,l,k,\nu',\eta)) + \lambda_\nu \sum_{\nu'=0}^{W} p_{\nu,\nu'}^{(0)} v(s,i,l,k,\nu',\eta) + \\ + (1-\delta_{i,0}) (\mathbf{S}_0)_\eta \sum_{\eta'=1}^{M} \beta_{\eta'} [v(s,i-1,l-1,k,\nu,\eta')(1-\delta_{l,0}) + v(s,i-1,0,k,\nu,\eta')\delta_{l,0}] + \\ + (1-\delta_{i,0}) \sum_{\eta'=1}^{M} (S)_{\eta,\eta'} v(s,i,l,k,\nu,\eta') + \gamma^+ v(s,i+1,i+1,k,\nu,\eta) + \\ + \gamma^+ (k-1) v(s,i+1,l,k,\nu,\eta) + \gamma^- (k-1) v(s,i,l,k-1,\nu,\eta) + \end{aligned} \tag{7}$$

$$+\gamma^{-}[((sI-S)^{-1}\mathbf{S}_0)_{\eta}(\boldsymbol{\beta}(sI-S)^{-1}\mathbf{S}_0)^{l-1}(1-\delta_{l,0})+\delta_{l,0}]]\times$$
$$\times(s+\lambda_{\nu}-S_{\eta,\eta}+k\gamma)^{-1}, l=\overline{0,i},\ i\geq 0,\ k=\overline{1,K},\ \nu=\overline{0,W},\ \eta=\overline{1,M}.$$

Let us explain formula (7) in brief. The denominator of the right hand side of (7) is equal to the total intensity of the events which can happen after the arbitrary time moment: catastrophe arrival, transition of the directing process of the MAP, transition of the directing process of the PH service process, and expiring the time till the moment of possible request arrival from sessions already admitted into the system. The first term in the square brackets in (7) corresponds to the case when a new session arrives. The second term corresponds to the case when transition of the directing process of the MAP occurs without new session generation. The third term corresponds to the case when service completion takes place. The fourth term corresponds to the case when the transition of the directing process of the PH service process occurs without the service completion. The fifth term corresponds to the case when the new request of the tagged session arrives into the system. In this case, the position of the last request of the tagged session in the system is reinstalled from l to $i+1$. The sixth term corresponds to the case when the new request from another session, which was already admitted to the system, arrives. The seventh term corresponds to the case when some non-tagged session terminates arrivals. The eighth term corresponds to the case when the expected new request of the tagged session does not arrive into the system and arrival of requests of the tagged session is stopped. This session will not more counted as arriving into the system and the tagged request finishes its sojourn time when the last request, who is currently the lth in the system, will leave the system. Number $((sI-S)^{-1}\mathbf{S}_0)_{\eta}$ defines the probability that catastrophe will not arrive during the residual service time conditional that the directing process of the PH service is currently in the state η. The number $\boldsymbol{\beta}(sI-S)^{-1}\mathbf{S}_0$ defines probability that catastrophe will not arrive during the service time of an arbitrary request.

Let us introduce column vectors

$$\mathbf{v}(s,i,l,k,\nu)=(v(s,i,l,k,\nu,1),\ldots,v(s,i,l,k,\nu,M))^T,$$
$$\mathbf{v}(s,i,l,k)=(\mathbf{v}(s,i,l,k,0),\ldots,\mathbf{v}(s,i,l,k,W))^T,$$
$$\mathbf{v}(s,i,l)=(\mathbf{v}(s,i,l,1),\ldots,\mathbf{v}(s,i,l,K))^T,$$
$$\mathbf{v}(s,i)=(\mathbf{v}(s,i,0),\ldots,\mathbf{v}(s,i,i))^T,\ \mathbf{v}(s)=(\mathbf{v}(s,0),\mathbf{v}(s,1),\ldots)^T.$$

System (7) of linear algebraic equations can be rewritten to the matrix form as

$$-(sI-\hat{Q}_{i,i})\mathbf{v}(s,i,l)+\hat{Q}_{i,i+1}\mathbf{v}(s,i+1,l)+\hat{Q}_{i,i-1}\mathbf{v}(s,i-1,l-1)(1-\delta_{0,l})+\hat{Q}_{i,i-1}v(s,i-1,0)\delta_{0,l}+ \quad (8)$$
$$+I_K\otimes\Gamma^{+}\mathbf{v}(s,i+1,i+1)+\gamma^{-}\mathbf{e}_{K(W+1)}\otimes((sI-S)^{-1}\mathbf{S}_0)(\boldsymbol{\beta}(sI-S)^{-1}\mathbf{S}_0)^{l-1}=\mathbf{0}^T_{KM_1},\ l=\overline{0,i},\ i\geq 0,$$

where

$$\hat{Q}_{i,i}=A_1+I_K\otimes(D_0\oplus S)(1-\delta_{i,0})+\tilde{E}\otimes((D_1\otimes I_M))+I_K\otimes(D_0\otimes I_M)\delta_{i,0}, i\geq 0,$$

$$\hat{Q}_{i,i+1} = \gamma^{+} C_{K-1} + E_K^{+} \otimes (D_1 \otimes I_M), i \geq 0,$$
$$\hat{Q}_{i,i-1} = I_{K(W+1)} \otimes \mathbf{S}_0 \boldsymbol{\beta}, i \geq 0, \hat{Q}_{0,-1} = O.$$

Let us introduce notation:

$\Omega(s)$ is three block diagonal matrix with non-zero blocks

$$\Omega_{i,j}(s),\ j = \max\{0, i-1\}, i, i+1, i \geq 0,$$

defined by

$$\Omega_{i,i}(s) = -I_{i+1} \otimes (sI - \hat{Q}_{i,i}), \Omega_{i,i-1} = D_3^{(i)} \otimes \hat{Q}_{i,i-1},$$
$$\Omega_{i,i+1} = D_1^{(i)} \otimes \hat{Q}_{i,i+1} + D_2^{(i)} \otimes I_K \otimes \Gamma^{+}.$$

Here the matrix $D_1^{(i)}$ of size $(i+1)\times(i+2)$ is obtained from the identity matrix I_{i+1} by means of supplementing from the right by the column $\mathbf{0}_{i+1}^T$. The matrix $D_2^{(i)}$ of the same size has the last column consisting of 1's and other columns consisting of 0's. The matrix $D_3^{(i)}$ of size $(i+1)\times i$ is obtained from the identity matrix I_i by means of supplementing from above by the row $(1,0,\ldots,0)$.

Vector $\boldsymbol{B}(s)$ is defined by

$$\boldsymbol{B}(s) = (\boldsymbol{B}_0(s), \ldots, \boldsymbol{B}_N(s), \ldots)^T$$

where

$$\boldsymbol{B}_i(s) = \gamma^{-}(\mathbf{e}_K \otimes \mathbf{e}_{M_1}, \mathbf{e}_K \otimes \mathbf{e}_{W+1} \otimes (sI - S)^{-1}\mathbf{S}_0, \mathbf{e}_K \otimes \mathbf{e}_{W+1} \otimes ((sI - S)^{-1}\mathbf{S}_0)\boldsymbol{\beta}(sI - S)^{-1}\mathbf{S}_0, \ldots,$$
$$\mathbf{e}_K \otimes \mathbf{e}_{W+1} \otimes ((sI - S)^{-1}\mathbf{S}_0)(\boldsymbol{\beta}(sI - S)^{-1}\mathbf{S}_0)^{i-1})^T, i \geq 0.$$

Using this notation we can rewrite the system (7) to the form

$$\Omega(s)\mathbf{v}(s) + \boldsymbol{B}(s) = \mathbf{0}^T. \quad (9)$$

It can be verified that the diagonal entries of the matrix $\Omega(s)$ dominate in all rows of this matrix. So the inverse matrix exists. Thus we proved the following assertion.

Theorem 3. The vector $\mathbf{v}(s)$ consisting of conditional Laplace-Stieltjes transforms LST $v(s,i,l,k,\nu,\eta), l = \overline{0,i}, i \geq 0, k = \overline{1,K}, \quad \nu = \overline{0,W}, \eta = \overline{1,M},$ is calculated by

$$\mathbf{v}(s) = -\Omega^{-1}(s)\boldsymbol{B}(s). \quad (10)$$

Corollary 2. The average sojourn time V_b of an arbitrary session is calculated by

$$V_b = -\sum_{i=0}^{\infty}\sum_{k=0}^{K-1} \boldsymbol{\pi}(i,k)\frac{(D_1 \otimes I_M)}{\lambda}\frac{\partial \mathbf{v}(s, i+1, i+1, k+1)}{\partial s}|_{s=0},$$

where column vectors $\frac{\partial \mathbf{v}(s,i+1,i+1,k+1)}{\partial s}\Big|_{s=0}$ are calculated as the blocks of the vector $\frac{d\mathbf{v}(s)}{ds}\Big|_{s=0}$ defined by

$$\frac{d\mathbf{v}(s)}{ds}|_{s=0} = \Omega^{-1}(0)(-\frac{d\boldsymbol{B}(s)}{ds}|_{s=0} + \mathbf{v}(0)),$$

where $\mathbf{v}(0) = -\gamma^{-}\Omega^{-1}(0)\mathbf{e}$.

Corollary 3. The average sojourn time $V_b^{(accept)}$ of an arbitrary admitted session is calculated by

$$V_b^{(accept)} = \frac{V_b}{1 - P_b^{(loss)}},$$

where $P_b^{(loss)}$ is probability of an arbitrary session rejection upon arrival.

5. Optimization problem and numerical examples

It is obvious that the most important from economical point of view characteristic of the considered model is the throughput T of the system because it defines the profit earned by information transmission. If the number K that restricts the number of sessions, which can be served in the system simultaneously, increases the throughput T of the system increases and the probability $P_b^{(loss)}$ of an arbitrary session rejection upon arrival decreases. So, it seems to be reasonable to increase the number K as much as possible until stability condition (1) is violated. However, such performance measures as the average sojourn time of an arbitrary request and jitter are also very important because they should fit requirements of Quality of Service. These performance measures become worse if the number K grows. Evidently, it does not make sense to admit too many sessions into the system simultaneously and provide bad Quality of Service (average sojourn time and jitter) for them. So, the system manager should decide how many sessions can be allowed to enter the system simultaneously to fit requirements of Quality of Service and to reach the maximally possible throughput.
Thus, one should solve, e.g., the following non-trivial optimization problem:

$$T = T(K) \to \max \tag{11}$$

subject to constraints (1) and

$$V_c^* \le V, \tag{12}$$

where V is the maximal admissible value of the sojourn time of the first request from non-rejected session and is assumed to be fixed in advance.
This optimization problem can be easy solved by means of computer, based on presented above expressions for the main performance measures of the system, by means of enumeration, i.e., increasing the value K until constraints (1) and (12) are violated. The

optimal value of K in the optimization problem (1), (11), (12) will be denoted by K^*. Corresponding computer program allows to validate the feasibility of such an optimization algorithm and to illustrate the dependencies of the system characteristics on the system parameters and the value of K. In what follows several illustrative examples are presented.
Before to start description of these examples, let us mention that numerous experiments show that the famous Little's formula holds good for the system under study in the form $\lambda L = V_c$, where L is the average number of requests in the system and V_c is the average sojourn time of an arbitrary request which is the first in a session.

5.1. Dependence of probabilities P_b^{loss} of an arbitrary session loss and P_c^{loss} of an arbitrary request loss on the number K of tokens and correlation in the sessions arrival process

The experiment has two goals. One is to illustrate quantitatively the dependence of probabilities P_b^{loss} of an arbitrary session loss and P_c^{loss} of an arbitrary request loss on the number K of tokens. The second goal is to show that for several different arrival processes having the same average rate but different correlation this dependence is quite different. This explains the importance of consideration of the model with the MAP arrival process of sessions, which can be essentially correlated in real telecommunication networks, instead of analysis of simpler model with the stationary Poisson arrival process of sessions.
We consider six different *MAP*s having the same fundamental rate $\lambda = 1$. The first *MAP* is the stationary Poisson arrival process. Variation coefficient of inter-arrival times is equal to 1. Four other *MAP*s have the variation coefficient equal to 2 but different coefficients of correlation of successive intervals between sessions arrival. These four *MAP*s are described as follows.

- MAP ($IPP - Interrupted\ Poisson\ Process$) flow with correlation coefficient equal to 0 is defined by the matrices

$$D_0 = \begin{pmatrix} -0.4 & 0.16 & 0.24 \\ 1.3 & -69.4 & 68.1 \\ 1.3 & 1.3 & -270 \end{pmatrix}; D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 100.2 & 167.2 & 0 \end{pmatrix}.$$

- *MAP* flow with correlation coefficient equal to 0.1 is defined by the matrices

$$D_0 = \begin{pmatrix} -2.66 & 0.12 & 0.12 \\ 0.13 & -0.5 & 0.08 \\ 0.14 & 0.08 & -0.32 \end{pmatrix}; D_1 = \begin{pmatrix} 2.3 & 0.08 & 0.04 \\ 0.09 & 0.18 & 0.02 \\ 0.5 & 0.01 & 0.04 \end{pmatrix}.$$

- *MAP* flow with correlation coefficient equal to 0.2 is defined by the matrices

$$D_0 = \begin{pmatrix} -3.16 & 0.12 & 0.12 \\ 0.1 & -0.45 & 0.09 \\ 0.12 & 0.11 & -0.39 \end{pmatrix}; D_1 = \begin{pmatrix} 2.84 & 0.06 & 0.02 \\ 0.02 & 0.21 & 0.03 \\ 0.02 & 0.04 & 0.1 \end{pmatrix}.$$

- *MAP* flow with correlation coefficient equal to 0.3 is defined by the matrices

$$D_0 = \begin{pmatrix} -5.11 & 0.08 & 0.07 \\ 0.029 & -0.446 & 0.04 \\ 0.06 & 0.08 & -0.35 \end{pmatrix}; D_1 = \begin{pmatrix} 4.85 & 0.09 & 0.02 \\ 0.007 & 0.333 & 0.037 \\ 0 & 0.05 & 0.16 \end{pmatrix}.$$

- The sixth *MAP* has correlation coefficient equal to -0.16 and the squared correlation coefficient equal to 1.896. It is defined by the matrices

$$D_0 = \begin{pmatrix} -3.607 & 0 \\ 0 & -0.617 \end{pmatrix}; D_1 = \begin{pmatrix} 0.347 & 3.26 \\ 0.478 & 0.139 \end{pmatrix}.$$

The service time distribution is Erlangian of order 2 with intensity of the phase equal to 16. The rest of the parameters are the following: $\gamma = 2,\ \theta = 0.9$.

Figures 1 and 2 illustrate the dependencies of probability P_b^{loss} of an arbitrary session loss and P_c^{loss} of an arbitrary request loss on the number K of tokens for the listed above different *MAP* s with the same fundamental rate but the different correlation.

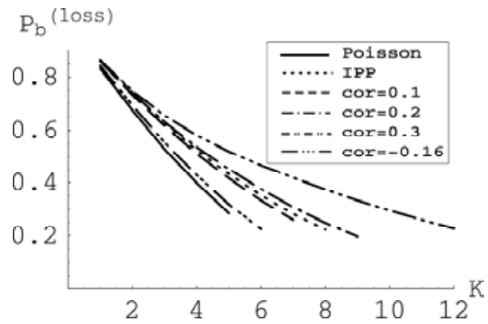


Fig. 1. Dependence of probability P_b^{loss} of arbitrary session loss on the number of tokens K

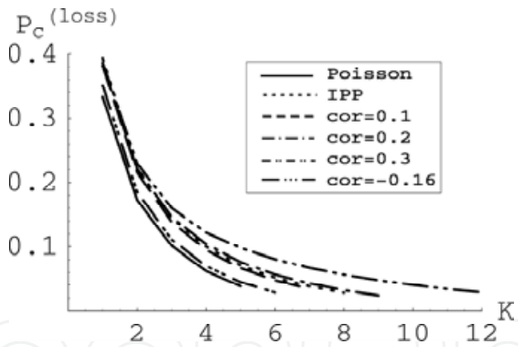


Fig. 2. Dependence of probability of an arbitrary request loss on the number of tokens K

One can pay attention that the curves corresponding to the different *MAP* s terminate at the different points, e.g., the curve corresponding to the stationary Poisson process terminates at the point $K = 5$, the curve corresponding to the *MAP* s having correlation coefficient 0.3 terminates at the point $K = 12$. The reason of termination is that the stationary distribution existence condition violates for K larger than 5 and 12 correspondingly.

It is worth to mention, that the previous analysis of different queues with the Batch Markovian Arrival Process given in many papers shows that usually the stability condition depends on the average arrival rate, but does not depend on correlation. It the model under study, stability condition (1) depends on correlation as well. This has the clear explanation: stability condition includes the stationary distribution of the corresponding *MAP*/*M*/*K*/0

queueing system which describes the behavior of the number of busy tokens. As it is illustrated in (Klimenok et al., 2005), this distribution essentially depends on the correlation in the arrival process.

Conclusion that can be made based on these numerical results is the following: higher correlation of the session's arrival process implies higher value of P_b^{loss} and P_c^{loss} but larger number of sessions which can be simultaneously processed in the system without overloading the system. IPP process violates this rule a bit. This is well known very special kind of arrival process. It has zero correlation. Intervals where arrivals occur more or less intensively alternate with time periods when no arrivals are possible. Such irregular arrivals make the IPP violating the conclusion made above. Note that the system with the negative correlation in the arrival process has characteristics close to characteristics of the system with the stationary Poisson process. While the more or less strong positive correlation changes these characteristics essentially.

5.2. Dependence of the throughput of the system on the number of tokens and correlation in the sessions arrival process

Let us consider the same system as in the previous experiment and consider optimization problem (11), (12) where the limiting value of the average sojourn time for the first request in non-rejected session is assumed to be $V = 40$. Figure 3 illustrates the dependence of the throughput T of the system on the number of tokens K. As it is expected, the throughput T is the increasing function of K for all arrival processes. However, the shape of this function depends on the correlation in the sessions arrival process. The lines corresponding to the different MAP s terminate when condition (12) is not hold true. So, as it is seen from Figure 3 the optimal value K^* of tokens is equal to 5 when the arrival process is the stationary Poisson or has the negative correlation or is equal to 0.1 and is equal to 6 for the rest of the arrival processes.

It is seen from Figures 1-3 that positive correlation has the negative impact on the system performance. Although the number of simultaneously processed requests can be larger, loss probability is higher and the throughput of the system is lesser.

Dependence of the average sojourn time V_c^* for the first request in non-rejected sessions on the number of tokens in these examples is presented on Figure 4.

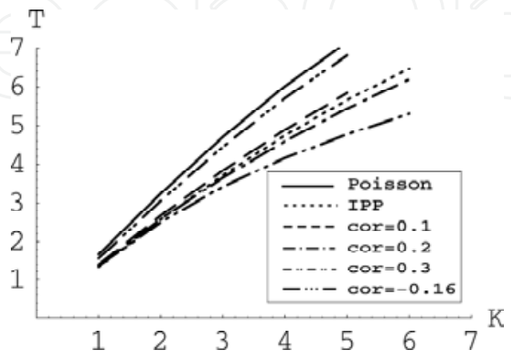


Fig. 3. Dependence of the throughput T of the system on the number of tokens under restriction $V_c^* < 40$

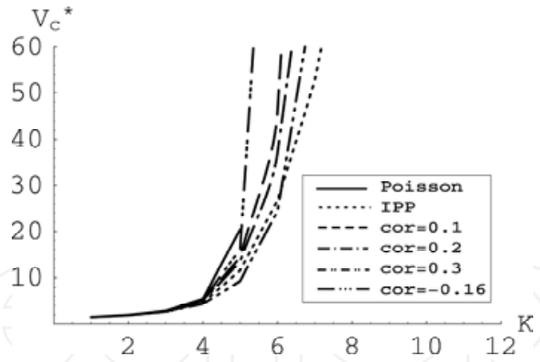


Fig. 4. Dependence of V_c^* on K

It is seen that the average sojourn time V_c^* sharply increases when the number of tokens K approaches the value $K=5$ or $K=6$, depending on correlation in the arrival process. For the model with the stationary Poisson arrival process stationary distribution does not exist for $K=6$.

5.3. Dependence of the optimal number of tokens on the session size, arrival and service rates

The goal of this experiment is to illustrate the dependence of the optimal number of tokens on the session size, average arrival and average service rates.

Firstly, let as clarify the impact of the session size. We assume that the MAP process of sessions is defined by the matrices

$$D_0 = \begin{pmatrix} -6.88 & 0.0008 \\ 0.0008 & -0.22 \end{pmatrix}; D_1 = \begin{pmatrix} 6.8 & 0.0792 \\ 0.016 & 0.2032 \end{pmatrix}.$$

This MAP has the average rate equal to 1.37, correlation coefficient 0.4 and the squared variation coefficient 9.4. As in the previous examples, the service time distribution is assumed to be Erlangian of order 2 with the intensity of the phase equal to 16.

On Figure 5, we vary the parameter θ, which characterizes the distribution of the number of requests in a session, in the interval $[0.1;0.8]$. This implies that the average session size varies in the interval $[1.111;5]$. Parameter V defining the limiting value of the average sojourn time for the first request in non-rejected sessions is assumed to be equal to 0.8.

On Figure 6, we vary the parameter θ in the interval $[0.8;0.98]$. This implies that the average session size varies in the interval $[5;50]$. Parameter V is assumed to be equal to 8.

As it is expected, the optimal number K^* is non-increasing function of θ. When θ increases from 0.1 to 0.8 the number K^* decreases from 8 to 1 under restriction $V_c^* < 0.8$. If we take θ greater than 0.8, restriction $V_c^* < 0.8$ is not fulfilled even only 1 session is allowed to enter the system. If the weaken this restriction to the restriction $V_c^* < 8$, four sessions can be processed in the system simultaneously for θ equal to 0.8. Situation when restriction $V_c^* < 8$ is not fulfilled even only 1 session is allowed to enter the system occurs for θ greater than 0.98.

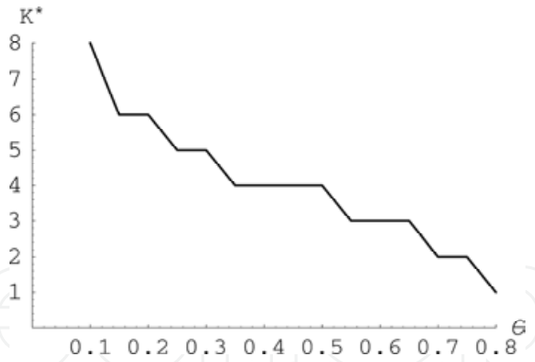


Fig. 5. Dependence of the optimal number K^* of tokens on the parameter θ under restriction $V_c^* < 0.8$

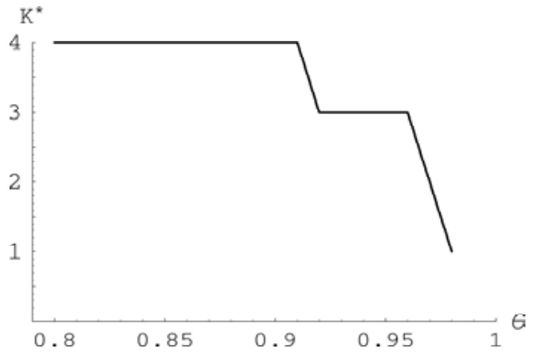


Fig. 6. Dependence of the optimal number K^* of tokens on the parameter θ under restriction $V_c^* < 8$

In the next example, we illustrate the impact of the average arrival rate. We consider the IPP process defined above and vary the average arrival rate in the interval $[1;11]$ by means of multiplication of the matrices D_0 and D_1 by the corresponding factor. The service time distribution is assumed to be Erlangian of order 2 with the intensity of the phase equal to 30. Parameter V defining the limiting value of the average sojourn time is assumed to be equal to 4. Figure 7 shows the dependence of the optimal number K^* of tokens on the average arrival rate λ.

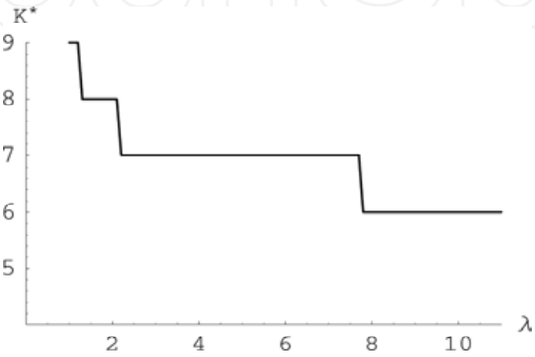


Fig. 7. Dependence of the optimal number K^* of tokens on the average arrival rate λ

As it is expectable, the optimal number K^* of tokens decreases when λ is increasing. The same dependence takes place for other MAP s, only the points of the jumps of the lines are different.

In the last example, we illustrate the impact of the average service rate. We consider the IPP process defined above having the average arrival rate $\lambda = 1$. The service time distribution is assumed to be Erlangian of order 2 with intensity of the phase varied to get the average service rate in the interval $[3.5;20]$. Parameter V defining the limiting value of the average sojourn time is assumed to be equal to 5.

Figure 8 shows the dependence of the optimal number K^* of tokens on the average service rate μ.

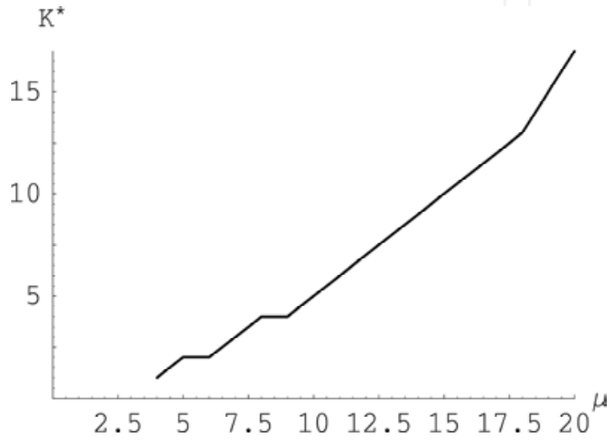


Fig. 8. Dependence of the optimal number K^* of tokens on the average service rate μ

As it is expectable, the optimal number K^* of tokens increases when μ is increasing. The same dependence takes place for other MAP s, only again the points of the jumps of the lines are different.

6. Conclusion

In this paper, the novel infinite buffer queueing model with session arrivals distributed in time is analyzed. Ergodicity condition is derived. Joint distribution of the number of requests in the system and number of currently admitted sessions is computed. The sojourn time distribution of an arbitrary request and arbitrary session is given in terms of the Laplace-Stieltjes Transform. Usefulness of the presented results is illustrated numerically. Validity of Little's formulas is checked by means of numerical experiment.

Results are planned to be extended to the systems with many servers, non-geometrical session size distribution, and heterogeneous arrival flow.

Acknowledgement

This work was supported by World Class Univ. R32-2008-000-20014-0 NRF and KRF-2007-521-D00330, Korea.

7. References

Asmussen, S.; Nerman, O. & Olsson, M. (1996) Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, Vol. 23, pp. 419-441.

Chakravarthy, S.R. (2001) The session Markovian arrival process: a review and future work, in: A. Krishnamoorthy, et al. (Eds.). *Advances in Probability Theory and Stochastic Process: Proc., Notable Publications*, pp. 21-49.

Danzig, D. (1955) Chaines de Markof dans les ensembles abstraits et applications aux processus avec regions absorbantes et au probleme des boucles. *Ann. de l'Inst. H. Pioncare,* Vol. 14, pp. 145-199.

Fisher, W. & Meier-Hellstern, K.S. (1993) The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation,* Vol. 18, pp. 149-171.

Heyman, D.P. & Lucantoni, D. (2003) Modelling multiple IP traffic streams with rate limits, *IEEE/ACM Transactions on Networking,* Vol. 11, pp. 948-958.

Kasten, H. & Runnenburg, J. Th. (1956) Priority in waiting line problems, *Mathematisch Centrum*, Amsterdam, Holland.

Klemm, A.; Lindermann, C.; & Lohmann, M. (2003) Modelling IP traffic using the session Markovian arrival process. *Performance Evaluation*, Vol. 54, pp. 149-173.

Klimenok, V.; Kim, C.S.; Orlovsky, D. & Dudin, A. (2005) Lack of invariant property of Erlang loss model in case of the MAP input. *Queueing Systems,* Vol. 49, pp. 187-213.

Kim, C.S.; Dudin, S. & Klimenok, V. (2009) The *MAP*/*PH*/1/*N* queue with time phased arrivals as model for traffic control in telecommunication networks. *Performance Evaluation*, Vol. 66, pp. 564-579.

Kist, A.A.; Lloyd-Smith, B. & Harris, R.J. (2005) A simple IP flow blocking model. Performance Challenges for Efficient Next Generation Networks. *Proceedings of 19-th International Teletraffic Congress,* 29 August – 2 September 2005, Beijing, pp. 355-364.

Lee, M.H.; Dudin, S. & Klimenok, V. (2007) Queueing Model with Time-Phased Session Arrivals. *In: Managing Traffic Performance in Converged Networks Proceedings 20th International Teletraffic Congress,* Ottawa, Canada, June 2007. Berlin: Springer. *Lecture Notes in Computer Science,* Vol. 4516. pp. 716-730.

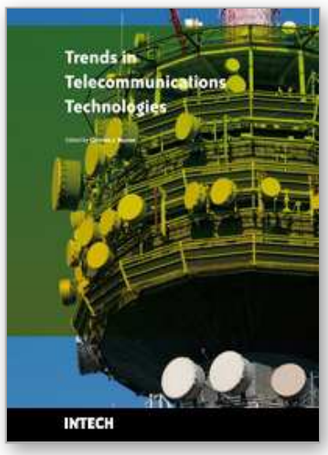
Lucantoni, D.M. (1991) New results on the single server queue with a session Markovian arrival process. *Communications in Statistics-Stochastic Models,* Vol. 7, pp. 1-46.

Neuts M.F., (1981) *Matrix-geometric solutions in stochastic models*. Baltimore, The Johns Hopkins University Press.

Panchenko, A. & Buchholz, P. (2007) A hybrid algorithm for parameter fitting of Markovian Arrival Process, in: Al-Begain K, Heindl A, Telek M. (Eds.). *14th Int. Conf.* "*Analytical and stochastic modelling technique and applications*". Prague, 2007, pp. 7-12.

Pattavina, A. & Parini, A., (2005) Modelling voice call inter-arrival and holding time distributions in mobile networks, in: Performance Challenges for Efficient Next Generation Networks - *Proc.of 19th International Teletraffic Congress*, Aug.-Sept. 2005, pp. 729-738.

Riska, A.; Diev, V. & Smirni, E. (2002) Efficient fitting of long-tailed data sets into hyperexponential distributions, *Global Telecommunications Conference (GLOBALCOM'02, IEEE),* 7-21 Nov. 2002, pp. 2513-2517.



Trends in Telecommunications Technologies

Edited by Christos J Bouras

ISBN 978-953-307-072-8

Hard cover, 768 pages

Publisher InTech

Published online 01, March, 2010

Published in print edition March, 2010

The main focus of the book is the advances in telecommunications modeling, policy, and technology. In particular, several chapters of the book deal with low-level network layers and present issues in optical communication technology and optical networks, including the deployment of optical hardware devices and the design of optical network architecture. Wireless networking is also covered, with a focus on WiFi and WiMAX technologies. The book also contains chapters that deal with transport issues, and namely protocols and policies for efficient and guaranteed transmission characteristics while transferring demanding data applications such as video. Finally, the book includes chapters that focus on the delivery of applications through common telecommunication channels such as the earth atmosphere. This book is useful for researchers working in the telecommunications field, in order to read a compact gathering of some of the latest efforts in related areas. It is also useful for educators that wish to get an up-to-date glimpse of telecommunications research and present it in an easily understandable and concise way. It is finally suitable for the engineers and other interested people that would benefit from an overview of ideas, experiments, algorithms and techniques that are presented throughout the book.

InTech Europe

University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China

Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen