8 QUEUING THEORY MODELS

8.1 KINDS OF QUEUES

Queuing theory (the theory of waiting lines) is a discipline of operational research, the subjects of which are mathematical models and quantitative analysis of processes involving waiting for the service of some technical equipment. Common to all these processes are the arrivals of people or objects requiring service and the attendant delays when the service mechanism is busy. The aim of the theory is the identification of these characteristics of the queue processes that facilitate technical and economic analysis of the given system and the attainment of optimal parameters of these systems.

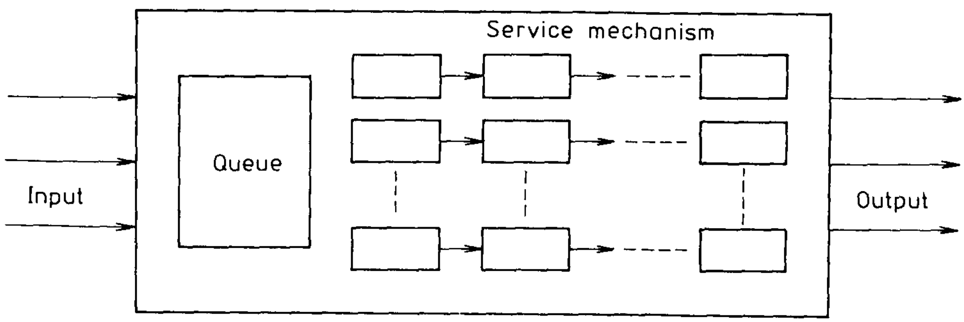

Fig. 8.1 Schematic representation of the queuing process

The first application of queue theory were to problems in the operation of telephone and telegraph lines. Now queue theory is applied in various fields, e.g. in the organisation of production, design operations, transport, services, maintenance, etc.

Problems using queue theory have the structure represented in Fig. 8.1. The mathematical description in queue models requires specification in this form: a mathematical description of the input process, its time dependence, queue discipline, service mechanism, and the output process resulting from the previous stages.

According to the given characteristics queue processes and their mathematical models can be classified in different ways. Stochastic models with random variables are frequently applied in queue theory.

The queue models are investigated with different input processes, queue discipline and service mechanism. Input is characterized by the pattern of arrivals into the system (regular or random patterns), individual or bulk arrivals, etc. A random pattern is described by its probability characteristics: a mean number of items arriving in a given unit of time, the probability that no unit arrives during a certain time interval, the mean number of items in the bulk arrivals, etc. It is important to make an analysis of the stationarity or non-stationarity of probability characteristics of the input process.

Queue discipline describes the order in which customers entering the system are eventually served. It can have different forms given by the input process and the ways of transition from queue to service. In the simplest case, the discipline is first come, first served (first in, first out – FIFO). However, some customers may become impatient and decide to leave the system before being served (reneging), or a cus-

Table 8.1 Kendall's notation of queuing systems

Letter	is substituted for	
	X	Y
M	Exponentially distributed (independent) interarrival time, (M = Markovian), Poisson input	Exponentially distributed service time
E_n	Erlangian distribution of order n for interarrival time (with parameters λ and n)	Erlangian distribution of order n for service time (with parameters μ and n)
K_n	Distribution χ^2 for interarrival time (with n degrees of freedom)	Distribution χ^2 for service time
D	Regular, deterministic interarrival time	Constant, deterministic service time
G	General distribution of interarrival time (no assumption for distribution)	General distribution of service time
GI	General independent distribution of interarrival time	

tomer seeing a long line may balk, i.e. not join the queue if it is too long. In more complicated cases a different queue discipline may be applied, e.g. last come, first served (last in, first out – LIFO) or a priority discipline if certain customers are served preferentially.

Similarly, there can be different types of service systems. Service can be offered along one or several parallel lines (servers in parallel), service time can be constant (for all customers) or it can be random. The probability characteristics of service time are investigated and its probability distribution can be stationary or time dependent.

A concise representation of queue systems was given by Kendall. In his notation, the systems are classified according to three main aspects: (1) the type of input stochastic process describing the arrival of customers; (2) the distribution of service time and (3) the number of servers. Information concerning these characteristics is signified by three symbols: $X/Y/c$, where X and Y are specified by capital letters and c is a natural number (or ∞) denoting the number of servers. For X and Y the letters M, D, G, E_n, K_n are substituted here (for X the couple GI may be substituted); an explanation of these symbols is given in Table 8.1.

Kendall's classification is not exhaustive, as some important characteristics are not included in the symbolic notation (e.g. the existence and length of the queue, queue discipline, etc.). Therefore, these data must be added in each case.

8.2 MARKOVIAN AND OTHER PROCESSES IN QUEUING MODELS

Markovian processes form a separate class of stochastic processes. A number of publications deal with this type of process (Walter, 1970; Zítek, 1969; Wagner, 1975) and its application to WRS problems (Votruba and Nacházel, 1971). In queue models the Markovian properties of processes are very important. The solution is simpler if the process is Markovian, or if it can be approximated by a Markovian process. In other cases, formidable computational difficulties occur and some sophisticated methods are used, e.g. the method of imbedded Markov chains.

In section 8.2 the relationship between Markovian properties of processes and the type of queue model is explained; in section 8.3 the models applicable to WRS are analyzed.

According to Kendall's notation of queue models (section 8.1) e.g. the system designated $M/M/1$ has the following features:

– Poisson input, i.e. exponentially distributed (independent) inter-arrival time
– exponentially distributed service time,
– single server.

Similarly, the system designated $M/E_n/1$ is characterized in the following way:

– Poisson input (see previous type),

– Erlangian distribution of order n for service time (with parameters μ and n),
– single server.

The theory of Markovian processes can be applied to queue systems with Poisson input, i.e. exponentially distributed inter-arrival time and exponentially distributed service time. In Kendall's notation these systems fall into the first row i.e. systems designated $M/M/n$. These systems have been intensively investigated and have been applied successfully in practice. Consequently, queue systems are often divided into two basic groups:

– Markovian systems (type $M/M/n$),
– other systems (non-Markovian, e.g. types $M/D/1$, $M/G/1$, $M/E_n/1$, $GI/M/1$, etc.).

8.2.1 Markovian Queuing Systems

System $M/M/1$

The model of this system is the simplest one. It can be described as follows: customers arrive in the system individually, and their arrivals are mutually independent and independent of the service mechanism. The inter-arrival time is an independent random variable with an identical exponential distribution. The customers who cannot be served because the single server is occupied, wait in a single, unlimited queue. The queue discipline is first come, first served, without priorities.

The solution gives the following parameters: the mean number of customers in the queue, the mean number of customers in the system, the probability distribution of the waiting time of any customer, the mean time spent by a customer in the system, the system utilization factor (fraction of the time the server is busy).

System $M/M/n$

When the requirements of customers exceed the capacity of a single server, a system with more servers, $n > 1$, can be used. (However, some other systems are possible, e.g. with a higher intensity of the service mechanism or with a reduction in the number of customers).

In the simplest form of this system, a single queue is assumed, common to all servers. The customers wait in the queue when all the servers are occupied. Whenever a server becomes available, the customer at the head of the queue is served. The solution of problems of this system is similar to that in the $M/M/1$ system. Special problems occur if the solution is to determine the number of servers n. In some alternatives a variable number of servers is possible providing flexibility in handling the requirements of customers.

Systems $M/M/1$ and $M/M/n$ can be relatively easily handled by the theory of Markovian processes, and various alternatives of these systems can be considered.

For example, different queue disciplines can be assumed, i.e. the rules determining the order in which customers entering the system are eventually served. Frequent types of queue discipline are FIFO (first in, first out), LIFO (Last in, first out; last come, first served) when the last customer is served when the server has just finished processing the previous one, random serving of customers in the queue, or *priority queue discipline* when the customers are classified into several types and priorities are assigned to these types in decreasing order of importance.

The theory of Markovian processes can be applied to systems $M/M/n$ with a finite queue. A loss of customers is assumed, i.e. the supposition that all customers will be served is not valid.

In solving problems of Markovian systems no major computational difficulties occur. The tasks can be reduced to the solution of a set of linear equations. This computational simplicity rests on the following assumptions:

– steady-state systems (i.e. we are not interested in the initial stages of the system before its stabilization),

– homogeneous Poisson input (exponential arrivals of customers),

– exponentially distributed service time.

These assumptions can be accepted in many practical cases and they are a good approximation of the real situation; however, there are many cases when they are not. Then a non-Markovian queue system with more general properties must be used.

8.2.2 Other Queuing Systems

In Kendall's notation in section 8.1 only the processes in the first row are Markovian. The applicability of other processes is more general, but their solution is more complicated. Often, a reduction to the Markovian case is attempted to allow solution by simpler methods (Zítek, 1969). A well-known procedure is called the *method of the imbedded Markov chain*[1]).

[1]) The principle of the imbedded Markov chain is as follows: Assume that the process being investigated is not Markovian, i.e. the condition

$$P\left\{X(t) = \frac{k}{X(s_1)} = j_1,\ X(s_2) = j_2 \ldots,\ X(s_n) = j_n\right\} = P\left\{X(t) = \frac{k}{X(s_1)} = j_1\right\} \tag{I}$$

is not met for some choice of numbers $n, t > s_1 > s_2 \ldots > s_n \geqq 0$ and $k, j_1, j_2, \ldots, j_n$. Then a certain sequence of times

$$0 \leqq t_1^* < t_2^* \ldots < t_m^* \ldots,\ \lim_{m \to \infty} t_m^* = \infty \tag{II}$$

can be formed in such a way that condition(I) is met if the numbers $t, s_1, \ldots, s_n$ are chosen from this time sequence (II). Therefore, a time sequence is sought where, for any $t = t_M^*$ and for natural n for any n-tuple (vector with n components) $s_i = t^*_{m_i}$ $(i = 1, 2, \ldots, n)$, $0 \leqq s_n < s_{n-1} \ldots < s_1 < t$, this condition is met.

Another method for transformation to Markovian-type processes is based on an approximate, *broader definition of the process.* Apart from reduction to the Markovian process, certain processes can be investigated directly, e.g. by the theory of *semi-Markovian stochastic processes.* Often the *Monte Carlo methods* are used for queue processes that are not Markovian. They are based on a simulation of the investigated systems, and they exemplify some of the experimental methods of probability theory. They can be used for any type of process but they require a good deal of computer time and capacity.

In the literature (Zítek, 1969) the following types of non-Markovian queue systems are listed.

System $M/D/1$

In system $M/M/n$ customer arrivals and service time are random. These assumptions may be an acceptable simplification of reality. In some queue systems model $M/D/1$ with constant service time may be more appropriate: each customer spends a fixed time interval in the serving stage.

System $M/E_n/1$

The advantage of this system as compared with the $M/M/1$ system is the Erlangian distribution of service time. This distribution is determined by two parameters, and therefore it can fit the observed frequency better than the exponential distribution determined by one parameter only. Some other modifications of this system are possible, e.g. the service can be composed of several independent, successive stages; it can include the assumption that more than one customer enters the system at the same time in the form of bulk arrivals.

System $M/G/1$

This system has been developed in a number of alternative forms that try to make the rules for customer arrivals and service time more flexible in order to produce a more general model. In computation the imbedded Markov chain is often used.

$$P\left\{X(t_M^*) = \frac{k}{X(t_{m_i}^*)} = j_i, i = 1, 2, \ldots, n\right\} = P\left\{X(t_M^*) = \frac{k}{X(t_{m_1}^*)} = j_1\right\}$$

for all integer non-negative numbers $j_1, j_2, \ldots, j_n, k$. The time sequence (II) is determined in such a way that the sequence of random numbers $X(t_m^*)$ for $m = 1, 2, \ldots$ forms a Markov chain. If such a sequence (II) has been found, the investigation is restricted to this chain $\{X(t_m^*)\}$ and used only the theory of Markov chains. The resulting properties of this process form the basis for the estimation of properties of the original process. This imbedded chain cannot reflect all the properties of the original process. It gives information concerning the momentary states of the process at times t_m^*. This is often sufficient, e.g. in investigation of the limiting behaviour of the process for $t \to \infty$.

System GI/M/1

This class of systems with one server is characterized by a generally independent distribution of inter-arrival time and exponential service time.

System GI/GI/1

This notation concerns general systems with one server and no assumptions about the inter-arrival time and service time distribution (with the exception of the assumption of the homogeneity of the process). In computation the method of the imbedded Markov chain is used and the aim is the distribution of steady-state waiting time.

8.3 QUEUING SYSTEMS MODELS

The basic queue model is the simple steady-state exponential channel. It is a single-server model with exponential inter-arrival and service times, and the queue discipline is first come, first served. The system can remain in some situation or it can move on to the adjoining situation. If n is the state of the system, i.e. n items are in the system (one customer is being served and $n - 1$ are in the queue) only transitions from state n to states $n - 1$, n, and $n + 1$ are possible.

On this assumption, the probability density function of the time that a customer spends in the queue $p(w)$ is given by (Walter *et al.*, 1973)

$$p(w) = \frac{\lambda}{\mu} e^{-w(\mu - \lambda)} \tag{8.1}$$

where λ is arrival rate per unit of time, i.e. probability rate, of transition from state n to state $n + 1$ ($1/\lambda$ is the mean time between arrivals);
μ – service rate per unit of time that the server is busy ($1/\mu$ is the mean service time)

The ratio $\varrho = \lambda/\mu$ of arrival rate to service rate is frequently called the traffic intensity.

The basic operating characteristics of the system are: the mean number of customers in the system and its variation, mean line length, mean number of customers served per busy period and mean time in the system.

The mean number of customers in the system with an infinite length of queue is given by

$$\bar{n} = \sum_{n=0}^{\infty} np_n = \sum_{n=0}^{\infty} n(1 - \varrho)\varrho^n = (1 - \varrho)\,\varrho \sum_{n=1}^{\infty} n\varrho^{n-1} \tag{8.2}$$

The expression $n\varrho^{n-1}$ is the first derivative of ϱ^n; the sum of derivatives is the derivative of the sum (the progression is convergent) and the following transcription is possible:

$$\bar{n} = (1 - \varrho)\,\varrho\,\frac{\mathrm{d}}{\mathrm{d}\varrho}\sum_{n=1}^{\infty}\varrho^n = (1 - \varrho)\,\varrho\,\frac{\mathrm{d}}{\mathrm{d}\varrho}\,\frac{\varrho}{1 - \varrho} = \frac{\varrho}{1 - \varrho} \tag{8.3}$$

In original values λ and μ, the expression (8.3) will be

$$\bar{n} = \frac{\lambda}{\mu - \lambda} \tag{8.4}$$

The mean number of items in the queue (mean line length) is

$$\bar{n}_f = \sum_{n=1}^{\infty}(n - 1)\,p_n = \sum_{n=1}^{\infty} np_n - \sum_{n=1}^{\infty} p_n = \bar{n} = (1 - p_0) = \frac{\varrho}{1 - \varrho} - (1 - 1 + \varrho) =$$

$$= \frac{\varrho^2}{1 - \varrho} \tag{8.5}$$

The mean number of items in the queue is equal to the mean number of items in the system reduced by the quantity $\varrho = 1 - p_0$, i.e. the fraction of the time the server is busy. For a single server the quantity $\bar{n}_f$ is not very important but it is in the case of several servers.

For the variance of the number of items in the system the following expression is valid:

$$\sigma_{\bar{n}}^2 = \bar{n}^2 + \bar{n} = \frac{\varrho}{(1 - \varrho)^2} \tag{8.6}$$

or, in original quantities,

$$\sigma_{\bar{n}}^2 = \frac{\lambda\mu}{(\mu - \lambda)^2} \tag{8.6'}$$

The mean time in the system can be computed using the following consideration. In the system, $\bar{n}$ items, on average, are waiting and an average of λ, items per time unit arrive in the system. The mean time in the system can be derived by the division of $\bar{n}$ by λ, i.e.

$$\bar{T}_s = \frac{\bar{n}}{\lambda} = \frac{1}{\lambda}\,\frac{\lambda}{\mu - \lambda} = \frac{1}{\mu - \lambda} \tag{8.7}$$

Similarly for the mean time in line

$$\bar{T}_f = \frac{\bar{n}_f}{\lambda} = \frac{1}{\mu - \lambda}\,\frac{\lambda}{\mu} \tag{8.8}$$

By these methods a simple queue model can be handled that uses a Poisson exponential distribution or Erlangian distribution. Erlangian distribution[1]) of k-th order is defined by the probability function (with parameters $\lambda > 0$ and $k \geqq 1$ integer):

$$f(x) = \lambda\, e^{-\lambda x}(\lambda x)^{k-1}/(k-1)! \quad \text{for } x \geqq 0$$
$$f(x) = 0 \quad \text{for } x < 0$$

It can be shown that exponential distribution is a special case of the Erlangian distribution for $k = 1$.

In the queuing theory, more complicated models occur and special methods have to be used (Dráb, 1973; Ventcel, 1966; Walter-Lauber, 1975).

The models of queuing systems can be applied for the solution of problems of various technological processes, the rational organization of the health service, traffic problems and of problems of water resources systems (see section 8.6).

8.4 UNRELIABLE SYSTEMS

So far, the queuing systems have been assumed to be reliable. This assumption is frequently unrealistic in practice. Therefore, much work has been devoted to the investigation of unreliable systems (Zítek, 1969; Gnedenko-Kovalenko, 1966; Klimov, 1966).

In models of queuing systems, unreliability is caused by the intermittent failure of servers that cannot, at some stage, serve the customers.

Often, the unreliable systems are modelled as *systems with a priority queue discipline*, as in the following example: from time to time a customer with absolute priority arrives (= failure) and the service of other customers continues only after this customer has left the system (i.e., after restarting the service).

In principle, the following systems with priority queue discipline are possible (characterized by the treatment of the customer that is being served at the moment when a customer with higher priority arrives, i.e., at the moment when the failure occurs):

a) the service of the customer is normally finished (a system with a weak priority),

b) the service is immediately interrupted (a system with strong or absolute priority),

ba) the customer whose service was interrupted immediately leaves the system (without being served)

---

[1]) The Erlangian cumulative distribution function is defined by

$$F(x) = 1 - e^{-\lambda x} \sum_{j=0}^{k-1} (\lambda x)^j/j! \quad \text{for } x \geqq 0$$
$$F(x) = 0 \quad \text{for } x < 0$$

bb) the customer with interrupted service returns to the queue and is then served
bba) the service continues at the point of interruption,
bbb) the service is restarted from the beginning.

Unreliable systems can similarly be classified by these types of failures.

Certain differences exist between unreliable systems and systems with priority discipline:

– in unreliable systems the server stops in the middle of carrying out the repair and no further failure is possible; it corresponds to the assumption that there is only one customer with absolute priority;

– the failure can occur at any time, even if the server is not busy and when no customer is being served. However, there are systems where failure is possible only when the server is busy or when the server is idle, or the service is in any case finished (this last case corresponds to the system with weak priority, the failure is modelled for a customer with weak priority).

In all unreliable systems, new data are added to the basic data describing the organization and operation of the system. These new data characterize the probability distribution of failure occurrence and the probability distribution of repair time. The probability of failure can depend on different factors, e.g., the time from starting the system after the last failure, the time spent by the customer in being served, the number of customers that have been served after the last failure, the total service time, the time that was necessary for the last repair, etc. Using these data, the probability of service without failure, the probability distribution of the possible failure, etc., can be determined. The problems are far more complicated in cases of unreliable systems with a larger number of servers.

8.5 QUEUING SYSTEMS SIMULATION

The principles of the simulation of queuing systems have been explained in the references (Buslenko and Shreyder, 1961; Zítek, 1969; Saaty, 1961; Votruba *et al.*, 1974; Kaufman and Cruon, 1961, etc.). Only the basic ideas of this approach are mentioned in this section.

Although some outstanding results have been achieved in the queuing theory, this theory is not (and probably never will be) able to solve the problems that arise in practical life. The models oversimplify the complex problems of reality, or if the models are a good approximation of reality they are mathematically intractable as there are no effective analytical methods for solution. If a rough approximation is not acceptable, simulation methods and mainly the Monte Carlo simulation models have to be used for this more complicated case. The main advantage of these methods is their universality; they can be used for any process and any model. However, they need a lot of computation and computer time and great memory capacity.

The application of simulation methods to the tasks of the queuing theory is based on a simple principle (Zítek, 1969): that the unknown probability of a random event is estimated by repeating many independent experiments, and the relative frequency of this event is the estimate of this probability. A similar method is used in the estimation of characteristics of distribution of the random variable, e.g., the mean value is approximated by the average of the observed values. Both of these tasks occur in the queuing theory; the probability of some event is looked for (e.g., service without queuing), i.e., the probability that x will get a value from some interval, or its mean value.

The practical possibility of the application of simulation models is determined at sufficient speed in repeating these experiments. Repetition of experiments is often necessary to obtain the probability (or mean values) with sufficient accuracy. These estimations are based on the law of large numbers.

The determination of the probability of one phenomenon is often not sufficient; a set of phenomena is required. For example, in searching for the probability distribution function of a random variable x, we want to know the values $F(x) = P(X \leqq x)$ for all real x values; in practice the values $F(x_j)$ are determined only for some samples x_j (in finite number) and the concept of the diagram of the whole function $F(x)$ is based on these samples. The more values $F(x_j)$ that are found, the more accurate is the approximation of function $F(x)$, but more experiment are required. In addition, we are interested in changes of function $F(x)$ related to changes in basic characteristics and parameters of the system. It is apparent that without computers the practical application of simulation procedures would not be possible.

8.6 APPLICATION OF QUEUING THEORY IN WRS

In section 7.4 the mathematical similarity of the inventory theory and the queuing theory was mentioned, resulting in the application of these theories in WRS. Now this similarity is clarified. In section 7.4.2 the general analytical solution of sets of equations (7.21) was not reached, i.e., the set

$$\begin{aligned}
R_0 &= p_0R_0 + p_0R_1 + \dots + p_0R_M \qquad (8.9)\\
R_1 &= p_1R_0 + p_1R_1 + \dots + p_1R_M + p_0R_{M+1}\\
&\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\\
R_{V-M-1} &= p_{V-M-1}R_0 + \dots + p_{V-M-1}R_M + p_{V-M-2}R_{M+1} + \dots + p_0R_{V-1}\\
R_{V-M} &= p_{V-M}R_0 + \dots + p_{V-M}R_M + p_{V-M-1}R_{M+1} + \dots + p_0(R_V + R_{V+1} + \dots)\\
&\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\\
R_{V-M+S} &= p_{V-M+S}R_0 + \dots + p_{V-M+S}R_M + p_{V-M+S-1}R_{M+1} + \dots +\\
&\quad + \dots + p_S(R_V + R_{V+1} + \dots)\\
&\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots
\end{aligned}$$

where (Fig. 7.8) V is the active storage of the reservoir, and M is the release from the reservoir.

The definitions of values p_i and R_i is as follows: if X_t is the inflow in the reservoir in the year t and Z_t is the storage before the inflow X_t then in time t the variable X_t is equal to i with probability p_i and the sum of the volume of water remaining in reservoir Z_t and the inflow X_t will equal i ($Z_t + X_t = i$), with probability R_i. By this infinite system of linear equations the probability distribution of Z_t is determined, given the distribution of $Z_t + X_t$ and X_t.

The main difficulties in the analytical solution of the system of these equations was caused by the boundary conditions $Z = 0$ and $Z = V - M$. Moran *et al.*, 1959, investigated the assumption of infinite storage in two cases:

(1) the phenomena near the top of the active storage in cases where the distribution of variable Z_t was such that it was very improbable that Z_t would be equal to values near zero, and they assumed an infinitely deep reservoir;

(2) in the opposite case where the probability of a full reservoir was very small, they assumed the reservoir to be infinite in the upward direction and a distribution of Z_t near zero values was considered.

Assuming the infinite active storage $V = \infty$ and steady state and discrete variables, the system of equations (8.9) has the following form:

$$\begin{aligned}
R_0 &= p_0 R_0 + p_0 R_1 + \ldots + p_0 R_M \qquad (8.10)\\
R_1 &= p_1 R_0 + p_1 R_1 + \ldots + p_1 R_M + p_0 R_{M+1}\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
R_S &= p_S R_0 + p_S R_1 + \ldots + p_S R_M + p_{S-1} R_{M+1} + \ldots + p_0 R_{M+S}
\end{aligned}$$

for $S = 2, 3 \ldots$.

Using Foster's method (Foster, 1953) it can be proved that this system has a solution different from zero, i.e., there exists a probability distribution (different from zero) near the level of the infinitely deep reservoir. The solution, however, is not easy, and some methods of the queuing theory can be used for it. The queue with bulk service is used (Bailey, 1954): the active storage is represented by the length of the queue and the service is assumed in time $\ldots t_{i-1}, t_i, t_{i+1}, \ldots$ so that intervals $v = (t_{i+1} - t_i)$ are independently distributed with probability distribution $\mathrm{d}B(v)$.

The customers' arrivals are random, forming a Poisson process with the mean value λ; in a time t_i, a bulk of customers, M (or the whole queue if its length is shorter than M) is served. Knowing the distribution $\mathrm{d}B(v)$ and the parameter λ, the probabilities p_i can be computed, i.e., the probabilities that at one interval, i, customers will arrive in the queue. The length of the queue just before the beginning of the service is a random variable of an infinite Markov chain, the transition and stationary probabilities of which are given by the set of equations (8.10).

Using this analogy with the queuing theory, the set of equations (8.10) can be solved. The solution by generating functions was first published by Bailey (1954) and was described in detail by Moran (1959).

The queuing theory can be used for the similar problem of the operating policy for release, where continuous variables are used and the input is a negative exponential distribution.

Let us investigate the possibility of the direct solution of a continuous alternative of the set of equations (8.10), i.e., the equation (7.22) of section 7.4.2. For an infinite V the set of equations (8.9) resp. the equations (7.22) become

$$g(x) = f(x) \int_0^M g(x)\,\mathrm{d}t + \int_M^{M+x} f(M + x - t)\, g(t)\,\mathrm{d}t \tag{8.11}$$

where $g(x)$ is the probability density function of variable Y_t, $f(x)$ is the probability density function of X_t. For this equation the Laplace transform can be used for the cumulative probability function of variable X_t and $V_t = X_t + Z_t$

$$F(y) = \int_0^y f(t)\,\mathrm{d}t, \qquad G(y) = \int_0^y g(t)\,\mathrm{d}t \tag{8.12}$$

and in simple form

$$G(y) = \int_0^y G(M + y - t)\,\mathrm{d}F(t) \tag{8.13}$$

This type of integral equation is well-known from the queuing theory. Lindley, 1952, derived the solution and proved that the steady-state solution exists if the mean input value is less than M, and if the initial state is zero. The distribution of variable Y_t converges on the stationary state of the process.

Smith, 1953, has shown that another useful analogy between infinite reservoirs and the queuing theory is possible, and it can be seen in the following case of the queuing theory. A queue is assumed with one server only and with regular interarrival time M. In Lindley's general theory these intervals are random, in this application it is possible to use fixed intervals. If no queue has been formed, the customer is served immediately, otherwise he must wait till all the customers in front of him are served.

The service time, denoted X_i, is a random variable with the cumulative distribution function $F(x)$, X_i being mutually independent and independent of queue characteristics. The time interval necessary for serving the whole queue that had formed immediately before the arrival time t_i, is denoted Z_i. This is the waiting time of a new customer in the queue between the time of his arrival and the beginning of his service. $Z_i + X_i$ is then the total time that a new customer spends in the system, i.e. the time interval between the arrival of a customer and the end of his service. The whole time interval necessary for serving the queue immediately before the time t_{i+1} is Z_{i+1}. It is equal to

$$\begin{aligned} Z_{i+1} &= Z_i + X_i - M \quad \text{if } Z_i + X_i - M > 0 \\ Z_{i+1} &= 0 \qquad\qquad\qquad\ \ \text{if } Z_i + X_i - M \leqq 0 \end{aligned}$$

This queuing problem can be applied for an infinite reservoir, assuming that X_i represents the random inflow X_t into reservoir and reservoir storage just before the withdrawal is represented by the waiting time of the last customer. The withdrawal M from reservoir is represented by the reduction in waiting time of the new customer during the time interval (t_i, t_{i+1}). In this model (Smith, 1953) the release is not realized at once, but in regular steps. The assumption of regular withdrawals does not change the underlying theory, as the equation (8.13) is valid for this example.

It is apparent that the way the problems of reservoir operation and queuing correspond to each other is quite different from the previous case, where the length of the queue taken as a discrete variable correspond to storage of the reservoir, and bulk service was assumed. In this case, the service of a single customer is assumed and storage of reservoir is modelled by the waiting time of a new customer in the queue.

The numerical solution of this case is that of equation (8.13). It was published by Lindley, 1952, for cases where the input variable has a Pearson type III probability distribution. The resulting probability distribution (Moran, 1959) is the distribution in this case. From a mathematical point of view, this latter case of the application of the queuing theory for WRS is a continuous analogue of the previous case, often called Bailey's queuing theory with bulk service. However, the correspondence between the WRS model and queuing theory is different in each case.

Gani and Prabhu, 1957, derived the solution of equations for the infinite reservoir for the case when storage in the reservoir has a negative exponential distribution.

Moran, 1959, Kendal, 1957, and others described a number of possible approaches, the latter used the operation of reservoirs with continuous time. Input is a continuous inflow, withdrawal is also continuous, and takes place until the reservoir is empty. Both finite and infinite reservoirs were investigated. In addition, Kendall, 1957, investigated a non-stationary case of distribution of storage: the initial state in time $t = 0$ is given by storage y and the probability distribution of time interval T in which the reservoir is emptied for the first time is looked for. This method approaches the theory of common risk. All these studies are stimulative and use sophisticated mathematical theory, but the stage of development is not sufficient for its practical application. In addition, most models use an unrealistic assumption that the inflow is an additive process with a Pearson type III distribution.

The numerical calculation in these models is similar to that in the inventory theory, illustrated by a simple example in section 7.4.1.

Another approach to the application of the queuing theory for reservoir operation was described by Chorafas, 1965. Binomial probability distribution is assumed for random inflow to the reservoir in time-sequenced stages. The release from the reservoir, in each stage, is predetermined, where there is an empty reservoir, or where spillage has occurred. These models yield relationships among the following quantities: mean values and variances of reservoir inflows, reservoir storage, the chosen values of reservoir release, probability of deficits and spillage.

These models can be described by

$$Z_i = Z_{i-1} + (\bar{x}_i + es_{x_i}) - M_i \tag{8.14}$$

where Z_i is content of reservoir at the end of stage i, Z_{i-1} is the remaining content of reservoir from the previous stage, $\bar{x}_i$ is the mean value of inflow at stage i, s_{x_i} is its standard deviation, e is a random number with a normal distribution N (0, 1) (zero mean and unit variance), and M_i is the reservoir release at stage i.

In solution the initial state is determined by the given or chosen storage of reservoir and release. In equation (8.14) other values are computed. The calculation is carried out several times on a computer with different input values. The results of such calculations are taken to be a sample of a set of physical reactions of the system that forms a response surface for various possible solutions. If the economic parameters are considered in the model, namely initial and operational costs, the benefits from different values of withdrawals, eventual losses due to deficits in water supply, the physical response surface can be transformed into an economical response surface or a net benefit one. These methods can be used to define operation of reservoirs with carry-over; some results, however, may be sub-optimal.

To summarize the application of the queuing theory in WRS: the idea of using the queuing theory for the operation of reservoirs is promising; it is, however, obvious that the people behind this idea are mathematicians and not water resource engineers. The studies use complicated and sophisticated mathematical operations, but the applicability of the results in the practice of WRS is limited for two main reasons: (i) the assumptions oversimplify the reality, and (ii) the difficulties in applying the calculations to real situations are neglected. Moran admitted this situation and stated that all the methods investigated can, in principle, be used for a system of reservoirs; however, the calculations involved for application to a single reservoir are so complicated that the recommended numerical method is the Monte Carlo procedure.

Kartvelishvili, 1963, criticises these applications because they use the simplified assumption that the river flow has a Pearson type III distribution, and he underlines the assumption that infinite active reservoir storage can be accepted in a limited number of cases. Buras, 1972, in his WRS monograph devotes little space to the application of the queuing theory in WRS.