SERVICE STRATEGIES IN TANDEM SERVER NETWORKS WITH FEEDBACK AND BLOCKING

Walenty Oniszczuk

Faculty of Computer Science, Bialystok University of Technology, Bialystok, Poland

Abstract: In this paper, we consider specialized tandem server networks with finite buffer capacities, feedback and intelligent service strategy, which are one of the key elements in ensuring quality of service in computer systems. Here, the two strategies of tasks service are presented and compared. Generally, in this paper two models of linked computer servers with blocking and with feedback service according to the HOL priority scheme are investigated. These kinds of models, describe behaviour of computer tandem networks, exposed to open Markovian queuing models with blocking. These models which are illustrated below are very accurate, derived directly from two-dimensional state graphs. In our examples, the performance is calculated and numerically illustrated by regulating intensity of the input flow and varying buffer capacities.

Keywords: feedback and blocking, service strategy, network performance analysis

1. Introduction

Queuing network models have been widely used to evaluate the performance of computer systems and communication networks. Queuing theory was developed to understand and to predict the behaviour of real life systems. Queuing networks models with finite capacity queues and blocking and feedback have been introduced and applied as more realistic models of systems with finite capacity resources and with population constrains [1, 2, 5, 12, 13]. Over the years high quality research has appeared in diverse journals and conference proceeding in the field of computer science, traffic engineering and communication engineering. However, there are still many important and interesting finite capacity queues under various blocking mechanisms and synchronization constraints to be analyzed [3, 4, 6, 9-11].

Most of specialized computer systems are connection oriented, which are also known as linked in series. There are many blocking models of linked in series networks that can be used to provide insight into the performance of those networks.

Zeszyty Naukowe Politechniki Białostockiej. Informatyka, vol. 9, pp. 93-104, 2012.

Blocking models, if they can be solved efficiently, are often used in network planning and dimensioning. Due to obvious resource constrains, realistic models have finite capacity buffers, where the queue length cannot exceed its arbitrary maximum threshold. When the queue length reaches its capacity, the buffer and the server are said to be full (blocking factors). Queuing network models (QNMs) with finite capacity queues and blocking provide powerful and practical tools for performance evaluation and predication of discrete flow systems in computer systems and networks.

Despite all the research done so far, there are still many important and interesting models to be studied. For example, finite capacity queues under various blocking mechanisms and synchronization constraints, such as those involving feedback service with priority scheduling, where in a feedback queue, a task with a fixed probability can return to the previous node immediately after its service at the current node.

The paper is organized as follows. Section 2, describes the analytical models of a tandem network with two proposed feedback service strategy in the first server. Models implementation and numerical examples are described in Section 3. And finally, conclusions are drawn in Section 4.

2. Models description and notations

The most common queuing models assume that the interarrival and service times are exponentially distributed. Equivalently, the arrival and service processes follow a Poisson distribution. That is, if the interarrival times are exponentially distributed then the number of arrivals at the system follows a Poisson distribution. Similarly, for the service process.

We consider an open queuing model of tandem networks with a single task class and three stations: a source, station A and B (see Figure 1). Tasks arrive from the source at station A according to the Poisson process with rate λ. The service rates at each station are μ_1^A, μ_2^A (for feed backed tasks) and μ^B, respectively. After service completion at station A, the task proceeds to station B. Once it finishes at station B, it gets sent back to station A for re-processing with probability σ. We are also assuming that tasks are leaving the network with $1-\sigma$ probability. Service at each station is provided by a single exponential server.

In the first model, a feed backed task is served at station A according to a non-preemptive priority scheme (Head-of-Line (HoL) priority discipline) independently of all other events, where a task cannot get into service at station A (it waits at station B - blocking factor) until the task currently in service is completed. Once finished, each re-processed task departs from the network. The successive service times at both

stations are assumed to be mutually independent and they are independent of the state of the network.

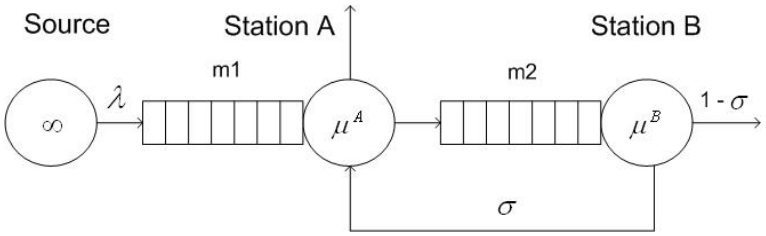

Fig. 1. Illustration of the three-station network model with feedback

Between station A and station B is a common waiting buffer with finite capacity $m2$. When this buffer is full, the accumulation of new tasks from the first station is temporarily suspended. Similarly, if the first buffer (with capacity $m1$) ahead of the first station is full, then the source station is blocked.

In theory, any Markov model can be solved numerically. In particular, solution algorithm for Markov queuing networks with blocking and priority feedback service is a five-step procedure:

a) Definition of the tandem station state space and choosing its state space representation.
b) Enumerating all the transitions that can possible occur among the states.
c) Definition of the transition rate matrix that describes the network evaluation that means generating the transition rate.
d) Solution of linear system of the global balance equations to derive the stationary state distribution vector (computing appropriate probability vector).
e) Computation, from the probability vector, of the average performance indices.

2.1 First service strategy for feed backed task: Head of Line

According to general assumptions, a continuous-time homogeneous Markov chain can represent this tandem network. The queuing network model reaches a steady-state condition and the underlying Markov chain has a stationary state distribution. The underlying Markov process of a queuing network with finite capacity queues has finite state space. If the numbers of tasks located simultaneously in the network in the first and second servicing stations are denoted by i and j plus an index k depicted

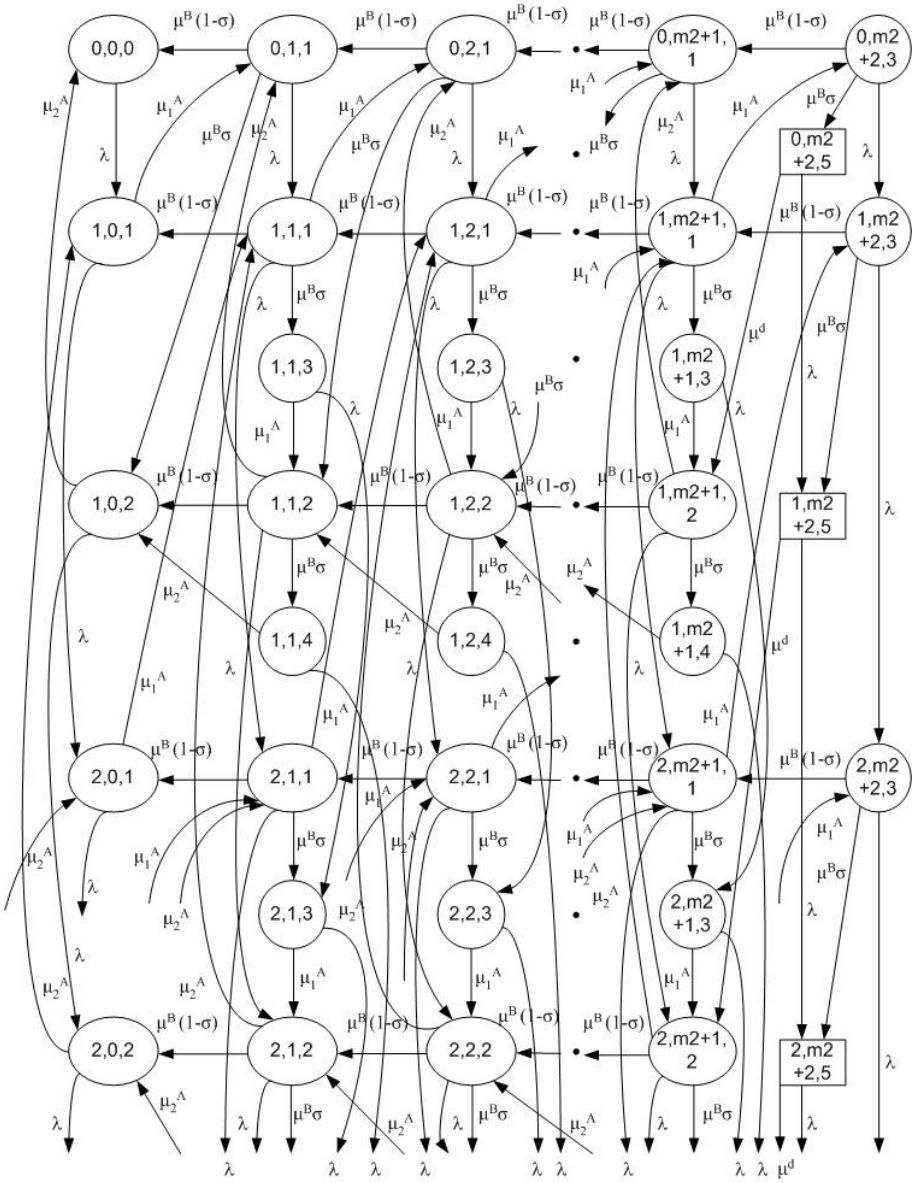

Fig. 2. Two-dimensional tandem network state diagram (first part)

the state of the each server, then a Markov model with two-dimensional state space is defined in this paper (see Figure 2 and Figure 3).

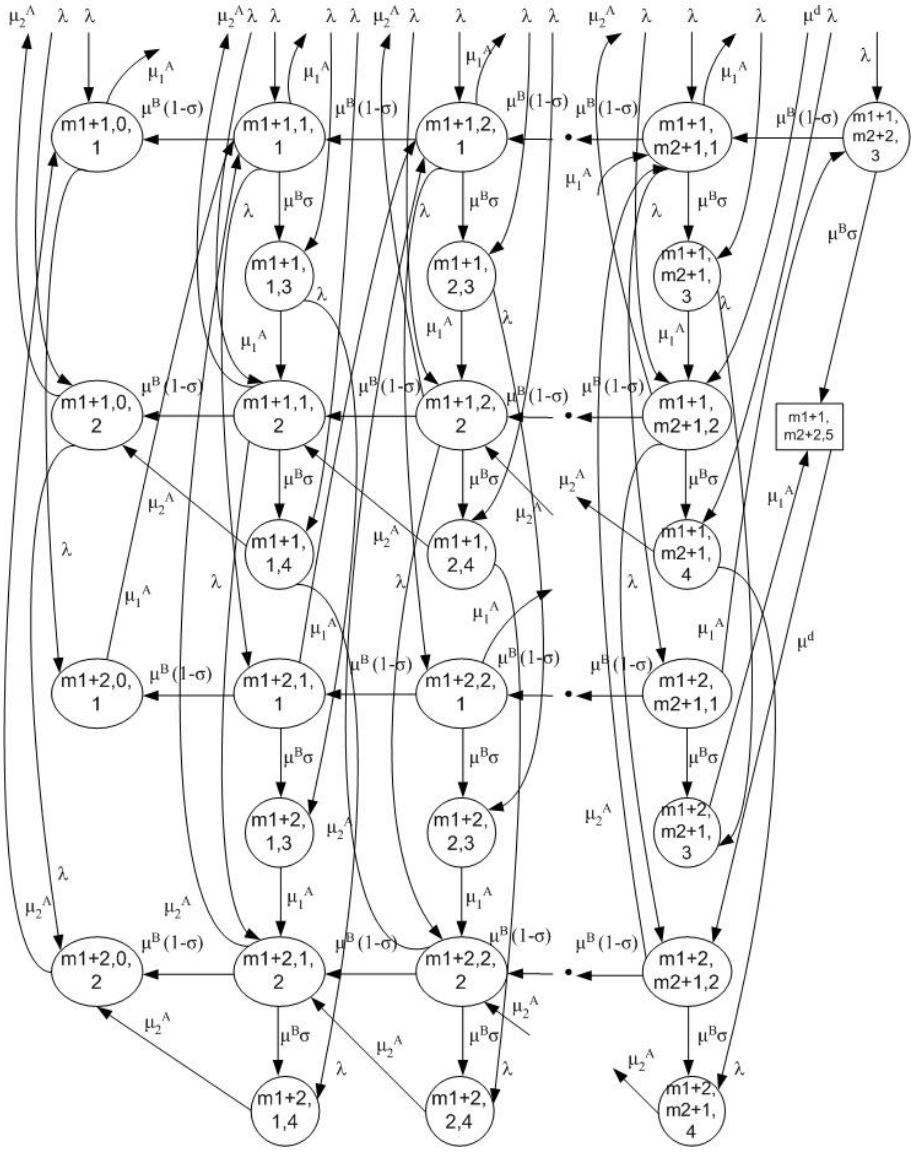

Fig. 3. Two-dimensional tandem network state diagram (second part)

In these diagrams, the index k has the following values: $0,1,2,3,4,5$. The value $k=0$ describes the idle network, $k=1$ servicing the ordinary tasks, $k=2$ servicing

priority tasks, $k = 3$ blocking tasks, $k = 4$ blocking tasks and priority service, $k = 5$ describes a deadlock.

Of course, in this special type of tandem network a deadlock may occur. For example, let us suppose that station A is blocked by station B, because buffer in B station is full. A deadlock will occur if the task in service at station B must be sent to station A upon completion of its service. We assume that a deadlock is detected instantaneously and resolved immediately, with some negligibly delay time by exchanging both the blocked tasks simultaneously with the mean rate equal to μ^d.

This kind of service strategy with HoL priority service for feed backed tasks is described more detail in [7]. In mentioned paper, the procedure of constructing the steady-state equations in the Markov model is shown. Of cause, the stationary probability vector for this model can be obtained using numerical methods for linear systems of equations. Based on this stationary probability vector we can calculate the quality of service (QoS) parameters and the measures of effectiveness for this model.

2.2 Second service strategy: no two priority services in succession

This service strategy principle include:

1. Procedure for feedback tasks **"no two priority services in succession"** (preventing a possible congestion in the first buffer),
2. Mechanisms for checking the current buffer occupancy (resource allocation policy by blocking operations),
3. Procedures for detecting and resolving a possible deadlock.

In this strategy, we assume that deadlocks are detected and resolved instantaneously without any delay, simply by exchanging the blocked tasks.

Notation: the state space of this tandem network model can be described by random variables (i, j, k), where i indicates the number of tasks at the first station, j indicates the number of tasks at second server and k represents the state of each server (see Figure 4 and Figure 5). Here, the index k may have the following values: $0, 1, 2, 3, 4$. If $k = 0$ - idle network, $k = 1$ - regular task service, $k = 2$ - priority task service, $k = 3$ - blocking one station and regular task service at the other one, $k = 4$ - blocking one station and priority task service at the other one. Based on our analysis of the state space diagrams, the steady-state equations in this Markov model ware constructed. More detail this procedure is shown in [8]. In this paper the procedures for calculation the measures of effectiveness and quality of service (QoS) parameters are presented.

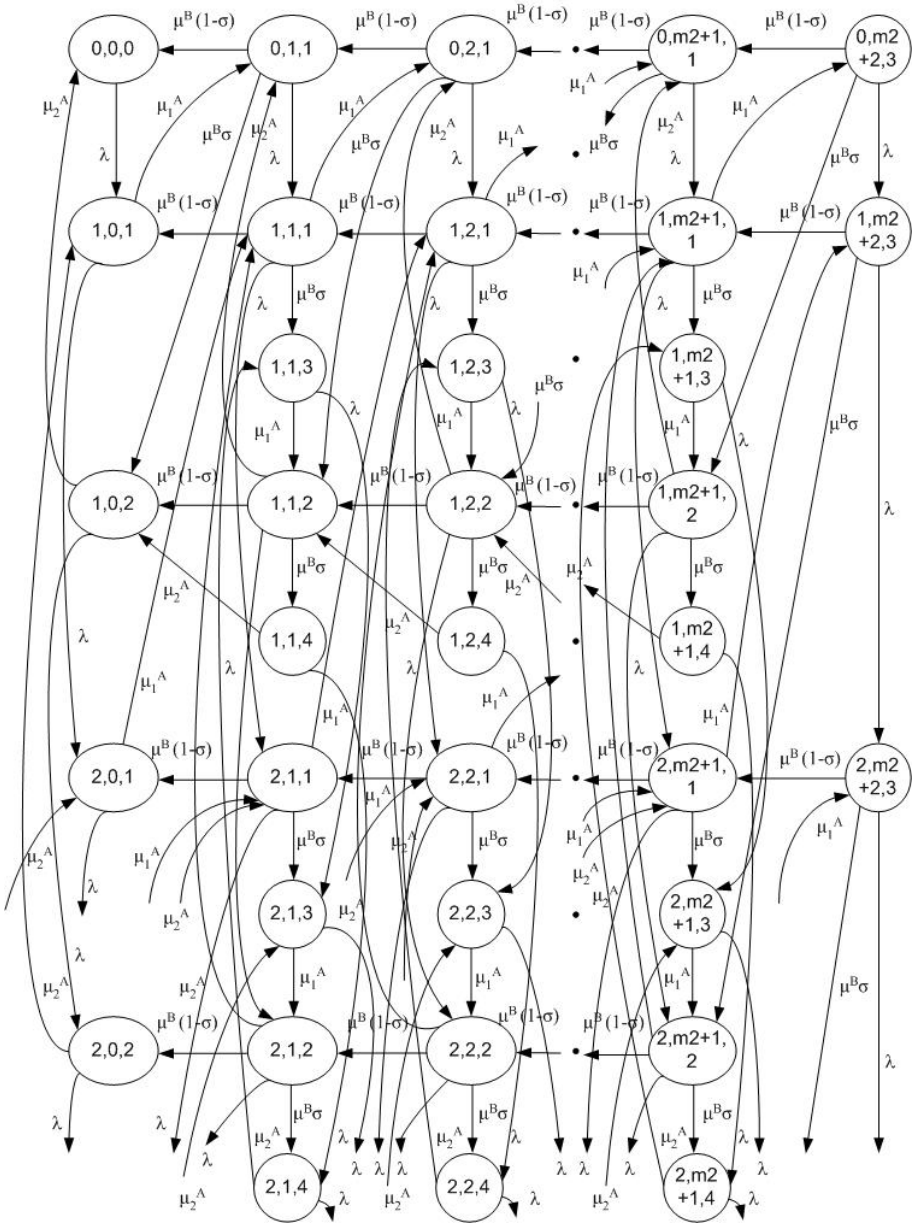

Fig. 4. Second strategy two-dimensional state diagram (first part)

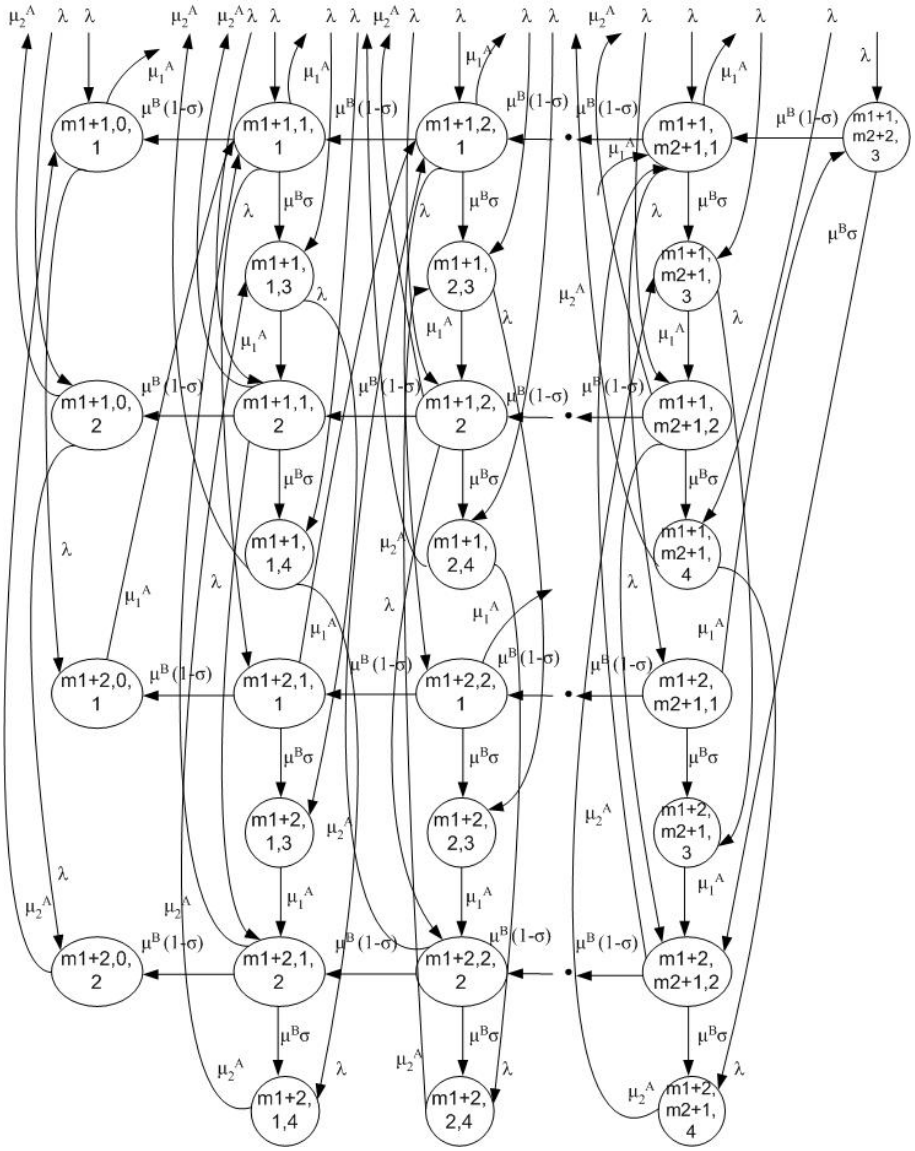

Fig. 5. Second strategy two-dimensional state diagram (second part)

3. Numerical examples

To demonstrate our analysis of two service strategies in a tandem server network with feedback and blocking presented in Section 2, we have performed numerous

calculations. These calculations were realized for many parameters combinations by varying the arrival rate (λ) from source station and by varying the buffer capacities $m1$ and $m2$.

For the first group of calculations the following parameters were chosen: the service rates in station A and station B are equal to $\mu_1^A = 3.0$, $\mu_2^A = 2.5$, $\mu^B = 2.0$. The inter-arrival rate λ from the source station to station A is changed from 0.5 to 4.0 with step of 0.5 and the feedback probability is chosen as $\sigma = 0.7$. The buffers capacities are set to $m1 = 10$ and $m2 = 8$.

Based on such parameters the following results were obtained and the majority of them are presented in Table 1, where λ is the inter-arrival rate from the source station to station A, $p - bB$ is the station B blocking probability, $n - B$ is the average number of tasks in the second station, $w - B$ is the mean waiting time in buffer B, $t - bB$ is the mean blocking time in B station, $utl - B$ is buffer $m2$ utilization coefficient and $\lambda 1$ is the effective input stream intensity from source station to first server (blocking factor).

Note: the columns with *italic* contains the results for second service strategy (no two priority services in succession), the standard columns - results for first Head of Line (HoL) service strategy.

Table 1. The measures of effectiveness and Quality of Service parameters.

λ	$\mathbf{p-bB}$	$p-bB$	$\mathbf{n-B}$	$n-B$	$\mathbf{w-B}$	$w-B$	$\mathbf{t-bB}$	$t-bB$	$\mathbf{utl-B}$	$\boldsymbol{\lambda 1}$
0.5	0.036	*0.043*	0.393	*0.403*	0.052	*0.053*	0.013	*0.015*	0.286	0.500
1.0	0.143	*0.183*	1.679	*1.999*	0.487	*0.607*	0.052	*0.066*	0.642	1.000
1.5	0.272	*0.339*	6.028	*7.939*	2.349	*3.144*	0.100	*0.122*	0.958	1.428
2.0	0.287	*0.345*	7.051	*8.490*	2.806	*3.386*	0.105	*0.124*	0.989	1.619
2.5	0.287	*0.345*	7.088	*8.497*	2.822	*3.389*	0.105	*0.124*	0.990	1.746
3.0	0.287	*0.345*	7.091	*8.498*	2.823	*3.389*	0.105	*0.124*	0.990	1.842
3.5	0.287	*0.345*	7.091	*8.498*	2.823	*3.389*	0.105	*0.124*	0.990	1.919
4.0	0.287	*0.345*	7.091	*8.498*	2.823	*3.389*	0.105	*0.124*	0.990	1.981

For the second group of experiments the following parameters were chosen: the service rates in station A and station B are equal to $\mu_1^A = 4.0$, $\mu_2^A = 3.5$, $\mu^B = 4.0$. The inter-arrival rate λ from the source station to station A is 2.0. The feedback probability σ is 0.2. Buffer capacities $m1$ and $m2$ are changed within the range from 1 to 10. For this series of experiments, the following results were obtained and the selected set of them are presented in Table 2.

In this table m is buffers capacities in A and B stations, $v1 - A$ and $v2 - B$ are mean number of tasks in first and second buffers, $q1 - A$ and $q2 - B$ are mean response

times at station A and B, $\rho - A$ and $\rho - B$ are utilization coefficients at both stations respectively.

Note: similarly as in the first experiment, the columns with *italic* contains the results for second service strategy (no two priority services in succession), the standard columns - results for first HoL service strategy.

Table 2. Selected measures of effectiveness.

m	*v1-A*	**v1-A**	*v2-B*	**v2-B**	*q1-A*	**q1-A**	*q2-B*	**q2-B**	ρ - *A*	ρ - *B*
1	*0.304*	0.304	*0.202*	0.199	*0.400*	0.400	*0.301*	0.300	*0.600*	*0.492*
2	*0.498*	0.449	*0.358*	0.350	*0.443*	0.443	*0.340*	0.338	*0.607*	*0.527*
3	*0.627*	0.628	*0.473*	0.460	*0.473*	0.473	*0.369*	0.365	*0.610*	*0.546*
4	*0.713*	0.715	*0.554*	0.537	*0.494*	0.494	*0.389*	0.385	*0.612*	*0.556*
5	*0.771*	0.775	*0.610*	0.590	*0.507*	0.509	*0.403*	0.398	*0.613*	*0.562*
6	*0.810*	0.815	*0.648*	0.626	*0.517*	0.518	*0.412*	0.407	*0.614*	*0.565*
7	*0.836*	0.843	*0.673*	0.649	*0.523*	0.525	*0.419*	0.413	*0.614*	*0.567*
8	*0.853*	0.862	*0.689*	0.664	*0.528*	0.530	*0.423*	0.417	*0.614*	*0.568*
9	*0.865*	0.874	*0.700*	0.674	*0.530*	0.533	*0.426*	0.419	*0.614*	*0.569*
10	*0.873*	0.883	*0.707*	0.680	*0.532*	0.535	*0.427*	0.421	*0.614*	*0.569*

The results of the experiments clearly show that the effect of the properly chosen service strategy in tandem network with feedback must be taken into account when analyzing performance such computer network. As noted above, feedback probability σ and blocking factor considerably change the performance measures in such networks.

4. Conclusions

An approach to compare the effectiveness of two service strategies in linked in series servers with blocking and feedback has been presented. Tasks blocking probabilities and some other fundamental performance characteristics of such network are derived, followed by numerical examples. The results confirm importance of a special treatment for the models with blocking and with HoL feedback service, which justifies this research. Moreover, our proposal is useful in designing buffer sizes or channel capacities for a given blocking probability requirement constraint. The results can be used for capacity planning and performance evaluation of real-time computer networks where blocking and feedback are present.

References

[1] S. Balsamo, V. De Nito Persone, R. Onvural, Analysis of Queueing Networks with Blocking, Kluwer Academic Publishers, Boston, 2001.

[2] M.C. Clo, MVA for product-form cyclic queueing networks with blocking, Annals of Operations Research, Vol. 79, pp. 83-96, 1998.

[3] A. Economou, D. Fakinos, Product form stationary distributions for queueing networks with blocking and rerouting, Queueing Systems, Vol. 30 (3/4), pp. 251-260, 1998.

[4] C.S. Kim, V. Klimenok, G. Tsarenkov, L. Breuer, A. Dudin, The BMAP/G/1-> ·/PH/1/M tandem queue with feedback and losses, Performance Evaluation, Vol. 64, pp. 802-818, 2007.

[5] J. B. Martin, Large Tandem Queueing Networks with Blocking, Queueing Systems, Vol. 41 (1/2), pp. 45-72, 2002.

[6] W. Oniszczuk, Modeling of dynamical flow control procedures in closed type queuing models of a computer network with blocking, Automatic Control and Computer Sciences, Vol. 39, Issue 4, pp. 60-69, 2005.

[7] W. Oniszczuk, Blocking and Deadlock Factors in Series Linked Servers with HOL Priority Feedback Service, Polish Journal of Environmental Studies, Vol. 16, No. 5B, pp. 145-151, 2007.

[8] W. Oniszczuk, An Intelligent Service Strategy in Linked Networks with Blocking and Feedback, Studies in Computational Intelligence N. 134 "New Challenges in Applied Intelligence Technologies", Springer-Verlag, Berlin, Heidelberg, pp. 351-361, 2008.

[9] W. Oniszczuk, Semi-Markov-based approach for analysis of open tandem networks with blocking and truncation, International Journal of Applied Mathematics and Computer Science, Vol. 19, No. 1, pp. 151-163, 2009.

[10] W. Oniszczuk, Analysis of linked in series servers with blocking, priority feedback service and threshold policy, International Journal of Computer Systems Science and Engineering, Vol. 5, No.1, pp.1-8, 2009.

[11] W. Oniszczuk, Loss Tandem Networks with Blocking Analysis - A Semi-Markov Approach, Bulletin of the Polish Academy of Sciences: Technical Sciences, Vol. 58, No. 4, pp. 673-681, 2010.

[12] R. Onvural, Survey of closed queuing networks with blocking, Computer Survey, Vol. 22 (2), pp. 83-121, 1990.

[13] H.G. Perros, Queuing Networks with Blocking. Exact and Approximate Solution, Oxford University Press, New York, 1994.

STRATEGIE OBSŁUGI W TANDEMACH SERWERÓW Z POWTÓRNĄ OBSŁUGĄ I BLOKADAMI

Streszczenie: W artykule poruszono zagadnienia związane z modelowaniem sieci serwerów z buforami o ograniczonej pojemności, powtórną priorytetową obsługą, które są ważnym elementem, w badaniu parametrów jakości obsługi w systemach komputerowych. Do badań i analizy wybrano dwie strategie powtórnej obsługi zadań w pierwszym z serwerów. Modele analityczne takich sieci stanowisk obsługi przedstawione są tutaj, jako otwarte markowowskie systemy kolejkowe z blokadami,. Tego typu modele w sposób najbardziej pełny odwzorowują ewolucję takich systemów w czasie. Zbudowano dwuwymiarowe grafy takich modeli tandemów oraz na przykładach pokazano jak zmieniają się ich miary wydajności i jakości obsługi, gdy zmienia się intensywność wejściowego strumienia i pojemność buforów.

Słowa kluczowe: powtórna obsługa, strategie obsługi, analityczne modele sieci