APPROACH FOR ANALYSIS OF FINITE $M_2|M_2|1|R$ WITH HYSTERIC POLICY FOR SIP SERVER HOP-BY-HOP OVERLOAD CONTROL

Alexander V. Pechinkin
Rostislav V. Razumchik
Institute of Informatics Problems of RAS
Vavilova, 44-2,
119333, Moscow, Russia
Email: apechinkin@ipiran.ru,
rrazumchik@ieee.org

KEYWORDS

SIP, hop-by-hop overload control, threshold, hysteretic load control, queueing system, matrix-analytic method.

ABSTRACT

Matrix-analytic method for the analysis of finite $M_2|M_2|1|R$ queueing system with bi-level hysteric policy that models signalling hop-by-hop load control mechanism for SIP server is presented. Algorithm for efficient computation of of joint stationary probability distribution and expressions for some performance characteristics are given. Illustrative numerical example is provided to demonstrate some aspects of the proposed method. Results of calculations using proposed method were compared with results obtained from simulation model, developed using GPSS software, and showed good accuracy.

INTRODUCTION

One of the benefits that network operators gained with the beginning of use of packet networks was the opportunity to provide new, additional services for end-users. Since that times number of services increased drastically and the amount of control information, needed to provide those services, that has to be transmitted hither and thither through network, also increased. This naturally led to several problems concerning overload issues in network nodes that handle the control information. It is known that at the moment SIP (Session Initiation Protocol) is the main signalling protocol carrying control information in NGN (Next Generation Network). The SIP protocol has a basic overload control mechanism (503 Service Unavailable response code) but, as stated in RFC 5390 (2008), it cannot be considered as satisfactory because in some cases it may even worsen an overload condition. In RFC 6357 (2011) one can find so called hop-by-hop overload control mechanism which may be considered as a good alternative method for resolution of overload problems in SIP servers' network. Hop-by-hop overload control does not require that all SIP entities in a network support it. It can be used effectively between two adjacent SIP servers if both servers support overload control and does not depend on the support from any other server or user agent. A thorough discussion on this subject one can find, for example, in Abaev et al. (2012a), Abaev et al. (2013a), Yang et al. (2009), Garroppo et al. (2011), Ohta et al. (2009), Shen et al. (2008). In paper Abaev et al. (2012a) there was also developed a method, based on hysteretic policy with thresholds, for realization for hop-by-hop overload control. There was constructed a suitable queuing model with bi-level hysteric policy and algorithm for its performance characteristics was obtained.

Roughly speaking the model of hop-by-hop overload control with bi-level hysteric policy can be described as follows. There are two communicating SIP servers, say sending and receiving. The receiving SIP server (modeled as $M|M|1|R$ queue) may be in three operational states ("normal", "overloaded", "blocking") that are defined with two thresholds. It monitors continuously the current utility of its processor and notifies the sending server when it detects overloading. Then it asks sending SIP server to reduce the number of packets it sends by a certain (desired) factor. If the utility of its processor continues to grow receiving SIP server asks sending SIP server to stop sending packets (i.e. its state changes to "blocking"). When the utilization goes down the state changes back to "overloading" (and sending server starts to send again a certain fraction of packets) and if it continues to go down receiving SIP server switches to "normal" state and sending server starts to send all packets destined to receiving server.

Several generalizations of the queueing model analyzed in Abaev et al. (2012a) were proposed. In Abaev et al. (2012b) consideration was given to $M_2|M|1|R$ with bi-level hysteric policy and and new matrix-method was proposed for calculation of joint stationary distribution and main performance characteristics. In Abaev et al. (2013b) and Abaev et al. (2013c), motivated by the fact that SIP messages may be served for constant time, finite queueing system with deterministic service ($M|D|1|R$) and bi-level hysteric policy was considered. There was developed new approach for calculation of joint stationary distribution, main performance charac-

Proceedings 27th European Conference on Modelling and Simulation ©ECMS Webjørn Rekdalsbakken, Robin T. Bye, Houxiang Zhang (Editors)
ISBN: 978-0-9564944-6-7 / ISBN: 978-0-9564944-7-4 (CD)

teristics, including one of the important characteristics of hysteric mechanism — mean return time of the system back to normal state. Among other recent papers that deal with analysis of queues with hysteric policy one can mention Chydzinski (2004), **?**, Avrachenkov et al. (2011), Bekker (2009), Bekker et al. (2007), Taremi et al. (2012), Choi et al. (2008). The utilized methods (including potential method) allow one to obtain different stationary performance characteristics under different assumptions about service time distribution and incoming flow.

In this paper we analyze queueing system $M_2|M_2|1|R$ with bi-level hysteric policy which is the generalization of model, presented in Abaev et al. (2012b) and develop new effective approach for calculation of its joint stationary distribution which also differs from approaches known from accessible literature, including the one cited above. The consideration of such system is motivated by the fact that SIP messages of different types (INVITE and NON-INVITE) may be served by central processor unit of the SIP server with different speeds, which was not taken into account in models considered earlier. One of the stumble rocks in the steady state analysis of such systems is the fast growth of state space dimension with the growth of threshold values. So though this system may be more adequate for modeling purposes and provide more accurate results, its analysis with classical methods becomes intractable starting from low values of thresholds. Driven by interest in the explicit solution of the considered problem we propose method that allows computation of performance characteristics and of joint stationary distribution of number of INVITE and NON-INVITE messages in the queue and system's state for relatively high values of thresholds.

The rest of the paper is organized as follows. In the next section we give a description of the considered system and introduce several auxiliary variables that constitute the essence of our approach. In the subsequent section using elimination method we derive system of equilibrium equations for stationary joint probabilities and comment on its solution. In the last section some illustrative numerical examples are provided. Conclusion contains short description of the obtained results and remarks on further study.

SYSTEM DESCRIPTION

Consider the queueing system with two poisson incoming flows of customers (say type 1 and type 2) with rate λ_1 and λ_2 respectively, finite queue of size $R < \infty$, and one server. If arriving customer sees R customers in the system, it is considered to be lost. Henceforth we denote by λ the sum of λ_1 and λ_2, i.e. $\lambda = \lambda_1 + \lambda_2$. Type 1 customers have relative priority over customers of type 2 (i.e. no service interruptions are allowed and type 2 customer enters server facility only when it becomes free and there are no type 1 customers in the queue). Customers of type 1 and type 2 are served exponentially with different service rates, say μ_1 and μ_2. The hysteric mechanism operates as follows. Choose arbitrary numbers L, H such that $0 < L < H < R$. When the system starts to work it is empty and as long as the total number of customers in the system remains below $(H-1)$, system is considered to be in "normal" state. When total number of customers exceeds $(H-1)$ for the first time, the system changes its state to "overload" and stays in it as long as the number of customers remains between L and $(R-1)$. Moreover when overloaded, system accepts only type 1 customers. Being in "overload" state, system waits till the number of customers drops down below L after which it changes its state back to "normal", or exceeds $(R-1)$ after which it changes its state to "blocking". In the "blocking" state systems does not accept new arriving customers until the total number of customers drops down below $(H+1)$, after which system's state changes back to "overload".

The operation of the considered queueing system can be completely described by Markov process $\mathbf{X}(t) = \{\xi(t); \eta(t); \nu(t); \theta(t)\}$ with four components: $\xi(t)$ — number of type 1 customers in the queue at time t, $\eta(t)$ — number of type 2 customers the queue at time t, $\nu(t)$ — state of the system at time t, $\theta(t)$ — type of customer being served at time t. Clearly $\mathbf{X}(t)$ is ergodic and thus stationary distribution exists. In the next subsection we introduce auxiliary variables that constitute new method that we propose for the analysis of the system.

Auxiliary variables

The idea of the method is, using the property of ergodic processes (see, for example, Lemma 1.4.1 and Corollary 1.4.1 in Bocharov et al. (2003)), to write out the system of equilibrium equations (SEE) in such a form that it can be solved iteratively with minor manipulations. But before we can write out SEE, several auxiliary variables are needed.

Due to the fact that type 1 and type 2 customers have different priorities and are served with different rates we need to introduce the following notation. Henceforth all matrices that appear are block partitioned matrices which have the structure

$$\mathrm{A}_n = \begin{pmatrix} \mathrm{A}_n^{(11)} & \mathrm{A}_n^{(12)} \\ \mathrm{A}_n^{(21)} & \mathrm{A}_n^{(22)} \end{pmatrix}.$$

Here index $^{(kl)}$, $k = 1, 2$, $l = 1, 2$, means that at some moment of time type k customer is being served and after any event (i.e. arrival or service completion) type l customer is being served.

By E_n henceforth we denote n-by-n identity matrix.

Let at some moment of time there be n, $n = \overline{H+1, R-1}$ customers in the system and i, $i = \overline{0, H-1}$, customers of type 2 in the queue, and the system is in "overload" state. Denote by A_n matrix of size $2H \times 2H$. The $(i,j)^{th}$ entry of A_n is the probability that at the moment of time when the total number

of customers in the system equals $(n-1)$ for the first time, there will be $j,\ j=\overline{0,H-1}$, customers of type 2 in the queue, and until that moment the total number of customers in the system remained below R.

For matrix A_{R-1} it holds

$$\mathrm{A}_{R-1}^{(l1)}=\frac{\mu_l}{\lambda_1+\mu_l}\mathrm{E}_H,\ \ l=1,2;\quad \mathrm{A}_{R-1}^{(12)}=\mathrm{A}_{R-1}^{(22)}=0.$$

Other matrices A_n can be found using the following relations

$$\mathrm{A}_n=\mathrm{U}+\mathrm{V}\mathrm{A}_{n+1}\mathrm{A}_n,\ \ n=\overline{H+1,R-2},$$

where

$$\mathrm{U}^{(l1)}=\tfrac{\mu_l}{\lambda_1+\mu_l}\mathrm{E}_H,\ \ l=1,2;\quad \mathrm{U}^{(l2)}=0,\ \ l=1,2;$$
$$\mathrm{V}^{(ll)}=\tfrac{\lambda_1}{\lambda_1+\mu_l}\mathrm{E}_H,\ \ l=1,2;\quad \mathrm{V}^{(12)}=\mathrm{V}^{(21)}=0.$$

Now let at some moment of time there be $n,\ n=\overline{H+1,R-1}$ customers in the system and $i,\ i=\overline{0,H-1}$, customers of type 2 in the queue, and the system is in "overload" state. Denote by Γ_n matrix of size $2H\times 2H$. The $(i,j)^{th}$ entry of Γ_n is the probability that at the moment of time when the total number of customers in the system equals R for the first time, there will be $j,\ j=\overline{0,H-1}$, customers of type 2 in the queue, and until that moment the total number of customers in the system remained above $(n-1)$.

Matrix Γ_{D-1} is diagonal and its blocks have the form

$$\Gamma_{R-1}^{(ll)}=\frac{\lambda_1}{\lambda_1+\mu_l}\mathrm{E}_H,\ \ l=1,2;\quad \Gamma_{R-1}^{(12)}=\Gamma_{R-1}^{(21)}=0.$$

Other matrices Γ_n can be found from the following relations

$$\Gamma_n=\mathrm{U}\Gamma_{n+1}+\mathrm{U}\mathrm{A}_{n+1}\Gamma_n,\ \ n=\overline{H+1,R-2},$$

where

$$\mathrm{U}^{(ll)}=\frac{\lambda_1}{\lambda_1+\mu_l}\mathrm{E}_H,\ \ l=1,2;\quad \mathrm{U}^{(12)}=\mathrm{U}^{(21)}=0.$$

Now let at some moment of time there be R customers in the system and $i,\ i=\overline{0,H-1}$, customers of type 2 in the queue. Clearly the system is in "blocking" state. Denote by D_n matrix of size $2H\times 2H$. The $(i,j)^{th}$ entry of D_n is the probability that at the moment of time when the total number of customers in the system equals $n,\ n=\overline{H,R-1}$ for the first time, there will be $j,\ j=\overline{0,H-1}$, customers of type 2 in the queue. Because in "blocking" state system does not accept newly arriving customers, for matrices D_n it holds

$$\mathrm{D}_n^{(l1)}=\mathrm{E}_H,\quad \mathrm{D}_n^{(l2)}=0,\ \ l=1,2,\ \ n=\overline{H,R-1}.$$

Further we will use the following notation. Let X_n, Y_n, Z_n and W_n be matrices of size $n\times(n-1)$, $n\times(n-1)$, $n\times(n+1)$ and $n\times(n+1)$ respectively, which have the following structure

$$X_n=\begin{pmatrix}1&0&0&\dots&0&0\\0&1&0&\dots&0&0\\0&0&1&\dots&0&0\\\vdots&\vdots&\vdots&\ddots&\vdots&\vdots\\0&0&0&\dots&1&0\\0&0&0&\dots&0&1\\0&0&0&\dots&0&1\end{pmatrix},$$

$$Y_n=\begin{pmatrix}0&0&0&\dots&0&0\\0&0&0&\dots&0&0\\0&0&0&\dots&0&0\\\vdots&\vdots&\vdots&\ddots&\vdots&\vdots\\0&0&0&\dots&0&0\\0&0&0&\dots&0&0\\0&0&0&\dots&0&1\end{pmatrix},$$

$$Z_n=\begin{pmatrix}1&0&0&\dots&0&0&0\\0&1&0&\dots&0&0&0\\0&0&1&\dots&0&0&0\\\vdots&\vdots&\vdots&\ddots&\vdots&\vdots&\vdots\\0&0&0&\dots&1&0&0\\0&0&0&\dots&0&1&0\end{pmatrix},$$

$$W_n=\begin{pmatrix}0&1&0&0&\dots&0&0\\0&0&1&0&\dots&0&0\\0&0&0&1&\dots&0&0\\\vdots&\vdots&\vdots&\vdots&\ddots&\vdots&\vdots\\0&0&0&0&\dots&1&0\\0&0&0&0&\dots&0&1\end{pmatrix}.$$

Let at some moment of time there be $n,\ n=\overline{L,H-1}$ customers in the system and $i,\ i=\overline{0,n-1}$, customers of type 2 in the queue, and the system is in normal state. Denote by A_n matrix of size $2n\times 2(n-1)$. The $(i,j)^{th}$ entry of A_n is the probability that at the moment of time when the total number of customers in the system equals $(n-1)$ for the first time, there will be $j,\ j=\overline{0,n-2}$, customers of type 2 in the queue, and until that moment the total number of customers in the system remained below H.

For matrix A_{H-1} it holds

$$\mathrm{A}_{H-1}^{(l1)}=\frac{\mu_l}{\lambda+\mu_l}\mathrm{X}_{H-1},\ \ \mathrm{A}_{H-1}^{(l2)}=\frac{\mu_l}{\lambda+\mu_l}\mathrm{Y}_{H-1},\ \ l=1,2.$$

Other matrices A_n can be found using the following relations:

$$\mathrm{A}_n=\mathrm{U}_n+\mathrm{V}_n\mathrm{A}_{n+1}\mathrm{A}_n,$$

where

$$\mathrm{U}_n^{(l1)} = \frac{\mu_l}{\lambda+\mu_l}\mathrm{X}_n, \quad \mathrm{U}_n^{(l2)} = \frac{\mu_l}{\lambda+\mu_l}\mathrm{Y}_n, \ \ l = 1,2;$$
$$\mathrm{V}_n^{(ll)} = \frac{\lambda_1}{\lambda+\mu_l}\mathrm{Z}_n + \frac{\lambda_2}{\lambda+\mu_l}\mathrm{W}_n, \ \ l = 1,2;$$
$$\mathrm{V}_n^{(12)} = \mathrm{V}_n^{(21)} = 0.$$

Let at some moment of time there be $n,\ n = \overline{L, H-1}$ customers in the system and $i,\ i = \overline{0, n-1}$, customers of type 2 in the queue, and the system is in normal state. Denote by Γ_n matrix of size $2n \times 2H$. The $(i,j)^{th}$ entry of Γ_n is the probability that at the moment of time when the total number of customers in the system equals H for the first time, there will be $j,\ j = \overline{0, H-1}$, customers of type 2 in the queue, and until that moment the total number of customers in the system remained above $(n-1)$.

For blocks of matrix Γ_{H-1} it holds

$$\Gamma_{H-1}^{(ll)} = \frac{\lambda_1}{\lambda+\mu_l}\mathrm{Z}_{H-1} + \frac{\lambda_2}{\lambda+\mu_l}\mathrm{W}_{H-1}, \ \ l = 1,2;$$
$$\Gamma_{H-1}^{(12)} = \Gamma_{H-1}^{(21)} = 0.$$

Other matrices Γ_n can be computed from the relations

$$\Gamma_n = \mathrm{U}_n\Gamma_{n+1} + \mathrm{U}_n\mathrm{A}_{n+1}\Gamma_n, \ \ n = \overline{L, H-2},$$

where

$$\mathrm{U}_n^{(ll)} = \frac{\lambda_1}{\lambda+\mu_l}\mathrm{Z}_n + \frac{\lambda_2}{\lambda+\mu_l}\mathrm{W}_n, \ \ l = 1,2;$$
$$\mathrm{U}_n^{(12)} = \mathrm{U}_n^{(21)} = 0.$$

Now let at some moment of time there be $n,\ n = \overline{L, H-1}$ customers in the system and $i,\ i = \overline{0, n-1}$, customers of type 2 in the queue, and the system is in "overload" state. Denote B_n — matrix of size $2n \times 2(n+1)$. The $(i,j)^{th}$ entry of B_n is the probability that at the moment of time when the total number of customers in the system equals $(n+1)$ for the first time, there will be $j,\ j = \overline{0, n}$, customers of type 2 in the queue, and until that moment the total number of customers in the system remained above $L-1$.

For matrix B_L is holds

$$\mathrm{B}_L^{(ll)} = \frac{\lambda_1}{\lambda_1+\mu_l}\mathrm{Z}_n, \ \ l = 1,2; \quad \mathrm{B}_L^{(12)} = \mathrm{B}_L^{(21)} = 0.$$

Other matrices B_n can be found from relations

$$\mathrm{B}_n = \mathrm{U}_n + \mathrm{V}_n\mathrm{B}_{n-1}\mathrm{B}_n, \ \ n = \overline{L+1, H-1},$$

where

$$\mathrm{U}_n^{(12)} = \mathrm{U}_n^{(21)} = 0; \quad \mathrm{U}_n^{(ll)} = \tfrac{\lambda_1}{\lambda_1+\mu_l}\mathrm{Z}_n, \ \ l = 1,2;$$
$$\mathrm{V}_n^{(l1)} = \tfrac{\mu_l}{\lambda_1+\mu_l}\mathrm{X}_n, \quad \mathrm{V}_n^{(l2)} = \tfrac{\mu_l}{\lambda_1+\mu_l}\mathrm{Y}_n, \ \ l = 1,2.$$

Now consider the following case. Let at some moment of time there be $n,\ n = \overline{L, H-1}$ customers in the system and $i,\ i = \overline{0, n-1}$, customers of type 2 in the queue, and the system is in "overload" state. Denote by D_n matrix of size $2n \times 2(L-1)$. The $(i,j)^{th}$ entry of D_n is the probability that at the moment of time when the total number of customers in the system equals $(L-1)$ for the first time, there will be $j,\ j = \overline{0, L-2}$, customers of type 2 in the queue, and until that moment the total number of customers in the system remained below n.

Matrix D_L is computed as follows

$$\mathrm{D}_L^{(l1)} = \frac{\mu_l}{\lambda_1+\mu_l}\mathrm{X}_L, \quad \mathrm{D}_L^{(l2)} = \frac{\mu_l}{\lambda_1+\mu_l}\mathrm{Y}_L, \ \ l = 1,2.$$

For other matrices D_n it holds

$$\mathrm{D}_n = \mathrm{U}_n\mathrm{D}_{n-1} + \mathrm{U}_n\mathrm{B}_{n-1}\mathrm{D}_n, \ \ n = \overline{L+1, H-1},$$

where

$$\mathrm{U}_n^{(l1)} = \frac{\mu_l}{\lambda_1+\mu_l}\mathrm{X}_n, \quad \mathrm{U}_n^{(l2)} = \frac{\mu_l}{\lambda_1+\mu_l}\mathrm{Y}_n, \ \ l = 1,2.$$

Let at some moment of time there be H customers in the system (i.e. the system is in "overload" state) and $i,\ i = \overline{0, H-1}$, customers of type 2 in the queue. Denote D — matrix of size $2H \times 2(L-1)$. The $(i,j)^{th}$ entry of D is the probability that at the moment of time when the total number of customers in the system equals $(L-1)$ for the first time, there will be j, $j = \overline{0, L-2}$, customers of type 2 in the queue.

Matrix D can be determined from the equation

$$\mathrm{D} = \mathrm{U}\mathrm{D}_{H-1} + \mathrm{U}\mathrm{B}_{H-1}\mathrm{D} + \mathrm{V}(\mathrm{A}_{H+1} + \Gamma_{H+1}\mathrm{D}_H)\mathrm{D},$$

where

$$\mathrm{U}^{(l1)} = \tfrac{\mu_l}{\lambda_1+\mu_l}\mathrm{X}_H, \quad \mathrm{U}^{(l2)} = \tfrac{\mu_l}{\lambda_1+\mu_l}\mathrm{Y}_H, \ \ l = 1,2;$$
$$\mathrm{V}^{(12)} = \mathrm{V}^{(21)} = 0; \quad \mathrm{V}^{(ll)} = \tfrac{\lambda_1}{\lambda_1+\mu_l}\mathrm{E}_H, \ \ l = 1,2.$$

Finally we introduce the last sequence of matrices. Let at some moment of time there be $n,\ n = \overline{1, L-1}$ customers in the system and $i,\ i = \overline{0, n-1}$, customers of type 2 in the queue, and the system is in normal state. Denote by A_n matrix of size $2n \times 2(n-1)$. The $(i,j)^{th}$ entry of A_n is the probability that at the moment of time when the total number of customers in the system equals $(n-1)$ for the first time, there will be $j,\ j = \overline{0, n-2}$, customers of type 2 in the queue. For matrix A_{L-1} it holds

$$\mathrm{A}_{L-1} = \mathrm{U}_{L-1} + \mathrm{V}_{L-1}(\mathrm{A}_L + \Gamma_L\mathrm{D})\mathrm{A}_{L-1},$$

where

$$\mathrm{U}_{L-1}^{(l1)} = \frac{\mu_l}{\lambda+\mu_l}\mathrm{X}_{L-1}, \ \ \mathrm{U}_{L-1}^{(l2)} = \frac{\mu_l}{\lambda+\mu_l}\mathrm{Y}_{L-1}, \ \ l=1,2;$$
$$\mathrm{V}_{L-1}^{(ll)} = \frac{\lambda_1}{\lambda+\mu_l}\mathrm{Z}_{L-1} + \frac{\lambda_2}{\lambda+\mu_l}\mathrm{W}_{L-1}, \ \ l = 1,2;$$
$$\mathrm{V}_{L-1}^{(12)} = \mathrm{V}_{L-1}^{(21)} = 0.$$

Other matrices A_n are computed from equations

$$\mathrm{A}_n = \mathrm{U}_n + \mathrm{V}_n\mathrm{A}_{n+1}\mathrm{A}_n, \ \ n = \overline{1, L-2},$$

where

$$\mathrm{U}_n^{(ll)} = \frac{\mu_l}{\lambda+\mu_l}\mathrm{X}_n,\ \ \mathrm{U}_n^{(l2)} = \frac{\mu_l}{\lambda+\mu_l}\mathrm{Y}_n,\ \ l=1,2;$$
$$\mathrm{V}_n^{(ll)} = \frac{\lambda_1}{\lambda+\mu_l}\mathrm{Z}_n + \frac{\lambda_2}{\lambda+\mu_l}\mathrm{W}_n,\ \ l=1,2;$$
$$\mathrm{V}_n^{(12)} = \mathrm{V}_n^{(21)} = 0.$$

STATIONARY JOINT PROBABILITY DISTRIBUTION

Let $\vec{p}_0 = p_0$ be stationary probability of the empty system. Denote by

- $\vec{p}_n,\ n=\overline{1,H-1}$, — row vector of size $2n$, whose first n elements $p_{ni},\ i=\overline{0,n-1}$, are stationary probabilities of the fact that there are total of n customers in the system, including i type 2 customers in the queue, system is in the normal state and type 1 customer is being served; last n elements $p_{ni},\ i=\overline{n,2n-1}$ are stationary probabilities of the fact that there are n customers in the system, including $(i-n)$ type 2 customers in the queue, system is in the normal state and type 2 customer is being served;
- $\vec{p}'_n,\ n=\overline{L,H-1}$, — row vector of size $2n$, whose first n elements $p'_{ni},\ i=\overline{0,n-1}$, are stationary probabilities of the fact that there are total of n customers in the system, including i type 2 customers in the queue, system is in "overload" state and type 1 customer is being served; last n elements $p'_{ni},\ i=\overline{n,2n-1}$ are stationary probabilities of the fact that there are n customers in the system, including $(i-n)$ type 2 customers in the queue, system is in "overload" state and type 2 customer is being served;
- $\vec{p}'_n,\ n=\overline{H,R-1}$, — row vector of size $2H$, whose first n elements $p'_{ni},\ i=\overline{0,H-1}$, are stationary probabilities of the fact that there are total of n customers in the system, including i type 2 customers in the queue, system is in "overload" state and type 1 customer is being served; last n elements $p'_{ni},\ i=\overline{H,2H-1}$ are stationary probabilities of the fact that there are n customers in the system, including $(i-H)$ type 2 customers in the queue, system is in "overload" state and type 2 customer is being served;
- $\vec{p}''_n,\ n=\overline{H+1,R}$, — row vector of size $2H$, whose first n elements $p''_{ni},\ i=\overline{0,H-1}$, are stationary probabilities of the fact that there are total of n customers in the system, including i type 2 customers in the queue, system is in "blocking" state and type 1 customer is being served; last n elements $p''_{ni},\ i=\overline{H,2H-1}$ are stationary probabilities of the fact that there are n customers in the system, including $(i-H)$ type 2 customers in the queue, system is in "blocking" state and type 2 customer is being served.

Finally, introduce the following matrices:

- $\mathrm{U}_n,\ n=\overline{1,H-1}$, of size $2n\times 2n$, where $\mathrm{U}_n^{(ll)} = (\mu_l+\lambda)\mathrm{E}_n,\ l=1,2,\ \mathrm{U}_n^{(12)} = \mathrm{U}_n^{(21)} = 0$;
- $\mathrm{U}'_n,\ n=\overline{L,H-1}$ of size $2n\times 2n$, where $(\mathrm{U}'_n)^{(ll)} = (\mu_l+\lambda_1)\mathrm{E}_n,\ l=1,2,\ (\mathrm{U}'_n)^{(12)} = (\mathrm{U}'_n)^{(21)} = 0$;
- $\mathrm{U}'_n,\ n=\overline{H,R-1}$ of size $2H\times 2H$, where $(\mathrm{U}'_n)^{(ll)} = (\mu_l+\lambda_1)\mathrm{E}_H,\ l=1,2,\ (\mathrm{U}'_n)^{(12)} = (\mathrm{U}'_n)^{(21)} = 0$;
- $\mathrm{U}''_n,\ n=\overline{H+1,R}$ of size $2H\times 2H$, where $(\mathrm{U}''_n)^{(ll)} = \mu_l\mathrm{E}_H,\ l=1,2,\ (\mathrm{U}''_n)^{(12)} = (\mathrm{U}''_n)^{(21)} = 0$;
- $\Lambda'_n = \lambda_1\mathrm{E}_{2H},\ n=\overline{H,R-1}$, of size $2H\times 2H$;
- $\mathrm{M}_n,\ n=\overline{H+2,R}$, of size $2H\times 2H$, where $M_n^{(l1)} = \mu_l\mathrm{E}_H,\ l=1,2,\ M_n^{(12)} = M_n^{(22)} = 0$;
- $\Lambda_0 = \lambda_1(1\ \ 0) + \lambda_2(0\ \ 1)$ of size 1×2;
- $\Lambda_n,\ n=\overline{1,H-1}$, of size $2n\times 2(n+1)$, where $(\Lambda_n)^{(ll)} = \lambda_1\mathrm{Z}_n + \lambda_2\mathrm{W}_n,\ l=1,2,\ (\Lambda_n)^{(12)} = (\Lambda_n)^{(21)} = 0$;
- $\mathrm{M}_n,\ n=\overline{L+1,H}$ of size $2n\times 2(n-1)$, where $(\mathrm{M}_n)^{(l1)} = \mu_l\mathrm{X}_n,\ (\mathrm{M}_n)^{(l2)} = \mu_l\mathrm{Y}_n,\ l=1,2.$

After making all preliminary remarks we can write out the system of equilibrium equations. It holds

$$\vec{p}_n\mathrm{U}_n = \vec{p}_n\Lambda_n\mathrm{A}_{n+1} + \vec{p}_{n-1}\Lambda_{n-1},\quad n=\overline{1,L-2}, \tag{1}$$
$$\vec{p}_{L-1}\mathrm{U}_{L-1} = \vec{p}_{L-1}\Lambda_{L-1}(\mathrm{A}_L+\Gamma_L\mathrm{D}) + \vec{p}_{L-2}\Lambda_{L-2}, \tag{2}$$
$$\vec{p}_n\mathrm{U}_n = \vec{p}_n\Lambda_n\mathrm{A}_{n+1} + \vec{p}_{n-1}\Lambda_{n-1},\quad n=\overline{L,H-2}, \tag{3}$$
$$\vec{p}_{H-1}\mathrm{U}_{H-1} = \vec{p}_{H-2}\Lambda_{H-2}, \tag{4}$$
$$\vec{p}'_H\mathrm{U}'_H = \vec{p}_{H-1}\Lambda_{H-1} + \vec{p}'_H(\mathrm{M}_H\mathrm{B}_{H-1} + \Lambda'_H(\mathrm{A}_{H+1} + \Gamma_{H+1}\mathrm{D}_H)), \tag{5}$$
$$\vec{p}'_n\mathrm{U}'_n = \vec{p}'_{n-1}\Lambda'_{n-1} + \vec{p}'_n\Lambda'_n\mathrm{A}_{n+1},\quad n=\overline{H+1,R-2}, \tag{6}$$
$$\vec{p}'_{R-1}\mathrm{U}'_{R-1} = \vec{p}'_{R-2}\Lambda'_{R-2}, \tag{7}$$
$$\vec{p}''_R\mathrm{U}''_R = \vec{p}'_{R-1}\Lambda'_{R-1}, \tag{8}$$
$$\vec{p}''_n\mathrm{U}''_n = \vec{p}''_{n+1}\mathrm{M}_{n+1},\ \ n=\overline{H+1,R-1}, \tag{9}$$
$$\vec{p}'_n\mathrm{U}'_n = \vec{p}'_{n+1}\mathrm{M}_{n+1} + \vec{p}'_n\mathrm{M}_n\mathrm{B}_{n-1},\quad n=\overline{L+1,H-1}, \tag{10}$$
$$\vec{p}'_L U'_L = \vec{p}'_{L+1}\mathrm{M}_{L+1}. \tag{11}$$

The probabilities must satisfy the normalization condition

$$p_0 + \sum_{n=1}^{H-1}\vec{p}_n\vec{1} + \sum_{n=L}^{R-1}\vec{p}'_n\vec{1} + \sum_{n=H+1}^{R}\vec{p}''_n\vec{1} = 1. \tag{12}$$

The main idea behind equations (1)–(11) is the elimination method which can be found in (Bocharov et al., 2003, Chapter 1). Their solution is straightforward. Divide (1)–(11) by p_0 and compute unknown variables consequently from each equation starting from (1). Then

make use of normalization condition (12) to find p_0 and multiply each of the previously obtained variables by p_0.

Let

$$p_n = \sum_{i=0}^{2n-1} p_{ni}, \quad n = \overline{1, H-1},$$

$$p'_n = \sum_{i=0}^{2n-1} p'_{ni}, \quad n = \overline{L, H-1},$$

$$p'_n = \sum_{i=0}^{2H-1} p'_{ni}, \quad n = \overline{H, R-1},$$

$$p''_n = \sum_{i=0}^{2H-1} p''_{ni}, \quad n = \overline{H+1, R},$$

Knowing joint stationary distribution one can calculate a number of performance characteristics. The system utilization is simply $(1 - p_0)$. The probability π_1 (π_2) that the arriving customer of type 1 (type 2) is lost equals

$$\pi_1 = \sum_{n=H+1}^{R} p''_n, \quad \pi_2 = \pi_1 + \sum_{n=L}^{R-1} p'_n.$$

Served load is $\lambda^* = (1-\pi_1)\lambda_1 + (1-\pi_2)\lambda_2$. The mean number Q_1 (Q_2) of customers of type 1 (type 2) in the queue can be calculated as follows:

$$Q_1 = \sum_{n=1}^{H-1}\sum_{i=0}^{n-1}(n-i-1)p_{ni} + \sum_{n=L}^{H-1}\sum_{i=0}^{n-1}(n-1-i)p'_{ni} + \\ + \sum_{n=H}^{R-1}\sum_{i=0}^{H-1}(n-1-i)p'_{ni} + \sum_{n=H+1}^{R}\sum_{i=0}^{H-1}(n-1-i)p''_n,$$

$$Q_2 = \sum_{n=1}^{H-1}\sum_{i=0}^{n-1} i p_{ni} + \sum_{n=L}^{H-1}\sum_{i=0}^{n-1} i p'_{ni} + \\ + \sum_{n=H}^{R-1}\sum_{i=0}^{H-1} i p'_{ni} + \sum_{n=H+1}^{R}\sum_{i=0}^{H-1} i p''_n.$$

Finally, mean time that customer of type 1 (type 2) has to wait in the queue until it receives service can be calculated using Little's law. That is, we have

$$V_1 = \frac{Q_1}{(1-\pi_1)\lambda_1}, \quad V_2 = \frac{Q_2}{(1-\pi_2)\lambda_2}.$$

NUMERICAL RESULTS

In this section we give illustrative numerical examples of application of developed results. In Abaev et al. (2012b) consideration was given to $M_2|M|1|R$ with bi-level hysteric policy for modeling of SIP server with load control. The question of interest is how much one underestimates or overestimates performance characteristics of SIP server if we model it with $M_2|M|1|R$ with bi-level hysteric policy (i.e. if we consider that all incoming messages are served by SIP server with the same rate). The model considered in this paper allows us to do it because it takes into consideration the type (e.g. INVITE or NON-INVITE) of message that arrives at the CPU of SIP server.

For the examples we use the following values of thresholds: $L = 10$, $H = 18$, $R = 25$. Let the mean service time for priority customers (e.g. INVITE messages) be equal 10ms, i.e. $\mu_1 = 0.1$. We assume that the total customers' arrival rate is 200 $(sec)^{-1}$, i.e. $\lambda_1 + \lambda_2 = 0.2\,(ms)^{-1}$. Recall that in "overload" state the system reduces input flow. Let q be the dropping probability of newly arriving customers when system is in "overload" state. Then $\lambda_1 = \lambda(1-q)$ and $\lambda_2 = \lambda q$.

In figure 1 and figure 2 one can see the behaviour of mean queue length (i.e. $Q_1 + Q_2$) and loss probability of type 2 customers (i.e. π_2) as functions of dropping probability q. for different values of mean service time $(\mu_2)^{-1}$ of type 2 customers.

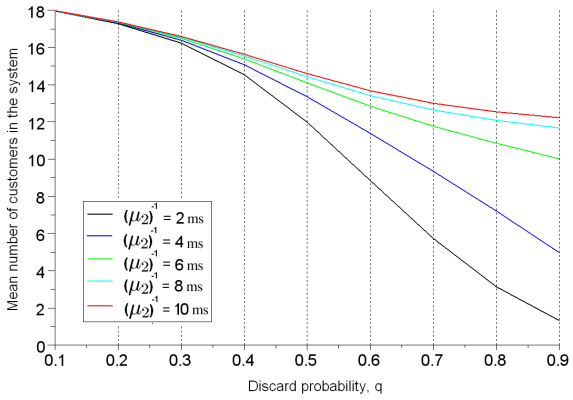

Figure 1: Mean number of customers in the queue

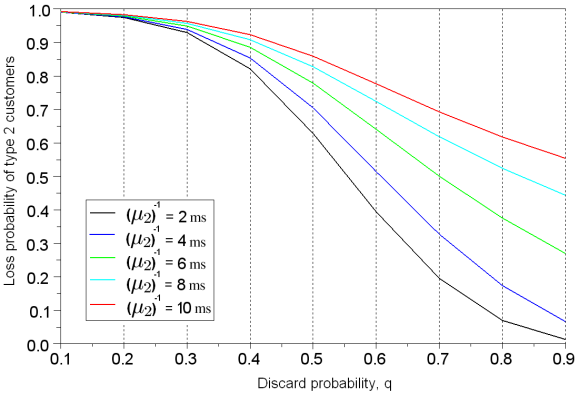

Figure 2: Loss probability of type 2 customers

From figures it is seen that starting from the certain value of q account of different service rates for different types of customers noticeably influences the values of mean queue length and especially of loss probability π_2. In order to check theoretical results there was built a simulation model using GPSS software. The comparisons of numerical and simulation results showed good accuracy.

SUMMARY

In this study matrix-analytic method for the analysis of finite queueing system $M_2|M_2|1|R$ with bi-level hysteric policy that models signalling hop-by-hop load control mechanism for SIP server is presented. Expressions for main performance characteristics are given. Illustrative numerical example is provided. The considered problem is only one of the many other possible formulations. Future work will be connected with the experimental verification of the proposed model using outcomes of real-world experiments. In addition attention will be paid to the analysis of stationary time-related performance characteristics of this system, posing and solution of optimization problems (in order to find appropriate values of thresholds), and the generalization of the proposed approach for cases with more general input flows and service times.

Notes and Comments. This work was supported in part by the Russian Foundation for Basic Research (grants 12-07-00108, 13-07-00223, 13-07-00284).

REFERENCES

Abaev, P., Gaidamaka, Yu., Pechinkin, A., Razumchik, R., Samouylov, K., Shorgin, S. 2012. Hysteretic control technique for overload problem solution in network of SIP servers. Computing and Informatics. (accepted).

Abaev, P., Pechinkin, A., Razumchik, R. 2013. Analysis of queueing system with constant service time for sip server hop-by-hop overload control. Lecture Notes in Communications in Computer and Information Science (LN CCIS). Pp. 1–10.

Abaev, P., Pechinkin, A., Razumchik, R. 2013. On mean return time in queueing system with constant service time and bi-level hysteric policy. Lecture Notes in Communications in Computer and Information Science (LN CCIS). Pp. 11–19.

Abaev, P., Gaidamaka, Yu., Pechinkin, A., Razumchik, R., Shorgin, S. 2012. Simulation of overload control in SIP server networks. Proc. of the 26th European Conference on Modelling and Simulation, ECMS 2012. Pp. 533–539.

Abaev, P., Pechinkin, A., Razumchik, R. 2012. On analytical model for optimal sip server hop-by-hop overload control. Proc. of the 4th International Congress on Ultra Modern Telecommunications and Control Systems ICUMT-2012. Pp. 303–308.

Avrachenkov, K., Dudin, A., Klimenok, V., Nain, P., Semenova, O. 2011. Optimal threshold control by the robots of web search engines with obsolescence of documents. Journal of Computer Networks. Vol. 55. No. 8. Pp. 1880–1893.

Bekker, R. 2009. Queues with Levy input and hysteretic control. Queueing Systems. Vol. 63. Issue 1. Pp. 281–299.

Bekker, R., Boxma, O.J. 2007. An M/G/1 queue with adaptable service speed. Sochastic Models. Vol. 23. Issue 3. Pp. 373–396.

Bocharov, P., D'Apice, C., Pechinkin, A., Salerno, S. 2003. Queueing Theory. Utrecht: VSP Publishing, 450 p.

Choi, D.I., Kim, T.S., Lee, S. 2008. Analysis of an $MMPP|G|1|K$ queue with queue length dependent arrival rates, and its application to preventive congestion control in telecommunication networks. European Journal of Operational Research. Vol. 187. Issue 2. Pp. 652–659.

Chydzinski, A. 2004. The oscillating queue with finite buffer. Performance Evaluation. Vol. 57. No. 3. Pp. 341–355.

Garroppo, R., Giordano, S., Niccolini, S., Spagna, S. 2011. A Prediction-Based Overload Control Algorithm for SIP Servers. IEEE Transactions on Network and Service Management. Vol. 8. No. 1. Pp. 39–51.

Hilt, V., Noel, E., Shen, C., Abdelal, A. 2011. Design Considerations for Session Initiation Protocol (SIP) Overload Control. RFC 6357

Ohta, M. 2009. Overload Control in a SIP Signalling Network. International Journal of Electrical and Electronics Engineering. Pp. 87–92.

Rosenberg, J. 2008. Requirements for Management of Overload in the Session Initiation Protocol. RFC 5390.

Shen, C., Schulzrinne, H., Nahum, E. 2008. Session Initiation Protocol (SIP) Server Overload Control: Design and Evaluation. Lecture Notes in Computer Science. Vol. 5310. Pp. 149–173.

Taremi, M., Reza Salehi Rad, M. 2012. Limit analysis of oscillating batch arrival $M^{[x]}|G|1$ sysmtes with finite capacity: EMC approach. Journal of Economic Theory. Vol. 6. No. 1. Pp. 29–36.

Yang, J., Huang, F., Gou, S. 2009. An Optimized Algorithm for Overload Control of SIP Signaling Network. 5th International Conference on Wireless Communications, Networking and Mobile Computing. Pp. 1–4.

AUTHOR BIOGRAPHIES

ALEXANDER V. PECHINKIN is a Doctor of Sciences in Physics and Mathematics and principal scientist at the Institute of Informatics Problems of the Russian Academy of Sciences, and a professor at the Peoples' Friendship University of Russia. He is the author of more than 150 papers in the field of applied probability theory. His email address is `apechinkin@ipiran.ru`.

ROSTISLAV V. RAZUMCHIK received his Ph.D. in Physics and Mathematics in 2011. Since then, he has worked as a senior researcher at the Institute of Informatics Problems of the Russian Academy of Sciences. His current research activities focus on stochastic processes and queuing theory. His email address is `rrazumchik@ieee.org`