ELSEVIER

Scopus

doi

Journal of Vibration Engineering

ISSN:1004-4523

Registered

- SCOPUS
- GOOGLE SCHOLAR
- DIGITAL OBJECT IDENTIFIER (DOI)
- IMPACT FACTOR 6.1

Our Website
www.jove.science

Queueing Theory and its Applications

Jiri Stodola

Department of Combat and Special Vehicle, Faculty of Military Technology, University of Defense Brno, Kounicova Street 65, 662 10 Brno, Czech Republic

Abstract

The paper focuses on the theory of queues (mass service), which deals with the operation of systems in which flows of requests occur repeatedly, which need to be performed as a sequence of operations (served); these requests are usually random in terms of their origin and moment of occurrence. An analytical method of the system under study is presented and explained using a practical example, using tools of probability theory, which allows obtaining relatively accurate results when determining the characteristics of a given system. The aim of using queues theory is to optimize service channels so that large queues of customers are not formed, waiting times are shortened and all service channels are used, while operating costs are naturally optimized. In conclusion, some possibilities of practical applications in the field of military logistics, or in the field of charging electric vehicles, in the case of their mass expansion, are presented.

Keywords: *Queue theory system, Requirements, Input and output flows, Service lines, Queues.*

1. Introduction

Queuing theory is a branch of applied mathematics (part of operations research) that deals with modeling, design, and management of situations in which customers and their service requirements (people, processes, things, machines, automata, etc.) are served through various channels. The foundations of queueing theory date back to the late 19th century and are related to the development of telephone exchanges (the first Connecticut exchange in 1878), their design, and sizing (Danish mathematician A. K. Erlang, working to reduce the waiting time for telephone calls). Current tasks are mainly optimizing the use of resources, which are customer waiting times and the capacity of service channels. Selected application areas for queueing systems are listed in Table 1.

Table 1: Selected application areas for queueing systems

Customer	System (line)	Operator
calling	switchboard	connection
buyer	checkout	payment
tourist	bus/train/plane/ship	ride/flight/cruise
pallet/container	lifting equipment	loading/unloading
classic vehicle	petrol pump	refueling
electric vehicle	charging station	battery recharge
traffic control	road/highway	number of vehicles
computer	printer	print job
electric car	charging station	battery recharge

After the service is completed, the served customers leave the service system, and it is possible to serve others on the line. The service facility consists of one or more service lines. Subsequently, customer queues (service requests) can arise, and in practice do arise. All the

examples given in Table 1 have something in common, namely that customers in the queue do not perform any activity, which leads to time and economic losses. When solving losses, or rather their mitigation, we can encounter two conflicting requirements. Customers want to shorten the service time, which would, however, lead to the need to build additional service facilities and to increase capacity. However, service providers need to reduce the number of service facilities to minimize costs [1]. It turns out that queue lengths can be influenced in two ways, namely either shortening the service time or increasing the number of service points. A simplified model of a mass service system is shown in Fig. 1 [2].

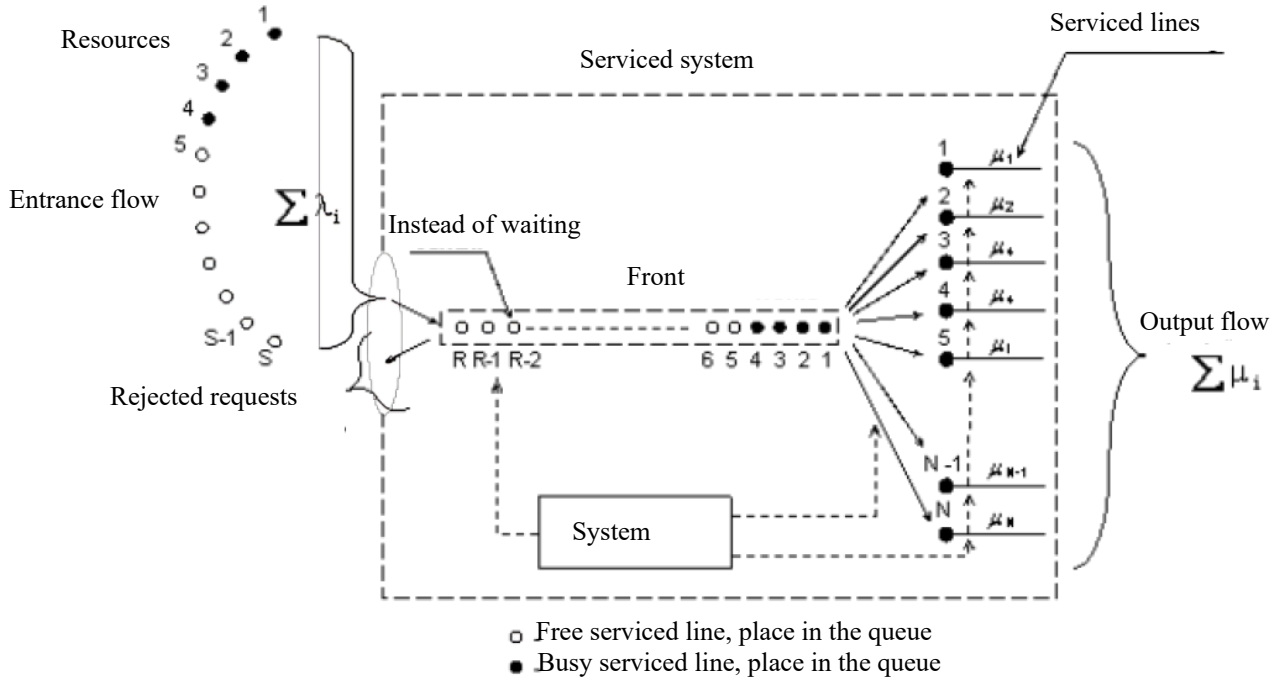

Fig. 1 Model of queueing theory service system

In practice, it is usually not possible to accurately estimate the behavior of a service system, so queuing theory can be used, which provides information about the likely behavior of the system in the future. Based on information that can be assigned a value, we can select the optimal variant of the system used. The following information is necessary for a mathematical description of a mass service system: customer arrival, service time, number of service lines and number of customers who cannot be served immediately at the time of arrival. Customer arrivals with the system can be regular with a fixed time interval or random. If a customer is not served immediately after arriving at the mass service system, a queue is created and the customer is: rejected (queue with losses), the queue is not limited (system with infinite queue length), the queue is limited (system with finite queue length) and the queue is admitted, but after a certain time, other customers are rejected (mixed systems). The set of rules called Kendall's [3, 4] classification (queueing mode) includes the following implementation: regular (customers enter the service in the order in which they arrived, so-called FIFO (First-In/First-Out); inverse, when customers enter the service in the opposite order to the order in which they arrived, so-called LIFO (Last-In/First-Out); random arrival of a customer from the queue to the service, so-called SIRO (Selection In Random Order); priority P-FIFO (Priority-First-In/First

Out), customers with priority are placed at the beginning of the queue, or they come in the event that the queue of customers with a higher priority is empty; finally SJF (Shortest Job First), when the request whose service is the fastest is served first. In terms of process classes, we can divide them into regular processes with independent increments, regenerative (renewal processes), recurrent (returning process), homogeneous (uniform, stationary process) and ordinary processes. Service requests in terms of the probability of occurrence of more than one request at a given moment is zero, therefore it is an ordinary stream of requests, when two events occur simultaneously. The parameter of the flow (density) of requests is introduced, if $\lambda = const.$ the stream of requests (process) is homogeneous (stationary). In this case, the probability of occurrence of a certain number of requests in the interval $< t; t + \Delta t >$ is independent of t, but depends only on the length of the time interval Δt. Queue theory is a relatively complex field of science, and it is not possible to systematically describe this theory in this paper, therefore we will focus only on discrete, stochastic controlled Markov processes, specifically on systems with two and four service lines. Markov analysis [5] is a quantitative technique, either discrete using the probability of change between individual states, or continuous using the intensities of change across states. It is used in cases where the future state depends only on the current state. The standard application is in the analysis of systems that can exist in multiple states. Markov analysis techniques are suitable for independent processes arranged in parallel or series, for load sharing systems, standby systems, e.g. in the event of a service line failure, etc. The inputs for Markov analysis include a list of different states (standby, operational, degraded, inoperative, etc.), possible transitions from one state to another that we model, and the intensity of change from one state to another is represented by the transition probability for discrete events, or the intensity of failures, or the intensity of repairs for continuous events. The Markov analysis technique focuses on the concept of “state” and the processes of transition between these states in time based on a constant probability of change [6, 7]. A stochastic transition probability matrix (the sum of each row and column is 1) is used to describe the transition to allow the calculation of different outputs. The system can also be represented by a Markov diagram, in which the circles represent the states and the arrows represent the transitions together with the original probability, Fig. 2 [6] an illustrative example is also given in [5]. The output of Markov analysis [8].is different probabilities of occurrence in different states, and therefore an estimate of the probability of unavailability and/or availability of the system.

2. Classification of mass systems

At the top level, queuing theory systems can be divided into open and closed. In an open system, requests are not returned to the system after they have been served. In a closed system, on the other hand, the requests served are returned to the source of the system's requests (the reason is e.g. maintenance of operational equipment, etc.). Another division of systems is based on the source of requests into finite (the set of potential requests is bounded) and infinite (the set is not bound in theory, but in practice it is large enough). Systems can be further divided into systems with losses and without losses. Loss is understood as the rejection of an incoming request, when Kendall's classification is used. In this simplified classification, systems are sorted into groups, namely: according to the type of stochastic process that describes the arrival of customers for service, according to the law of distribution of service duration, and according to the service lines that are available to customers. This classification is formally denoted

$A/B/n$, where A is the distribution function of the interval between transitions, B is the distribution function of the service time and n is the number of service lines (a natural number). For a simple stationary mass service system with two lines, A means the Poisson process of transitions, i.e. the exponential distribution of mutually independent intervals between transitions, B means the exponential distribution of the service time and $n = 2$, respectively $n = 4$ means the number of lines. The general conditions of the $A/B/n$ system can be solved under the following assumptions: customers line up in the order in which they arrive, they enter the service in the order in which they arrive, and both service lines are equivalent (the customer does not care which line will provide him with service). After the customer is served, the line immediately starts serving the next customer, if there is a customer in the queue. After the customer is served, the customer immediately leaves the system. The service lines operate independently of each other and of customer arrivals. The service time on both lines for all customers follows the exponential distribution law of a random variable with the same parameter μ.

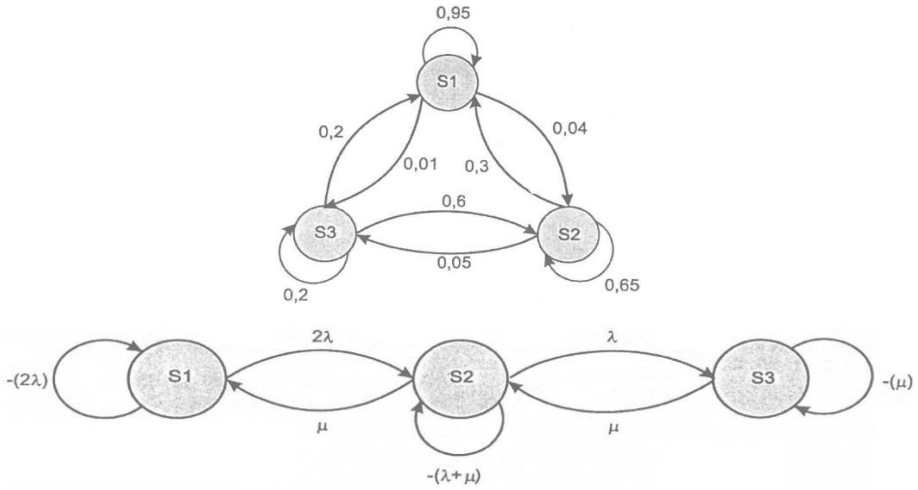

Fig. 2 Example of a Markov diagram of a system and transitions between states

Legend:

λ … probability of state 1 (e.g. number of customers per time unit), μ … probability of state 2 (e.g. average number of customers served)

3. System, transition and state probability $A/B/n$

For stationary Markov processes, the transition probabilities of equations (1) and (2) apply [8, 9]. The number of served and waiting customers in the queue at time t is X_t. If X_t= 0, the system is empty and no service is performed. For $X_t \geq n$ is n customers are served and the others are waiting in the queue. Given the random nature of the system, the sequence of discrete random variables $X_t; t > 0$ forms a homogeneous Markov process of growth and decay with a set of states $S = (0, 1, 2, \dots, k)$, for which the following holds:

- if there are k customers present in the system at time t, where $k \leq n$, then all customers are served and their number increases by one in the time interval $(t, t + \Delta t)$ with probability $\lambda\Delta t + 0(\Delta t)$ and decreases by one with probability $k\mu\Delta t + 0(\Delta t)$. λ and μ are the corresponding intensities of the probability of transition from state 1 to state 2.

- if there are k customers present in the system at time t, where $k > n$, then only n customers are served, who are waiting in the queue. The number of customers increases by one in the time interval $(t, t + \Delta t))$ with probability $\lambda\Delta t + 0(\Delta t)$and decreases by one with probability $n\mu\Delta t + 0(\Delta t)$.

An important characteristic is the probability p that a customer will find all lines occupied upon arrival and will have to wait, assuming that there are at least n customers in the system, given by equation (3). The probability p_0 that a customer will find all lines free upon arrival and will not have to wait is given by equation (4). The limiting probability that there will be k customers on the line is given by equation (5). The mean number of customers in the queue δ (mean queue length) is given by equation (6). The utilization of the line service is informed by the mean number of occupied lines ε, equation (7) applies. The mean number of occupied lines φ depends only on the input flow parameter λ and the service parameter μ, equation (8) applies. The mean number of customers in the system φ is equal to the mean number of customers in the queue δ and the mean number of occupied lines ε. The mean time (downtime) of one customer $\varphi_1$in the system, equation (9) applies. The mean time τ spent by one customer in the system is equal to the meantime $\varphi_1$spent by the customer in the queue and the meantime spent by the customer in the service μ, equation (10) applies. The symbols for the classification of queue theory models are in Table 2. The relevant relations are given in Table 3.

Table 2: Meaning of symbols for the classification of queue theory models

Symbol	Meaning	May contain
A	Type of stochastic process describing the arrival of customers to the service	M – Poisson process of input, meaning Exponential distribution of the interval between inputs Er – Erlang distribution of the interval between inputs of the request D – deterministic distribution, i.e. regular request inputs G – general case, any distribution
B	The law of distribution of service durations	M – exponential distribution of service duration Er – Erlang service duration distribution D – deterministic distribution, i.e. constant service time G – any distribution of service duration
n	Number of parallel operators positions (operator capacity)	1, 2, …∞ (positive integer)

Table 3: Relations for calculating probabilities and mean service times

Equation no.		**Form of equation**	**Note**
(1)	Transition probability p_{ij}	$p_{ij}(t+u) = \sum_k p_{ik}(t) \cdot p_{kj}(u)$	For $t \geq 0$ and $u > 0$
(2)	Transition probability p_j	$p_j(t+u) = \sum_k p_k(t) \cdot p_{kj}(u)$	For $t \geq 0$ and $u > 0$
(3)	Probability of all lines being occupied p	$p = \sum_{k=n}^{\infty} p_k = \sum_{k=n}^{\infty} p_n \rho^{k-n} = p_n \frac{1}{1-\rho}$	The system has n customers
(4)	Probability of all lines being free p_0	$p_0 = \frac{1}{\sum_{k=0}^{n-1} \frac{\beta^k}{k!} + \frac{n^n}{n!} \sum_{k=n}^{\infty} \rho^k}$	

(5)	Probability of k customers on a line p_k	$p_k = \rho^{(k-n)} p_n = \frac{\beta^k}{n!\, n^{(k-n)}} p_0$	
(6)	Mean number of customers in the queue δ	$\delta = p_n \frac{\rho}{(1-\rho)^2}$	
(7)	Mean number of occupied lines ε	$\varepsilon = p_0 \beta \left(\sum_{k=0}^{n-1} \frac{\beta^k}{k!} + \frac{n^n}{n!} \sum_{k=n}^{\infty} \rho^k \right) = \beta$	
(8)	Mean number of occupied lines φ	$\varphi = \sum_{k=0}^{\infty} k \cdot p_k = \delta + \varepsilon$	φ depends on the input flow parameter λ and the service parameter μ
(9)	Mean time (downtime) of one customer φ_1	$\varphi_1 = \frac{p}{n\mu - \lambda}$	The average number of customers φ is equal to the average. Number of people waiting in the queue δ and the average number of busy lines ε

4. Solved practical example

An example is the operation of a mass service system with two service devices, which is, for example, a fuel dispensing point (gas station) with two dispensers that allow the service of two customers (vehicles) simultaneously. One vehicle arrives at the gas station every 80 seconds. The service time of one vehicle lasts on average 150 seconds. Service requirements (customer arrivals) form a Poisson process with a mean value of intensity $\mathrm{E(X)} = \lambda$, which is the mean number of significant events (vehicles that arrive at the gas station per unit of time) and a variance $\mathrm{D(X)} = \lambda$. The service time has an exponential distribution with a mean value $\mathrm{D(X)} = \frac{1}{\mu}$, which is the mean number of customers that the line can serve per unit of time [10]. The task is to find the distribution of the number of elements in the system (served and waiting), i.e. $p_0,\ p_1,\ p_2, \ldots, p_n$ for $\mathrm{n} \geq 2$. To solve the requirements of the assignment, it is first necessary to determine the parameters μ and λ. According to the above relations, the probability of the states of this mass service system and its probability characteristics can be determined in Table 4. A comparison of the results for the variants of two and four dispensing points is in Table 5.

Table 4: Solution of a practical example for two dispensing points

Evaluated indicator	Relationship	Result
Average number of customers per time unit (1hour) λ; equation (1)	$\lambda = \frac{60 \cdot 60}{80}$	45
Average number of customers μ, that is the dispensing point, can serve in 1 hour; equation (2)	$\frac{1}{\mu} = \frac{1}{\frac{2.5}{60}} = \frac{60}{2.5}$	24
Average number of occupied dispensing points β; equation (3)	$\beta = \frac{\lambda}{\mu} = \frac{45}{24}$	1.88
Probability ρ that all dispensing points are occupied	$\rho = \frac{\lambda}{n \cdot \mu} = \frac{45}{2 \cdot 24}$	0.94
Probability p that a vehicle will have to wait in a queue	equation (3)	0.91
Probability p_0 that there will be no vehicle at the dispensing point	equation (4)	0.032
Probability p_2 that there will be exactly 2 vehicles at the dispensing point	equation (5)	0.057
Probability p_4 that there will be exactly 4 vehicles at the dispensing point	equation (5)	0.050
Average number of vehicles δ waiting in a queue	equation (6)	13.6

Mean number of occupied lines ε	equation (7)	
Mean number of occupied lines φ	equation (8)	
Average downtime of one vehicle φ_1 in the queue	equation (9)	0.30
Average time τ of one vehicle at the filling station	equation (10)	0.34
Average number of waiting vehicles ω in total	$\omega = \beta + \delta$	15.5

Table 5: Comparison of solution results for two and four dispenser variants

n	ρ	p_0	p_2	p_4	δ		β	ω	φ_1	τ
2	0.94	0.03	0.057	0.05	13.6		1.88	15.5	0.30	0.34
4	0.47	0.15	0.193	0.08	0.1		1.88	2.9	0.003	0.045

5. Conclusion

The theory of fronts has a very wide application in almost all areas of society. It is especially applicable in modern armies and their logistics. The application in military logistics concerns the theory and practice of planning, implementation of movements and comprehensive material and technical support. Planning and implementation of material movements to a designated place at the optimal time and with optimal costs, as well as maintaining stocks, form a link between industrial enterprises that produce military equipment and military logistics. Logistics includes design, construction and development, identification, acquisition, storage, movement, distribution, transportation and disposal, transportation of people, acquisition or provision of services, catering in field conditions, sorting of the wounded by type of injury in combat conditions or in mass accidents, de-icing of aircraft at low temperatures, etc. [11]. Therefore, logistics plays a very important role in the armed forces of the state, because it ensures the smooth operation of the army. It can be divided into two basic areas: production (acquisition) logistics – deals with research, development, construction and procurement of material, including quality control, codification, standardization and interoperability; operational logistics (in service); is a connecting element between the procurement of material and its delivery (storage, distribution, etc.); consumption logistics (operational) – its task is to supply and support units, including the receipt of delivered material, its storage and maintenance in an appropriate condition (maintenance systems). The above-mentioned military logistics activities show that many of the above-mentioned functions have a very close relationship to mass service systems. The solved example [12] is one of many applications in the field of logistics. What is interesting about the above example is the comparison of the behavior of the mass service system in the event of changes in input parameters. For example, when the number of service lines (e.g. fuel dispensers) is doubled, the average number of vehicles waiting in the queue does not decrease by half, as we would expect, but decreases many times, for example being solved by 136 times. Similarly, the average downtime of waiting vehicles, for example being solved, is reduced many times). In the text and the example being solved, an analytical solution was used, which allows obtaining the characteristics of the system by simply substituting it into the equations. However, this analytical solution is only applicable to simple mass service systems. The solution for complex systems allows simulations using appropriate software tools. Based on the data obtained during the simulations, it is possible to derive approximately the characteristics of the simulated system. This solution has the great advantage that the computer simulation takes place for a very short time (on the order of seconds or units of minutes), while the actual analysis takes place for a relatively very long time. In practice, software tools such as [13] are used to visualize, model, simulate, and optimize logistics systems or subprocesses in them. These tools can be used to optimize, investigate, and plan material flows or create layouts for production sites. Another tool is MPL (Modeling Programming Language) [14], which is an advanced object-oriented visual modeling language that allows you to model and

formulate complex optimization models in a clear, concise, and effective way. MPL operates based on an algebraic modeling language that allows you to relatively easily create optimization models using algebraic equations. The models serve as the basis for creating a mathematical matrix that can be directly transferred to the solver who performs the optimization, e.g. using a suitably prepared simple MS Excel file. Algebraic modeling languages have proven to be an effective method for creating optimization models, the reason being that they are relatively easy to learn, quickly formulated, and require less programming work. Software tools include heuristics, which are previous experiences that have yielded results and solutions. Heuristics are used especially in situations where classical exact methods are too slow or cannot find a solution. The goal of heuristics is to find a solution in a reasonable time that is good enough for the solver given the nature of the problem. An important property of heuristic methods is that the solution found may not be the best of all solutions, but it may approach the optimal solution [15]. A solution using a heuristic method is admissible, but it cannot be called optimal. Heuristic methods include, for example, the northwest corner method, the minimum matrix method, the traveling salesman method, etc. The current task of queue theory methods is a model for charging electric vehicles. With the increase in the number of electric vehicles, the number of charging stations is also increasing, but currently this is not sufficient and at the same time, current technology does not allow charging as fast as refueling, which results in the formation of queues at charging stations.

Acknowledgement

Presented work has been prepared with support of the project Military Autonomous and Robotic Assets (DZRO VAROPS), University of Defense Brno, Czech Republic

References

1. Danek, A., Broncek, M., Janosec, J., Jurak, J. (2000). Road vehicle repair. College of Mining. Technical University of Ostrava. ISBN 80-7078-779-1. 126 p. (In Czech).
2. Toman, V. (2009). Possibilities of using crowd-service models in services and sales of goods. University of Economics in Prague. 38 p. (In Czech).
3. Kendall, D. G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. In The Annals of Mathematical Statistics Vol. 24, 19, pp. 338–354.
4. Kendall's classification of operating systems. [online] [cit. 2026-02-09] Available at: http://access.feld.cvut.cz/view.php?nazevclanku=kendallova-klasifikace-obsluhovych-systemu&cisloclanku=20051 11601 (In Czech).
5. Valis, D., Breznicka, A., Stodola, J. (2021). Risk management. University of Defense in Brno. ISBN 978-80-7582-350-2. (In Czech). 114 p.
6. Bhat, U N. (2015). An introduction to queueing theory: modeling and analysis in applications. USA: Birkhäuser, 2 ISBN 978-0-8176-8421-1. 339 p.
7. Chan, W C. (2014). An elementary introduction to queueing systems. Singapore: World Scientific, ISBN 978-981-4612-00-5. 116 p.
8. Gauniuc, P. A. (2017). Markov chains: from theory to implementation and experimentation. Hoboken, John Wiley. ISBN 978-1-119-38755-8.
9. Said, D., Cherkaoui, S., Khoukhi, L. (2015). Multi-priority queuing for electric vehicles charging at public supply stations with price variation. Wirel. Commun. Mob. Comput. 20, vol. 15, pp. 1049–1065. DOI: 10.1002/wcm.2508
10. Shone, R., Glazebrook, K., Zografos, G. K. (2019). Resource allocation in congested queueing systems with time-varying demand: An application to airport operations. European Journal of Operational Research. Vol. 276, no. 2. ISSN 0377-2217. Pp. 566-581.

11. Schwarz, J. A., Selinka, G., Stolletz., R. (2016). Performance analysis of time dependent queueing systems: Survey and classification. Omega Oxford. Vol. 63. ISSN 0305-0483. pp. 170-189.
12. Vild, M. (2023). Use of priority mass service models. DP SKODA AUTO University College o.p.s. (In Czech).
13. Bangsow, S. (2015). Tecnomatix Plant Simulation. Springer Verlag. ISBN 978-331-919-502-5. 731 p.
14. ISO/IEC JTC1. 19501:2005 (2005). Information Technology – Open Distributed Processing – Unified Modeling Language (UML) version 1.4.2 432 p.
15. Hillier, F.S., Liebermann, G.J. (2015). Introduction to Operations Research. (10th ed.) New York: McGraw-Hil. ISBN 007-113-989-3