Optimization of Buffers Capacity in Tandem Queueing Systems with Batch Markovian Arrivals Process†

김제숭*, 이석준**
(Chesoong Kim, Seokjun Lee)

Abstract Tandem queueing systems well suit for modeling many telecommunication systems. Recently, very general BMAP/G/1/N/1 → ●/PH/1/M-1 type tandem queues were constructively studied. In this paper we illustrate application of the obtained results for optimization of a buffer pool design.

Key Words : Stochastic Models, Queueing Networks, Queueing Systems, Markov Models, and Optimization.

1. Introduction

Open queueing networks and tandem queues as their important special case are widely used in capacity planning and performance evaluation of computer and communication systems, service centers, manufacturing systems, etc. Theory of tandem queues is well developed, for references see, e.g. (Gnedenko and Koenig 1983). Apart from approximate analysis of the tandems used, e.g. (Heindl 2001; Heindl 2003; Ferng and Chang 2001a; Ferng and Chang 2001b), there are papers presented, that use analytical methods for such investigation (Breuer et al. 2004; Gomez Corral 2002a; Gomez-Corral 2002b; Klimenok et al. 2005).

This paper exploits results on BMAP/G/1/N/1 → ●/PH/1/M-1 type tandems with losses and blocking at the second phase, see (Breuer et al. 2004; Klimenok et al. 2005), to calculate performance characteristics of the tandem queues in case of finite buffer at the first phase (i.e. $N \leq +\infty$). The analytical results were implemented in the Sirius-C software package. This allows authors to perform optimization of some real-life networks or their fragments. In the following sections we outline major analytical results necessary to calculate system characteristics used in optimization criterion as well as numerical examples of optimization in these tandem systems.

2. Mathematical Model

We consider tandem queue consisting of two queues (phases). The first queue is of

† This research has been supported by Sangji University Research Fund 2006. Also this work was partially supported by the Korea Research Foundation(KRF- 2006-521-D00569)
* Dept of Industrial Engineering, Sangji University
** School of Business Administration, Sangji University

the BMAP/G/1/N type, i.e., it has a single server, finite buffer of capacity N, general service time distribution function B(t) having the finite initial moment $b_1\int_0^\infty t dB(t)$. The input flow is described by the BMAP(Batch Markovian Arrival Process). The BMAP was introduced by D. Lucantoni (Lucantoni 1991) as extension of the models of versatile flows by M. Neuts (Neuts 1989) and N flows by (Ramaswami 1980). The class of the BMAP includes many previously considered input flows such as, e.g., the stationary Poisson (M), Erlangian (Ek), Hyper-Markovian (HM), Phase-Type (PH), Markov Modulated Poisson Process (MMPP), etc. As opposed to recurrent (GI) flows and PH flow in particular, the BMAP flow is correlated one. It makes it extremely useful for modeling the real flows in modern telecommunication networks. The BMAP is defined by means of the directing (underlying) process $\tilde{v}_t, t \geq 0$, which is a continuous time Markov chain with the state space $\{0,1,\cdots,W\}$. Arrival of customers occurs in batches at the epochs when the process $\tilde{v}_t, t \geq 0$, has jumps. The intensities of jumps from one state into another one, which are accompanied by arrival of a batch consisting of k customers, are combined into the matrices $D_k, k \geq 0$, of size $(W+1)\times(W+1)$. The matrix generating function of these matrices is $D(z)=\sum_{k=0}^{\infty} D_k z^k, |z|\leq 1$. The matrix $D(1)$ is an infinitesimal generator of the process $\tilde{v}_t, t \geq 0$. The vector $\vec{\theta}$ of this process stationary distribution satisfies equations $\vec{\theta}D(1) = \vec{0}, \vec{\theta}\vec{e} = 1$. Here and in the sequel $\vec{0}$ are zero row vectors and $\vec{e}$ is column vector of appropriate size consisting of units. In case the dimensionality of the vector is not clear from context, it is indicated as a subscript, e.g. $\vec{e}_{\overline{W}}$ denotes the unit column vector of dimensionality $\overline{W} = W + 1$.

The average intensity λ (fundamental rate) of the BMAP is defined as

$$\lambda = \vec{\theta}D'(z)\big|_{z=1}\vec{e}$$

The intensity λ_g of groups arrival is defined as

$$\lambda_g = \vec{\theta}(-D_0)\vec{e}$$

Variance v of intervals between the groups arrival is calculated as:

$$v = 2\lambda_g^{-1}\vec{\theta}(-D_0)^{-1}\vec{e} - \lambda_g^{-2}$$

The correlation coefficient c_{cor} of intervals between the successive groups arrival is calculated as

$$c_{cor} = \left[\lambda_g^{-1}\vec{\theta}(-D_0)^{-1}(D(1)-D_0)(-D_0)^{-1}\vec{e} - \lambda_g^{-2}\right]/v$$

For more information about the BMAP and related research see (Lucantoni 1991) and the overviewing paper by S. Chakravarthy (Chakravarthy 2001). The second queue has a finite buffer of capacity $M-1, M \geq 1$. The single server is characterized by the PH-type service time distribution having an irreducible representation (β, S). Here β is the stochastic row-vector of dimension K and S is $K \times K$ matrix having the negative diagonal and non-negative non-diagonal entries, such as the column vector $S_0 = -S\vec{e}$ is non-negative and has at least one positive entry. The average service time is defined as

$\beta(-S)^{-1}\vec{e}$. In case the entering batch of customers finds insufficient number of places in a buffer (or the buffer is already full), the suitable number of customers from the batch joins a queue while the rest (or the whole group) leaves the system forever (is lost at the first phase). In case the customer completes the service at first phase and meets the buffer before the second phase is busy, we consider two cases:

- this customer leaves the system forever and is considered to be lost at the second phase.
- this customer waits for a buffer to have empty cell and prevents, during waiting period the other customers from being served at the first phase (i.e. service is blocked).

For optimization tasks we calculate the steady state distribution of the number of customers and loss probability in the tandem system.

2.1 Embedded Markov Chain

Following (Klimenok et al. 2005) for tandem with losses at the second phase, we consider the process $\varsigma_t^{(1)} = \{\tilde{i}_t, \tilde{j}_t, \tilde{v}_t, \tilde{\eta}_t\}, t \ge 0, 0 \le \tilde{i}_t \le N+1$, $0 \le \tilde{j}_t \le M, 0 \le \tilde{v}_t \le W, 1 \le \tilde{\eta}_t \le K$, where $\tilde{i}_t$ is the number of customers at the first phase, $\tilde{j}_t$ is the number of customers at the second phase, $\tilde{v}_t$ is the state of the BMAP directing process, $\tilde{\eta}_t$ is the state of the directing process of the PH service at epoch $t, t \ge 0$.

Following (Breuer et al. 2004) for tandem with blocking at the second phase we consider the process $\varsigma_t^{(2)} = \{\tilde{i}_t, \tilde{j}_t, \tilde{v}_t, \tilde{\eta}_t, \tilde{\chi}_t\}, t \ge 0, 0 \le \tilde{i}_t \le N+1$, $0 \le \tilde{j}_t \le M, 0 \le \tilde{v}_t \le W, 1 \le \tilde{\eta}_t \le K$, $\tilde{\chi}_t = 0,1$ where $\tilde{i}_t$, $\tilde{j}_t$, $\tilde{v}_t$, $\tilde{\eta}_t$ have the same meanings as in the tandem with losses, and $\tilde{\chi}_t = 0$ if the first server is working or waiting for a customer at moment t or $\tilde{\chi}_t = 1$ if it is blocked.

The processes $\varsigma_t^{(1)}, t \ge 0$, and $\varsigma_t^{(2)}, t \ge 0$, are non-Markovian. Thus, to investigate these processes, first we consider the embedded (at the service completion epochs $t_n, n \ge 1$ at the first phase) Markov chains.

Due to fact that, in tandem with blocking a customer, which causes a blocking, is not counted either at the second or at the first phases, the multi-dimensional embedded Markov chains will have the same components

$$\varsigma^{(1)}, \varsigma_t^{(2)} = \{i_t, j_t, v_t, \eta_t\}, n \ge 1$$

where

$$i_n = \tilde{i}_{t_n+0}, 0 \le i_n \le N, \quad j_n = \tilde{j}_{t_n-0}, 0 \le j_n \le M$$

$$v_n = \tilde{v}_{t_n}, 0 \le v_n \le W, \eta_n = \tilde{\eta}_{t_n-0}, 1 \le \eta_n \le K$$

Note that if $j_n = 0$, and then value of the component $\eta_n = 0$ is not defined. Introduce the stationary state probabilities of these Markov chains as:

$$\pi(i,0,\eta) = \lim_{n\to\infty} P\{i_n = i, j_n = 0, v_n = v\}$$

$$\pi(i,j,v,\eta) = \lim_{n\to\infty} P\{i_n = i, j_n = j, v_n = v, \eta_n = \eta\}$$

$$0 \le i \le N, \quad 0 \le j \le M,$$

$$0 \le v \le W, 1 \le \eta \le K$$

Vectors of stationary state probabilities are defined as follows:

$$\pi(i,0) = \{\pi(i,0,0), \pi(i,0,1), \cdots, \pi(i,0,W)\}, i = \overline{0,N}$$

$$\pi(i,j) = \{\pi(i,j,0,1), \cdots, \pi(i,j,0,K),$$

$$\pi(i,j,1,1), \cdots, \pi(i,j,1,K),$$

$$\pi(i,j,W,1), \cdots, \pi(i,j,W,K)\}, i \ge 0, j\rangle 0$$

$$\pi_i = \{\pi(i,0), \pi(i,1), \cdots, \pi(i,M)\}$$

Conditions for this stationary distribution existence as well as definition of transition probabilities matrices can be found in (Breuer et al. 2004; Klimenok et al. 2005).

2.2 Steady State Distribution at Arbitrary Epochs

Having calculated vectors $\pi_i, i = \overline{0,N}$, by means of algorithms suggested in (Breuer et al. 2004; Klimenok et al. 2005), we could calculate steady-state distributions at arbitrary epochs. For the tandem with losses at the second phase, these stationary probabilities and their vectors we define as follows:

$$p^{(1)}(i,0,v) = \lim_{t\to\infty} P\{\tilde{i}_t = i, \tilde{j}_t = 0, \tilde{v}_t = v\}$$

$$p^{(1)}(i,j,v,\eta) = \lim_{t\to\infty} P\{\tilde{i}_t = i, \tilde{j}_t = j, \tilde{v}_t = v, \tilde{\eta}_t = \eta\}$$

$0 \le i \le N+1$, $0 \langle j \le M$, $0 \le v \le W, 1 \le \eta \le K$

Enumerate the states of process $\varsigma_t^{(1)}, t \ge 0$, in the lexicographic order and form the probability vectors $p^{(1)}(i,j), i = \overline{o,N+1}, j = \overline{1,M}$, of corresponding probabilities and $\vec{p}_i^{(1)} = \{p^{(1)}(i,0), \cdots, p^{(1)}(i,M)\}$. For the tandem with blocking at the second phase these stationary probabilities and their vectors we define as follows:

$$p^{(2)}(i,0,v) = \lim_{t\to\infty} P\{\tilde{i}_t = i, \tilde{j}_t = 0, \tilde{v}_t = v\}$$

$$p^{(2)}(i,j,v,\eta) = \lim_{t\to\infty} P\{\tilde{i}_t = i, \tilde{j}_t = j, \tilde{v}_t = v, \tilde{\eta}_t = \eta\}$$

$$p^{(2)}(i,M,v,\eta,\chi) = \lim_{t\to\infty} P\{\tilde{i}_t = i, \tilde{j}_t = j, \tilde{v}_t = v, \tilde{\eta}_t = \eta, \tilde{\chi}_t = \chi\}$$

$0 \le i \le N+1$, $0 \langle j \le M$, $0 \le v \le W, 1 \le \eta \le K$, $\chi = 0,1$

Enumerate the states of process $\varsigma_t^{(2)}, t \ge 0$, in the lexicographic order and form the probability vectors $p^{(2)}(i,j), i = \overline{o,N+1}, j = \overline{1,M-1}$ and $p^{(2)}(i,M,0), p^{(2)}(i,M,1)$ of corresponding probabilities and denote $\vec{p}_i^{(2)} = \{p^{(2)}(i,0), \cdots, p^{(2)}(i,M-1), p^{(2)}(i,M,0) + p^{(2)}(i,M,1)\}$, $i = \overline{o,N+1}$.

We refer to the papers (Breuer et al. 2004; Klimenok et al. 2005), where the analytical formulas for calculation of steady-state probability vectors $\vec{p}_i^{(r)}, i = \overline{0,N+1}, r = 1,2$, are presented.

2.3 Tandem Performance Characteristics

The most important characteristics for investigated systems are:

- the loss probability at the first phase $P_{loss\,1}^{(1)}$, $P_{loss\,1}^{(2)}$ for tandems with losses and blocking respectively;
- the loss probability at the second phase $P_{loss\,1}^{(2)}$ for the tandem with losses only.

These loss probabilities can be calculated as follows:

$$P_{loss1}^{(1)} = 1 - \lambda^{-1} \sum_{i=0}^{N} \vec{p}_i^{(1)} \sum_{k=0}^{N+1-i} (k+i-N-1) \tilde{D}_k \vec{e}$$

$$P_{loss1}^{(2)} = 1 - \lambda^{-1} \sum_{i=0}^{N} \vec{p}_i^{(2)} \sum_{k=0}^{N+1-i} (k+i-N-1) \tilde{D}_k \vec{e}$$

$$P_{loss\,1}^{(2)} = \sum_{i=0}^{\infty} \pi(i,M) \vec{e}$$

where

$$\tilde{D}_k = \begin{bmatrix} D_k & 0 \\ 0 & I_M \otimes D_k \otimes I_K \end{bmatrix}, k \ge 0$$

$\otimes$-denotes Kroneker product of matrices, ..

denotes identity matrix of appropriate dimension. or the system with blocking it is not possible to lose the customer at the second phase, thus we consider $P_{loss\,2}^{(2)}=0$ in numerical examples for optimization criterion without loss of generality.

3. Optimization Criterions and Numerical Example

We consider two strategies of buffers capacity planning:

- buffer capacities at the first and at the second phases are independent;
- memory is shared between first and second phases, thus total buffers capacity at both phases is constant.

To perform optimization task, we use the following cost criterion

$$C=C_1N+C_2(M-1)+C_{loss1}\lambda P_{loss1}+C_{loss2}\lambda(1-P_{loss1})P_{loss2}$$

where $C_{loss\,1}$ is penalty for a customer loss at the first phase, $C_{loss\,2}$ is penalty for a customer loss at the second phase, C_1 is cost of the cell at the first buffer; C_2 is cost of the cell at the second buffer.

For the case of shared buffer capacities we used the following criterion

$$C=C_{loss1}\lambda P_{loss1}+C_{loss2}\lambda(1-P_{loss1})P_{loss2}$$

3.1 Parameters of Tandems for Numerical Examples

We use the following parameters for numerical examples. Input flow is a BMAP-flow of intensity, $\lambda=10$, intensity of groups $\lambda_g=5$, with correlation $c_{cor}=0.199999$, and variation $c_{var}=12.2732$.

Matrices defining the BMAP are as follows

$$D_0=\begin{bmatrix}-6.74538 & 5.45412\times10^{-6}\\ 5.45412\times10^{-6} & -0.219455\end{bmatrix}$$

$$D_1=D_3=\begin{bmatrix}2.01021 & 0.0134084\\ 0.036728 & 0.0291068\end{bmatrix}$$

$$D_2=\begin{bmatrix}2.68027 & 0.0178778\\ 0.0489707 & 0.038809\end{bmatrix}$$

$$D_k=0,\ k\ge 4$$

Service time distribution at the first phase is degenerate with mean $T=0.07$. Phase-type service time distribution at the second phase has the following parameters

$$S=\begin{bmatrix}-10 & 0\\ 0 & -40\end{bmatrix},\ \vec{\beta}=[0.7\quad 0.3]$$

These parameters make mean service time at the second phase equal to 0.0775.

3.2 First Numerical Example

In this example we consider independent buffer capacities at the first and at the second phases. Cost parameters in the criterion are specified as follows. Cost C_1 of a buffer cell at the first phase is 1, cost C_2 of a buffer cell at the second phase is 2, penalty $C_{loss\,1}$ for customer loss at the first phase is 9, penalty $C_{loss\,2}$ for the customer loss at the second phase is 15. Figures 1 and 2 present values of the optimization criterion for various values of buffer size at the first and the second phases.

It can be seen that, for the fixed above values of the cost parameters, the value of the criterion is smaller for the model with customer's loss at the second phase. Optimal value of the criterion is equal to 39.334 at point N = 5, M = 7.

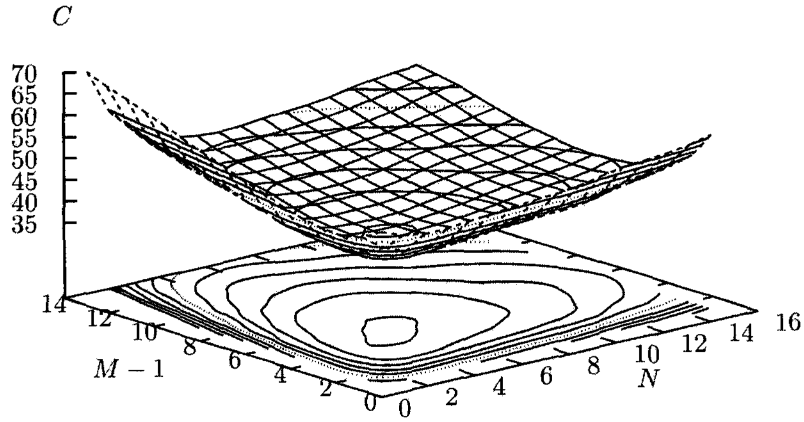

Figure 1: Dependency of criterion value on the buffers capacity for tandem with losses

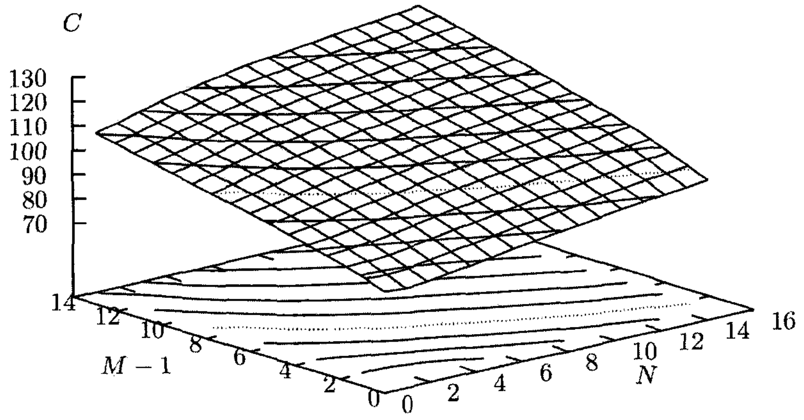

Figure 2: Dependency of criterion value on the buffers capacity for tandem with blocking

3.3 Second Numerical Example

In this numerical example we consider the same costs as in the previous example and shared buffer capacities, i.e. it is assumed, that the total capacity M + N of two buffers is 16 and we share it between two phases. The result is presented on Figure 3. Optimal value of the criterion is equal to 17.90 and M and N are equal to 7 and 9 respectively.

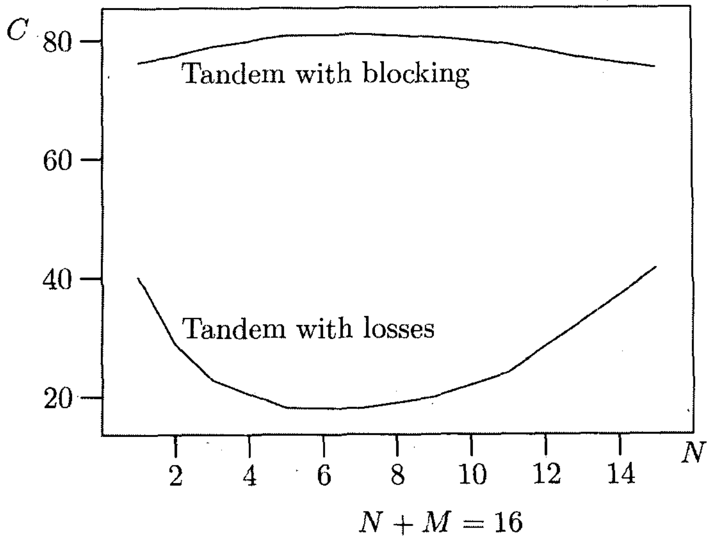

Figure 3: Dependency of criterion value on the first buffer capacity for N +M = 16

4. Conclusion

This paper gives examples of optimization of the tandem queueing models, which can be used to support operational maintenance of the telecommunication networks or their fragments. It can also be used to verify approximating results during investigation of the real-life network objects. This paper demonstrates the definite advantages of the ability to analyze networks with correlated traffic, which is the case in modern telecommunication systems. Given examples show that application of such models can save up to 200% of operational costs. This work addresses practical applicability of the analytical formulas and is aimed to help to further application of the queueing networks in the investigation of modern telecommunication systems like cellular phones networks, Internet service providers' networks and many others under considerations of the complex heterogeneous nature of the traffic.

References

[1] Breuer, L., Dudin A.N., Klimenok V.I. and G.V.Tsarenkov. 2004, "A Two-Phase BMAP/G/1/N → PH/1/M-1 System with Blocking", *Automation and Remote Control*, Vol. 1, No. 1, pp. 104-115.

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

[3] Ferng, H.W. and J.F. Chang. 2001a, "Connection wise end-to-end performance analysis of queueing networks with MMPP inputs", *Performance Evaluation*, Vol. 43, pp. 39-62.

[4] Ferng, H.W. and J.F. Chang. 2001b, "Departure processes of BMAP\G\1 queues." *Queueing Systems*, Vol. 39, pp. 109-135.

[5] Gnedenko, B. W. and D. Konig. 1983. *Handbuch der Bedienungstheorie*. Akademie

Verlag. Berlin.

[6] Gomez-Corral, A. 2002a. "A tandem queue with blocking and Markovian arrival process." *Queueing Systems*, Vol. 41, pp. 343-370.

[7] Gomez-Corral, A. 2002b. "On a tandem G-network with blocking." *Advances in Applied Probability*, Vol. 34, No.3, pp. 626-661.

[8] Heindl, A. 2001. "Decomposition of general tandem networks with MMPP input." *Performance Evaluation*, Vol. 44, pp. 5-23.

[9] Heindl, A. 2003. "Decomposition of general queue networks with MMPP inputs and customer losses." *Performance Evaluation*, Vol. 51, pp. 117-136.

[10] Klimenok, V.I., Breuer, L., Tsarenkov, G.V. and A.N. Dudin. 2005. "Tandem queue BMAP\G\1\N → •\PH\1\M with losses." *Performance Evaluation*, Vol. 61, pp. 17-40.

[11] Lucantoni, D.M. 1991. "New results on the single server queue with a batch Markovian arrival process", *Commun. Statist.-Stochastic Models*, Vol. 7, pp. 1-46.

[12] Neuts, M.F. 1989. *Structured stochastic matrices of M\G\1 type and their applications*. Marcel Dekker. New York.

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

[14] Ramaswami, V. 1980. "The N\G\1 queue and its detailed analysis." *J. Appl. Prob.* Vol. 12, pp. 222-261.

김 제 숭 (Chesoong Kim)

Che-Soong Kim received his Master degree and PhD in Engineering from Department of Industrial Engineering at Seoul National University in 1989 and 1993, respectively. He is currently Professor of the Department of Industrial Engineering at Sangji University. His current research interests are in stochastic process, queueing theory, with particular emphasis on computer and wireless communication network, queueing network modeling and their applications, and reliability analysis. He has published around 50 papers in internationally refereed journals such as *Reliability Engineering & System Safety, International Transactions in Operational Research, International Journal of Reliability, Quality and Safety Engineering, Mathematical & Computer Modeling, Australian & New Zealand Journal of Statistics, Statistics & Probability Letters, Automation & Remote Control, Queueing Systems, Operations Research Letter and Annals of Operations Research.* Prof. Kim has been the recipient of a number of grants from the *Korean Science and Engineering Foundation* (KOSEF) and *Korea Research Foundation* (KRF).

이 석 준 (Seokjun Lee)

Seokjun Lee received his B.S. and M.S. degree in Industrial Engineering at Sangji University, and his PhD in School of Business Administration at Sangji University in 2007. His current field of scientific interests is in e-commerce, queueing network modeling and their applications, collaborative filtering on the recommender system and data mining.