Marek MŁYŃCZAK

QUEUING SYSTEM WITH VARIABLE SERVER NUMBER

Description of main characteristics of Mass Maintenance Systems is given. Problems of queuing system effectiveness due to loss of time for both arrivals while waiting for service and for servers waiting for arrivals are discussed. System with changeable number of servers is proposed. Calculations are made in order to find out what is influence of main queuing system parameters on the total operational cost regarding time losses. It is shown that decision about system structure depends mainly on system service index and server initial cost.

***Keywords**: mass maintenance syetem, system structure, operational cost.*

1. Introduction

A model of transportation systems applies usually Queuing Models (also: Waiting Lines, Mass Maintenance System) as a tool to modeling, improving and quality assessment. These models take into account various undesired events disturbing correctly designed process. Queues in real operation process arise as an effect of event randomness and shortage of dynamic adaptation due to external demands. Process is defined as a function assigning to operation states set of operation times and creates a set of random time intervals corresponded to states separated by events. Transportation processes are described as a set of states of transportation process which superior function is to perform randomly arising transportation services. Key elements of the Mass Maintenance Systems (MMS) are: customers (service demands) and service places (servers). Depending on necessities, availability or opportunities, one may permit in the system for creating queuing for service or resource releasing. Working of the system consists on: accepting a customer for free service place or position it in the queue, if it is possible, perform the proper service and remove it from the system. System works properly if customers are not rejected, do not wait too long or if servers are not idle (do not wait for customers). From the customer point of view, quality of the maintenance system is high if on demand at least one server is free. From the system management point of view, server effectiveness is the best if it is busy continuously, even independently if customers queue for service. Adaptation of the system to such variations of demands is difficult as well technically as organizationally but minimizing of waiting intervals both customers and servers may in longer period decrease operational loses [1,5,7].

2. Queuing systems characteristic

Maintenance system has to accept the customer, get him service and release it [1,3,5]. If necessary in the system may be crested queue and than the system contains (Fig. 1):

- arriving in time service requests (arrivals- failed vehicle with repair demand, customer for shopping, airplane collecting passengers, ship coming for cargo),
- service stands offering action (servers- vehicle diagnostic place, fuel distributor, salesman, loading place),
- queue to place customers waiting for service.

Classification of MMS's takes into account several criterions: the way of arrivals (batch, singly), time distribution between arrivals, number of servers, distribution of service time, possible queue, its regulation and capacity.

According to known notations (Kendall, Lee) [1], system is described by the set of symbols: A/B/C/D/E/F, where: A, B describe arrivals stream and distribution of service time, C is a number of servers, D is a queue regulation (way of entering the service system from the queue), E is a total number of customers staying in the system (total number of servers and queue capacity).

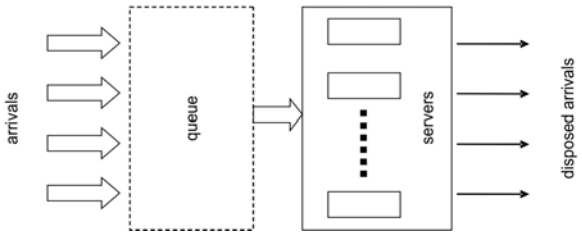

Fig. 1. Elements of mass maintenance system

The main modeling objective is a possibility of analyzing and assessing of system performance, where the most important assessment characteristics are the probability of acceptance refusal, expected number of busy servers or queue length. Analysis and assessment of system parameters is possible analytically by the way of Markov Chains. It is necessary anyway accepting strong limitations and assumptions regarding arrival stream and distribution of service time. Arrival stream is required to be Poisson and service time distribution should be exponential. In that case system assessment is possible analytically. In other situation (time distribution of interarrivals and service not exponential) more effective are simulation methods, though there are some approximate methods giving analytical solution by a little less strong assumptions (semi Markov method) [1,4,6].

Queuing systems are classified according to parameters (arrival stream, service time) and their structure. There are single and multiserver systems, open and closed, and series and parallel. There are very few examples of the systems having changeable number of servers, i.e. systems having possibility of opening and closing servers depending on queue parameters [6,7]. In system with losses (queues not allowed) one may observe number of lost arrivals in given period.

3. System with changeable number of servers

In M/M/m/∞ system arrival stream is Poisson and service time is exponentially distributed. Arrival intensity λ, service intensity μ and number of servers m are the parameters of that system. System allows queuing. Mean size of the queue is given as:

$$\bar{v} = \frac{\dfrac{\rho^{m+1}}{(m-\rho)^2(m-1)!}}{\sum_{i=0}^{m-1}\dfrac{\rho^i}{i!} + \dfrac{\rho^m}{(m-1)!(m-\rho)}};\ \rho = \frac{\lambda}{\mu};\ \frac{\rho}{m} < 1 \qquad (1)$$

and average number of busy servers is $\bar{m}_{nz} = 1 - \rho$, and probability of idle state P_0 is probability that there is no arrivals in the system: $P_0 = \dfrac{1}{\sum_{i=0}^{m-1}\dfrac{\rho^i}{i!} + \dfrac{\rho^m}{(m-1)!(m-\rho)}}$

Especially, considering single server system (m=1, M/M/1/∞), the above formulas are simplified and in steady state are: $\bar{v} = \dfrac{\lambda^2}{\mu(\mu-\lambda)}$ and $P_0 = 1 - \dfrac{\lambda}{\mu}$.

System with the ability of adaptation to changeable conditions may work in this way that depending on given criterion (queue length, idle time) system open or close server that criterion is maintained on required level. Time of server awaking after idleness (tuning time) may also be taken into account but this special case has limited application [2]. The problem of changeable server number is significant regarding three important operational costs: cost due to waiting time for service, cost of lost time while server is idle, initial cost due to construction/opening next server.

It is shown in Fig. 2 comparison for various parameters of arrival intensity (λ=0,2-0,9), service intensity (μ=1) and idle state probability.

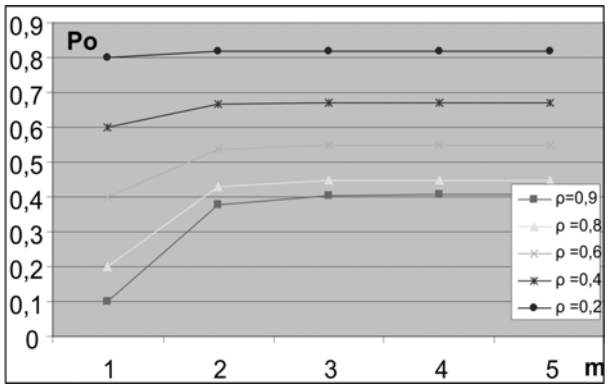

Fig. 2. Server idleness probability due to server number and index of relative service intensity of the system

4. Operational cost analysis for multi server system

The largest change in the probability of server idleness is seen in the range between one and two servers. Hence in systems M/M/m where inter arrival periods and service times are highly variable (variation index in exponential distribution is equal to 1), by relative service intensity of the system approaching 1, probability of meeting zero arrivals in system approaches 0. Actuation of second and following servers raises probability of free server and on the other hand elongates server idle time. In that case, according to instantaneous or periodic arrival intensity or queue length, if many arrivals wait then the system puts working a new server, while there is no arrival waiting, system gets back to previous state (decreases number of servers).

Introducing cost as a quality criterion for the system operation, the target function is described as:

$$Kc(m) = tk(m) * Ko + tb(m) * Kb + Kp \qquad (2)$$

where: $Kc(m)$ – total system operating cost, Ko – unit cost of arrival waiting, Kb – unit cost of server idleness,
$tk(m)$- average arrival waiting time:

$$tk(m) = \frac{\bar{v}}{\lambda} = \frac{\dfrac{\rho^{m+1}}{(m-\rho)^2(m-1)!}}{\lambda\left(\sum_{i=0}^{m-1}\dfrac{\rho^i}{i!} + \dfrac{\rho^m}{(m-1)!(m-\rho)}\right)}$$

$tb(m)$- average server idleness time:

$$tb(m) = \frac{P_o}{\lambda} = \frac{1}{\lambda\left(\sum_{i=0}^{m-1}\dfrac{\rho^i}{i!} + \dfrac{\rho^m}{(m-1)!(m-\rho)}\right)}$$

Kp – initial server cost.

Analytical determination of the minimum cost function is complex because of existence in above formulas of sum dependent on m, therefore in the range of largest variability, the shape of cost function was obtained numerically for single and double server system (Fig. 3). In calculation, unit cost for waiting and idleness are equal.

Analysis of the above diagrams (Fig. 3a) says that if initial cost is neglected, double server system is in the whole range of system service index ρ=0,1 - 0,9 "cheaper" in operation. Single server cost function has minimum at ρ=0,5, and for two

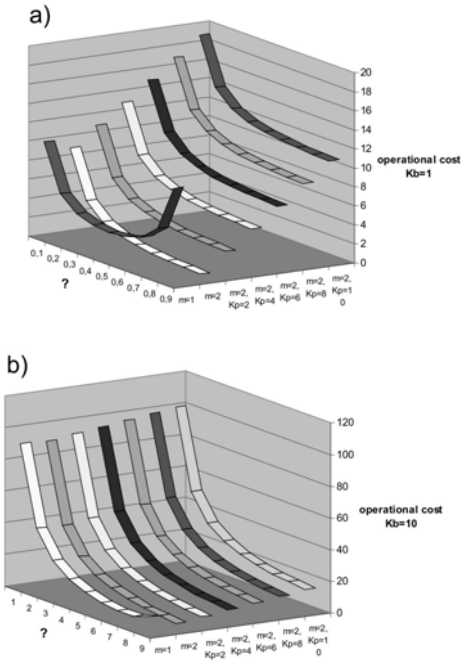

Fig. 3. Operational cost at one and two servers, regarding initial cost in the range of 2 to 10 units and for idleness cost Kb=1: (a) and Kb=10: (b)

servers is monotonically decreasing. Taking into account initial cost one gets that single server system is cheaper if initial cost does not exceed 2 cost units. For more expensive servers (Kp >10), single server system has lower cost in the whole range of ρ variability.

Next stage of the analysis is the determination of influence of the ratio of idleness cost to arrival waiting cost (Fig. 3b). It may be reasonably assumed, that idleness cost of the server (it serves for many arrivals for long time) should be higher than waiting time of the single arrival. In the numerical example idleness server cost is assumed 10 times the cost of waiting time of the arrival. Obtained results show that total operation cost raises only a little due to initial cost of the server Kp and even for single server system total cost is on the same level like for many servers.

5. Summary

Markov chains applied to queuing systems introduce to model of the real system strong assumptions about exponential service time which make this model not very realistic. Analytical outcomes for M/M/m systems let us only in insignificant level for its optimization due to complicated form of equations. Numerical analysis shows that the most effective organizational actions in multiserver system are valid in the range between one and two servers (total operational cost is the most sensitive for changes in server number 1 to 2).

6. References

[1] Filipowicz B.: *Modele stochastyczne w badaniach operacyjnych*. WNT. Warszawa, 1996.
[2] *Handbook of reliability engineering.* ed. Ushakov I. John Wiley&Sons. Inc. New York, 1994.
[3] Karpiński J., Firkowicz S.: *Zasady profilaktyki obiektów technicznych*. PWN. Warszawa, 1981.
[4] König D., Stojan D.: *Metody teorii obsługi masowej*. WNT. Warszawa, 1979.
[5] Leszczyński J.: *Modelowanie systemów i procesów transportowych*. Oficyna Politechniki Warszawskiej. Warszawa, 1999.
[6] Son J.H, Kim M.H.: *An analysis of the optimal number of servers in distributed client/server environments*. Decision supports systems, no 36, pp. 297-312. Elsevier Science Ltd., 2004.
[7] Yamashiro M.: *A system where the number of server changes depending on the queue length*. Microelectronic reliability, vol.36, no. 3, pp. 389-391. Elsevier Science Ltd., 1996.

Dr inż. Marek MŁYŃCZAK

Wroclaw University of Technology
Faculty of Mechanical Engineering
Institute of Machines Design and Operation
Department of Logistics and Transportation Systems
Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland
E-mail: marek.mlynczak@pwr.wroc.pl