GENERALIZED PRIORITY SERVICE AND ITS APPLICATION IN ANALYSIS OF ONE-CHANNEL DATA TRANSMISSION SYSTEMS

Isaac M. Dukhovny and Vladimir I. Pankratov
Institute of Information Transmission Problems, USSR Academy of Sciences
Moscow, USSR

§ 1. INTRODUCTION
1. In working out complex communication and control systems with several input information flows, a question naturally arises on the priority rule, i.e. the order in which calls are serviced. In studying the models of appropriate queueing operation, the following priority disciplines are used: first-come-first-served, non-preemptive priority and different preemptive priorities. A number of works deal with investigations of the influence of these service disciplines on the quality of one-channel and multichannel systems at various analytical limitations, on the queue length, on the structure of input flows and on the service-time distribution function (Cobham, A. [5]; Kesten, H. and Runenngen, J. [10]; Miller, R.G. [2]; Jaiswal, N.K. [3]; Gaver, D.P. [6]; Chang, W. [5]; Basharin, G.P. [2]; Klimov, G.P. [13], and others).

2. All the above works assume that the priority class of a call is determined by an out-of-the-system mechanism and served irrespective of the specificity or state of the service device. However, there are many applications in which the selection discipline is more anatural consequence of the service specificity than of the service urgency. Thus, in many cases and, particularly, in industry, the reswitching of service from one class of calls to another very often leads to additional losses (or rebuffering), and, besides, frequent reswitching may have a negative influence on the reliability of various systems. For example, often reswitching of an operator from one type of information to another leads to many errors in the information being transmitted [14]. In such cases, none of the above disciplines can be used, as they allow frequent reswitchings.

3. There are very few works investigating in detail the disciplines that take into account the specificity of service devices. We may name only an article by B. Avi-Itzhak, W.L. Maxwell, and L.W. Miller, 1965 [1] and the work by L.Takacs [13], published three years later. The authors [1] considered a one-channel system with two Poisson input flows, an absolutely reliable device and service discipline which they called alternating priority, the essence of which is as follows: if call i (i = 1, 2) arrives at a moment when call j (j ≠ i) is being served, it will be served only after all the calls j, as well as calls i that arrived earlier, are served. Out of all known disciplines that do not permit the device to idle if calls are waiting, the alternating priority requires the least number of reswitching. Basing on the assumption that service-time distribution functions for both types of calls can be differentiated, the authors [1] determined the main characteristics of the system operation.

L.Takacs [13] investigated the same system by Kendall method and obtained similar characteristics at arbitrary service-time distribution laws. It is evident that the case with two input flows is rather degenerated both theoretically and practically, if this discipline is used. One of the authors [7] studied the system with alternating priorities and arbitrary number of flows n. Some of the results concerning systems with cyclic service are presented in [5].

4. The present report proposes and analyses a one-server system with generalized non-preemptive priorities, taking into account probable unreliability of a serve device. (The practical significance of studying service systems with several input flows and a "unreliable" device is obvious, though, as far as we know, there are very few works on this subject). The content of the report is as follows. § 2 presents the mathematical definition of the study, introduces necessary notations and points out the general approach to the solution. § 3 offers functional equations describing the system's operation and determines the main service characteristics. The results of the theoretical study of §§ 2-3 are used in § 4 to numerically analyse the functioning of one system transmitting data of a numerical complex in which processing of input data leads to the formation of several information flows to be transmitted along one channel.

Due to the lack of space, the report will be presented in an abstracted form.

§ 2. PROBLEM NOTATIONS. SOLUTION.
Let us consider a one-server system at which n independent stationary non-ordinary Poisson call flows U_i, i = 1, 2, ..., n, arrive with arbitrary service-time distributions. As to the serve device, the following assumptions are made. If at moment T the device started servicing a call of the i-th flow and the service time is ≥ t, then the probability of the device failing in the interval (T, T + t) is C_i(t), i = 1, 2, ..., n. After the device is renewed, the renewal time is an arbitrary value with distribution function D(t). The interrupted call of the i-th flow is served to the end. If at some moment the device becomes free and for time t no calls arrive, the probability of the device failing in the interval (T, T + t) is E(t). The renewal time is an arbitrary value with distribution function F(t).

The selection of calls to be served is determined by the generalized non-preemptive priority discipline which assumes the input flows to be divided into two classes: the first one includes U_1, U_2, ..., U_l and the second U_{l+1}, ..., U_n flows, where l is an integer, 1 ≤ l ≤ n. The calls of the first class have non-preemptive priority over those of the second class. Among the first-class calls the alternating priority is observed [1, 7, 13], while ordinary non-preemptive priority is the law for the second-class calls.

For the sake of convenience, the discipline of non-preemptive priority is denoted as f_1, alternating priority f_2 and generalized non-preemptive
priorities $f_{31}(1)$. The name of discipline $f_{31}(1)$ justifies itself, since it is obvious that $f_{31}(1) = f_1$ and $f_{31}(1) = f_3$.

The notations are as follows:

- $\lambda_i$ - the intensity of "calling moments" of the $i$th flow, $i = 1,n$; $\sum_{i=1}^{n} \lambda_i = \Lambda$.
- $V_i$ - the number of calls in an arbitrary "calling moment" of the $i$th flow.

The notations are as follows:

- $A_i$ - the intensity of "calling moments" of the $i$th flow, $i = 1,n$.
- $A_i = \sum_{j=1}^{n} A_{ij}$.

- $\Phi_i(z) = \sum_{j=1}^{n} \frac{V_i}{j} -$ the generating function of the $i$th flow at a "calling moment", $i = 1,n$ (M is mathematical expectation).

- $B_i(t)$ - the service-time distribution function for a call of the $i$th flow; $\beta_i(t) = \int_{0}^{t} e^{-s \theta_i(t)} ds$.

- $P_{im}(k_1, \ldots, k_n)$ - the probability that the $m$th call (the calls are numbered in the order of their service) is a call of flow $U_i$ and on leaving the device, having been served, leaves queue $k = (k_1, \ldots, k_n)$. $k_j > 0$, $j = 1,n$.

- $W_m(t)$ - the distribution function of waiting-time for the $m$th call, if it is a call of the $i$th flow, $i = 1,n$.

- $P_{im}(x_1, \ldots, x_n) = \sum_{k_1 = 0}^{\infty} \sum_{k_2 = 0}^{\infty} \ldots \sum_{k_n = 0}^{\infty} P_{im}(k_1, k_2, \ldots, k_n) x_1^{k_1} \ldots x_n^{k_n}$.

The investigation method used to solve the problem is based on the construction of the Markoff imbedded chain, with the use of introducing an additional event, and the analysis of appropriate functional equations. The method of introducing an additional event is described for the first time in the works by Kesten and Runnenberg [10] and Dantzig [9], it is better described by G.P. Klimov [13].

The advantage of this method is that it assigns a simple probability sense to the Laplace-Stieltjes transforms, generating function, etc., which very often occur in the queueing theory. In order to construct the imbedded Markoff chain describing our system's operation, the problem is simplified and reduced, similarly to [11], to the same problem for a system that does not fail when occupied.

For this purpose, $H_i(t)$, the arrival time distribution function for a call of the $i$th priority, $i = 1,n$, is introduced. If $P_{ki}(t)$ is the probability that during time $t$, needed to serve a call of the $i$th flow, the device fails $k$ times, and

$$P_i(z,t) = \sum_{k=0}^{\infty} e^{-k \lambda_i} P_{ki}(t), i = 1,n; |z| < 1,$$

then

$$h_i(s) = \int_{0}^{\infty} e^{-st} P_i(s,t) dB_i(t), i = 1,n;$$

Relations (2.3) is proved in the same way as in 11, from (2.3) it follows that when $t_1(t) = 1 - e^{-\alpha t}$, $t \geq 0$, $i = 1,n$,

$$h_i(s) = \beta_i(s + c_i - c_i \delta(s)).$$

§ 3. FUNCTIONAL EQUATIONS. THE MAIN THEOREM

Functional equations describing the operation of the system under consideration in the transitional condition take the form:

$$x_{i} P_{im}(x) = \left\{ \begin{array}{l} \left[ \sum_{j=1}^{n} [P_{jm}(0^{+}x) - P_{jm}(0^{+}x)] + \left[ P_{im}(0^{+}x) - P_{im}(0^{+}x) \right] \right] + \\
\sum_{j=1}^{n} [P_{jm}(0^{+}x) - P_{jm}(0^{+}x)] + \left[ P_{im}(0^{+}x) - P_{im}(0^{+}x) \right] \right\} x_{i} P_{im}(x), \quad i = j; \quad m \geq 1, |x| < 1,$$  

$$h_i(s) = \delta_i (s - \lambda_i \Phi), \quad i = j; \quad m \geq 1, |x| < 1,$$  

Then

$$P_{im}(1^{x_1}, \ldots, 1^{x_n}) = P_{im}(1^{x_1}, \ldots, 1^{x_n}) w_{im}(1^{x_1}, \ldots, 1^{x_n}),$$

$$v_{im}(s) = w_{im}(s) h_i(s), \quad Res > 0.$$  

where

$$w_{im}(s) = \int_{0}^{\infty} e^{st} dW_i(t), \quad w_{im}(s) = \int_{0}^{\infty} e^{st} dW_i(t).$$

In (3.1)-(3.3) the following notations are used:

$$x = (x_1, \ldots, x_n); \quad (u^x) = (u^x_1, \ldots, u^x_n); \quad (u^x v^y) = (u^x_1 v^y_1, \ldots, u^x_n v^y_n); \quad (u^x v^y) = (x^y_1, \ldots, x^y_n); \quad (u^x v^y) = (x^y_1, \ldots, x^y_n), \quad i = j; \quad m = 1.$$

Here, $n$-dimensional vectors are considered.

Functions $P_i(x_1, \ldots, x_n), i = 1,n$, are given by:

$$P_i(x_1, \ldots, x_n) = \frac{1 - e^{-\lambda_i}}{1 - e^{-\lambda_i} \varphi_i} \varphi_i P_i(x_i) + \frac{e^{-\lambda_i}}{1 - e^{-\lambda_i} \varphi_i} \varphi_i \left\{ \varphi_i \left( \frac{\varphi_i}{\varphi_i - \lambda_i} \right) - \varphi_i \left( \frac{\varphi_i}{\varphi_i - \lambda_i} \right) \right\}.$$  

Here

$$e(s) = \int_{0}^{\infty} e^{st} dE(t), \quad \varphi(s) = \int_{0}^{\infty} e^{st} d\varphi(t), Res > 0, \frac{1}{\varphi_i} = \lambda_i.$$
where 
\[ [\lambda \Phi]_{k} = \sum_{j=0}^{n} \lambda_{j} \phi_{j}(x_{j}) \]
\[ \lambda \Phi = \sum_{j=0}^{n} \lambda_{j} \phi_{j}(x_{j}) \]
\[ \lambda_{j} \Phi = \sum_{j=0}^{n} \lambda_{j} \phi_{j}(x_{j}) \]
\[ k = 1, \ldots, n \]
\[ [\lambda \Phi]_{0} = 0 \]

Let us explain the probability sense of function \( f_{i}(x_{1}, \ldots, x_{n}) \) when each call is either "red" or "blue", an arbitrary call is considered "red" with probability \( \frac{x_{i}}{x_{j}} \) of the colour of other calls. Then, if \( f_{i}(x_{1}, \ldots, x_{n}) \) is the probability that, after a call has been served the system is free, the next call to be served is a call of the i-th priority, and all the calls that arrived in the system before this call had been served are "red". The proof of relation (3.4) is omitted. The detailed procedure of introducing an additional event when deriving equations (3.1) - (3.3) is given in [7, 11].

The main results of the present work are presented by the following theorem.

**Theorem**

a) \( P_{im}(x_{1}, \ldots, x_{n}), W_{im}(t), V_{im}(t) \) for \( |x_{j}| \leq 1, i = 1, m \) may be determined by recurrent formulae (3.1)-(3.3), where functions \( h_{1}(s) \) are given by (3.4); 
\[ h_{1}(s) = \frac{e^{-st}}{s} P_{im}(s, t) dC_{n}(t), i = 1, \ldots, m \]
\[ \frac{e^{-st}}{s} P_{im}(s, t) dC_{n}(t), i = 1, \ldots, m \]
b) \( h_{1}(s) \) is the Laplace-Stieltjes transform from distribution function \( H_{i}(t) \) of the time interval covering the beginning and end of the servicing of a call of the i-th flow.

c) if \( \beta = \sum_{j=1}^{n} \lambda_{j} M_{j} h_{j} < 1 \) and \( \varphi_{i} = \int_{0}^{\infty} dF(t) < \infty \)
\[ \lim_{t \to \infty} P_{im}(x_{1}, \ldots, x_{n}) = P_{im}(x_{1}, \ldots, x_{n}) \]
\[ \lim_{t \to \infty} W_{im}(t) = W_{im}(t) \]
\[ \lim_{t \to \infty} V_{im}(t) = V_{im}(t) \]

b) \( h_{1}(s) \) is the Laplace-Stieltjes transform from distribution function \( H_{i}(t) \) of the time interval covering the beginning and end of the servicing of a call of the i-th flow.

c) if \( \beta = \sum_{j=1}^{n} \lambda_{j} M_{j} h_{j} < 1 \) and \( \varphi_{i} = \int_{0}^{\infty} dF(t) < \infty \)
\[ \lim_{t \to \infty} P_{im}(x_{1}, \ldots, x_{n}) = P_{im}(x_{1}, \ldots, x_{n}) \]
\[ \lim_{t \to \infty} W_{im}(t) = W_{im}(t) \]
\[ \lim_{t \to \infty} V_{im}(t) = V_{im}(t) \]

b) \( h_{1}(s) \) is the Laplace-Stieltjes transform from distribution function \( H_{i}(t) \) of the time interval covering the beginning and end of the servicing of a call of the i-th flow.

c) if \( \beta = \sum_{j=1}^{n} \lambda_{j} M_{j} h_{j} < 1 \) and \( \varphi_{i} = \int_{0}^{\infty} dF(t) < \infty \)
\[ \lim_{t \to \infty} P_{im}(x_{1}, \ldots, x_{n}) = P_{im}(x_{1}, \ldots, x_{n}) \]
\[ \lim_{t \to \infty} W_{im}(t) = W_{im}(t) \]
\[ \lim_{t \to \infty} V_{im}(t) = V_{im}(t) \]

d) \( P_{im}(x_{1}, \ldots, x_{n}) \) for all \( |x_{j}| \leq 1, i = 1, \ldots, m \) is given by
\[ P_{im}(x_{1}, \ldots, x_{n}) = \frac{\phi_{i}(x_{1}, \ldots, x_{n})}{h_{j}(s)} \]
where
\[ \phi_{i}(x_{1}, \ldots, x_{n}) = \frac{\phi_{i}(x_{1}, \ldots, x_{n})}{h_{j}(s)} \]

where
\[ \omega_{i}(s) = \int_{0}^{\infty} dW_{i}(t), \quad \omega_{i}(s) = \int_{0}^{\infty} dV_{i}(t), \quad i = 1, \ldots, n \]

Functions \( \psi_{k}(x_{1}, \ldots, x_{n}), k = 1, \ldots, n \), which are part of (3.7), satisfy the following functional equations:

\[ \sum_{i=1}^{n} \left( \phi_{i}(x_{1}, \ldots, x_{n}) - \sum_{j=1}^{n} \lambda_{j} h_{j}(x_{1}, \ldots, x_{n}) \right) \psi_{k}(x_{1}, \ldots, x_{n}), \quad i = 1, \ldots, n \]

Here, \( \omega_{i}(s) = \int_{0}^{\infty} dW_{i}(t) = \int_{0}^{\infty} dV_{i}(t) \) for all \( |x_{j}| \leq 1, i = 1, \ldots, m \) the first two moments of the waiting time are found. The mean waiting time has this comparatively simple formula:

\[ \frac{\omega_{k}}{\lambda_{k} h_{k}(M_{k} + M_{k} + M_{k}) + \lambda_{k} h_{k}(M_{k} + M_{k})}, \quad k = 1, \ldots, n \]

Values \( \rho_{k} = \frac{\phi_{k}(x_{1}, \ldots, x_{n})}{\phi_{k}(x_{1}, \ldots, x_{n})} \) for \( |x_{j}| \leq 1, i = 1, \ldots, n \) are given by recurrent formula

\[ \rho_{k} = \left( d_{k} + \sum_{i=k+1}^{n} \rho_{i} \right)^{-1}, \quad \rho_{k} = \left( d_{k} + \sum_{i=k+1}^{n} \rho_{i} \right)^{-1} \]

where
\[ d_{k} = \left( \frac{\rho_{k}}{\lambda_{k} h_{k}(M_{k} + M_{k}) + \lambda_{k} h_{k}(M_{k} + M_{k})}, \quad k = 1, \ldots, n \right) \]
The result $\lfloor \frac{2}{3} \rfloor$ is obtained from (3.15), if $n = 2$, $\lambda = 2$ and the device is absolutely reliable.

§ 4. APPLICATION AND NUMERICAL RESULTS

The above queueing model with generalized non-preemptive priorities was used to study a one-channel system transmitting data from periphery to the computing centre. Each transmission point generates three Poisson flows of different urgency ($n = 3$). Technologically, all the arriving calls are divided into urgent (1st priority), operative (2nd priority) and non-urgent (3rd priority). The influence of various transmission modes, i.e. the service discipline and reliability parameters of data transmitting apparatuses, on the waiting time in the system periphery was studied.

Apart from the non-preemptive priority discipline $f_1$, it is expedient to analyse the characteristics of a system with alternating priorities $f_3$, because this discipline simplifies the selection and transmission of information and minimizes the time of sorting out and preliminary processing of data in the central computer due to grouping of monotype messages of the arriving information. But in case of discipline $f_3$ it may turn out that the time of delay of urgent messages is greater than the given value. This may occur, for instance, in case of a great intensity of a flow of non-urgent data due to the appearance of long cycles of non-urgent data transmission. Then it is possible to obtain satisfactory results as to the time of data delay, using the generalized non-preemptive priority discipline $f_3$ (1). Besides, by comparing the mean delay time $W$ of data from the overall flow

$$W = \sum_{k=1}^{n} \lambda_k \omega_{k_1}$$

and accounting for the above quantitative aspects, the optimal service discipline is selected in data transmission systems.

The analysis of characteristics of delay time of each flow was made in a data transmission system with the following ratio of priority messages $p_1$, ..., $p_n$ in the overall flow: $p_1 = 2\%$; $p_2 = 40\%$; and $p_3 = 18\%$. Since the intensity of the overall flow $\lambda$ has its value for each peripheral point, the calculation of the delay time characteristics depended on the total traffic

$$\rho = \lambda_1 \beta_{11} + \lambda_2 \beta_{21} + \lambda_3 \beta_{31}$$

where $\lambda_k = \Lambda \rho$, $k = 1, n$.

$\beta_{k1}$ is the mean time of transmission of data of the $k$th priority ($\beta_{11} = 2.0$ min, $\beta_{21} = 1.5$ min, $\beta_{31} = 3.5$ min).

In calculations, the time distribution of the transmission of items from each flow was assumed exponential. No results obtained under different time distributions are given.

Fig. 4.1 shows the dependence of the mean value and coefficient of the delay time variation for items of the $k$th priority on the total traffic of system $\rho$ in case of the non-preemptive priority discipline. The changes of variation $V_k$ coefficients are given only for items of the 1st priority, since for other flows the differences are insignificant. The characteristics under study are dash lined, the reliability of data transmission apparatus and connection channel are taken into consideration. The mean time of reliable operation of servicing complex $1/\lambda_0$ is assumed to be 100 hours, while the mean recovery time $\beta_{01} = 38$ min. The reliable time and recovery time distributions are considered exponential and Erlang of the third order, respectively.

The same characteristics of delay time of messages in case of the alternating priority discipline $f_3$ are given in Fig. 4.2.
Discipline $f_3$ (Fig. 4.2) results in great delays of urgent messages in peripheral points with heavy total traffic. Therefore, the generalized non-preemptive priority discipline $f_{31}(1=2)$ was considered, i.e., alternation of priorities only for the flows of the 1st and 2nd category of urgency. (Fig. 4.3).

The comparison of these service disciplines by criterion $W$ (4.1) shows that in this case the minimum value of the mean delay time is observed if discipline $f_{31}(1=2)$ is employed. The numerical values of $W$ for some values of $p$ are given in Table 4.1.

<table>
<thead>
<tr>
<th>$p$</th>
<th>Criterion $W$ (in min)</th>
<th>$f_1$</th>
<th>$f_3$</th>
<th>$f_{31}(1=2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>1.185</td>
<td>1.222</td>
<td>1.179</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>2.463</td>
<td>2.630</td>
<td>2.438</td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>5.166</td>
<td>5.852</td>
<td>5.089</td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>17.030</td>
<td>22.262</td>
<td>16.807</td>
<td></td>
</tr>
</tbody>
</table>

The absolute increase in the mean delay time due to the limited reliability of the data transmission apparatus and connection channel is shown in Figs. 4.1 - 4.3. Table 4.2 presents the results of calculation of the relative increase in the mean delay time of messages of various priority when the service discipline used is $f_1$.

<table>
<thead>
<tr>
<th>Relative increase in $W_{x1}$ for $f_1$ %</th>
<th>k = 1</th>
<th>k = 2</th>
<th>k = 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>68.5</td>
<td>68.5</td>
<td>68.6</td>
</tr>
<tr>
<td>0.3</td>
<td>23.7</td>
<td>23.9</td>
<td>24.2</td>
</tr>
<tr>
<td>0.5</td>
<td>14.9</td>
<td>15.3</td>
<td>15.8</td>
</tr>
<tr>
<td>0.7</td>
<td>11.1</td>
<td>11.8</td>
<td>13.2</td>
</tr>
<tr>
<td>0.9</td>
<td>9.1</td>
<td>10.2</td>
<td>16.4</td>
</tr>
</tbody>
</table>

When designing data transmission systems, the requirements to the apparatus reliability are determined by the influence of the processes of refusal and recovery on the delay time. Fig. 4.4 illustrates the dependence of the mean delay time of call $W$ on the mean time of reliable operation $1/A_0$ and mean recovery time of the servicing system.

In this example, an increase in the mean time of reliable operation by more than 500 hours practically does not affect $W$ irrespective of the organisation of the repair service.

Since the mean value is not the only characteristic of the random value, to evaluate the degree of spread of the delay time, the dispersion and variation coefficient $V_k(k=1,n)$ of the delay time were calculated, depending on the reliability parameters of the serve device. Fig. 4.5 shows the dependence of the variation coefficient of urgent items (k=1) on the mean recovery time. The large variation coefficients point to the need of taking into account the delay time variation when working out the requirements of the time of supplying information to the computing centre.
If the mean delay time $W$ is limited by some value $W_{\text{lim}}$, a non-dimensional value
$$\tilde{W} = 1 - \frac{W}{W_{\text{lim}}}$$
(4.3)
will serve as a criterion for the quality of the system operation.

The servicing system will satisfy the given requirements if $\tilde{W} > 0$. The dependence of $\tilde{W}$ on traffic, for different mean recovery times $\beta_0$, is given in Fig. 4.6 ($W_{\text{lim}} = 10 \text{ min}$). The plots, corresponding to various $\beta_0$, lay on the abscissa values of possible total traffic above which $\tilde{W}$ exceeds the given value of $W_{\text{lim}}$. For peripheral points, where the total traffic exceeds the allowed one, emergency equipment is required.

All the calculations of the delay time characteristics were made on digital computers by programmes compiled by the authors.

REFERENCES


