Queueing Model of a Single-Level Single-Mediator with Cooperation of the Agents

Moon Ho Lee, Aliaksandr Birukou, Alexander Dudin, Valentina Klimenok, Olga Kostyukova and Chang-hui Choe

***Abstract*— Performance characteristics of an organizational structure with single-level single-mediator and cooperation of multiple agents are computed of terms of queueing network of three-like topology with cooperation of the servers. This network is analytically investigated by means of the matrix analytic methods. Results can be used for the logical and technical design and optimal resources sharing in multi-agent systems.**

I. Introduction

Queueing theory (QT) investigates situations when some restricted resource should be efficiently shared between competitive flow of requests in an optimal way. So, definitely, it should be useful in quantitative investigation and comparison of different organizational structures of Multi-Agent Systems (MAS). Possibility of QT applications for MAS was discussed, e.g., in [1],[2],[3],[4]. In particular, in [3] operation of MAS is described in terms of queueing networks. In [1], the $M/M/1$ queueing system was used for utility prediction for a range of possible MAS.

In this paper we make attempt to analyze performance characteristics of the MAS system where the single mediator serves as dispatcher for several independent heterogeneous agents which handle the user queries. If all links to the system agents are busy at the epoch of query arrival to mediator, the query is stored in a buffer. Later it will be picked up from this buffer according to the First In - First Out discipline. If several links to the system agents are free at the query arrival epoch, the mediator assigns all corresponding agents to provide the service to this query and transmits the query via communication channel (link) to all these agents. This model is novel from point of view of queueing theory because the standard assumption is that one query is served exactly by one server (agent). Our motivation of assumption that the query is dispatched to all free agents is the following. Because the agents are autonomous, with some probability the agent can decline the offer to handle a query. Also, we assume that agents can get simultaneously queries from another mediators. The buffers for postponed queries in each server (agent) are assumed to be finite. The the agent can reject the query from mediator just because capacity of his buffer is exhausted. In such situation, when the query can be rejected by an agent (by his desire or due to the buffer overflow), parallel sending of query to all currently available agents has to increase chance of an arbitrary query to be successfully processes in MAS. Besides this reliability aspect, parallel handling of query by several agents can decrease response time because the response time in this case is the minimum of durations of handling the query by all involved agents.

The first, third and sixth authors are with Institute of Information and Communication, Chonbuk National University, Chonju, 561-765, Korea (e'mail: moonho@chonbuk.ac.kr, dudin_alex@yahoo.com,nblue95@chonbuk.ac.kr).

The second author is with Department of Information and Communication Technology, University of Trento, Italy (e'mail: birukou@dit.unitn.it)

The third and fourth authors are with Department of Applied Mathematics and Computer Science, Belarusian State University, Minsk 220030, Belarus, (e'mail: dudin@bsu.by, vklimenok@yandex.ru, chernovsy@tut.by).

The fifth author is with Institute of Mathematics, National Academy of Science of Belarus, 220000, Belarus (e'mail: kostyukova@im.bas-net.by).

II. Mathematical Model

The structure of the considered MAS system is presented in Figure 1.

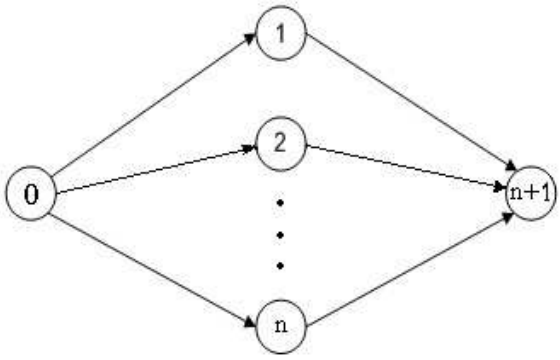

Fig. 1. The structure of the MAS system

The service (queries processing) in the network presented in Figure 1 is actually performed in the links of the graph. The node number $n+1$ is considered as some virtual destination query. The node number 0 is considered as the place of queries arrival to MAS and buffering in the case of busyness of all links to agents. Nodes $1, 2, \ldots, n$ can be considered as the place when the user queries are buffered in the case if the assigned agent is busy. Taking into account this consideration, we conclude that the operation of the MAS having structure given in Figure 1 is described by the queueing network represented in Figure 2.

This queueing network consists of two interacting parts. In the sequel, the left part of the network (mediator part) will be referred to as the queueing system number 0. It consists of one buffer with a infinite capacity and n possibly heterogeneous servers (links to agents). We refer to these servers as server number $1, \ldots,$ server number n.

The right part (autonomous agents part) consists of n independent service systems referred below to as the queueing

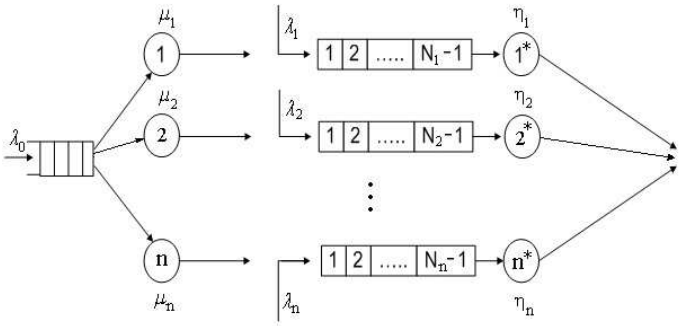

Fig. 2. Queueing network model for the MAS system operation

system number $1^*, \ldots,$ system number n^*. Each of these systems has a finite buffer and a single server (agent). The capacity of the buffer of the system number k^* is equal to $N_{k^*} - 1$, so the maximal total number of queries presenting in this system is equal to $N_{k^*}, k = \overline{1,n}$.

We assume that the customers (queries) arrive to the queueing system number 0 according to the stationary Poisson process with intensity λ_0. If some of servers $1, \ldots, n$ are idle at the arrival epoch, the customer starts the service in all these servers simultaneously. We assume that service times in these servers are mutually independent random variables having exponential distribution with parameter μ_k for the server number $k, k = \overline{1,n}$. If all servers $1, \ldots, n$ are busy at the arrival epoch, the customer is buffered at the buffer to the queueing system number 0. We assume that this buffer has infinite capacity. The buffered customers are picked up from the buffer when any of servers $1, \ldots, n$ completes the service of previous customers according to the *FIFO* (first in - first out) discipline.

After the service in the server k, the customer moves for the service in the queueing system number $k^*, k = \overline{1,n}$. If the server of that system (agent of MAS) is idle at the arrival epoch, it starts processing of the arriving customer with probability $q_k^{(1)}$. Service times of successive customers in the server k^* are independent random variables having exponential distribution with parameter $\eta_k, k = \overline{1,n}$. After the service, customer leaves the system number k^* (reaches destination node $n+1$).

If the server k^* is busy at a customer arrival epoch from the queueing system number 0, the arriving customer with probability $1 - q_k^{(2)}$ is rejected and with supplementary probability it should be placed into the buffer of capacity $N_{k^*} - 1, k = \overline{1,n}$. If this buffer is already full at arrival epoch, the customer is lost in the queueing system number k^*.

Besides processing the transit customers from the queueing system number 0, server of the system number k^* can also process another customers. These customers arrive to server k^* according to the stationary Poisson process with intensity $\lambda_k, k = \overline{1,n}$. Service times of these customers also have exponential distribution with parameter η_k. In the case if the buffer is full at the arrival epoch, the customer is considered to be lost. No priority for any kind of customers is assigned.

Thus, operation of the queueing network presented in Figure 2 is completely described.

Our purpose is derivation of the necessary and sufficient condition for the existence of the stationary mode of the queueing network operation and stationary analysis of distribution of the number of customers in the nodes of this queueing network.

III. Stationary state distribution of the network

Behavior of the queueing network under study can be described by the multi-dimensional continuous time Markov chain

$$\xi_t = \{j_t, i_t^{(1)}, \ldots, i_t^{(k)}\}, t \geq 0,\ i_t^{(k)} = \overline{0, N_k}, k = \overline{1,n},$$

where the component $i_t^{(k)}$ is equal to the number of customers in queueing system $k^*, k = \overline{1,n}$, at the moment $t, t \geq 0$. It includes the customers in the corresponding buffer, if any, and a customer in the server. Component j_t describes the state of the n-server queueing system number 0. The state $j, j \geq 1$, of the component j_t corresponds to the state of the queueing system number 0 when there are j customer in a buffer (sure, all the servers of this system are busy).

If the queue in this system is absent, the state of the component j_t is described by the group of n numbers $\{l_1, \ldots, l_n\}$ where the entry l_k has value 0 if the kth server is idle and value 1 if the kth server is busy at epoch $t, t \geq 0$. Denote by $\mathcal{L}$ the set of all such states. It is evident that it consists of 2^n states.

Aiming to simplify denotations and use benefits of the matrix analytic methods, we enumerate the components of the process $\xi_t = \{j_t, i_t^{(1)}, \ldots i_t^{(n)}\}, t \geq 0$, in the lexicographic order. Then, we refer to the whole set of states $\{j, i_t^{(1)}, \ldots, i_t^{(n)}\},\ i_t^{(k)} = \overline{0, N_k},\ k = \overline{1,n},$ as to the state j of the process $\xi_t, t \geq 0$,

$$j = \{\underbrace{0, \ldots, 0}_{n}\}, \{\underbrace{0, \ldots, 0, 1}_{n}\}, \ldots, \{\underbrace{1, \ldots, 1}_{n}\}, 1, 2, 3 \ldots.$$

For use in the sequel, we introduce the following notation.

- I is identity matrix of dimension $K = \prod_{k=1}^{n} (N_k + 1)$;
- I_k is identity matrix of dimension $N_k + 1,\ k = \overline{1,n}$;
- O is zero square matrix of dimension K;
- $O_{l,m}$ is zero matrix of dimension $Kl \times Km$;
- $\otimes$ is the symbol of Kronecker product of the matrices;
- $\oplus$ is the symbol of Kronecker sum of the matrices;
- ^{T} denotes transposition of a matrix or vector;
- $\mathbf{e}_k$ is the column vector of dimension $N_k + 1$ consisting of all 1's;
- $\mathbf{e}_K$ is the column vector of dimension K consisting of all 1's;
- $\mathbf{0}_k$ is the row vector of dimension $N_k + 1$ consisting of all 0's; $k = \overline{1,n}$;
- $\mathbf{0}_K$ is the row vector of dimension K consisting of all 0's;
- $\mathbf{f}_k^{(i)}$ is the column vector of dimension $N_k + 1$ having the form $(\underbrace{0, \ldots, 0}_{i}, 1, 0, \ldots, 0)^T, i = \overline{0, N_k},\ k = \overline{1,n}$;
- $\tilde{\mathbf{e}}_k^{(i)}, i = \overline{0, N_k}$, is the column vector of dimension K defined by formula $\tilde{\mathbf{e}}_k^{(i)} = \mathbf{e}_1 \otimes \ldots \otimes \mathbf{e}_{k-1} \otimes \mathbf{f}_k^{(i)} \otimes \mathbf{e}_{k+1} \otimes \ldots \otimes \mathbf{e}_n$;

- $\tilde{I}_k$,$\tilde{I}_k$,I_k^+,I_k^-, I_k^0 are the square matrices of dimension $N_k+1,\ k=\overline{1,n}$, having the following structure:

$$\tilde{I}_k = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix},$$

$$\hat{I}_k = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix},$$

$$I_k^+ = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix},$$

$$I_k^- = (I_k^+)^T,$$

$I_k^0 = I_k - \hat{I}_k$;
- $J_k = q_k^{(1)} I_k^0 I_k^+ + (1-q_k^{(1)}) I_k^0 I_k + q_k^{(2)} \hat{I}_k I_k^+ + (1-q_k^{(2)}) \hat{I}_k I_k$,
- $\mathcal{A}_k = \lambda_k \tilde{I}_k + \eta_k \hat{I}_k$;
- $\mathcal{A} = \bigoplus_{k=1}^{n} \mathcal{A}_k = \mathcal{A}_1 \oplus \ldots \oplus \mathcal{A}_n$;
- $\mathcal{B}_k = \lambda_k I_k^+ + \eta_k I_k^-$;
- $\mathcal{B} = \bigoplus_{k=1}^{n} \mathcal{B}_k$;
- $\mathcal{H} = \mathcal{B} - \mathcal{A}$;
- $\mathcal{C}_k = I_1 \otimes \ldots \otimes I_{k-1} \otimes \mu_k J_k \otimes I_{k+1} \otimes \ldots \otimes I_n$;
- $\mathcal{C} = \bigoplus_{k=1}^{n} \mu_k J_k = \sum_{k=1}^{n} \mathcal{C}_k$;
- $\mathcal{M}_k = I_1 \otimes \ldots \otimes I_{k-1} \otimes \mu_k I_k \otimes I_{k+1} \otimes \ldots \otimes I_n$;
- $\mathcal{E} = \bigoplus_{k=1}^{n} \mu_k I_k = \sum_{k=1}^{n} \mathcal{M}_k$;
- $\mathcal{D}_0 = \lambda_0 I$;
- $\mathcal{D}_1 = -\lambda_0 I - \mathcal{E} + \mathcal{H}$;
- $\hat{\mathcal{D}}_1 = -\lambda_0 I + \mathcal{H}$;
- $\mathcal{D}_2 = \mathcal{C}$;
- $$Q^{\{0,\ldots,0\},\{0,\ldots,0\}} = -\lambda_0 I + \mathcal{H},\ Q^{\{0,\ldots,0\},\{1,\ldots,1\}} = \lambda_0 I;$$
- $$Q^{\{0,\ldots,0\},\{l_1,\ldots,l_n\}} = O, \{l_1,\ldots,l_n\} \in \mathcal{L},$$
$$\{l_1,\ldots,l_n\} \notin \{\{0,\ldots,0\},\{1,\ldots,1\}\};$$
- $$Q^{\{l_1,\ldots,l_{k-1},1,l_{k+1},\ldots,l_n\},\{l_1,\ldots,l_{k-1},0,l_{k+1},\ldots,l_n\}} = \mathcal{C}_k;$$
- $$Q^{\{l_1,\ldots,l_n\},\{1,\ldots,1\}} = \lambda_0 I,\ \{l_1,\ldots,l_n\} \neq \{1,\ldots,1\};$$
- $$Q^{\{l_1,\ldots,l_n\},\{l_1,\ldots,l_n\}} = -\lambda_0 I + \mathcal{H} - \sum_{j:l'_j=1, l'_j \in \{l_1,\ldots,l_n\}} \mathcal{M}_j;$$
- $Q_{0,0}$ is the blocking matrix: $Q_{0,0} = (Q^{\{l_1,\ldots,l_n\},\{l'_1,\ldots,l'_n\}})_{\{l_1,\ldots,l_n\},\{l'_1,\ldots,l'_n\} \in \mathcal{L}}$, which can be decomposed as
$$Q_{0,0} = \begin{pmatrix} \tilde{Q}_{0,0} & \tilde{\mathcal{D}}_0 \\ V & \mathcal{D}_0 \end{pmatrix}$$
where the matrix $\tilde{\mathcal{D}}_0$ is a matrix column consisting of 2^n-1 matrices $Q^{\{l_1,\ldots,l_n\},\{1,\ldots,1\}},\ \{l_1,\ldots,l_n\} \in \mathcal{L}$, $\{l_1,\ldots,l_n\} \neq \{1,\ldots,1\}$, and the matrix V is the matrix row consisting of 2^n-1 square matrices of dimension K. More exact description of the matrix V is the following. All the matrices of dimension K, which constitute the matrix V of dimension $K \times K(2^n-1)$, are O except the blocks in positions $2^n - 2^{n-k}$ that are equal to $\mathcal{C}_k, k = \overline{1,n}$.
- $$Q_{1,0} = \begin{pmatrix} O_{1,2^n-1} & \mathcal{D}_2 \end{pmatrix},$$
- $$Q_{0,1} = \begin{pmatrix} O_{2^n-1,1} \\ \mathcal{D}_0 \end{pmatrix}.$$

Denote by Q the block matrix which is the generator of the Markov chain $\xi_t, t \geq 0$.

Lemma: The generator Q has the following block structure:

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & O_{2^n,1} & O_{2^n,1} & O_{2^n,1} & O_{2^n,1} & \dots \\ Q_{1,0} & \mathcal{D}_1 & \mathcal{D}_0 & O & O & O & \dots \\ O_{1,2^n} & \mathcal{D}_2 & \mathcal{D}_1 & \mathcal{D}_0 & O & O & \dots \\ O_{1,2^n} & O & \mathcal{D}_2 & \mathcal{D}_1 & \mathcal{D}_0 & O & \dots \\ O_{1,2^n} & O & O & \mathcal{D}_2 & \mathcal{D}_1 & \mathcal{D}_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}. \quad (1)$$

Proof is implemented by means of analysis of the possible transitions of the Markov chain $\xi_t, t \geq 0$, during infinitesimally small time interval. It is almost straightforward and so it is omitted. Several comments are as follows.

Diagonal entry of the diagonal matrix $\mathcal{A}_k$ defines, up to the sign, the sum of intensities of transition of the n-dimensional process $\{i_t^{(1)},\ldots,i_t^{(n)}\}, t \geq 0$, from the corresponding state under the fixed value $j, j \geq 1$, of the process $j_t, t \geq 0$. The entries of the matrix $\mathcal{B}_k$ define the intensity of the corresponding transitions of the process $\{i_t^{(1)},\ldots,i_t^{(n)}\}, t \geq 0$, between its states.

Under the fixed value $\{l_1,\ldots,l_n\}$ of the process $j_t, t \geq 0$, a diagonal entry of the matrix $Q^{\{l_1,\ldots,l_n\},\{l_1,\ldots,l_n\}}$ defines, up to the sign, the sum of intensities of transitions of the n-dimensional process $\{i_t^{(1)},\ldots,i_t^{(n)}\}, t \geq 0$, from the corresponding state and non-diagonal entries define the intensity of the corresponding transitions of this process between its states.

The entries of the matrix $Q^{\{l_1,\ldots,l_n\},\{l'_1,\ldots,l'_n\}}$ define the intensity of the corresponding transitions of the process $\{i_t^{(1)},\ldots,i_t^{(n)}\}, t \geq 0$, between its states when the state of the process $j_t, t \geq 0$, is changed from $\{l_1,\ldots,l_n\}$ to $\{l'_1,\ldots,l'_n\}$.

In computer realization, it is useful to check correctness of calculation of the generator by summing up the entries of each row and comparing the sum with 0. Formally this well-known property of generator can be formulated as

$$Q\mathbf{e} = \mathbf{0}^T.$$

Theorem 1. Stationary distribution of the Markov chain $\xi_t, t \geq 0,$ exists if and only if the following inequality holds true:

$$\lambda_0 < \sum_{k=1}^{n} \mu_k. \tag{2}$$

In the sequel, we assume that parameters of the network satisfy stability condition (2).

Let us denote the stationary probabilities of the states of the Markov chain $\xi_t, t \geq 0,$ by

$$p(j, i_1, \ldots, i_n) = \lim_{t\to\infty} P\{j_t = j, i_t^{(1)} = i_1, \ldots, i_t^{(k)} = i_k\},$$

$$j = \{l_1, \ldots, l_n\} \in \mathcal{L}, 1, 2, \ldots;\ i_k = \overline{0, N_k}, k = \overline{1, n}.$$

According to the lexicographic enumeration of the components of the Markov chain $\xi_t, t \geq 0,$ which was already exploited above, we combine probabilities $p(j, i_1, \ldots, i_n),\ i_k = \overline{0, N_k}, k = \overline{1, n},$ into probability row vectors $\mathbf{p}_j, j = \{l_1, \ldots, l_n\} \in \mathcal{L}, 1, 2, \ldots$ and the macro-vector

$$\vec{\mathbf{p}} = (\mathbf{p}_{\{0\}}, \mathbf{p}_1, \mathbf{p}_2, \ldots)$$

where

$$\mathbf{p}_{\{0\}} = (\mathbf{p}_{\{0,\ldots,0\}}, \mathbf{p}_{\{0,\ldots,0,1\}}, \ldots, \mathbf{p}_{\{1,\ldots,1\}}).$$

Theorem 2. Stationary probability vectors $\mathbf{p}_{\{l_1,\ldots,l_n\}}, \{l_1, \ldots, l_n\} \in \mathcal{L}, \quad \mathbf{p}_1, \mathbf{p}_2, \ldots$ are calculated in the following way:

- the vector $\mathbf{p}_{\{l_1,\ldots,l_n\}}$ is computed as the block number $\sum_{k=1}^{n} l_k 2^{n+1-k} + 1$ in the block vector $\mathbf{p}_{\{1,\ldots,1\}}\mathcal{F}_1, \quad \{l_1, \ldots, l_n\} \in \mathcal{L}, \ \{l_1, \ldots, l_n\} \neq \{1, \ldots, 1\}$;
- the vectors $\mathbf{p}_j, j \geq 1,$ are computed by

$$\mathbf{p}_i = \mathbf{p}_{\{1,\ldots,1\}} R^i, i \geq 1,$$

where

$$\mathcal{F}_1 = -V(\tilde{Q}_{0,0})^{-1},\ \mathcal{F} = \mathcal{D}_1 + \mathcal{F}_1\tilde{\mathcal{D}}_0;$$

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

$$R^2\mathcal{D}_2 + R\mathcal{D}_1 + \mathcal{D}_0 = O;$$

- the vector $\mathbf{p}_{\{1,\ldots,1\}}$ is the unique solution to the following system of linear algebraic equations

$$\mathbf{p}_{\{1,\ldots,1\}}[\mathcal{F} + R\mathcal{D}_2] = \mathbf{0}_K,$$

$$\mathbf{p}_{\{1,\ldots,1\}}[\mathcal{F}_1 + (I - R)^{-1}]\mathbf{e}_K = 1.$$

Proof of the theorem can be done essentially following to M.F. Neuts' book [5]. However, it is not straightforward due to the necessity to carefully take into account complex behavior of the Markov chain $\xi_t, t \geq 0,$ in the boundary states when not all servers are busy.

This theorem gives a straightforward algorithmic way for calculation the stationary probability vector $\vec{\mathbf{p}}$ which well suits for realization on computer. The problem of solving the matrix equation is extensively addressed in literature, see, e.g., [5],[6]. All other required operations are routine ones. Infrastructure of software "SIRIUS++" and "SIRIUS-C", which is developed for performance evaluation, capacity planning and optimization of telecommunication networks in Belarusian State University and is described in [7],[8], is suitable for organization of the variety of operations with matrices.

IV. Calculation of the Network Performance Measures

Having the stationary probability vectors been computed, we can calculate different performance measures of the queueing network. Formulae for calculation of some of them are given below.

- Average queue length L_0 at the system number 0 is calculated by

$$L_0 = \sum_{i=1}^{\infty} i\mathbf{p}_i\mathbf{e}_{\mathbf{K}} = \mathbf{p}_{\{1\}}R(I - R)^{-2}\mathbf{e}_{\mathbf{K}};$$

- Average number of customers $\tilde{L}_1$ at the system number 0 is calculated by

$$\tilde{L}_0 = [\mathbf{p}_{\{2\}} + \mathbf{p}_{\{3\}} + 2\mathbf{p}_{\{1\}} + \sum_{i=1}^{\infty}(i + 2)\mathbf{p}_i]\mathbf{e}_{\mathbf{K}} =$$

$$= \{\mathbf{p}_{\{2\}} + \mathbf{p}_{\{3\}} + \mathbf{p}_{\{1\}}[R(I - R)^{-2} + 2(I - R)^{-1}]\}\mathbf{e}_{\mathbf{K}};$$

- Average waiting time $W_1^{(0)}$ of a customer in the queueing system number 0 is calculated by

$$W_1^{(0)} = \lambda_1^{-1}L_0,$$

- Average sojourn time $\tilde{W}_1^{(1)}$ of a customer in the queueing system number 0 is calculated by

$$\tilde{W}_1^{(0)} = \lambda_1^{-1}\tilde{L}_0;$$

- Joint distribution of the number of customers in the systems $k^*, k = \overline{1, n},$ is defined by the row vector $\boldsymbol{\theta}$ computed by formula

$$\boldsymbol{\theta} = \mathbf{p}_{\{1,\ldots,1\}}[\mathcal{F}_1 + (I - R)^{-1}];$$

- Stationary distribution of the number of customers in the system k^* is given by the vector $\boldsymbol{\theta}^{(k)}$ having components

$$\boldsymbol{\theta}_i^{(k)} = \boldsymbol{\theta}\tilde{\mathbf{e}}_k^{(i)}, i = \overline{0, N_k}, k = \overline{1, n};$$

- Average number of customers L_k in the system k^* is calculated by

$$L_k = \sum_{i=1}^{N_k} i\boldsymbol{\theta}_i^{(k)}, k = \overline{1, n};$$

- Average total number of customers L in the network is calculated by

$$L = \tilde{L}_0 + \sum_{k=1}^{n} L_k;$$

- Probability $\tilde{P}_{loss}^{(k)}$ that arbitrary own customer arriving to the system k^* is rejected because the buffer is full is calculated by formula

$$\tilde{P}_{loss}^{(k)} = \boldsymbol{\theta}\tilde{\mathbf{e}}_k^{(N_k)}, \quad k = \overline{1, n};$$

- Probability $\tilde{P}_{loss}$ that arbitrary customer arriving to this network is rejected because the corresponding buffer is full is calculated by formula

$$\tilde{P}_{loss} = \frac{\sum\limits_{k=1}^{n} \lambda_k \tilde{P}_{loss}^{(k)}}{\sum\limits_{k=1}^{n} \lambda_k};$$

- Probabilities $P_{loss}^{(k)}$ that arbitrary customer arriving to the system k^* will be rejected due to desire of agent or because the buffer is full and is calculated by formula

$$P_{loss}^{(k)} = \boldsymbol{\theta}_{N_k}^{(k)} + (1-q_1^{(k)})\boldsymbol{\theta}_0^{(k)} + (1-q_2^{(k)})(1-\boldsymbol{\theta}_0^{(k)} - \boldsymbol{\theta}_{N_k}^{(k)}),$$
$$k = \overline{1,n}.$$

- Probability P_{loss} that arbitrary customer arriving to this network will not get service by any agent is computed by formula

$$P_{loss} = \mathbf{p}_{\{1,\dots,1\}}(I + R(I-R)^{-1})\mathbf{e}_{\mathbf{K}} \sum_{k=1}^{n} \frac{\mu_k}{\mu} P_{loss}^{(k)} +$$
$$+ \sum_{(\{l_1,\dots,l_n\})\in\mathcal{L}} \mathbf{p}_{\{l_1,\dots,l_n\}}\mathbf{e}_{\mathbf{K}} \times$$
$$\times \left(1 - \prod_{r=1,l_r=0}^{n} (1 - P_{loss}^{(k_r)})\right),$$

$\mu = \sum\limits_{k=1}^{n} \mu_k$.

- Average waiting time $W_1^{(0)}$ of a customer in the queueing system number 0 is calculated by

$$W_1^{(0)} = \lambda_0^{-1} L_0;$$

- Average average sojourn time $\tilde{W}_1^{(0)}$ of a customer in the queueing system number 0 is calculated by

$$\tilde{W}_1^{(0)} = \lambda_0^{-1} \tilde{L}_0;$$

These formula follow from well-known Little's formulae.

- Distribution function $W^{(k)}(x)$ of the waiting time in the queueing system number k^* is calculated by

$$W^{(k)}(x) = P_{loss}^{(k)} + \sum_{i=0}^{N_k-1} \boldsymbol{\theta}_i^{(k)} E_k^{(i)}(x),$$

where

$$E_k^{(0)}(x) = 1,$$

$$E_k^{(i)}(x) = \int_0^x \eta_k \frac{(\eta_k t)^{i-1}}{(i-1)!} e^{-\eta_k t} dt, i \geq 1,\ k = \overline{1,n};$$

- Conditional distribution function $\bar{W}^{(k)}(x)$ of waiting time for customers who are not rejected in the queueing system number k^* is calculated by

$$\bar{W}^{(k)}(x) = \frac{W^{(k)}(x) - P_{loss}^{(k)}}{1 - P_{loss}^{(k)}}, k = \overline{1,n};$$

- Mean sojourn time $\tilde{W}_1^{(k)}$ in the system number k^* and the conditional mean sojourn time $\tilde{\tilde{W}}_1^{(k)}$ for customers who are not rejected are calculated by

$$\tilde{W}_1^{(k)} = \sum_{i=0}^{N_k-1} \frac{i+1}{\eta_k} \boldsymbol{\theta}_i^{(k)},$$

$$\tilde{\tilde{W}}_1^{(k)} = \frac{\tilde{W}_1^{(k)}}{1 - P_{loss}^{(k)}}, k = \overline{1,n}.$$

- Average sojourn time V of a customer in the queueing network is approximately computed by

$$V = \tilde{W}_1^{(0)} + \mathbf{p}_{\{1,\dots,1\}}(I + R(I-R)^{-1})\mathbf{e}_{\mathbf{K}} \sum_{k=1}^{n} \frac{\mu_k}{\mu} \tilde{\tilde{W}}_1^{(k)} +$$
$$+ \sum_{(\{l_1,\dots,l_n\})\in\mathcal{L}} \mathbf{p}_{\{l_1,\dots,l_n\}}\mathbf{e}_{\mathbf{K}} \min_{r=\overline{1,n},\ l_r=0} \tilde{\tilde{W}}_1^{(r)}.$$

The first summand in the right side of this formula is average sojourn time in the system number 0. The second summand gives average sojourn time in the system, to which the customer will be assigned when he is assigned to a single server. The last summand takes into account that the average sojourn time after getting service in the system number 0 is equal to the minimum of the average sojourn times in systems, involved to its service. We speak here about the approximate computation only because expectation of minimum of several randoms, generally speaking, is not equal to the minimum of expectations.

The formula becomes exact if we replace the value $\min\limits_{r=\overline{1,n},\ l_r=0} \tilde{\tilde{W}}_1^{(r)}$ by the expectation of the minimum of random variables having distribution being the weighted sum of Erlangian distributions with the jump at point zero. Because such distribution is a partial case of a slight modification of PH (Phase type) distribution with a jump at point zero, the following lemma can be exploited.

Lemma 2. Let $\xi_k, k = \overline{1,m}$, be m independent identically distributed random variables having PH distribution defined by the irreducible representation $(\boldsymbol{\beta}, S)$.

Then a random $\xi = \min\limits_{k=\overline{1,m}} \xi_k$ has PH distribution defined by the irreducible representation $(\boldsymbol{\beta}^{\otimes m}, S^{\oplus m})$.

V. Conclusion

The process of user query processing in MAS is described in terms of the queueing network. Tree-like structure of the network topology allows to get the steady state-distribution of the network state in the elegant exact analytic form via the application of the tool of multi-dimensional Markov chains. Main performance measures of the network are calculated.

The results are extendable to the cases where the input and service processes have more complicated nature, e.g., they are modeled by the Markovian Arrival Process (MAP) and Markovian Service Process (MSP) correspondingly. Modifications to the network where the customer is blocked in the case of the full intermediate buffer, where the service can be provided with the error and where the servers are subject to breakdowns and recovering be can investigated analogously.

VI. Acknowledgments

This research was supported in part by Ministry of Information and Communication (MIC) Korea, under the IT Foreign Specialist Inviting Program (ITSIP), ITSOC, International Cooperative Research by Ministry of Science and Technology, KOTEF, and 2nd stage Brain Korea 21.

References

[1] Horling B., Lesser V.: Using Queueing Theory to Predict Organizational Metrics. AAMAS'06 May 8-12 2006, Hakodate, Japan. 1098-1100.

[2] Gnanasambandam N., Lee S.C., Kumara S.R.T.: An Autonomous Performance Control Framework for Distributed Multi-Agent Systems: a Queueing Theory Based Approach. AAMAS'05 July 25-29 2005, Utrecht, Netherlands. 1313-1314.

[3] Gnanasambandam N., Lee S.C., Gautam N., Kumara S.R.T., Peng W., Manikonda V., Brinn M., Greaves M.: Reliable MAS Performance Prediction Using Queueing Models. IEEE First Symposium on Multi-Agent Security and Survivalibility, 2004. 55-64.

[4] Gnanasambandam N.: Survivability of Multi-Agent Systems. AAMAS'05 July 25-29 2005, Utrecht, Netherlands. 1376.

[5] Neuts M.F.: Matrix-Geometric Solutions in Stochastic Models - An Algorithmic Approach. Johns Hopkins University Press (1981).

[6] Latouche G., Ramaswami V.: Introduction to Matrix Geometric Methods in Stochastic Modelling. Philadelphia, USA: ASA-SIAM Series in Statistics and Applied Probability (1999)

[7] Dudin A.N., Tsarenkov G.V., Klimenok V.I.: Software "SIRIUS++" for performance evaluation of modern communication networks. Modelling and Simulation 2002. 16th European Simulation Multi-conference, Darmstadt, (2002) 489–493.

[8] Dudin A.N., Klimenok V.I., Tsarenkov G.V., Semenova O.V., Birukov A.A.: Software "SIRIUS-C" for synthesis of optimal control by queues. Proceedings of 11-th International Conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA 2004), 13-16 June 2004, Magdeburg, Germany. (2004) 123-129