The finding of the queuing theory models for evaluation throughput of the IRS radio network in the Czech Republic.

Filip Večeřa¹ and Lukáš Pavlík²
¹² Tomas Bata University in Zlín Faculty of Applied Informatics, Nad Stráněmi 4511, 760 05 Zlín, Czech Republic

Abstract. The queuing theory is a discipline that uses mathematical procedures to evaluate the usability of the utilization system - a radio-communication network. The paper describes the basic parameters of queuing systems and explains the individual variables used in queuing theory models. The functional principles of individual queuing theory models are graphically depicted in the article. The article also includes basic knowledge of the functions of the PEGAS radio communication network, which is used by IRS in the Czech Republic. Not least, the article searches for the most convenient model of queuing theory for the possible evaluation of the transmissivity of the Czech IRS radio network.

1 Network throughput

One of the basic parameters of communication links - radio networks is their permeability, which is a quantity that says what amount of data can be transmitted over a network per unit of time. In a real network, this number can actually change and the network must respond to these changes. The aim of routing is usually to deliver resources as quickly as possible to the destination, and the throughput plays an important role. It is also associated with a data rate parameter that is specified in bits or bytes per second. The throughput can be defined as a basic indicator of network capacity. It is determined by the number of transactions made per unit of time. The throughput is influenced by the network load variability (the number of sent requests to system servers). A typical throughput curve grows to the level of 100% utilization of channel resources [1, 2].

2 Queuing Theory

The queuing theory was created as a mathematical discipline at the beginning of the 20th century. Its foundations were laid by Danish mathematician and scientist Agner Krarup Erlang (1878-1929). The impulse was the development of telephone exchanges for which the proposal was initially intended. The queuing theory uses probability theory, mathematical statistics, and random function theory. It explores the functioning of systems in which serial operation applications are repeated, which usually occur randomly (the number of incoming requests and the amount of time the request spends in the system is random). The aim of the queuing theory is to recognize the terms by which the system works and to optimize the system to avoid unnecessary queuing of queues or leaving the system and vice versa to avoid unnecessary system operation costs [3-5].

2.1 Solutions for queuing systems

Two-base modes can be used to solve tasks in relation to the queuing system. Either using analytical stochastic models or using simulation models.

Analytic models - based on known model parameters using probability theory tools or other mathematical branches, we calculate or estimate those parameters of the model of interest (e.g.: average number of requests in the queue, etc.). Simulation - Validated Parameters of a Model Using a Situation Using Software. Based on these simulations, we estimate the model parameters that interest us.

The aim of the queuing system model is to coordinate the functions of the particular parts of this system. It means the coordination of the incoming requests - (access control system) with own the probability of the behaviour.

Another important role is played by the serving system, which has its limited boundaries and also has random behaviour over time [5, 6].

2.2 Parameters of queuing systems

The operating system serves to satisfy emerging requirements in agreed quality. In general, the quality of the operating system expresses its level of readiness to meet the request in full or in part - in terms of operating time.

Basic operating system parameters (Fig. 1.):
- Input flow - (Unit Input - Source - request)
- Queue - (rules for sorting units and waiting for a free channel - server)
- Servers - Channels - (service line network - operator organization)
- Output flow - (output of units from the system)

Corresponding author: fveceraj@utb.cz

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).
On the input side of the operating system is the input flow consisting of resource requirements. Some of them are satisfied (immediately, or in a certain time - they are queued) and form the output flow. The output flow is often the input flow of the following up operating system. Unsatisfied requests form an outgoing flow [3, 6].

Fig. 1. Basic parameters of the operating system.

2.2.1 Input Flow - Source

The input flow (stream) is made up of requests coming from the request sources (units) to the service line (server). This number of requests may be finite, or theoretically non-finite. Mostly the units come into the system at random times, such that the time between arrivals (time spacing) of the individual units is a random variable. If the time spacing is constant, we talk about a deterministic flow, if it is random, it is a stochastic flow. There may also be a situation where some requirements come to the operator's location at a fixed interval, and some units at a variable interval, in which case we talk about mixed flow. The number of applications per unit of time indicates the flow density [3, 6, 7].

The source of a potential set of units that can enter the system is called resource units. If it is strictly limited, we talk about a closed system - the units after the service are returned to the source, otherwise, it is an open system - the units are not returned to the source after serving [8].

2.2.2 Queue - waiting area and operator time

The queue is composed of pending requests waiting for the server. The method that determines the form of queuing to queue forwarding is called the queue order. The basic types of the queue to attendance include:

- FIFO (first-in / first-out) - the one who comes first is served first
- LIFO (last-in / first-out) - the one who comes last is served first
- PRI (priority or HVF - high value first) - depending on importance, the highest priority request is selected when the service channel is released
- SIRO (selection in random order) - requests are handled in random order

Queue discipline can be absolutely impatient - a system request whose all operator channels are occupied does not enter and resigns. Furthermore, it can be impatient - requests wait regardless of time until the operator is implemented or the half-impatient - requests wait in the queue for a certain period of time and then exit the system if its operator has not yet started.

Waiting space is the space between the unit source and the service channels (servers). A queue is created in this space. Space can be:

- zero - a request that cannot be served immediately is rejected, the queue cannot be created,
- non – zero - a) if the latency is nonzero and unlimited, then the operating situation selects a queue of any length.
- non – zero - b) if the waiting area is non-zero and limited (the most common case in practice), then the request is rejected when the system has a maximum allowable length [7, 8].

The operator duration can be constant or random. In the case of constant time, the operator duration is still the same. The fluctuating duration of the operator is described by some probability distribution. The most often used is the exponential division [6, 8, 9].

2.2.3 Servers - Operating Organization

The Servers (service lines - channels) realize requests of input flow resources. Depending on the number of servers in the operator station, we distinguish between single-server systems (Fig. 2) and multi-servers systems (Fig. 3). The servers may be arranged in parallel - it is sufficient for the unit to be serviced by one, any operator server. The queue can be common to all operator servers and the request always comes to a server that is being released or a separate queue can be created before each operator server. The servers can be arranged in series too - the unit must pass through all servers. In a serial configuration, the queue can be created before the first operator server or queue can be ahead of each operator server [6, 8, 9].

Fig. 2. Example of a series single-server system [9].

Fig. 3. Example of a parallel multi-servers system.

2.2.4 Output Flow

It consists of requests leaving the queuing system, which is characterized by the same criteria as the input flow.
and is assumed to have the same characteristics. This is important when the output flow is routed to another operating system [3].

3 Classification of queuing systems

Due to the diversity of queuing systems, D.G. Kendall has written a system that captures and classifies standard types of queuing models in a compact manner. It usually contains a sequence of five characters. Kendall A / B / X / Y / Z.

- A - indicates the type of probability distribution describing the intervals between system request arrivals. For exponential (mutually independent) distributions (Poisson process of inputs), the symbol M is used, for constant (deterministic) intervals the symbol D is used, for Erlang distribution the symbol Ek, for normal distribution the symbol N, for non-specified distribution with some mean value a standard deviation G.
- B - indicates the type of probability distribution describing the duration of the operator (operator length).
- X - the number indicating the count of parallel operator servers (service lines).
- Y - the number indicating the queuing system capacity (count of units that may be present in the system) - if this capacity is not limited, the symbol ∞ is used.
- Z - the queue mode (FIFO, LIFO, SIRO, PRI).

Only A / B / X symbols are used in narrower three-character notation, and the queue mode is assumed to be FIFO, and the system capacity and resource requirement are unlimited. A wider six-character notation adds an additional attribute in front of the Z symbol, which indicates the frequency of the request source. If this symbol is not present, it is assumed that the source of requirements is not limited [8, 9].

3.1 Quantities used in queuing systems models

- \( \lambda \) - mean input intensity (the average number of requests that enter the system per unit time)
- \( \mu \) - mean intensity of service - output (average number of requests - units that can be served per unit of time) -> average throughput

If the queue is not to grow and rise, then it is \( \mu > \lambda \), or \( \beta = \lambda / \mu < 1 \) (the operator must be faster than arrivals)

- \( \rho \) - system utilization coefficient (traffic intensity - load)
- \( \rho_n \) - the probability that there are just „n“ requirements in the system
- \( n_s \) - average number of requests in the system
- \( n_f \) - average queue length (queue requests)
- \( t_r \) - average system residence time requirement
- \( t_f \) - average waiting time
- \( C \) - number of operator servers - channels

- \( C_o \) - the average number of occupied servers - channels

3.2 The basic models of queueing theory application to Radio Networks

The \( M / M / 1 / FIFO / \infty \) model, which is also used as a benchmark for other models, is one of the basic and most widely used queuing systems (Fig. 4). From the point of view of the complexity of analysis, this is the simplest and most generalized model of queueing theory. In this model is the breakdown time between the arrival times and the operator time is exponential distribution (independent) and is an open system (unlimited number of requests) with one server. The queue mode for this model is simple, and in the event of insufficient capacity, the server waits in the queue for the operator to enter in the order they arrived (FIFO) [10].

System load intensity:

\[
\rho = \frac{\beta}{c} = \frac{\lambda}{c \mu}
\]  

(1)

To avoid system congestion, it must apply:

\[
\rho < 1, \ \beta < c
\]  

(2)

Average number of occupied servers:

\[
\bar{n} = \frac{\lambda}{\mu - \lambda}
\]  

(3)

Average system requests:

\[
\bar{\pi} = \frac{\lambda}{\mu - \lambda} \quad \bar{\pi} = \bar{n} + \beta
\]  

(4)

Average queue requests:

\[
\pi_f = \frac{\lambda^2}{\mu (\mu - \lambda)}
\]  

(5)

Average system time:

\[
\bar{t} = \frac{1}{\mu} \quad \bar{t} = \frac{1}{\mu} \quad \bar{t} = \frac{1}{\mu - \lambda}
\]  

(6)

Average waiting time in queue:

\[
\bar{t}_f = \frac{\lambda}{\mu} \quad \bar{t}_f = \frac{\lambda}{\mu (\mu - \lambda)}
\]  

(7)

**Fig. 4.** Graphical representation of model \( M / M / 1 \) FIFO / \( \infty \) [10].

Other frequent radio model includes the \( M / M / 1 / FIFO / 0 \) model, which in most respects is the same as the previous model, with the absence of a queue being a peculiarity. In the case of multiple requests, operating system requests are rejected at the same time. This system model is referred to as a loss system. The \( M / M / 1 / FIFO / 0 \) model (Fig. 5) is used for simple simplex
radio networks where only one request occupies one channel resource and other requests are immediately rejected.

Another model that has similar parameters and is widely used for radio networks is the $M / M / c / FIFO / n$. This model has a change in the number of servers as well as in the final length of the queue compared to the previous model.

According to Kendall’s notation $n$ denotes the maximum number of requests that can be found in the system - one request is handled and $n-1$ requests wait in the queue for service (Fig. 6). Due to the queue capacity is final, the requirements that come when the system is fully occupied are lost and leave the system unattended. The requests have been blocked and such a system is referred to as a loss system. The probability of lost requests (referred to as blocking probability - $P_B$) is one of the most important parameters for systems with losses.

\[ P_B = \frac{(1-\rho)\rho^n}{1-\rho^{n+1}} \]  

Current sophisticated radio systems use a priority model, where some communications take precedence over others (dispatcher broadcast, emergency calling etc.). The $M / M / c / PRI / n$ model works with priorities.

Priority mechanisms are implemented in the network nodes to minimize the impact of network traffic on the delay of multimedia service packets. These mechanisms reduce packet delays that require priority operation. When implementing priorities, it is necessary to divide network traffic into so-called priority classes. The division is performed with a focus on the service’s delay sensitivity. In the model with priorities, we assume that the arrivals of the requirements of both priorities in the system correspond to the Poisson distribution (independent of each other). The intensities of arrivals are denoted as $\lambda 1$ and $\lambda 2$, the operation time corresponds to the exponential distribution with intensity $\mu$ and is the same for all requirements. To satisfy the stability condition, we consider $\rho 1 + \rho 2 <1$, where $\rho 1$ represents the operating load constituted by the requirements of the first priority class and $\rho 2$ is the operating load formed by the requirements of the second priority class.

Priority mechanisms can be broken down by the way they handle the packets being served. We distinguish between pre-emptive priority and non-pre-emptive priority. Pre-emptive priority ensures that when it comes to a system with a busy service line occupied that has a higher priority than the currently serviced packet, lower priority packet service is interrupted or terminated (pre-emptive priority can be further divided into end-of-service mechanisms and without completion of service). If a packet with the same or lower priority than the packet in the service comes to the system with the occupied server, then the incoming packet is queued and another operation depends to its the priority and mode of the queue. The non-pre-emptive priority ensures that incoming packets are queued in the case of busy servers, regardless of the priority of the packets currently serviced. Subsequently, when the operator server is released, the highest priority packet is served, which is next in order of queue mode (each priority class has its own logical queue and how the operator server is released, the packet at the beginning of the top priority queue enters the service server) [10].

4 The queuing theory models in the basic radio-communication interfaces of the IRS

The units of the Integrated Rescue System (IRS) of the Czech Republic use the PEGAS radio-communication network as the basic communication interface in the group communication, individual call and direct simplex communication modes. Some IRS units also use an analogue radio network, both in repeater mode and direct simplex mode.

4.1 Radio-communication interface throughput in the PEGAS radio communication network
The throughput of the PEGAS radio-communication network in system mode is dependent on the number of channel resources at each base station to which individual terminals are logged. For this reason, we would have to determine the throughput of individual base stations - cells. In order to determine the permeability of the PEGAS radio network as a whole, it does not depend on the selected communication mode whether we use the so-called "Talkgroups" (TKG), individual calls, or data transmissions (AVL - vehicle location, status message, etc.). The system mode allows as many communications as the base station contains channel resources. In nowadays most of the base stations are equipped with 2nd and 3rd generation technology, ie 12 and 16 channel resources. One channel source is always used as a "dedicated channel" for continuous communication - authentication and cryptography of communication between the base station and terminals. Another channel source is used for data communication (statuses, AVL - localization of vehicles, etc.), the remaining channel sources are already intended for individual group communications (TKG, individual calls, etc.). However, for the remaining channel resources, the emergency call priority is determined, which, if all channel resources are used, terminates the service of existing users in the system and enables the user of the emergency call. If all channel resources are occupied, the requests of other system users are rejected.

From the point of view of the throughput of the PEGAS radio-communication network as a whole, group communications (TKG, AVL, statuses, individual calls, small conferences, emergency calls, broadcast, etc.) can be considered as a resource. The intervals of the resource inputs (group communications) are random - exponential (M). Operator duration (group communication activation) is also random (M). The number of channels resources depends to base station generation, ie 8/12/16. The type of queue is set according to the priority levels and within the given level, according to the order in which they entered the system. Queue length is unlimited.

The available information suggests that for the classification of the one base station in system mode of the IRS - PEGAS radio-communication network throughput it is the best to choose the model of queuing theory in the format $M / M / 16 / PRI / \infty$.

- There are 16 service lines (channel sources) in the system.
- Intervals between request arrivals can be described by exponential distribution with parameter $\lambda$ - (n / hour).
- Operator time is a random variable with exponential distribution with parameter $\mu$ - (a / hour).
- Unlimited system capacity, unlimited resource requirements and PRI queue mode.

The radio-communication network PEGAS can be in three states:

- The system is empty (status 0) – no requests.
- The system performs customer service (ie non-priority request - state 1).
- The system is malfunctioning (ie it performs priority customer service - state 2).

5 Conclusion

The aim of the article is to introduce and evaluate the use of individual models of the queuing theory to evaluate the throughput of radio networks and to find the most suitable model applicable to the radio communication network PEGAS of the Czech IRS. The throughput is one of the most important parameters of any operating system, and radio networks are no exception. From the above-mentioned parameters, it was concluded that it is best to apply a queuing theory model to parameters of exponential occurrence of demand inputs and random processing times for the requirements. The number of servers is dependent on the channel resources of specific base stations and the queue type is prioritized to operate requests. Most modern radio networks are currently being used by rescue units on this model.

To initiate practical research, real-time measurement of specific base stations and the use and collection of relevant data for future mathematical validation is essential.

This paper was supported by the TBU IGA project: Determining the Impact of Selected Cyber Threats on the Organization's Information Environment, registered under IGA/FAI/2019/004.

References

1. J. Vaňková, M. Černý, RVP, Computer Network Parameters, 24, 9 (2012)
2. B. Sosinsky, Networking Bible, 56, 32 – 146, (2009)
3. V. Kapoun, Development of Access Network, 7, 205, (2002)
4. Š. Voráčková, M. Pěnička, J. Veselý, Queuing Theory, 8, 2-7, (2008)
5. H. Brožová, M. Holuška, Teorie hromadné obsluhy, 14, 8 – 21, (2002)
6. J. Friebelová, Teorie hromadné obsluhy, 1-11 (2006)
7. J. Dvořák, Operational Analysis, Queuing system, 14, 10 – 22, (2000)
8. B. Leitner, Stochastic methods of operational analysis, 34, 219-250, (2016)
9. M. Žižka, Teorie front, 9, 3-15, (2007)
10. Z. Rýzner, Queueing theory utilization in packet network design and optimization process, 2, 10, 31, 33, (2011)