Queuing System with Variable Server Number

Marek Młyńczak
Wroclaw University of Technology, Poland
marek.mlynczak@pwr.wroc.pl

ABSTRACT

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 as well as for servers waiting for arrivals are discussed. Model of system with changeable number of servers is proposed. Calculations are made in order to find out what is the influence of main queuing system parameters on the total operational cost regarding time and arrivals losses. It is shown that decision about system structure depends mainly on system service index, server initial cost, number of lost arrivals and time of server emptiness.

Key Words: queuing system, server number, structure, costs

1. Introduction

A model of transportation systems applies usually Queuing Models (also: Waiting Lines, Mass Maintenance System) as a tool used to modeling, improving and quality assessment. These models take into account various undesired events that disturb 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: arrivals, 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 highest if it is busy continuously, even despite customers queuing 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 system characteristics

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 the 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.

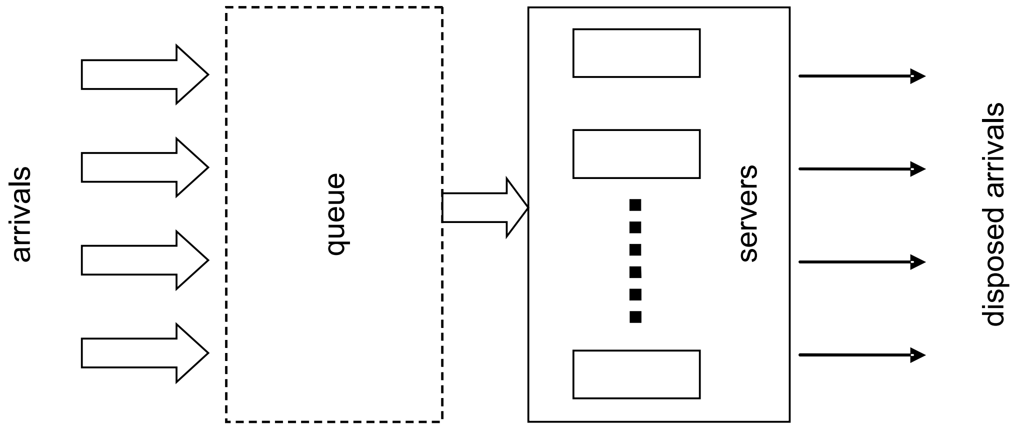

Fig. 1 Elements of mass maintenance system

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,
- possibility of creating a queue,
- queue regulation (way of servicing waiting arrivals),
- queue capacity.

According to known notations (Kendall, Lee) [1], system is described by the set of symbols (1):

$$A/B/C/D/E/F \qquad (1)$$

where:

A,B- describe arrivals stream and distribution of service time,
C - number of servers,
D - queue regulation (way of entering the service system from the queue),
E - total number of customers staying in the system (total number of servers and queue capacity).

The main modeling objective is to give a possibility of analyzing and assessing of system performance, where the most important assessment characteristics are the probability of arrivals acceptance or 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 probability 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 inter arrivals 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 ones. There are very few examples of the systems having changeable number of servers, i.e. systems having possibility of opening and closing servers depend 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

Arrival stream is Poisson and service time is exponentially distributed in M/M/m/∞ system (M- means Markovian). 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 (2):

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

where: $\rho = \frac{\lambda}{\mu}$ and $\frac{\rho}{m} < 1$.

Average number of busy servers is $\overline{m}_{nz} = 1 - \rho$, and probability of idle state P_0 (3) is the probability that there is no arrivals in the system:

$$P_0 = \frac{1}{\sum_{i=0}^{m-1} \frac{\rho^i}{i!} + \frac{\rho^m}{(m-1)!(m-\rho)}}. \quad (3)$$

Especially, considering single server system (m=1) the above formulas are simplified and in steady state we have mean size of the queue (4) and probability of empty system (5):

$$\overline{v} = \frac{\lambda^2}{\mu(\mu - \lambda)} \quad (4)$$

$$P_0 = 1 - \frac{\lambda}{\mu} \quad (5).$$

System with the ability of adaptation to changeable conditions may work in this way that depending on given criterion (critical queue length or idleness time of server) system may open or close servers so, that efficiency 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 four important operational costs:

- cost due to waiting time of an arrival for service,
- cost of lost time while server is idle,
- initial cost due to complexity of server (cost of opening a new server),
- operational cost of operating server (readiness of crew, energy, supply, maintenance).

Comparison of various parameters: arrival intensity (λ=0,2-0,9), service intensity (μ=1) and idle state probability is shown in Fig. 2. Lines drawn in Fig. 2 correspond to service intensity of the system ρ.

4. Operational cost analysis for multi server system

The largest change in the probability of server idleness (Fig. 2) 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 (raises system quality and in consequence the profit) and on the other hand elongates server idle time (brings losses). In that case, according to instantaneous or periodic arrival intensity, if many arrivals are lost, the system puts working a next server, while there are no arrival losses, system gets back to previous state (decreases number of servers).

Introducing cost as a criterion for the system operation, the target function $K(m)$ (6) is described as a difference between profit brought by assurance of service to arrivals and costs incurred in the given period due to:

- initial investments for setting up servers proportional to expected number of arrivals,
- server operating time while working (being busy)
- lost arrivals meeting busy all servers:

$$K(m) = t_{serv}(m) \cdot k_{serv} - m \cdot n \cdot k_{invest} - \\ - t_{serv}(m) \cdot k_{oper} - n_{lost}(t) \cdot k_{lost} \quad (6)$$

where:

m – number of servers in the system,

t_{serv} – time of keeping arrivals in servicing,

n_{lost} – number of lost arrivals,

k_{serv} – profit gained from served arrival over time unit,

k_{invest} – unit cost of initial investment per server,

k_{oper} – cost of server operation over time unit,

k_{lost} – unit cost of lost arrival.

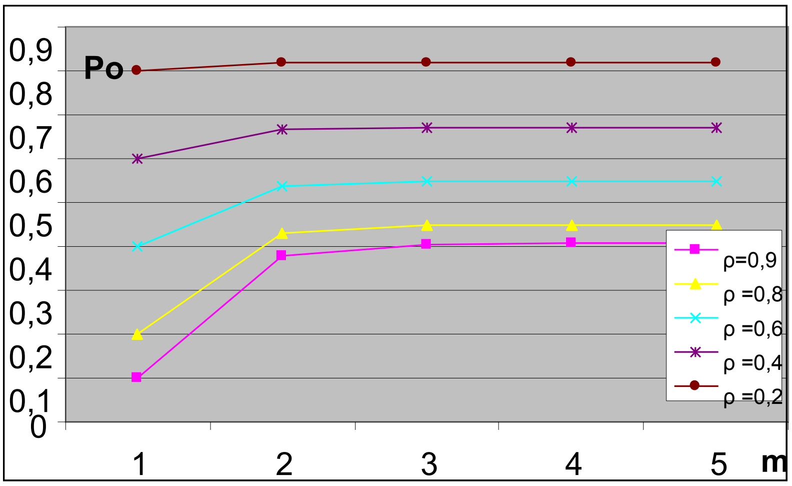

Fig. 2 Server idleness probability due to server number and index of relative service intensity of the system

Analytical determination of the minimum cost function regarding number of servers in the system is complex because of existence in the formulas above an operation of summation dependent on *m*:

- *tk(m)*- average waiting time of an arrival while queuing (7):

$$tk(m) = \frac{\bar{v}}{\lambda} = \frac{\dfrac{\rho^{m+1}}{(m-\rho)^2 (m-1)!}}{\lambda \left[\sum_{i=0}^{m-1} \dfrac{\rho^i}{i!} + \dfrac{\rho^m}{(m-1)!(m-\rho)} \right]} \tag{7}$$

- *tb(m)*- average server idleness time (8):

$$tb(m) = \frac{P_o}{\lambda} = \frac{1}{\lambda \left[\sum_{i=0}^{m-1} \dfrac{\rho^i}{i!} + \dfrac{\rho^m}{(m-1)!(m-\rho)} \right]} \tag{8}$$

Therefore the shape of cost function *K*(*m*) was obtained numerically assuming multi server system (Fig. 3). Simulation was limited to the system with number of servers varied between 1 and 5 and ρ=(0,1-2,1) what means that the test looks at time between arrivals from very long, regarding service time, to very short ones (Fig 3).

Analysis of the diagram says that single server system is profitable in operation if arrival stream is relatively rare (time between arrivals is considerably less than service time). Almost all arrivals are served and only one server in the system is well used. On the other hand, for very dense arrival stream that system is not efficient and cost of lost arrivals decreases total profit.

System with large number of servers is efficient at dense arrival stream but very costly when servers are usually empty (e.g. *m*=5, *ρ*=0,1).

5. Conclusions

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). Design of the multiserver system may be supported by simulation. Advantage of that method is based on expanding the range of conditions influencing total operating cost like initial, investment cost, cost of lost time while servers are empty and lost of arrivals while servers are busy. Shown simulation shown for M/M/*m* system is easy transformable for any non Markovian queuing system.

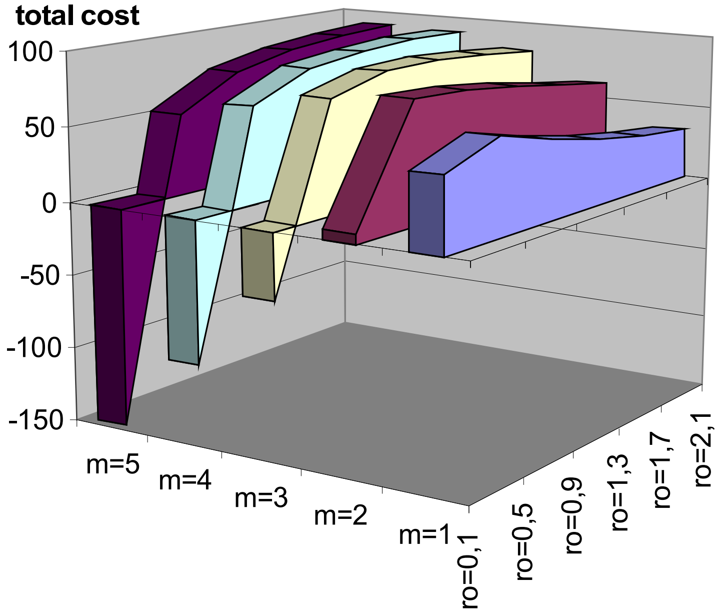

Fig. 3 Total operational cost of the system consisting of 1-5 servers, regarding ρ.

References:

[1] B. Filipowicz, *Stochastic Models in Operation Research* (in Polish). *Modele stochastyczne w badaniach operacyjnych*. WNT. Warszawa, 1996.

[2] *Handbook of reliability engineering*. ed. I. Ushakov. John Wiley&Sons. Inc. New York, 1994.

[3] J. Karpiński, S. Firkowicz S, *Principles of Technical Objects Prophylactic Maintenance* (in Polish). *Zasady profilaktyki obiektów technicznych*. PWN. Warszawa, 1981.

[4] D. König, D. Stojan, *Principles of Mass Maintenance Theory* (in Polish). *Metody teorii obsługi masowej.* WNT. Warszawa, 1979.

[5] J. Leszczyński, *Systems and Transportation Processes Modeling* (in Polish). *Modelowanie systemów i procesów transportowych*. Oficyna Politechniki Warszawskiej. Warszawa, 1999.

[6] J.H. Son, M.H. Kim, *An analysis of the optimal number of servers in distributed client/server environments*. Decision supports systems, no 36. Elsevier Science Ltd., 2004. pp. 297-312

[7] M. Yamashiro, *A system where the number of server changes depending on the queue length*. Microelectronic reliability, vol.36, no. 3. Elsevier Science Ltd., 1996. pp. 389-391