Method development for analysis of information components of the monitoring system with a limited delay

To cite this article: O J Kravets and I A Panfilov 2020 IOP Conf. Ser.: Mater. Sci. Eng. 734 012095

View the article online for updates and enhancements.
Method development for analysis of information components of the monitoring system with a limited delay

O J Kravets¹ and I A Panfilov²

¹ Voronezh State Technical University, 14, Moskow ave., Voronezh, 394026, Russian Federation
² Reshetnev Siberian State University of Science and Technology, Krasnoyarsk Russian Federation

E-mail: olegkravets@tdt.edu.vn, crook_80@mail.ru

Abstract. Methods of modeling and analysis of information components of a monitoring system are examined. They are based upon methods of the queueing theory. In particular, a limited waiting for service requests in a multi-linear system with a general queue is considered. Various types of failures have also been taken into consideration. Particular emphasis is made on deriving the analytic dependence of loss probabilities from factors that limit a waiting queue (i.e. number of positions available within a queue and maximum stay time in the given queue) in case of a simple stream of calls which are incoming for service, and the demonstrative law of channel occupancy distribution.

1. Analysis of data flow in monitoring systems

Monitoring systems integrate an elaborate complex of communication products of diverse nature. The management system is one of such products. As a rule, probabilistic approach is applied in studying of such systems. Probabilistic analysis of monitoring management systems will be outlined in details below. Stochastic nature of data inflow predetermines utilization of queueing theory for development of analytical models. Evaluation of interaction efficiency between monitoring system users (callers) via central server is of a particular interest for us. It is anticipated that the center of monitoring management comprises the complex of software and hardware tools, interconnected via a switched line (as well as through allocated lines when sufficient resource capacity is available).

While modeling systems for management of monitoring systems (i.e. subsystems for information collection and access to a server) a “request” is conventionally defined as user’s request for service (i.e. establishing a connection/link with a server).

Considering that service quality is mainly determined by throughput capacity and quality of communication equipment, given analysis will be carried out using following definitions of general concepts in terms of the queueing theory:

User workstation (UW) (otherwise known as an abonent station) is a source of a request.

Request will be understood as a request for establishment of a virtual link between a user workstation and a central workstation; a stream of requests from one UW is formed during the process of incurring requests for establishing a virtual link during transmission of maintenance messages. A tool for filling of such requests (i.e. a virtual link) is considered as a maintenance device. The key element of given
description is that before servicing all requests are arriving into a storage circuit with First-Come, First-Served (FCFS) algorithm for issuing serving requests. Assuming that the storage circuit has an unlimited capacity, we are able to evaluate characteristics of the system which are required for a standard waiting time. Such storage circuit will be considered as a dummy activity “waiting for a demand”, completion of which equals to the arrival of a certain request into a servicing device. Complete stream of requests incoming into the storage circuit is generated by addition of request streams from UW. Finally, the monitoring management system with access to the central server (CS) with N servicing devices will also be analyzed in this paper.

2. The model as a multi-linear system with a common queue
By having the information collection subsystem defined in terms of the queueing theory, we have come up with the multi-linear system with the general queue G/G/N/∞, which is depicted on figure 1.

![Figure 1. Representation of the system: SD is a service device](image)

Let us highlight following types of failures in the system functioning that emerge during either scheduled or non-scheduled increase of number of users.

1) Malfunctioning of one local subsystem which has an impact on the intensity of incoming streams.
2) Failure of one out of several channels without recovery, which prompts us to analyze the system which comprises N-1 devices.
3) Failure of one out of several channels with consequent recovery of functionality.

As long as the capacity of the storage circuit is assumed infinite, during non-routine situations the evaluation of functioning efficiency of the system is reduced to a simple estimate of waiting time for a service request in the storage circuit. In given study the probability of service expectation, which exceeds permitted time, will be used as such an estimate, because the average waiting time evaluation does not result in the estimate of number of requests that can remain within the storage circuit during time, which exceeds time allowed by operational regulations of communications protocols. Let us provide some results, which are related to the possibility of calculating such a probability.

Considering the fact that the analytical analysis of the G/G/N system is overwhelmingly complex and in many instances simply impossible to perform (i.e., only analytical expressions that connect service waiting time with other parameters of the system can be derived in this instance), only particular cases, related to both specific types of distributions of incoming requests stream and servicing time [1,2] will be reviewed further.

**Incoming stream of requests.** Arrival moments for requests from individual users could be distributed according to different laws: regular, deterministic or laws, which are similar to Poisson law. Concurrently, superposition of both internal and external streams occurs within each servicing device, which leads to formation of a cumulative stream. In this case, well-known stream theorems serve as a theoretical justification for a broad utilization of Poisson distribution [3].

**Servicing time.** In case of a study of information-telecommunication systems where a virtual link is a servicing device, servicing time for a single request consists of time for:
- Searching for an available channel;
- Verifying occupancy of the channel;
- Sending controlling signals indicating seizure of the channel;
- Servicing the request by itself (transaction – “communication”);
- Sending controlling requests indicating vacation of the channel.

Let us assume that N of servicing devices are located within the system and each of them is able to service simultaneously only one request. It was demonstrated that the duration of serving of i-th type requests is distributed according to exponential law with a parameter \( \mu_j = \{\mu_{1,j}, \mu_{2,j}, \ldots, \mu_{m,j}\} \), i.e. average time for execution of an operation for establishing a link between the i-th user and the j-th user is \( \tau_j = 1/\mu_{j} \), and communication time of the j-th user with k-th user is equal to the constant value \( T_j \) for every request [4]. Here \( 1 \leq j \leq L \); \( 1 \leq i \leq m \), where m is a number of different request types, residing within a system. Therefore, after completion of the transaction by the j-th type of user the request of the i-th type UW calls the k-th with a probability \( P_{i\rightarrow j} \) for an execution of a request of the i-th type. Thus, it can be assumed that servicing time is distributed according to the exponential law. Generally, a number of available publications supports this assumption [5-7].

Reduction to M/M/N/∞ system. After the pattern of distribution has been elucidated, it was determined that original description of the queueing system G/G/N/∞ can be further reduced to the system of M/M/N/∞ type, for which the following analytical expression of a probability of waiting time exceeding permitted time, otherwise known as nominal losses, can be derived:

\[
P(\tau_{\text{wait}} > \tau_{\text{perm}}) = NE_N(a)(N - a(1 - E_N(a)))e^{-(N-a)\lambda\tau_{\text{perm}}}
\]

where

\[
E_T(a) = \frac{a^N/N!}{\sum_{j=0}^{N} a^j/j!}
\]

\( \lambda \) is a parameter of a servicing time stream; \( a \) is a parameter of an incoming stream.

Evaluation of functioning efficiency. According to abovementioned types of failures of system functioning following tasks present themselves:

1) How does reduction of one component of an incoming stream of requests influence servicing efficiency (reduction a).
2) How does irreparable failure of servicing device influence servicing efficiency (reduction N).
3) What is an optimal value N, which ensures required probability of nominal losses under pre-set permissible waiting time.

Analytical expressions (1), (2) already contain required data for solving tasks 1, 2, however, solution of the task 3 is associated with a necessity of solving equation (1) numerically for N.

Typical dependencies of probability \( P(\tau_{\text{wait}}>\tau_{\text{perm}}) \) of nominal losses from N for different incoming streams are laid out on figure 2.

![Figure 2. Typical dependencies of a probability of nominal losses from number of servicing devices for different incoming streams](image-url)
3. Analysis of information streams for a management system during development.

Let us review a management system of information-telecommunication subsystem of a monitoring system. In this case, either users or a central node are a source of requests, a management system is a servicing device, a storage circuit consists of computational network equipment, which facilitates accrual, storage of requests and their release into the management system by utilizing multi-channel tools.

Let us construct an analytical description of a corresponding system for queue servicing (Figure 3), in which intervals between arrival of requests and servicing times are independent from each other and distributed equally. Our goal is to construct such a modification of a conventional description of equivalent systems, under which the load upon management system for cell communication is reduced and, consequently, its reserve productive capacity is increased. Efficiency of management system functioning will be evaluated by utilizing a probability of incoming call losses, which in this instance is understood as a fraction of incoming calls that were unable to wait (exceeded allocated waiting time) until being serviced by the system during specified time.

![Figure 3. Representation of the management system](image)

In case of a failure, a running servicing operation is interrupted. It can be assumed, that the malfunction of objects equates to emergence of a request, which has an absolute precedence over requests that are being serviced.

Let us examine a homogeneous queueing system with a quantity of places in the queue, which is limited by a value \( n \), and waiting time in the queue, which is limited by a constant value \( \tau \). If during time \( \tau \) a queueing incoming call has not started to be serviced, than it receives a denial of service and the source of the incoming call repeats the request. Incoming calls from the queue are serviced in the order of their arrival.

The method, which was outlined in [8], can be utilized for characterization description properties of given queueing system. In this paper the servicing system is described by a certain multidimensional Markov process. Let us assign indices 1, 2, …, \( N \) to communication channels and indices 1, 2, …, \( n \) – to places in the queue.

Let us examine a stochastic process:

\[
\vartheta(t) = \{\varepsilon_1(t), \ldots, \varepsilon_n(t), \ldots, \eta_1(t), \ldots, \eta_n(t)\}
\]

(3)

which comprises two stochastic processes. Here \( \varepsilon_i(t) \) signifies time, which should pass from a moment \( t \) until a channel with an index \( i \), which has been occupied by an incoming call, is cleared. If at the moment \( t \) the channel with the index \( i \) is free and there are no incoming calls awaiting in the servicing system, then \( \varepsilon_i(t)=0 \). \( \eta_j(t) \) denotes time that should pass from time moment \( t \) until the place in the queue with the index \( j \) is cleared by the incoming call, which occupied it until moment \( t \). If the place in the queue with the index \( j \) is free, then \( \eta_j(t)=0 \).

With the passage of time, non-zero components of the vector \( \vartheta(t) \) depreciate by the factor equal to the value of time period, which has passed since moment \( t \). let us examine, how components of the vector \( \vartheta(t) \) change at the moment \( t'(t'>t) \), when the first incoming call entered into the servicing system.
In case of \( \min_{i} \varepsilon_{i}(t_{-0}) = 0 \), the entered incoming call occupies a free channel, if several free channels are available than a single one can be selected with an equal probability.

Under these conditions, the component \( \varepsilon_{i}(t_{+0}) \) increases from zero to the value, which is equal to the duration of servicing of a newly entered incoming call. In case of \( \min_{i} \varepsilon_{i}(t_{-0}) \neq 0 \) free channels are absent and the incoming call occupies one of places in the queue, when the equality \( \min_{i} \varepsilon_{i}(t_{-0}) = 0 \) is exactly satisfied. In such case the component \( \varepsilon_{i}(t_{+0}) \) increases from zero to the value that is smallest of components for which the inequality \( \min_{i} \varepsilon_{i}(t_{-0}) \leq \varepsilon \) is satisfied. If \( \min_{i} \varepsilon_{i}(t_{-0}) > \varepsilon \), then the examined incoming call will leave the servicing system without occupying a channel. This means that not a single channel will be cleared during time \( \tau \), while the component \( \varepsilon_{i}(t_{+0}) \) will be incremented by \( \tau \). If \( \min_{i} \varepsilon_{i}(t_{-0}) \neq 0 \), i.e. free places in the queue are not available, then the examined incoming call will leave the serving system (in other words, the incoming call will be lost). In this case, emergence of the incoming call will not influence components of the vector \( \vartheta(t) \).

It is clear, that in the moment of time \( t > t \) the state of the multidimensional process, which has been described using the vector \( \vartheta(t) \), is fully determined by its state in the moment of time \( t \). Therefore, the stochastic multidimensional process \( \vartheta(t) \) is in fact the unidimensional Markov process.

In order to define a multidimensional stochastic process \( \vartheta(t) \), which consists of two stochastic processes \( \varepsilon(t) \) and \( \eta(t) \), let us introduce a function of a combined distribution of probabilities that \( s \) channels and \( k \) places in the queue are occupied within a servicing system in a time moment \( t \):

\[
P_{s,k}(t,x_{1},...,x_{n},y_{1},...,y_{k})
\]

provided that \( \varepsilon_{i}(t)<x_{1}, \ldots, \varepsilon_{i}(t)<x_{s}, \eta_{i}(t)<y_{1}, \ldots, \eta_{i}(t)<y_{k} \). It is obvious that the distribution function will be written as \( P_{s,0}(t,x_{1},...,x_{n}) \) when \( s<N \) for the value \( y_{i}=0 \). This means that \( s+k \) incoming calls are residing inside the servicing system. The probability of \( \vartheta(t)=0 \) is denoted as \( P_{0}(t) \).

By the virtue of the assumption about the exponential rule of distribution for servicing of incoming requests, the following inequality can be written as:

\[
P\{t, x_{i} < \varepsilon_{i} < x_{i} + \Delta t\} < (a\Delta t)^{s}
\]  \( (4) \)

Hence, it follows from this inequality that the distribution function \( P_{s,0}(t,x_{1},...,x_{n}) \) is continuous within corresponding regions of \( s \)-dimensional space. The given statement is obvious for \( P_{s,0}(t,x_{1},...,x_{n},y_{1},...,y_{k}) \), because in this case the duration of waiting cannot exceed the value \( \tau \). From here it follows that densities of probability distributions \( p_{s,0}(t,x_{1},...,x_{n}) \) and \( p_{N,k}(t,x_{1},...,x_{n},y_{1},...,y_{k}) \) do exist. For brevity, in differential equations they will be denoted as \( p_{s,0}(t) \) and \( p_{N,k}(t) \), correspondingly.

As \( 1 \leq s \leq N, k=0 \) differential form of state equations for the servicing system will be the same as equation for the servicing system with limited waiting time (henceforth, the sign \( \delta \) denotes partial derivation) [9]:

\[
\frac{\delta p_{0,0}(t)}{\delta t} = -\lambda p_{0,0}(t) + Np_{1,0}(t)
\]

\[
\frac{\delta p_{s,0}(t)}{\delta t} - \frac{\delta p_{s,0}(t)}{\delta x_{1}} - \cdots - \frac{\delta p_{s,0}(t)}{\delta x_{s}} = -\lambda p_{s,0}(x_{1},...,x_{s}) + (N-s)p_{s+1,0}(x_{1},...,x_{s}) + \frac{+\lambda a}{N-s+1}\sum_{i=1}^{s}p_{s-1,0}(x_{1},...,x_{i-1},x_{i+1},...,x_{s})e^{-ax_{i}}
\]  \( (6) \)

Let us examine the case when \( s=N, 1 \leq s \leq n \). At a time moment \( t+\Delta t \) the servicing system can be found in the state \( \eta(t+\Delta t)=y_{k}, \xi(t+\Delta t)=x_{N} \) only under following assumptions (let us enumerate only following conditions, from which the system can transition into the state \( \eta(t+\Delta t)=x_{4} \)):

1) During period \( \Delta t \) new incoming call have not arrived and waiting time for incoming calls has not lapsed. The probability of this event is equal to:
(1 - λΔt)P_u^* p_{N,k}(t, x_1 + Δt, ..., x_N + Δt, y_1 + Δt, ..., y_k + Δt) + o(Δt).

where $P_u^*$ is a probability that time of residence in the queue has not lapsed for waiting incoming calls during period $Δt$.

2) At a time moment $t + k + 1$ places in the queue were occupied in the servicing system, during period $Δt$ one of N devices became free and available. The probability of this event is equal to:

$$NaΔtp_{N,k+1}(t, x_1 + Δt, ..., x_N + Δt, y_1 + Δt, ..., y_k + Δt) + o(Δt)$$

3) At a time moment $t + k - 1$ places in the queue were occupied in the servicing system, during period $Δt$ an incoming call, which will occupy a place in the queue for a time period $y_k$, has arrived in the system. The probability of this event is calculated as follows:

$$\frac{λaΔt}{Na - λP_u^*} \sum_{j=1}^{k} p_{N,k+1}(t, x_1 + Δt, ..., x_N + Δt, y_1 + Δt, ..., y_j - 1 + Δt, y_j + 1 + Δt, ..., y_k + Δt) e^{-αy_j} + o(Δt)$$

4) At a time moment $t + k + 1$ waiting incoming calls were residing within the servicing system; over a period $Δt$ waiting time has lapsed for one of the incoming calls. The probability of such an event equals to:

$$(1 - λΔt)P_u^* p_{N,k+1}(t, x_1 + Δt, ..., x_N + Δt, y_1 + Δt, ..., y_k + Δt) + o(Δt).$$

$P_u$ denotes the probability of waiting time for one of $k+1$ incoming calls lapsing during time period $t$, correspondingly, $P_u^* = 1 - P_u$. After completing series of transformations and proceeding to limit when $Δt \to 0$ the system of differential equations can be obtained:

$$\frac{δp_{N,k}(t)}{δt} - \frac{δp_{N,k}(t)}{δy_1} - ... - \frac{δp_{N,k}(t)}{δy_k} = -λ(1 - e^{-γ(Na - λ)})p_{N,k}(x_1, ..., x_N, y_1, ..., y_k) +$$

$$+ \frac{λa(1 - e^{-γ(Na - λ)})}{Na - λe^{-γ(Na - λ)}} \sum_{j=1}^{k} p_{N,k-1}(x_1, ..., x_N, y_1, ..., y_j - 1, y_j + 1, ..., y_k)e^{-αy_j} +$$

$$+ λe^{-γ(Na - λ)}p_{N,k+1}(x_1, ..., x_N, y_1, ..., y_k, 0)$$

The system of N+n+1 equations (6) and (7) with N+n+1 unknown functions $p_{s,0}(t,x_1,...,x_s)$, $0 \leq s \leq N$, $p_{N,k}(t,x_1,...,x_N,y_1,...,y_k)$, $1 \leq k \leq n$ is a sought-for system of differential equations.

The stationary mode of the servicing system is of particular interest. Let us assume that while $t \to \infty$ the stochastic process comes arbitrarily close to a certain stationary mode, which is independent of initial values $ψ(0)$, i.e., the distribution of the vector $ψ(t)$ converges to the stationary asymptotic distribution.

In stationary mode both $p_{s,0}$ and $p_{N,k}$ functions do not depend on time $t$. In these circumstances equations (6) and (7) are simplified, the function $p_{s,0}(t)$ turns into the constant value $p_{s,0}$ (let us denote it as $p_{0,0}$).

The distribution of servicing durations for incoming calls is exponential, therefore when $t=0$:

$$p_{s,0}(x_1, ..., x_s) = α_{N,k} e^{-α(x_1, ..., x_N, y_1, ..., y_k)}$$

where
\[ a_{s,0} = \frac{\lambda^s(n-s)}{N} p_0 \]  
\[ a_{N,k} = \frac{\lambda^{N+1} (1 - e^{-\tau(Na-\lambda)})^k}{N! (Na - \lambda)^k} p_0 \]

After series of transformations, the following expression can be obtained:

\[ p_0^{-1} = \sum_{i=0}^{N} \frac{y_i^i}{i!} + \sum_{j=1}^{N} \left[ \frac{y(1-e^{-\tau(N-y)})}{N-y e^{-\tau(N-y)}} \right]^j \]

During the first stage of engineering of servicing systems with limited waiting time the question arises at to specifications of waiting devices (that are required) in order to assure the pre-defined quality of servicing, and, conversely, what will the servicing quality be for a specific number of servicing places.

Meanwhile, in a number of occasions it might be required to compare the servicing system with limited waiting time against the servicing system with losses. The probability of losses of incoming calls is a convenient criterion for such a comparison.

An incoming call, which has entered the servicing system, will be rejected in the case when both all N channels and n places in the queue are occupied. Besides that, the waiting incoming call, which was not serviced during time \( \tau \), will also be lost. The probability of losses of incoming calls is equal to the sum of the probability \( P_I \) of N channels and n places in the queue being occupied and the probability \( P_{II} \) of the incoming call not being serviced during time \( \tau \): \( P = P_I + P_{II} \). Following expression can be derived for the loss probability for incoming calls:

\[ P_I = \frac{\sum_{i=0}^{N} \frac{y_i^i}{i!} \left( 1 - e^{-\tau(N-y)} \right)^i}{\sum_{i=0}^{N} \frac{y_i^i}{i!}} \]

where \( y \) is intensity of incoming load;  
N – number of devices;  
N – number of places in the queue;  
\( \tau \) – limited waiting time for being serviced;  
1/a – average duration of channel occupancy.

Formula (10) enables us to calculate dependence of loss probability on factors that limit the waiting queue (number of places in the queue and maximum time of residence in the queue) in case of a simple stream of incoming calls which are arriving for servicing and exponential rule of channel occupancy distribution.

4. Conclusions
As it can be seen from abovementioned, the introduction of limited waiting ensures load reduction on the management system, thus, it facilitates increase of the stand-by performance which is required for servicing new objects of the system.

References
[1] Stefan Traub and Hongyan Yang 2019 Tax Competition and the Distribution of Income Scandinavian Journal of Economics
[2] Vivek S Pai et al. 1998 Locality-Aware Request Distribution in Cluster-Based Network Servers ACM SIGPLAN Notices 33(11)
[3] Thompson W J 2001 Poisson Distributions Computing in Science and Engineering 3(3) 78–82
[4] Shareiko L A 1986 An integrated approach to modeling and optimization of information
transmission systems (M.: Nauka) pp 92-9

[5] Rutgaizer V M and Sheviakhov Iu E 1987 Distribution According to One's Labor Problems of Economic Transition 30(7) 31-46

[6] Bukhalo O P and Dmitriv Z V 1988 Generator of pseudorandom numbers distributed according to a poisson law Measurement Techniques 31(10) 937-40

[7] Dallas A C 1979 On the exponential law Metrika 26(1) 105-8

[8] Gómez-Aguilar J F and Atangana Abdon 2019 Power and exponentials laws: Theory and application Journal of Computational and Applied Mathematics

[9] Kovalenko I N 1960 The study of a multi-line service system with a queue and a limited time spent in the system Ukrainian Mathematical Journal 12(4) 471-6

[10] Zhongshun Shi, Zewen HUANG and Leyuan Shi 2017 Two-stage scheduling on batch and single machines with limited waiting time constraint Front. Eng. Manag. 4(3) 368–74