Parallel Waiting Queues in Real-Time Computer Systems

By **Paul Kühn**

UDC 681.32.004

A Report) from the Institute for Switching and Data Technics, University of Stuttgart, Germany*

1. Introduction

Real-time conditions arise in computer systems where a quick response is required as for automatic control, centralized reservation, banking transactions, man-machine communication, and shared computer use. In many cases there is a great number of users and each of them will communicate with the computer system either on-line or off-line via buffering facilities.

The great number of requests, their individual differences, and the real-time conditions cause in general a complex computer structure in hard- and software. For example, let us consider the flow of data in a typical multi-thread real-time system [1]; *Fig. 1.*

At the line control equipment messages are received which were sent from remote terminals, concentrators, or computers. After having analyzed the messages according to their urgency, real-time requests are led to the new input queue, non-real-time requests are stored and will be put into the non-real-time queue later on when it is convenient. During programme executing an interrupt may occur if there is either a message of higher priority or if some data are needed from the peripheral memories. Interrupted requests are waiting in the work-in-progress queue until the main scheduling routine gives service to them. After total service of a request the outgoing message waits in the output queue until the input/output schedule gives way to transfer it to the request origin.

During service the various requests have to pass different bottlenecks within the system which turn out to be critical with respect to response time and throughput requirements. Bottlenecks can be formed by limited storage capacity in core memory, peripheral files, and buffer divices, by limited utilization of terminals, communication lines, multiplex channels, and access mechanisms as well as by limited speed of precessing units, memory actions, and transfer media.

For better utilization of such systems they are not only fed by high priority real-time tasks but also by low priority batch programmes. To run such systems economically proper structures and organizations have to be found. The main features a system designer is concerned with are

1. Waiting and response times should be as small as possible.

2. Optimum utilization of the capacity of the service system, i. e. maximization of throughput without delaying the important requests. A bottleneck in the processing unit, for example, can be reduced by introduction of parallel processing; the total response times with respect to all requests can be minimized by an organization (discipline) which serves the short jobs with a higher priority. Taking the statistical behaviour of arrival times and service times into account, dimensions have to be laid such that there are no longer critical factors in the system.

A great difficulty for this procedure is the fact that data needed to implement a system are not available until the system is implemented. In many cases, however, systems have to be laid out for some future conditions so that investigation of bottleneck mechanisms is worth while in order to find proper structures and efficient disciplines to meet the prescribed requirements.

The best known tool for the investigation of such systems is the simulation technique. A somewhat simplified mechanism is drawn from reality and compressed into a mathematical model. This model can be simulated on a digital computer. As a result answers can be found as

- distributions of queue sizes
- distributions of waiting and response times

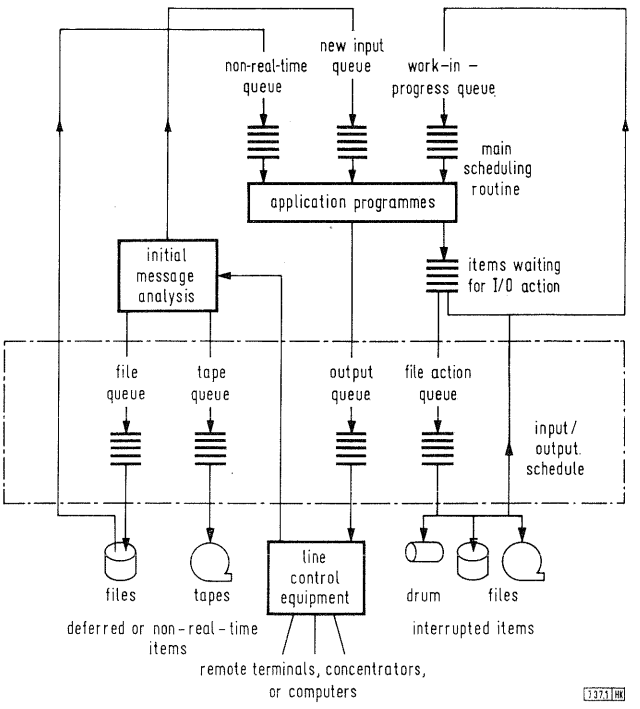


Fig. 1. The flow of data in a real-time computer system.

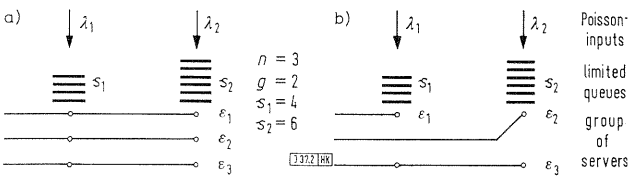


Fig. 2. Basic multiqueueing models.
a) full available servers,
b) limited available servers.

*) According to a paper read at the "International Computing Symposium 1970" organized by the European Chapters of the Association for Computing Machinery (ACM) in cooperation with Gesellschaft für Mathematik und Datenverarbeitung, Bonn, Germany, 21.–22. 5. 1970.

- utilization of various devices
- throughput of the model.

Parameters for the simulation programme can be changed easily so that an iterative process for optimization of the system can be performed.

Simulation techniques, however, involve a great disadvantage: for each new combination of parameters a relatively long computer run is necessary so that simulation proves to be unhandy for complex computer structures. For such cases mathematical analysis can form an alternative or, at least, a supplementation to simulation. Knowing the relations between input and output of separated subsystems the variation of parameters for simulation can be limited drastically.

The single server system with one queue is the most investigated queueing model. Extensions to that model are many server systems with one queue, tandem queueing systems without or with different input feeds, feedback queueing systems, and parallel waiting line systems without or with feedback.

This paper focuses on systems with parallel waiting queues. A distinction will be made between systems where requests of a certain input queue have access to all servers (full availability) or only to a limited number out of all serves (limited availability).

Fig. 2 shows two basic structures of a 3-server system with full (a) and limited availability (b). The number of servers is $n=3$. Each model has $g=2$ limited input queues with s_j, $j=1, 2$, waiting places. Requests arrive at each queue according to a Poisson-process with mean arrival rates λ_j, $j=1, 2$. The service times are negative exponentially distributed with mean $h_i=1/\varepsilon_i$ for the i-th server, $i=1, 2, 3$. (Markovian assumptions).

The queue disciplines are "first-come, first-served" service (D1), "random-served" service (D2), and "last-come, first-served" service (D3) within the queues. The service of a certain queue occurs with probability p_j, $j=1, 2$. For special values of p_j the cases of nonpre-emptive priorities with fixed priority class waiting capacity, of cyclic order service, and of service according to the queue lengths are obtained.

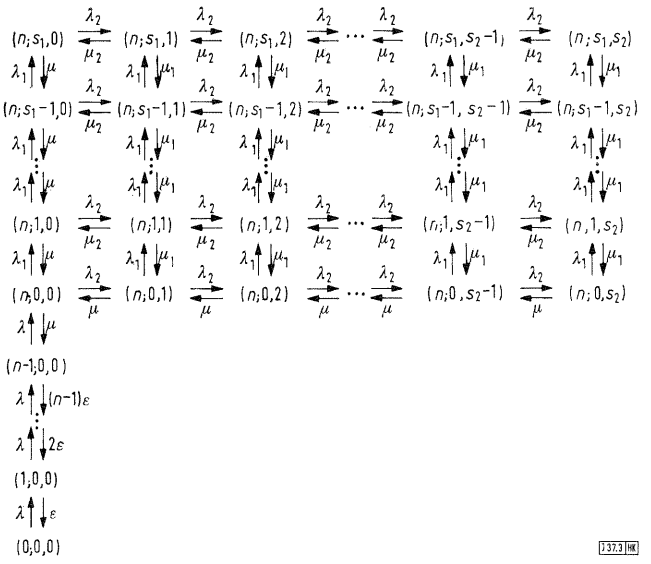


Fig. 3. State space and transition coefficients for the full available n-server system with two queues.

The characteristic values to be calculated are the probabilities of waiting and loss (overflow), the mean queue lengths, and the mean waiting times. The most important information, however, is given by the distribution function of waiting times (d.f.w.t.), which will be dealt with separately.

The aim of this paper is to demonstrate calculation methods for dimensioning of such systems with respect to a prescribed grade of service.

2. Stationary probabilities of state and characteristic values

2.1. *Full available servers*

For simplicity all the n servers are supposed to have the same mean service time h. A state $(x; z_1, z_2)$ is defined by "x servers are busy and z_j waiting places are occupied within the j-th queue, $j=1, 2$". *Fig. 3* shows the state space with the transitions in the general case. The following abbreviations are used:

$$\varepsilon = 1/h\,, \qquad (1\text{a}) \qquad \lambda = \lambda_1 + \lambda_2\,, \qquad (1\text{d})$$

$$\mu = n \cdot \varepsilon\,, \qquad (1\text{b}) \qquad A_j = \lambda_j \cdot h,\ j = 1, 2, \qquad (1\text{e})$$

$$\mu_j = \mu \cdot p_j,\ j = 1, 2, \qquad (1\text{c}) \qquad A = \lambda \cdot h\,. \qquad (1\text{f})$$

The total state space consists of two subspaces, a one-dimensional subspace with the states $(x; 0, 0)$, $x=0, 1, \ldots, n-1$, when no item is waiting, and a two-dimensional subspace $(n; z_1, z_2)$, $z_j=0, 1, 2, \ldots, s_j$, $j=1, 2$, when all the servers are busy.

In the general case the service times of the various servers have different means h_i, $i=1, 2, \ldots, n$. Then we have an n-dimensional subspace when no item is waiting. This case will be considered in an example of limited availability.

For each state $(x; z_1, z_2)$ a probability $p(x; z_1, z_2)$ for its existence is defined. In the stationary case the probabilities of state $p(x; z_1, z_2)$ are independent of time t and the initial conditions. For the probabilities of state the Kolmogorov-forward-equation [2] holds which reads in the stationary case as shown with Eqns. (2a) – (2j) on page 578.

Finally, a normalizing condition for the probabilities of state is given by

$$\sum_{x=0}^{n-1} p\,(x; 0, 0) + \sum_{z_1=0}^{s_1} \sum_{z_2=0}^{s_2} p\,(n; z_1, z_2) = 1. \qquad (2\text{k})$$

In general, the inhomogeneous linear equation system must be solved as a whole. In the above case the solution for the probabilities of state of the one-dimensional subspace can be given explicitly. Therefore, only the two-dimensional subsystem must be treated as a whole. Because of the large number of unknowns the solution must be carried out on a digital computer by iterative methods (overrelaxation method). Below, some special cases will be discussed for which the solution can be given either in terms of recursion formulas or explicitly.

The stationary probabilities of state are not used directly to judge a service system. For this purpose characteristic values are calculated from the probabilities of state. It should be noted that the probabilities of state are not affected by the various queue disciplines, whereas the characteristic values and, above all, the distribution function of waiting times depend on the order a waiting request is served from a certain queue.

Eqns. (2a) — (2j)

$$x \varepsilon\, p(x;0,0) = \lambda\, p(x-1;0,0) \qquad x = 1,2,\ldots,n \quad (2a)$$

$$\lambda\, p(n;0,0) = \mu\, p(n;1,0) + \mu\, p(n;0,1) \quad (2b)$$

$$(\lambda+\mu)\, p(n;z_1,0) = \lambda_1\, p(n;z_1-1,0) + \mu\, p(n;z_1+1,0) + \mu_2\, p(n;z_1,1) \qquad z_1 = 1,2,\ldots,s_1-1 \quad (2c)$$

$$(\lambda_2+\mu)\, p(n;s_1,0) = \lambda_1\, p(n;s_1-1,0) + \mu_2\, p(n;s_1,1) \quad (2d)$$

$$(\lambda+\mu)\, p(n;0,z_2) = \lambda_2\, p(n;0,z_2-1) + \mu\, p(n;0,z_2+1) + \mu_1\, p(n;1,z_2) \qquad z_2 = 1,2,\ldots,s_2-1 \quad (2e)$$

$$(\lambda_1+\mu)\, p(n;0,s_2) = \lambda_2\, p(n;0,s_2-1) + \mu_1\, p(n;1,s_2) \quad (2f)$$

$$(\lambda+\mu)\, p(n;z_1,z_2) = \lambda_1\, p(n;z_1-1,z_2) + \lambda_2\, p(n;z_1,z_2-1) + \mu_1\, p(n;z_1+1,z_2) + \mu_2\, p(n;z_1,z_2+1) \qquad z_1 = 1,2,\ldots,s_1-1;\ z_2 = 1,2,\ldots,s_2-1 \quad (2g)$$

$$(\lambda_2+\mu)\, p(n;s_1,z_2) = \lambda_1\, p(n;s_1-1,z_2) + \lambda_2\, p(n;s_1,z_2-1) + \mu_2\, p(n;s_1,z_2+1) \qquad z_2 = 1,2,\ldots,s_2-1 \quad 2h)$$

$$(\lambda_1+\mu)\, p(n;z_1,s_2) = \lambda_1\, p(n;z_1-1,s_2) + \lambda_2\, p(n;z_1,s_2-1) + \mu_1\, p(n;z_1+1,s_2) \qquad z_1 = 1,2,\ldots,s_1-1 \quad (2i)$$

$$\mu\, p(n;s_1,s_2) = \lambda_1\, p(n;s_1-1,s_2) + \lambda_2\, p(n;s_1,s_2-1) \quad (2j)$$

The definitions for the most important characteristic values are given by Eqns. (3a) – (8).

a) The probability of waiting at arrival (1-items) W_1

$$W_1 = \sum_{z_1=0}^{s_1-1} \sum_{z_2=0}^{s_2} p\,(n;z_1,z_2) \qquad \text{(D1, 2), (3a)}$$

$$W_1 = \sum_{z_1=0}^{s_1} \sum_{z_2=0}^{s_2} p\,(n;z_1,z_2) \qquad \text{(D3) . (3b)}$$

b) The probability of waiting successfully (1-items) W_1^*

$$W_1^* = W_1 \qquad \text{(D1, 2), (4a)}$$

$$W_1^* = \sum_{z_1=0}^{s_1-1} \sum_{z_2=0}^{s_2} p\,(n;z_1,z_2) \qquad \text{(D3). (4b)}$$

c) The probability of waiting at arrival and being pushed out (1-items) W_1^{**}

$$W_1^{**} = 0 \qquad \text{(D1, 2), (5a)}$$

$$W_1^{**} = \sum_{z_2=0}^{s_2} p\,(n;s_1\,z_2) \qquad \text{(D3). (5b)}$$

d) The probability of loss or overflow (1-items) B_1

$$B_1 = \sum_{z_2=0}^{s_2} p\,(n;s_1,z_2) \qquad \text{(D1, 2, 3). (6)}$$

e) The mean queue length of the first queue 1 Ω_1

$$\Omega_1 = \sum_{z_1=0}^{s_1} \sum_{z_2=0}^{s_2} z_1 \cdot p\,(n;z_1,z_2) \qquad \text{(D1, 2, 3). (7)}$$

f) The mean waiting time for waiting at arrival (1-items) t_{W1}

$$t_{W_1} = \frac{\Omega_1}{W_1 \cdot \lambda_1} \qquad \text{(D1,2,3). (8)}$$

The corresponding values for 2-items are obtained in the same way as shown above for 1-items by substituting the index numbers.

2.1.1. *Generalizations*

By a more specified description of the busy sources and servers the model can be extended in two ways:

2.1.1.1. *Allowing a finite number of sources*

In this case the arrival rates λ_j, $j=1,2$, have to be replaced by the arrival rates of the remaining non-busy sources of type j.

2.1.1.2. *Allowing different mean service times* for the different servers

This is the case when the servers symbolize different fast computers or different fast transmission lines. An example will be given below in case of limited available servers.

2.1.2. *Special cases*

Above, no special assumptions were made for the probabilities p_j, $j=1,2$, which stand for selection of the j-th queue when a server becomes idle (inter-queue discipline). Some special values for these probabilities will be considered now.

2.1.2.1. *Priority type discipline*

The case of nonpre-emptive priority type discipline with a fixed number of waiting places for items of each priority class is obtained if

$$\begin{aligned} &p_1 = 1, \quad p_2 = 0 \quad \text{for } z_1 > 0,\ z_2 \geqq 0\,, \\ &p_1 = 0, \quad p_2 = 1 \quad \text{for } z_1 = 0,\ z_2 > 0\,. \end{aligned} \qquad (9)$$

This case differs from the usual nonpre-emptive priority models [3], where the items of all priority classes have a fixed maximum number of waiting places. The above model holds for all applications, where each priority class is associated with an own storage array, or where the items are associated with a certain incoming queue, as in real-time dialogue systems.

For the priority type discipline a simple recursion algorithm can be given to calculate the probabalities of state. All transitions of Fig. 3 with coefficient μ_2 vanish. Taking $p(n;0,z_2)$ as unknown, all the probabilities $p(n;z_1,z_2-1)$, $z_1=1,\ldots,s_1$, can be expressed in terms of $p(n;0,z_2)$. The equilibrium for the state $(n;s_1,z_2-1)$ allows to calculate the unknown

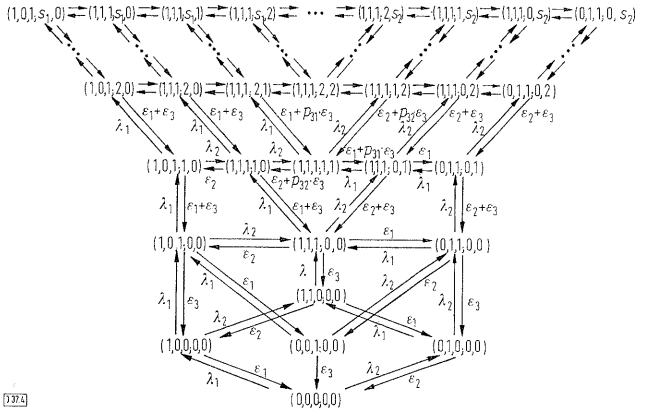


Fig. 4. State space and transition coefficients for the limited available 3-server system with two queues.

$p(n; 0, z_2)$. This is true for all $z_2 = 1, 2, \ldots, s_2$. By this method all probabilities of state can be expressed by $p(0; 0, 0)$. (Note that all $p(x, 0; 0)$, $x = 1,2 \ldots$, n follow from Eqn. (2a) by recursion). Eqn. (2k), finally, allows to calculate $p(0; 0, 0)$, which is the parameter the recursion starts with.

2.1.2.2. *Cyclic type discipline*

In some applications queues are served alternatively. If

$$p_1 = p_2 = \frac{1}{2} \quad \text{for } z_1, z_2 > 0\,, \tag{10}$$

we obtain approximate probabilities of state from the system (2a–k). Below, an approximate solution for the d.f.w.t. will be given, too.

2.1.2.3. *Queue length type discipline*

For this discipline we define

$$p_j = \frac{z_j}{z_1 + z_2}\,, \quad j = 1,\ 2, \quad z_1 + z_2 > 0\,. \tag{11}$$

This discipline serves the longer queue with a greater probability. Additionally, it can be shown, that Eqn. (11) holds also for the disciplines D1, D2, D3 with respect to all waiting items. In this case the probabilities of state can be given explicitly:

$$p\,(0; 0, 0)^{-1} = \sum_{x=0}^{n-1} \frac{A^x}{x!} + \frac{A^n}{n!} \cdot \sum_{z_1=0}^{s_1} \sum_{z_2=0}^{s_2} \left(\frac{A_1}{n}\right)^{z_1} \cdot \left(\frac{A_2}{n}\right)^{z_2} \cdot \frac{(z_1+z_2)!}{z_1!\, z_2!}$$

$$p\,(x; 0, 0) = p(0; 0, 0) \cdot \frac{A^x}{x!} \tag{12}$$

$$p(n; z_1, z_2) = p(0; 0, 0) \cdot \frac{A^n}{n!} \cdot \left(\frac{A_1}{n}\right)^{z_1} \cdot \left(\frac{A_2}{n}\right)^{z_2} \cdot \frac{(z_1+z_2)!}{z_1!\, z_2!}$$

$$x = 0, 1, 2, \ldots, n; \qquad z_j = 0, 1, 2, \ldots, s_j; \qquad j = 1, 2.$$

These formulas can be extended easily to the general case of an arbitrary number of queues.

2.2. *Limited available servers*

In Fig. 2 the simplest model of a service system with limited available servers and two input queues is shown. It is assumed that the mean service times h_i, $i = 1, 2, 3$, are not identical. The interqueue discipline for the i-th server is described by the probabilities p_{ij}, $j = 1, 2$, which stand for service of the j-th queue, if server i terminates its occupation. In this example only for server 3 exists an interqueue discipline probability p_{3j}, $j = 1, 2$.

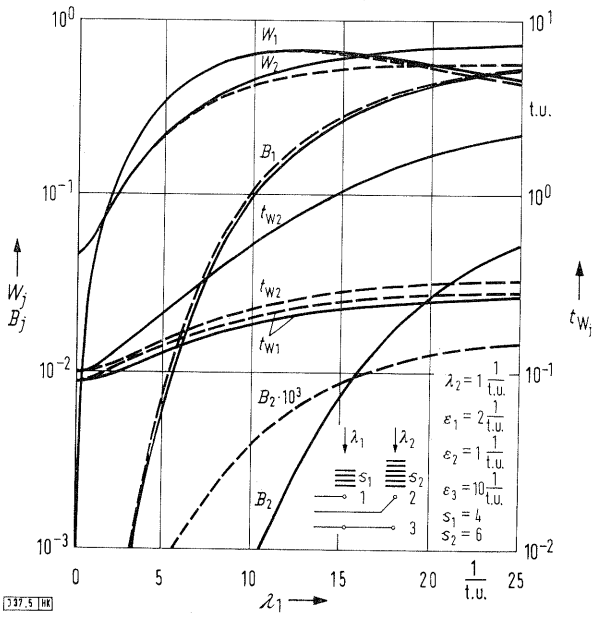


Fig. 5. Characteristic values for a 3-server system with limited availability and two input queues.
——— priority type discipline (queue 1 has nonpre-emptive priority),
------ queue length type discipline.

Let be $(x_1, x_2, x_3; z_1, z_2)$ the state defined by

$$x_i = \begin{cases} 0 & \text{server number } i \text{ is idle} \\ 1 & \text{server number } i \text{ is busy} \end{cases}, \quad i = 1, 2, 3,$$

z_j waiting places are occupied within the j-th queue, $j = 1, 2$.

In *Fig. 4* the state space for the model of Fig. 2b is shown. Note that queues can be built up already when there are still servers which are not busy.

The equations of state can be written in a similar way as in the case of full available servers. In any case, the whole linear equation system is to be solved by an iterative method. From the stationary probabilities of state the corresponding characteristic values can be obtained by a somewhat more complicated summation as in case of full availability.

Computer programmes have been developed for calculation of arbitrary systems with full and limited availability from the input datas $g, n, s(j)$, type of server arrangement, type of interqueue discipline, $\lambda(j)$, $\varepsilon(i)$.

Fig. 5 shows the characteristic values W_j, B_j, t_{Wj}, $j = 1, 2$, for the example of the 3-server system with limited availability according to Fig. 2b.

The example was chosen such that 1-items represent real-time requests whereas 2-items stand for non-real-time batch requests. Server j serves only j-requests, $j = 1, 2$, where server 3 (e. g. a fast remote computer for safety and overload reasons) serves both queues. The third server, which is considerably faster than servers 1 and 2, serves queue 1 (real-time-requests) with nonpre-emptive priority (solid curves), or both queues according to the queue length type discipline (dashed curves). The queue discipline within the queues was assumed as D1 or D2.

The curves of Fig. 5 are given for a constant non-real-time input rate and a variable real-time input rate. By means of such curves one can study the influence of both inputs to each other as well as the influence of system parameters and service disciplines on the characteristic values, i. e. how to dimension a system to meet a prescribed grade of service.

3. Distribution function of waiting times

3.1. *Abstract of the general theory*

For Markovian queues with only one queue and full available servers *R. Syski* [2] has given a method to calculate the d.f.w.t. Further investigations on such systems were made by the author [4]. This theory can be extended to multiqueueing systems. In the following a short summary of the method to calculate the d.f.w.t. for multiqueueing systems will be given. The results will be applied to two examples of a full available n-server system with different queue disciplines.

A j-test item arrives at the j-th queue and starts a special waiting process. This process terminates when the j-test item is either served or being pushed out. A random variable $\zeta_j(t)$ is defined as a random occupation pattern, containing only those items in the system after the waiting time t, which may influence the waiting time of the j-test item, the j-test item being excluded. The special definition of $\zeta_j(t)$ depends on the system structure and on the queue discipline. It can be shown, that the $\zeta_j(t)$-process is a Markovian process.

A (complementary) conditional d.f.w.t. for the j-test item which starts waiting from the occupation pattern i will be defined by

$$w_j(t \mid i) = P\{t_j > t \mid \zeta_j(0) = i\}, \quad i \notin H_j. \tag{13}$$

In Eqn. (13) t_j means the waiting time of the j-test item, H_j means the set of absorbing states for the j-test item, i. e. all those occupation patterns, where the j-test item will either be served or pushed out.

The differential equation system (Kolmogorov-backward-equation) for the complementary conditional d.f.w.t. is given by the following

theorem:
$$\frac{\mathrm{d}w_j(t \mid i)}{\mathrm{d}t} = -q_j(i)\, w_j(t \mid i) + \sum_{\substack{k \neq i \\ k \notin H_j}} q_j(i, k)\, w_j(t \mid k), \quad i \notin H_j, \tag{14a}$$

$$\lim_{t \to 0+} \frac{\mathrm{d}w_j(t \mid i)}{\mathrm{d}t} = -q_j(i)\, w_j(0 \mid i) + \sum_{\substack{k \neq i \\ k \notin H_j}} q_j(i, k)\, w_j(0 \mid k), \quad i \notin H_j. \tag{14b}$$

In Eqn. (14a, b) $q_j(i)$, $q_j(i, k)$ are the conditional transition coefficients for the $\zeta_j(t)$-process. The initial conditions $w_j(0 \mid i)$, $i \notin H_j$, are calculated from the linear equation system (14b), where the expression

$$-\lim_{t \to 0+} \frac{\mathrm{d}w_j(t \mid i)}{\mathrm{d}t} = \varepsilon_j(i), \quad i \notin H_j, \tag{14c}$$

denotes the conditional transition coefficient for termination of the $\zeta_j(t)$-process at the instant of reaching the state i.

It should be noted that the differential equation system (14a) holds for the quantities $w_j(t \mid i)$, $w_j^*(t \mid i)$, and $w_j^{**}(t \mid i)$ which refer to j-test items waiting from initial state i, waiting from initial state i successfully, and waiting from initial state i in vain, respectively. The difference lies only in the quantities $\varepsilon_j(i)$, $\varepsilon_j^*(i)$, and $\varepsilon_j^{**}(i)$ of Eqns. (14b) and (14c).

Finally, the total d.f.w.t. for all j-items is given by

$$W_j(> t) = \sum_{i \notin H_j} P_j(i)\, w_j(t \mid i), \tag{15}$$

where $P_j(i) = P\{\zeta_j(0) = i\}$. $P_j(i)$ can be calculated from the stationary probabilities of state. Eqn. (15) holds also for the corresponding quantities W_j^*, w_j^*, and W_j^{**}, w_j^{**}.

To treat Eqns. (14a–c) we use the Laplace-transformation [5] and obtain from the above theorem:

$$[s + q_j(i)]\, W_j(s \mid i) - \sum_{\substack{k \neq i \\ k \notin H_j}} q_j(i, k) \cdot W_j(s \mid k) = w_j(0 \mid i), \quad i \notin H_j, \tag{16}$$

where $W_j(s \mid i)$ denotes the Laplace-transform of $w_j(t \mid i)$, and s a complex variable. By this method the waiting time problem can be reduced to an eigenvalue problem. Eqn. (16), written in matrix notation, is the starting point to investigate the eigenvalues. For systems with only one queue the eigenvalues have been investigated [4] for several queue disciplines. This method can also be applied to multiqueueing systems. The eigenvalues prove in any case to be negative-real. The location of the eigenvalues on the negative-real axis of the s-plane reflects the influences of the various queue and interqueue disciplines on the d.f.w.t.

The solutions of Eqn. (16), $W_j(s \mid i)$, are rational functions of s. Partial fraction expansion of the $W_j(s \mid i)$ and the inverse Laplace-transformation lead finally to the quantities $w_j(t \mid i)$ and, hence, to $W_j(>t)$.

The conditional mean waiting times for waiting of a j-item from an initial state i, $t_{\mathrm{W}_j}(i)$, from an initial state i successfully, $t_{\mathrm{W}_j}^*(i)$, and from an initial state i in vain, $t_{\mathrm{W}_j}^{**}(i)$, can be obtained by integration of $w_j(t \mid i)$, $w_j^*(t \mid i)$, and $w_j^{**}(t \mid i)$, respectively. Using the definition of Laplace-transformation it can be shown that

$$t_{\mathrm{W}_j}(i) = W_j(0 \mid i), \qquad i \notin H_j. \tag{17}$$

For $t_{\mathrm{W}_j}^*(i)$ and $t_{\mathrm{W}_j}^{**}(i)$ the corresponding relations hold. The conditional mean waiting times can be calculated from Eqn. (16) at $s = 0$. The total mean waiting time of a waiting j-item, $t_{j\mathrm{W}}$, is obtained by integration of $W_j(>t)/W_j$, i. e.

$$t_{\mathrm{W}_j} = \frac{1}{W_j} \sum_{i \notin H_j} P_j(i)\, t_{\mathrm{W}_j}(i). \tag{18}$$

Eqn. (18) holds also for $t_{\mathrm{W}_j}^*$ and $t_{\mathrm{W}_j}^{**}$, when $t_{\mathrm{W}_j}(i)$ is replaced by $t_{\mathrm{W}_j}^*(i)$ and $t_{\mathrm{W}_j}^{**}(i)$, and W_j is replaced by W_j^* and W_j^{**}, respectively.

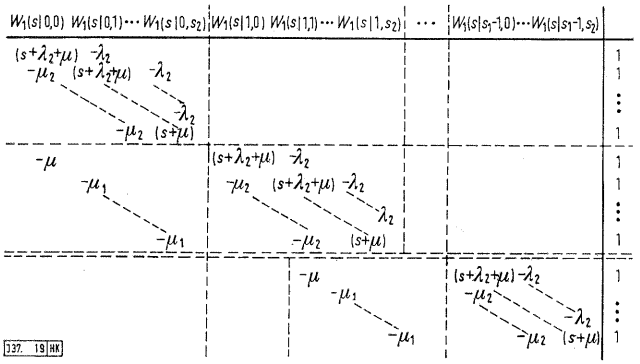
3.2. *Applications*

The above theorem will be demonstrated for two queue disciplines, D1 and D3. For example, a double-queue many server system with full availability will be considered. In both cases it is assumed

that the interqueue discipline probabilities p_j, $j=1, 2$, are constants.

3.2.1. *First-come, first-served service*

For 1-items a conditional d.f.w.t. $w_1(t \mid z_1, z_2)$ is defined, where $(n; z_1, z_2)$ is the state the j-test item found on its arrival. For this discipline all customers which were accepted once will wait successfully. From the equation system (14b, c) follows that $w_1(0 \mid z_1, z_2)=1$ for $z_1=0, 1, \ldots, s_1-1$, $z_2=0, 1, \ldots, s_2$. The equation system (16), in matrix notation, is shown by Eqn. (19):



From the special structure of this matrix one can see that the $W_1(s \mid z_1, z_2)$ can be determined sectionwise, as shown by the dashed lines. In any case, only a system of (s_2+1)st order is to solve instead of a system of $s_1 \cdot (s_2+1)$st order.

Having calculated all the quantities $w_1(t \mid z_1, z_2)$ the total d.f.w.t. for 1-items is given according to Eqn. (15) by

$$W_1(>t) = \sum_{z_1=0}^{s_1-1} \sum_{z_2=0}^{s_2} p(n; z_1, z_2)\, w_1(t \mid z_1, z_2). \quad (20)$$

Items of type 2 can be treated in a similar way as items of type 1.

For priority type discipline Eqn. (19) assumes a somewhat simpler form, because all μ_2 vanish. Then all the unknowns can be calculated recursively, starting with $W_1(s \mid 0, s_2)$, which can be determined directly, up to $W_1(s \mid 0, 0)$, then starting with $W_1(s \mid 1, s_2)$ up to $W_1(s \mid 1, 0)$, and so on.

3.2.2. *Last-come, first-served service*

As an example we only want to consider the d.f.w.t. for 1-items which wait successfully. Let $w_1^*(t \mid i_1, i_2)$ be the conditional d.f.w.t. for waiting items of type 1. Clearly, $w_1^*(t \mid 0, i_2)$ is the d.f.w.t. for the whole waiting time of 1-items, because new 1-items are allowed to occupy the first waiting place in their queue. The quantities $w_1^*(t \mid i_1, i_2)$, $i_1>0$, are the d.f. for partial waiting times for waiting for a server from the (i_1+1)st waiting place in the 1st queue. i_2 denotes the number of waiting 2-items at the instant when the 1-test item becomes (i_1+1)st in its queue.

The system for the determination of d.f.w.t. is given by Eqn. (21).

Before starting the solution of Eqn. (21) the initial values $w_1^*(0/i_1, i_2)$ must be calculated from a linear equation system according to Eqns. (14b, c).

The matrix shows that the system has to be solved as a whole, no sectionalization is possible.

According to Eqn. (15) the total d.f.w.t. for successfully waiting 1-items is given by

$$W_1^*(>t) = \sum_{i_2=0}^{s_2} P_1(0, i_2)\, w_1^*(t \mid 0, i_2)\,, \text{ where} \quad (22)$$

$$P_1(i_1, i_2) = \begin{cases} \sum_{z_1=0}^{s_1} p(n; z_1, i_2) & \text{for } i_1 = 0, \\ 0 & \text{for } i_1 > 0. \end{cases}$$

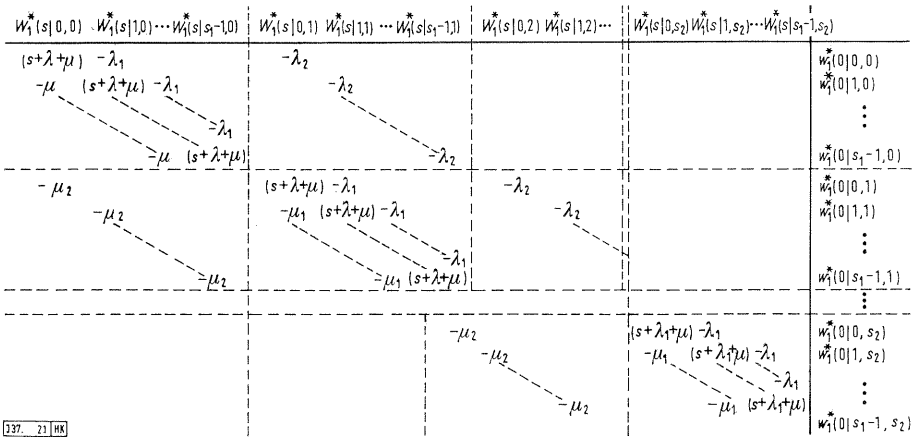
For priority type discipline again all μ_2 vanish in Eqn. (21). Then a sectionalization of the total matrix is possible: beginning at the lower right, only systems of s_1-st order are to solve. Additionally, the eigenvalues of the subsystems can be given explicitly, because the matrices of the subsystems can be reduced to difference matrices [4].

For random service (D2) a similar structured matrix can be derived as above for last-come, first-served service. Assuming other rules for the interqueue discipline, e.g. according to Eqn. (11), in general a more detailed pattern of occupations has to be chosen to differ the items according to the dispatching rule.

3.3. *Approximate calculation of the d.f.w.t. for cyclic type service*

For serving both queues alternatively the probabilities of state can be calculated approximately by p_j according to Eqn. (10). Assuming the discipline D1 within the queues in the following a very simple method is given for calculation of the d.f.w.t. approximately.

For demonstration let us consider 1-items as test items. *Fig. 6* shows the two possible cases when a 1-test item arrives at queue 1.



Eqn. (21)

For both cases a lower and an upper bound for the conditional d.f.w.t. $w_1(t \mid z_1, z_2)$ can be given. Since the termination process of the fully occupied server group is also Poissonian, the probability $p_x(t)$ that during the time interval $(0, t)$ x terminations occur is

$$p_x(t) = \frac{(\mu t)^x}{x!} \cdot \mathrm{e}^{-\mu t}. \qquad (23)$$

For the lower bound of the total d.f.w.t. we have

$$\underline{W}_1(> t) = \sum_{z_1=0}^{s_1-1} \sum_{z_2=0}^{s_2} p(n; z_1, z_2)\, \underline{w}_1(t \mid z_1, z_2),$$

where
$$\underline{w}_1(t \mid z_1, z_2) = \begin{cases} \sum_{i=0}^{2z_1} p_i(t), & z_1 \leqq z_2, \\ \sum_{i=0}^{z_1+z_2} p_i(t), & z_1 > z_2. \end{cases}$$

The upper bound is given by

$$\overline{W}_1(> t) = \sum_{z_1=0}^{s_1-1} \sum_{z_2=0}^{s_2} p(n; z_1, z_2)\, \overline{w}_1(t \mid z_1, z_2),$$

where
$$\overline{w}_1(t \mid z_1, z_2) = \sum_{i=0}^{2z_1+1} p_i\,(t).$$

The exact d.f.w.t. must be included between both the lower and the upper limit.

Now an interpolation is made between both limit curves such that the interpolated curve yields the right mean waiting time. From both limit functions the lower and the upper mean waiting times $t_{\underline{W}_1}$ and $t_{\overline{W}_1}$ are calculated by integration of $\underline{W}_1(> t)/\overline{W}_1$ and $\overline{W}_1(> t)/W_1$ respectively. On the other hand, from Eqn. (8) we know the value t_{W_1}. The linear interpolation

$$W_1(> t) \cong \frac{1}{1+\alpha}[\alpha\, \underline{W}_1(> t) + \overline{W}_1(> t)], \qquad (24)$$

where $\alpha = \dfrac{t_{\overline{W}_1} - t_{W_1}}{t_{W_1} - t_{\underline{W}_1}}$,

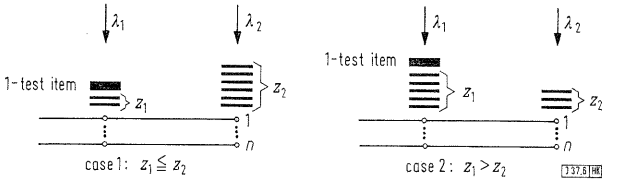


Fig. 6. For derivation of an approximate formula for the distribution function of waiting times in case of cyclic type service.

yields just the mean value t_{W_1}. For practical estimations this method allows a simple calculation of the d.f.w.t. for cyclic type service.

4. Conclusion

Service systems with parallel waiting queues and full or limited available servers have been investigated. Under the Markovian assumptions the equations for the stationary probabilities of state and the conditional distribution functions of waiting time have been derived. The solvability of the linear equation systems and the linear differential equation systems has been discussed for several queue and interqueue disciplines. Curves of the characteristic values are given for an example of a 3-server system with limited available servers.

References

[1] *Martin, J.*: Design of real-time computer systems. Englewood Cliffs, N. J.: Prentice Hall Inc., 1967.

[2] *Syski, R.*: Markovian queues. Symposium on Congestion Theory, University of North Carolina, 1965.

[3] *Wagner, W.*: Über ein kombiniertes Warte-Verlust-System mit Prioritäten. Dissertation. (PHD-Thesis) University of Stuttgart, 1968.

[4] *Kühn, P.*: Über ein kombiniertes Warte-Verlust-System mit verschiedenen Abfertigungsdisziplinen. To be published in Arch. elektr. Übertr., 1970.

[5] *Doetsch, G.*: Anleitung zum praktischen Gebrauch der Laplace-Transformation. München: R. Oldenbourg, 1961.

(Eingangsdatum: 3. Juli 1970)

Dissertationen

Die Wendelleitung als Laufzeitverzögerungs- und Resonanzleitung. Von Gerhard Rotter. Dissertation TH Aachen (14. 2. 1970). Berichter: Prof. Dr.-Ing. H. Döring; Mitberichter: Prof. Dr. rer. nat. H. Lueg.

Die Arbeit untersucht die Eigenschaften von koaxial abgeschirmten Wendelleitungen im Bereich kleiner Frequenzen bis zu einigen 100 MHz, wobei die vorgelegte Theorie sich auf Wendeln sehr geringer Steigung beschränkt. Aufgrund der Existenz eines Wellentyps mit kleiner Phasen- bzw. Gruppengeschwindigkeit hat die Wendelleitung den Vorteil, daß Leitungsbauelemente wie Laufzeitverzögerungs- und Resonanzleitungen verkürzt aufgebaut werden können.

Ausgehend von einer Berechnung der Feldstruktur dieses Wellentyps werden die Leitungskenngrößen in Abhängigkeit von der Frequenz und den verschiedenen geometrischen Abmessungen angegeben. Bei den aus Wendelleitungen aufgebauten Leitungsresonatoren ist vor allem die $\lambda/4$-Resonanz von Interesse, die im Schrifttum auch als Spulenresonanz bezeichnet wird. Die Resonatorfelder, insbesondere die an den Leitungsenden entstehenden Streufelder, werden diskutiert und unter ihrer Berücksichtigung Resonanzfrequenz sowie Güte bestimmt. Eine Untersuchung der kapazitiven Abstimmung dieser Resonatoren schließt sich an. In einer Reihe von Messungen konnten die vorgelegten theoretischen Ergebnisse bestätigt sowie geeignete Meßmethoden zur Bestimmung der Kenngrößen angegeben werden. (3813)

Bau und Erprobung einer strahlungsgeheizten Hochtemperaturanlage zur Kristallzüchtung. Von Eike Schwarz. Dissertation Universität (TH) Karlsruhe, 1970. Berichter: Prof. Dr. H. Friedburg; Mitberichter: Prof. Dr. H. G. Kahle.

Die Arbeit beschreibt den Aufbau und die Erprobung einer Kristallzüchtungsapparatur, mit der in oxidierender Atmosphäre Einkristalle im Temperaturbereich zwischen 1400 und 2000 °C nach dem Czochralski'schen Ziehverfahren gezüchtet werden können, ohne die mit der üblichen Verwendung eines Tiegels aus Fremdmaterial zur Aufnahme der Schmelze verbundenen Nachteile aufzuweisen.

Bei dem angewandten Verfahren wird die Schmelztemperatur mit einer Hochleistungs-Strahlungsheizung erzeugt, die mit einer konventionellen Widerstandsheizung kombiniert ist. Das Prinzip der Strahlungsheizung besteht darin, die von einer Xenonlampe ausgehende Strahlung so stark zu bündeln, daß eine begrenzte Schmelze in dem im Brennfleck befindlichen Tiegel entsteht.

Nach einer Untersuchung des Einflusses der verschiedenen Ziehparameter auf Größe und Temperatur der Schmelze wurden Einkristalle aus $CaWO_4$ gezogen. Die bisher durchgeführten Untersuchungen an den Kristallen lassen erkennen, daß das angewandte Verfahren zu einer mit sonstigen Züchtungsmethoden vergleichbaren Kristallqualität, aber mit geringerer Möglichkeit der Verunreinigung führt. (3627)