Chapter 11
Applied Queueing Systems

11.1 Bandwidth Sharing of Finite-Capacity Links with Different Traffic Classes

Traditional telephone networks were designed to implement a single type of communication service, i.e., the telephone service. Today's telecommunication networks implement a wide range of communication services. In this section we introduce Markov models of communication services that compete for the bandwidth of a finite-capacity communication link.

11.1.1 Traffic Classes

There are several important features of traffic sources of communication services that allow for their classification. Here assume that the traffic sources require the setting up of a connection for a finite period of time during which data communication is carried out between the parties of the communication service. We classify the traffic sources based on the bandwidth of the data transmission during a connection. The simplest case is where data are transmitted with a fixed bandwidth during a connection. This case is commonly referred to as constant bit rate (CBR). A more general traffic behavior is obtained when the bandwidth of data transmission varies during a connection. This case is commonly referred to as variable bit rate (VBR). The most common form of bandwidth variation is when the bandwidth alternates between 0 and a fixed bandwidth. These VBR sources are referred to as ON-OFF sources, and we restrict our attention to the ON-OFF case. The most complex traffic sources adjust their bandwidth according to the available capacities of the network resources. There are two classes of this kind of source. *Adaptive* traffic sources set up a connection for a given period of time and transmit data according to the available bandwidth in the network. If the network resources are occupied during the connection of an adaptive traffic

L. Lakatos et al., *Introduction to Queueing Systems with Telecommunication Applications*,
DOI 10.1007/978-1-4614-5317-8_11, © Springer Science+Business Media, LLC 2013

source, then the source transmits data with a low bandwidth, and the overall amount of transmitted data during a connection is low. *Elastic* traffic sources set up a connection for transmitting a given amount of data. The bandwidth of the data transmission depends on the available bandwidth in the network. If the network resources are occupied during the connection of an elastic connection, then the period of the connection is extended in such a way that the source transmits the required amount of data.

In this section we assume that the traffic sources demonstrate a memoryless time-homogeneous stochastic behavior and, consequently, the arrival processes are Poisson processes and the connection times are exponentially distributed, except for the elastic class, where the amount of data to transmit is exponentially distributed. Additionally, the traffic sources are characterized by their bandwidth parameters. In the case of CBR and ON-OFF VBR sources, the bandwidth parameter is the bandwidth of the active period. In the case of adaptive and elastic sources, the bandwidth parameters are the minimal and maximal bandwidth at which the source can transmit data.

Consequently, in the case of the different kinds of traffic sources, a class k traffic source is characterized by the following parameters:

- CBR connection: connection arrival intensity λ_k, bandwidth requirement c_k, parameter of exponentially distributed connection holding time μ_k;
- VBR connection: connection arrival intensity λ_k, bandwidth requirement in ON state c_k, parameters of exponentially distributed connection holding time, ON time, and OFF time μ_k, α_k, and β_k, respectively.
- Adaptive connection: connection arrival intensity λ_k, minimal bandwidth $c_{\min}^{(k)}$, maximal bandwidth $c_{\max}^{(k)}$, parameter of exponentially distributed connection holding time μ_k;
- Elastic connection: connection arrival intensity λ_k, minimal bandwidth $c_{\min}^{(k)}$, maximal bandwidth $c_{\max}^{(k)}$, parameter of exponentially distributed amount of transmitted data δ_k.

These parameters define the arrival process and the bandwidth needs of the traffic sources but they do not define completely the service procedure as the common resource (the finite capacity link) is shared among the traffic types and classes. In the case of traditional telephone services, the procedure for a new telephone call is obvious: accept as many calls as possible with the given finite-capacity link. In the case of different traffic classes, more complex procedures are required to properly utilize the resources and to provide the desired service features to each traffic class. The set of rules concerning the acceptance or rejection of a new connection is referred to as call admission control (CAC). CAC defines the acceptance or rejection of a new connection of all types under all possible traffic conditions. We will see some typical CACs and their properties.

The most common performance parameters of interest in these kinds of traffic models are

- Per-class connection-dropping probability (at arrival connection arrival),
- VBR connection-dropping probabilities (during ongoing connection at an OFF to ON transition),
- Per-class mean bandwidth of adaptive and elastic connections,
- Sojourn time of elastic connections.

Different dimensioning methods apply for different traffic classes. In the following sections we investigate the simple Markov models of these traffic classes, which form the bases of the complex dimensioning methods used in practice.

11.1.2 Bandwidth Sharing by CBR Traffic Classes

One of the first generalizations of traditional telecommunication models is due to the coexistence of communication services with different bandwidth requirements. When a link is utilized by different kinds of CBR connections with the previously detailed Markovian properties, then the overall system behavior can be described by a CTMC. The main problem of analyzing the performance parameters through this CTMC is the potentially very high number of states. If a finite-capacity link of bandwidth C is utilized by I different kinds of CBR connections, then a state of the CTMC should represent the number of ongoing connections of each class, and the number of states is proportional to the product $\prod_{i=1}^{I}(\frac{C}{c_i}+1)$.

To overcome this practical problem, an efficient numerical procedure was proposed by two researchers independently [50, 80]; the procedure is often referred to as the Kaufman–Roberts method. It is based on the fact that a large CTMC, which represents the number of ongoing connections of each class, satisfies the local balance equations

$$\lambda_i \; p(n_1,\ldots,n_i-1,\ldots,n_I) = n_i \mu_i \; p(n_1,\ldots,n_i,\ldots,n_I),$$

where $p(n_1,\ldots,n_i,\ldots,n_I)$ denotes the stationary probability of the state where the number of class i connections is n_i for $i = 1,\ldots,I$. The local balance equation represents that the stationary state-transition rate due to an arriving class i connection is in balance with the stationary state-transition rate due to a departing class i connection. The main idea of the Kaufman–Roberts method is to unify those states of a large Markov chain that represent the same bandwidth utilization of a link. In the state $(n_1, n_2, \ldots, n_I)$, the bandwidth utilization is $c = \sum_{i=1}^{I} n_i c_i$. Summing up the local balance equations for the states where the bandwidth utilization on the right-hand side is c we have

$$\sum_{i=1}^{I} \lambda_i \; P(c-c_i)\, \mathcal{I}_{\{c_i \geq c\}} = \sum_{i=1}^{I} n_i \mu_i \; P(c)$$

$$\sum_{i=1}^{I} \frac{\lambda_i c_i}{\mu_i} \; P(c-c_i)\, \mathcal{I}_{\{c_i \geq c\}} = \underbrace{\sum_{i=1}^{I} n_i c_i}_{c} \; P(c)$$

$$\sum_{i=1}^{I} \frac{\lambda_i c_i}{\mu_i c} \; P(c-c_i)\, \mathcal{I}_{\{c_i \leq c\}} = \; P(c),$$

where $P(c)$ denotes the sum of the stationary probabilities of the states where the bandwidth utilization is c, that is, $P(c) = \sum_{(n_1,n_2,\ldots,n_I):\sum_{i=1}^{I} n_i c_i = c} p(n_1, n_2, \ldots, n_I)$. The last equation is the core of the Kaufman–Roberts method, which computes the relative (nonnormalized) probabilities of the link utilization levels first and then normalizes probabilities as follows.

1. Let $\tilde{P}(0) = 1$, and for $c = 1, 2, \ldots, C$ compute

$$\tilde{P}(c) = \sum_{i} \frac{\lambda_i c_i}{\mu_i c} \; \tilde{P}(c-c_i)\, \mathcal{I}_{\{c_i \leq c\}}.$$

2. Compute $\tilde{P} = \sum_{c=0}^{C} \tilde{P}(c)$.
3. Normalize the probabilities by $P(c) = \tilde{P}(c)/\tilde{P}$.

There is an implicit technical assumption that is necessary for the application of the Kaufman–Roberts method. There must be a bandwidth unit such that each c_i is an integer multiple of this bandwidth unit. (The method remains applicable if C is not an integer multiple of the bandwidth unit.) Fortunately, in important applications such a bandwidth unit exists.

Having the stationary probabilities of the utilization levels we can compute the loss probabilities. If the CAC allows all connections entering the link as long as the available bandwidth is not less than the bandwidth of the entering connection, then the loss probability of class i connections is

$$b_i = \sum_{c > C - c_i} P(c).$$

It is a straightforward consequence of the CAC that connections with higher bandwidth requirements have a higher loss probability. If a kind of fairness is required among the different classes such that each class experiences the same loss probability, then the CAC needs to be modified. Let us assume that the traffic class with the highest bandwidth is class I. If the CAC is modified such that each incoming connection is rejected when the available bandwidth is less than c_I, then the distribution of the link utilization changes, but each class is accepted and rejected at the same time at the different link utilization levels, and consequently they have the same loss probability. If the CAC depends only on the link utilization level (as in the case of a modified CAC with identical dropping probabilities), then the Kaufman–Roberts method remains applicable. In this case the main iteration step of the procedure changes to

$$\tilde{P}(c) = \sum_i \frac{\lambda_i c_i}{\mu_i c} \tilde{P}(c - c_i)\, \mathrm{CAC}(i, c - c_i),$$

where $\mathrm{CAC}(i, c)$ is one if a class i connection is accepted at link utilization c, and zero otherwise. The link utilization-level-dependent CAC can also be generalized to probabilistic CACs. In this case the main iteration step of the procedure remains the same as for the deterministic one, and $\mathrm{CAC}(i, c)$ indicates the probability that a class i connection is accepted at link utilization c.

11.1.3 Bandwidth Sharing with VBR Traffic Classes

When a link is utilized by different kinds of VBR connections and each of them is characterized by the previously described Markovian properties, the overall system behavior can be described by a CTMC. The states of this CTMC represent the number of ongoing VBR connections of each class and the number of connections in the ON phase, $(n_1, m_1, n_2, m_2, \ldots, n_I, m_I)$. Note that $n_i \geq m_i$, $i = 1, \ldots, I$, and $\sum_{i=1}^{I} m_i c_i \leq C$, where the second inequality means that the utilized bandwidth should not exceed the link capacity. The state space of this CTMC is even larger than that for CBR connections, which represents only the number of ongoing connections of each class, but unfortunately there is no more efficient computation method available for this model than to solve the CTMC. This is due to the fact that this CTMC does not satisfy the local balance equations. At any rate, the numerical solution of this CTMC is still possible for a limited number of VBR classes and connections.

Generally, the CAC for VBR connections is more complex than that for CBR connections. A conservative CAC does not allow more VBR connections than the link can serve, assuming that all VBR connections are in the ON state. That is, $\sum_{i=1}^{I} n_i c_i \leq C$ holds for each state. Unfortunately, a conservative CAC results in a very low resource utilization, especially when the length of the ON period is short with respect to the length of the OFF period. In these cases, it is worth allowing more VBR connections than a conservative CAC in order to increase the link utilization. The drawback of nonconservative CACs is that accepted ongoing VBRs can be dropped due to insufficient capacity with positive probability at an OFF to ON phase transmission. In practice, it is usually required that the dropping probability of ongoing VBR connections be much lower than that of newly arriving ones.

The possible state transitions of Markov chains are

(a) $(n_1, m_1, \ldots, n_i, m_i, \ldots, n_I, m_I) \to (n_1, m_1, \ldots, n_i + 1, m_i + 1, \ldots, n_I, m_I)$ at rate λ_i if $\mathrm{CAC}(i, \{n_1, m_1, \ldots, n_i, m_i, \ldots, n_I, m_I\}) = 1$;

(b) $(n_1, m_1, \ldots, n_i, m_i, \ldots, n_I, m_I) \to (n_1, m_1, \ldots, n_i - 1, m_i - 1, \ldots, n_I, m_I)$ at rate $m_i \mu_i$;

(c) $(n_1, m_1, \ldots, n_i, m_i, \ldots, n_I, m_I) \to (n_1, m_1, \ldots, n_i - 1, m_i, \ldots, n_I, m_I)$ at rate $(n_i - m_i)\mu_i$;

(d) $(n_1, m_1, \ldots, n_i, m_i, \ldots, n_I, m_I) \to (n_1, m_1, \ldots, n_i, m_i - 1, \ldots, n_I, m_I)$ at rate $m_i \alpha_i$;

(e) $(n_1, m_1, \ldots, n_i, m_i, \ldots, n_I, m_I) \to (n_1, m_1, \ldots, n_i, m_i + 1, \ldots, n_I, m_I)$ at rate $(n_i - m_i)\beta_i$ if $\sum_{j=1}^{I} m_j c_j + c_i \le C$;

(f) $(n_1, m_1, \ldots, n_i, m_i, \ldots, n_I, m_I) \to (n_1, m_1, \ldots, n_i - 1, m_i, \ldots, n_I, m_I)$ at rate $(n_i - m_i)\beta_i$ if $\sum_{j=1}^{I} m_j c_j + c_i > C$,

where $\mathrm{CAC}(i, \{n_1, m_1, \ldots, n_I, m_I\})$ denotes the CAC decision in state $(n_1, m_1, \ldots, n_I, m_I)$ for an incoming class i connection, and state transitions with a zero rate are impossible. According to the bandwidth limit of a link,

$$\mathrm{CAC}(i, \{n_1, m_1, \ldots, n_I, m_I\}) = 0 \text{ if } \sum_{j=1}^{I} m_j c_j + c_i > C.$$

The transitions represent the following events:

(a) New class i connection arrival.
(b) Departure of a class i connection that is in the ON phase.
(c) Departure of a class i connection which is in the OFF phase.
(d) A class i connection switches from ON to OFF phase.
(e) A class i connection switches from OFF to ON phase.
(f) A class i connection is lost due to insufficient bandwidth for OFF to ON phase transition.

With the stationary probabilities of this CTMC, denoted by $p(n_1, m_1, \ldots, n_I, m_I)$, the dropping probability of class i incoming and ongoing connections can be computed as follows:

$$b_i^{\text{new}} = \frac{\text{number of class } i \text{ incoming connections dropped upon arrival}}{\text{number of class } i \text{ incoming connections}}$$
$$= \sum_{n_1, m_1, \ldots, n_I, m_I} p(n_1, m_1, \ldots, n_I, m_I)(1 - \mathrm{CAC}(i, \{n_1, m_1, \ldots, n_I, m_I\})),$$

$$b_i^{\text{ongoing}} = \frac{\text{number of class } i \text{ dropped ongoing connections}}{\text{number of class } i \text{ incoming connections}}$$
$$= \sum_{\mathcal{S}_i} \frac{(n_i - m_i)\beta_i}{\lambda_i} p(n_1, m_1, \ldots, n_I, m_I),$$

where $\mathcal{S}_i$ denotes the set of states for which $\sum_{j=1}^{I} m_j c_j + c_i > C$. The link utilization is

$$\rho = \sum_{n_1, m_1, \ldots, n_I, m_I} p(n_1, m_1, \ldots, n_I, m_I) \sum_{j=1}^{I} m_j c_j.$$

If the state space of the CMTC is such that a stationary analysis is feasible, then the computation of the performance parameters is straightforward, but the inverse problem, the design of a CAC that satisfies blocking probability constraints and maximizes link utilization, is still an interesting research problem.

11.1.4 Bandwidth Sharing with Adaptive Traffic Classes

In the case of adaptive traffic classes, connections can adapt their bandwidth to the available bandwidth of the link between the class-specific bandwidth limits $c_{\min}^{(i)}$ and $c_{\max}^{(i)}$. If the link is not completely utilized, then each connection receives its maximal bandwidth. If the sum of the maximal bandwidth needs is larger than the link capacity, then the link is completely utilized and a bandwidth reduction affects the bandwidth of all classes according to the following rule. If the actual bandwidth of a class i connection is c and c is less than $c_{\max}^{(i)}$, then for any other class j the bandwidth is c if $c \le c_{\max}^{(j)}$ or $c_{\max}^{(j)}$ if $c > c_{\max}^{(j)}$. This means that the bandwidth of each class is reduced to the same level c if the class-specific maximal bandwidth $c_{\max}^{(j)}$ is not less than c. Consequently, the main features of bandwidth sharing with adaptive traffic classes are as follows:

- The departure rate of connections is proportional to the number of active connections and is independent of the instantaneous bandwidth of the connections.
- The bandwidth of the connections varies according to the link capacity and the number of active connections.
- An arriving class i connection is rejected when the minimal required bandwidth $c_{\min}^{(i)}$ cannot be granted.
- The transmitted data of a connection depends on the instantaneous bandwidth during the connection.

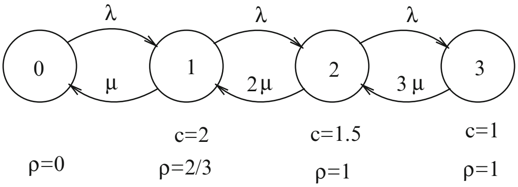


Fig. 11.1 Markov chain of the number of adaptive connections on a finite-capacity link

Example 11.1. We demonstrate the behavior of adaptive connections on a finite-capacity link in the case of a single adaptive class with link bandwidth $C = 3$ Mbps, bandwidth limits $c_{\min} = 1$ Mbps, $c_{\max} = 2$ Mbps, and connection arrival and departure rates λ [1/s] and μ[1/s], respectively. Due to the memoryless arrival and departure processes, the number of active connections $X(t)$ is a Markov chain and it is depicted in Fig. 11.1. The figure also indicates the bandwidth of the ongoing connections. If there are three ongoing connections, then the arriving connections are rejected because in the case of four connections the common bandwidth $c = 3/4$ Mbps would be smaller than the minimal bandwidth requirement $c_{\min} = 1$ Mbps.

The main performance measures of this system are the mean bandwidth of connections

$$\bar{c} = \sum_{i=0}^{3} i c(i) p_i = 2p_1 + 2 \cdot 1.5 p_2 + 3 \cdot 1 p_3,$$

the link utilization

$$\rho = \sum_{i=0}^{3} \rho_i p_i = 2/3 p_1 + 1 p_2 + 1 p_3,$$

and the blocking probability

$$b = p_3,$$

where p_i, ρ_i, and $c(i)$ denote the stationary probability, the utilization, and the bandwidth of a connection in state i, respectively.

Example 11.2. The approach applied to the single class model can be used for the analysis of models with multiple adaptive classes. In the case of two adaptive classes with link bandwidth $C = 5$, bandwidth limits $c_{\min}^{(1)} = 1.5$, $c_{\max}^{(1)} = 3$, $c_{\min}^{(2)} = 1$, $c_{\max}^{(2)} = 2$, connection arrival and departure rates λ_1, λ_2 and μ_1, μ_2, the Markov chain describing the number of active connections of class 1 and 2 is depicted in Fig. 11.2. The figure indicates the bandwidth of the ongoing connections by bold characters. Arriving connections of both classes are rejected in states $(0,5)$, $(1,2)$, $(2,1)$, $(3,0)$, and additionally arriving connections of class 1 are rejected in states $(0,3)$, $(0,4)$. Considering only the minimal bandwidth constraints and the link bandwidth, the state $(1,3)$ would be feasible $(1.5 + 3 \cdot 1 < 5)$, but the identical bandwidth sharing of the classes makes this state infeasible because it violates the minimal bandwidth requirement of class 1 $(5/4 < c_{\min}^{(1)})$. In contrast, in the state $(1,1)$ the bandwidth is

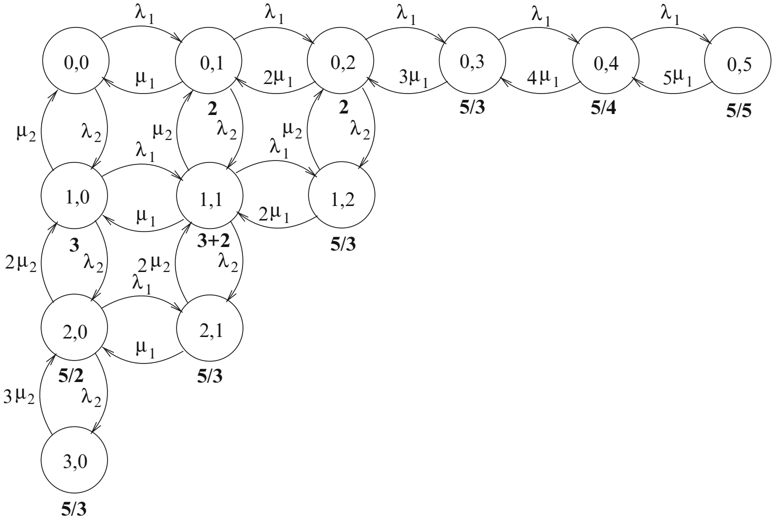


Fig. 11.2 Markov chain of the number of adaptive connections with two adaptive classes

unevenly divided. This is possible because a class 2 connection obtains its maximal bandwidth and the remaining bandwidth is utilized by the class 2 connection. The performance measures can be computed in a similar way as in the case of a single adaptive class.

11.1.5 Bandwidth Sharing with Elastic Traffic Classes

In the case of elastic traffic classes, the connections can adapt their bandwidth to the available bandwidth similar to the adaptive class, but the amount of data transmitted through a connection is fixed. Thus, during a period when the bandwidth is low, the sojourn time of the elastic connections is longer. The bandwidth of elastic connections is also bounded by class-specific bandwidth limits $c_{\min}^{(i)}$ and $c_{\max}^{(i)}$, and the bandwidth sharing between traffic classes follows the same role as in the case of adaptive connections. The main features of bandwidth sharing with elastic traffic classes are as follows:

- The departure rate of a connection of class i depend on the instantaneous bandwidth of the class i connections. Thus, the length of the connections varies according to the link capacity and the number of active connections.
- An arriving class i connection is rejected when the minimal required bandwidth $c_{\min}^{(i)}$ cannot be granted.
- The amount of transmitted data of a connection is a class-specific random variable that does not depend on the instantaneous bandwidth during the connection.

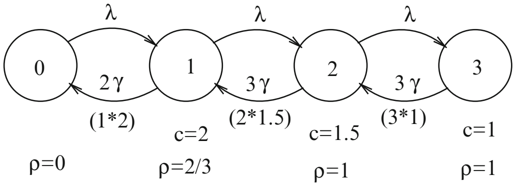


Fig. 11.3 Markov chain of the number of elastic connections on a finite-capacity link

Example 11.3. We demonstrate the behavior of elastic connections with the same model as in Example 11.2 but assuming that the connections are elastic. That is, the link bandwidth is $C = 3$ Mbps, the bandwidth limits are $c_{\min} = 1$ [Mb/s] and $c_{\max} = 2$ [Mb/s], the connection arrival rate is $\lambda[1/\text{s}]$, and the amount of transmitted data of an elastic connection is exponentially distributed with the parameter $\gamma[1/\text{Mb}]$. Due to the memoryless arrival process and the exponential distribution of the amount of transmitted data of the elastic connections, the number of active connections $X(t)$ is a Markov chain (Fig. 11.3). The figure indicates the bandwidth of the ongoing connections (parameter c) and the computation of the departure rate of connections in brackets. For example, in state 2 there are two ongoing connections with bandwidth 1.5 [Mb/s]. The rate at which one of them completes the data transmission is $1.5[Mb/s] \times \gamma[1/Mb] = 1.5\gamma[1/s]$ and the sum of the two identical departure rates is $2 \times 1.5[1/s] = 3[1/s]$. Apart from these differences, the bandwidth sharing, the link utilization, and the rejection of arriving connections are the same as in the case of adaptive connections.

11.1.6 Bandwidth Sharing with Different Traffic Classes

In the previous sections we discussed the bandwidth sharing of a finite-capacity link by traffic classes of the same type. All of the discussed traffic classes have a memoryless stochastic behavior, and thus the performance of the models can be analyzed by CTMCs. Unfortunately, practical limitations arise when the size of the state space gets large, which is often the case in practically interesting situations. The case where only CBR-type connections are present at a link allows for the efficient analysis method referred to as the Kaufman–Roberts method. If any other types of connections appear, then this method will no longer be applicable. The Markov-chain-based framework of the previous sections is also applicable to the analysis of bandwidth sharing by traffic classes of different types. Interested readers may find further details in [68, 77, 78, 81].

11.2 Packet Transmission Through Slotted Time Channel

In this section we focus on a peculiar detail of modeling slotted time systems with discrete-time Markov chains (DTMCs) – the definition of time slots, more precisely, the positioning of the beginning of a time slot on continuous-time axes. The modeler has some freedom in this respect, and consequently different DTMC models can be obtained for describing the same system behavior. It turns out that these different models result in the same performance parameters if the performance parameters are independent of the slot definition, which is the case with the majority of the practically important queueing parameters. Below we evaluate two models of a simple packet transmitter, which can be seen as discrete-time counterparts of an M/M/1 queue.

Consider a packet transmitter with the following properties:

- Packet arrival process: in each time slot, 1 packet arrives with probability p and 0 packets arrive with probability $1 - p$ independently of past history.
- Service (packet transmission) process: if there is at least one packet to transmit, then 1 packet is transmitted with probability q and 0 packets are transmitted with probability $1-q$ independently of past history, which means that the service time of a packet is geometrically distributed with parameter q [Pr service time $= k = (1 - q)q^{k-1}$].
- Service discipline: FIFO.
- Buffer size: infinite.

Let X_n denote the number of packets in a system at the beginning of the nth time slot. X_n is a DTMC with a special birth-and-death structure and infinite state space. Depending on the definition of the beginning of a time slot, the following two cases arise.

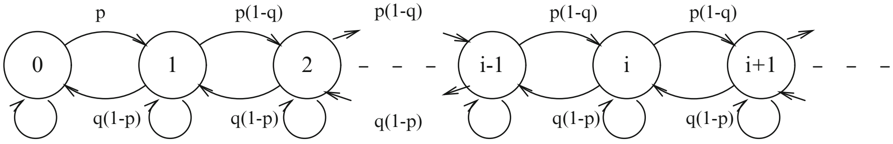
- **Case I.**: A time slot starts with packet transmission (if any), and after that packet, arrivals can happen.
- **Case II.**: A time slot starts with packet arrival (if any), and after that packet, transmission can happen.

These two cases result in different Markov chains, as detailed below.

Case I. In case I, the X_n Markov chain can be described by the following evolution equation:

$$X_{n+1} = (X_n - V_{n+1})^+ + Y_{n+1},$$

where the random variable Y_{n+1} is the number of packet arrivals during time slot $n + 1$ and the random variable V_{n+1} is the number of packets that can be transmitted during time slot $n + 1$. Y_n and V_n are Bernoulli distributed with parameters p and q, respectively. The state-transition graph of this Markov chain is



The stationary distribution of this Markov chain is

$$p_0 = \frac{q-p}{q}, \; p_i = \left(\frac{p(1-q)}{q(1-p)}\right)^i \frac{q-p}{(1-q)q}, \; i \geq 1.$$

The numerator of the stationary probabilities already indicates that the condition of stability is $p < q$. This result can also be obtained from the evolution equation $\mathbf{E}(V) > \mathbf{E}(Y) \to p < q$ and from the Foster criterion (Theorem 3.42) $q(1-p) > p(1-q) \to p < q$. The basic performance measures can be computed from the stationary distribution. The utilization is

$$\rho = 1 - p_0 = 1 - \left(1 - \frac{p}{q}\right) = \frac{p}{q},$$

the mean of the stationary number of packets in the queue is

$$\mathbf{E}(X) = \sum_{i=1}^{\infty} i p_i = \frac{p(1-p)}{q-p},$$

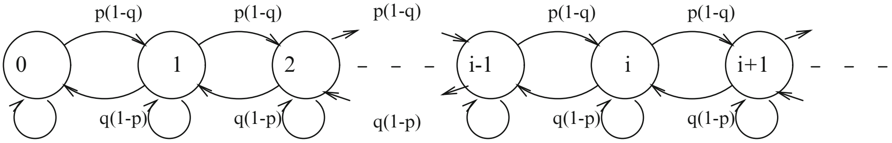
and the mean of the stationary system time of a packet is

$$\mathbf{E}(T) = \frac{1}{q} p_0 + \sum_{i=1}^{\infty} p_i \left(\frac{i+1}{q} - 1\right) = \frac{1-p}{q-p}.$$

Case II. In this case the evolution equation has the form

$$X_{n+1} = (X_n - V_{n+1} + Y_{n+1})^+,$$

and the transition graph of the Markov chain is



Due to the different transition probabilities around state 0, we have a different stationary distribution

$$p_i = \left(\frac{p(1-q)}{q(1-p)}\right)^i \frac{q-p}{(1-p)q}, \; i \geq 0.$$

The computation of some performance measures is identical in this case. for example, the condition of stability is $\mathbf{E}(V) > \mathbf{E}(Y) \to p < q$ based on the evolution equation and $q(1-p) > p(1-q) \rightarrow p < q$ based on the Foster criterion. The computation of some other performance measures is different in case II. For example the utilization is computed as

$$\rho = 1 - p_0(1-p) = 1 - \frac{(q-p)(1-p)}{(1-p)q} = \frac{p}{q}$$

because the server can be utilized by a packet that arrives with probability p when it is idle at the beginning of a time slot. The mean of the stationary system time can be computed as

$$\mathbf{E}(T) = \sum_{i=0}^{\infty} p_i \left(\frac{i+1}{q} - 1 \right) = \frac{1-q}{q-p},$$

and the results are identical with those in case I. In contrast, the mean of the stationary number of packets in the queue is

$$\mathbf{E}(X) = \sum_{i=1}^{\infty} i p_i = \frac{p(1-q)}{q-p},$$

which is different from the results of case I. It reflects the fact that a different number of packets is in the system before and after an arrival.

The evaluated performance measures validate the intuitive expectations that there are performance measures that are dependent on the definition of time slot and others that are independent of that definition. A modeler can choose the time slot freely if the required performance measures are time slot definition independent, but the time slot definition should be related to the required performance measures otherwise.

11.3 Analysis of an Asynchronous Transfer Mode Switch

11.3.1 Traffic Model of an Asynchronous Transfer Mode Switch

In this section we consider the behavior of an asynchronous transfer mode (ATM) [28] switch with N input and N output ports and set up a traffic model of this behavior. An ATM switch transmits packets of fixed size (53 bytes), referred to as a cell. The input and output ports work in a slotted synchronized manner. The length of a time slot is the transmission time of a cell.

We assume that the arrival processes of cells to the input ports of the switch are independent and memoryless. The processes of cell arrival at input ports are

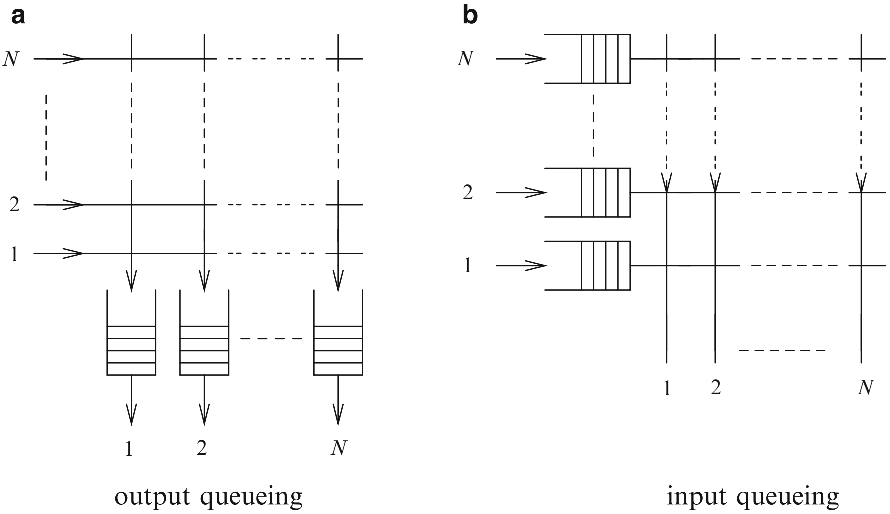


Fig. 11.4 Input and output buffering in packet switching

characterized by a vector $q = \{q_i\}$, $i = 1, \ldots, N$, where q_i is the probability that 1 cell arrives at input port i in a time slot. Consequently, the probability that 0 cells arrive at input port i in a time slot is $1 - q_i$.

An incoming cell is directed to one of the N output ports. We assume that a cell from input port i is directed to output port j with probability w_{ij} independent of the past and the state of the system. The matrix composed by these probabilities $W = \{w_{ij}\}$ is referred to as a traffic matrix. The traffic is a stochastic matrix, that is, $w_{ij} \geq 0$ and $\sum_j w_{ij} = 1$.

The bandwidth of the input and output ports are identical. If more than one cell is directed to a given output port in a given time slot, then only one of them can be transmitted and the others are buffered. As is quantified below, the location of the buffers where the colliding cells are stored has a significant effect on the performance of the switch. We consider two cases: buffering at the input ports and buffering at the output ports. Figure 11.4 depicts the structure of these cases.

Real systems contain buffers both at the input and at the output ports. The performance characteristics and the design of the switch determine the proper model of the system. In the worst case, one cell arrives at each input port and each cell is directed to the same output. If the switch is designed such that it can transfer all of the N cells to the buffer of the given output port in a single time slot, then the output buffer model describes the system properly. If the switch is designed such that it can transfer only one of the conflicting cells to the output buffer and the remaining $N-1$ cells are left at the input buffers, then the input buffer model is the proper model of the system.

Between the input and output buffer models the output buffer seems to provide better performance because in the case of an input buffer model it can happen

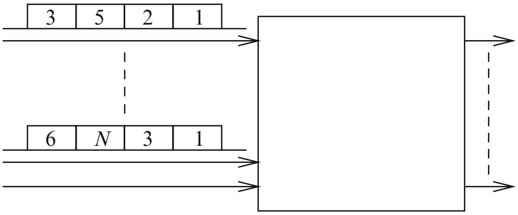


Fig. 11.5 Head-of-line blocking with input buffering

that a given input port is blocked due to the conflict of the first cell in the queue, while other cells waiting in the buffer are directed to idle output ports (Fig. 11.5). This phenomenon is often referred to as head-of-line blocking. This very intuitive qualitative comparison of the two buffer models will be quantified below for some special symmetric configurations.

11.3.2 Input Buffering

In this section we consider the simplest input buffering case, where $N = 2$. If two cells at the heads of the two input ports are directed to the same output port, then one of them is chosen with even probability $(1/2)$ and the chosen one is transferred to the output port and the other one is left in the buffer of the input port. With the preceding modeling assumptions the number of cells in the two input buffers is a DTMC. We assume that the time slots are such that if a cell arrives at an idle buffer and does not collide with any other cell, then it leaves the input port in the same time slot.

Due to the fact that the system state is described by two discrete variables (the number of cells at the two input buffers) it is worth it to depict the state space as a two-dimensional one (Fig. 11.6). The state space can be divided into four parts: both queues are idle, queue 1 is idle and queue 2 is busy, queue 2 is idle and queue 1 is busy, both queues are busy. The state-transition probabilities follow the same structure in these four parts.

Figure 11.7 shows the environment of (0, 0). It is the state where both buffers are idle, and state transitions starting from this state are depicted. In this and the following figures S denotes the probability of conflict. Conflict occurs when two cells from the head of the two buffers are directed to the same output. Its probability is $S = w_{11}w_{21} + w_{12}w_{22}$, where the first term stands for the case where both cells go to input 1 and the second term stands for the case where they go to input 2. Starting from (0, 0), there are the following cases:

- The next state is (1, 0) if there is a conflict and the cell from input 2 is chosen for transmission.
- The next state is (1, 0) if there is a conflict and the cell from input 1 is chosen for transmission.

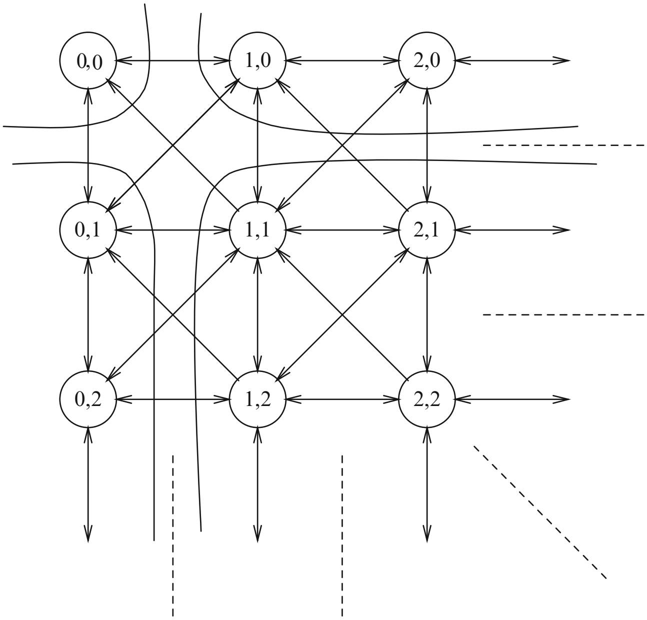
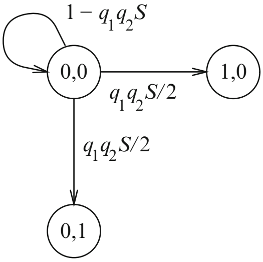


Fig. 11.6 Markov chain with input buffering

Fig. 11.7 Idle buffers



- If there is no conflict (zero or one cell arrives or two cells arrive but the cells are directed to a different output), then the next state is (0, 0).

Figure 11.8 shows the state transitions when buffer 2 is idle and there is at least one cell in buffer 1. Denoting the starting state by $(x, 0)$, $x \geq 1$, the following state transitions can occur:

- $(x, 0) \rightarrow (x - 1, 0)$ if
  - No new cells arrive,
  - A cell arrives at input 2.
- $(x, 0) \rightarrow (x, 0)$ if
  - A cell arrives at input 1 and no cell arrives at input 2;

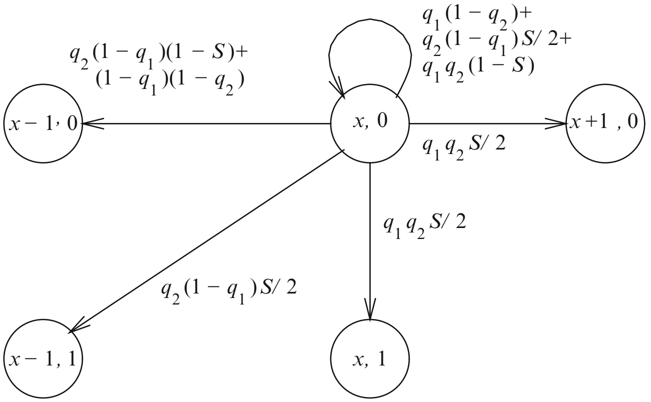


Fig. 11.8 Buffer 2 is idle

- A cell arrives at input 2 and no cell arrives at input 1, it is in conflict with the one at the head of buffer 1, and the cell in buffer 2 is chosen for transmission;
- Cells arrive at both inputs and there is no conflict (the cells at the head of the buffer are directed to different outputs).

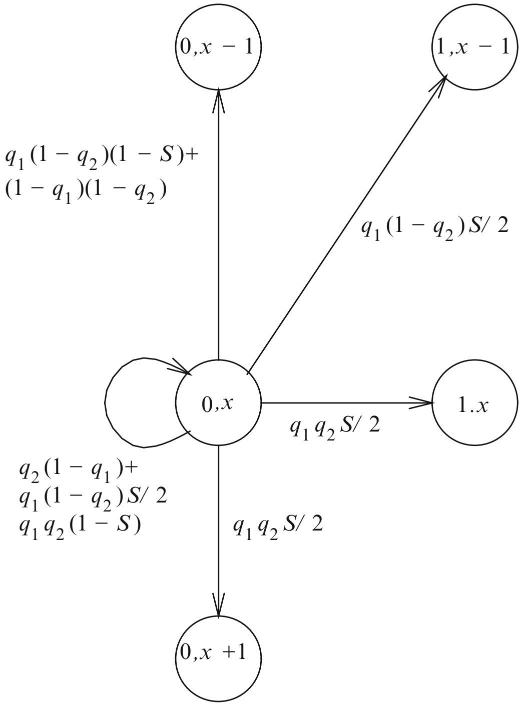
- $(x, 0) \to (x+1, 0)$ if
  - Cells arrive at both inputs, there is a conflict, and the cell in buffer 2 is chosen for transmission.
- $(x, 0) \to (x-1, 1)$ if
  - A cell arrives at input 2 and no cell arrives at input 1, there is a conflict, and the cell in buffer 1 is chosen for transmission.
- $(x, 0) \to (x, 1)$ if
  - Cells arrive at both inputs, there is a conflict, and the cell in buffer 1 is chosen for transmission.

States where buffer 1 is idle and buffer 2 is not idle is depicted in Fig. 11.9. The state transitions of these cases follow a similar pattern as those in Fig. 11.8 by replacing the role of the buffers.

Figure 11.10 presents a case where cells are waiting in both buffers. The figure does not show transition $(x, y) \to (x, y)$ whose probability is 1 minus the sum of the depicted transition probabilities. The following state transitions are possible.

- $(x, y) \to (x-1, y-1)$ if
  - No new cells arrive and there is no conflict.
- $(x, y) \to (x, y-1)$ if
  - A cell arrives at input 1, no cells arrive at input 2, and there is no conflict;

Fig. 11.9 Buffer 1 is idle



  - No new cells arrive, there is a conflict, and the cell in buffer 2 is chosen for transmission.

- $(x, y) \to (x + 1, y - 1)$ if
  - A cell arrives at input 1, no cells arrive at input 2, there is a conflict, and the cell in buffer 2 is chosen for transmission.

- $(x, y) \to (x - 1, y)$ if
  - A cell arrives at input 2, no cells arrive at input 1, and there is no conflict;
  - No new cells arrive, there is a conflict, and the cell in buffer 1 is chosen for transmission.

- $(x, y) \to (x, y)$ if
  - A cell arrives at input 1, no cells arrive at input 2, there is a conflict, and the cell in buffer 1 is chosen for transmission (with probability $q_1(1 - q_2)S/2$);
  - A cell arrives at input 2, no cell arrives at input 1, there is a conflict, and the cell in buffer 2 is chosen for transmission (with probability $q_2(1 - q_1)S/2$);
  - New cells arrive at both buffers, and there is no conflict [with probability $q_1q_2(1 - S)$].

- $(x, y) \to (x + 1, y)$ if

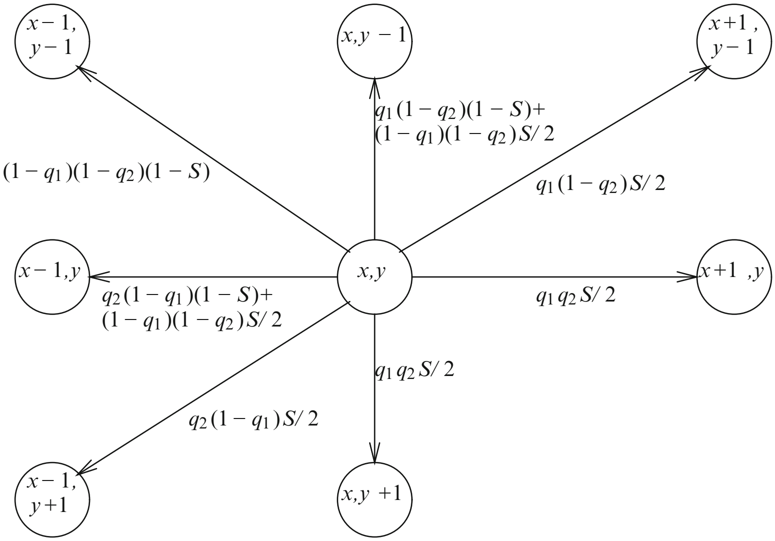


Fig. 11.10 There are cells in both buffers

  - New cells arrive at both buffers, there is a conflict, and the cell in buffer 2 is chosen for transmission.

- $(x, y) \to (x-1, y+1)$ if
  - A cell arrives at input 2, no cells arrive at input 1, there is a conflict, and the cell in buffer 1 is chosen for transmission.
- $(x, y) \to (x, y+1)$ if
  - New cells arrive at both buffers, there is a conflict, and the cell in buffer 1 is chosen for transmission.

11.3.3 Output Buffering

The analytical description of the switch with output buffering is easier than that with input buffering because in this case the number of cells in a buffer depends only on the properties of the arriving cells and is independent of the number of cells in the other buffer. Consequently, it is possible to analyze one output buffer in isolation.

Figure 11.11 presents the Markov chain of buffer 1 with output buffering. There are two possible state transitions if the buffer is idle:

- 0 → 1 if cells arrive at both inputs and both cells are directed to output 1.
- 0 → 0 otherwise.

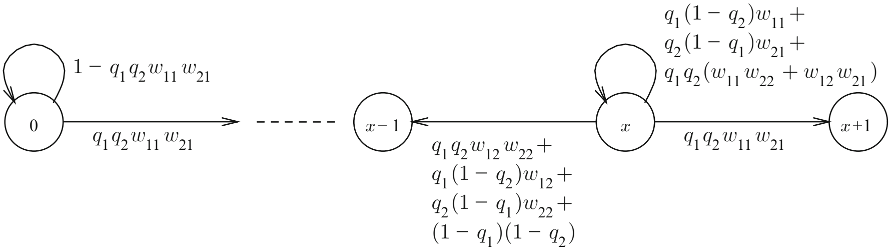


Fig. 11.11 Markov chain of buffer 1 with output buffering

There are three possible state transitions if the buffer is not idle:

- $x \to x - 1$ if a cell arrives at output 1.
- $x \to x$ if one cell arrives at output 1.
- $x \to x + 1$ if two cells arrive at output 1.

The probabilities of these state transitions are provided in Fig. 11.11.

11.3.4 Performance Parameters

In this section we compute some performance parameters in the case of input and output buffering assuming that the buffers are finite.

Mean Number of Cells in Buffers

Let P_{ij}, $i, j \geq 0$, be the steady-state probability of state (i, j) of a Markov chain describing a switch with input buffers. The mean number of cells in buffers 1 and 2 can be computed as

$$E_1 = \sum_{i \geq 0} \sum_{j \geq 0} i P_{ij},$$

$$E_2 = \sum_{i \geq 0} \sum_{j \geq 0} j P_{ij}.$$

Similarly, let $P_i^{(1)}$ and $P_i^{(2)}$, $i \geq 1$, be the steady-state probability of having i cells in buffers 1 and 2, respectively, in Markov chains describing a switch with output buffers. The mean number of cells in buffers 1 and 2 can be computed as

$$E_1 = \sum_{i \geq 1} i P_i^{(1)},$$

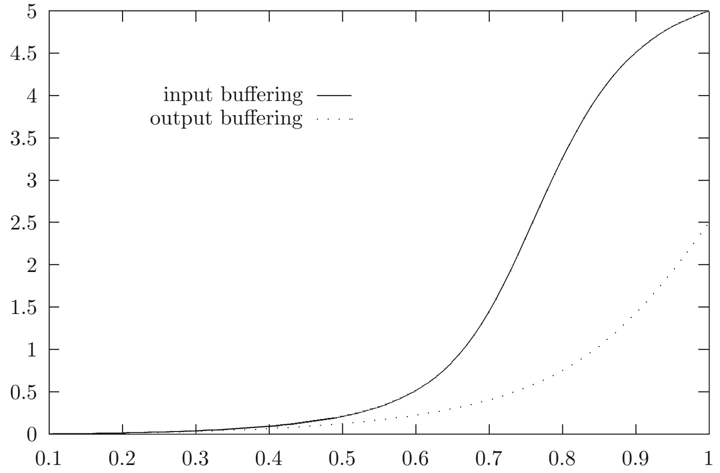


Fig. 11.12 Average number of cells with input and output buffering

$$E_2 = \sum_{i \geq 1} i P_i^{(2)}.$$

Figure 11.12 plots the average buffer content as a function of the arrival probability, $q = q_1 = q_2$, for the input and output buffer models, where the buffer length is limited to 5 and $w_{11} = w_{21} = 0.5$.

Throughput

The throughput (δ) is the mean number of cells the switch transmits in a time slot. In the case of input buffering, we can compute the throughput following the same division of the states of the Markov chain. Denoting the stationary probability of the four parts by $P_{00}, P_{x0}, P_{0y}, P_{xy}$, the throughput is

$$\begin{aligned}\delta = {} & P_{00}[1 \times (q_1(1-q_2) + q_2(1-q_1) + q_1 q_2 S) + 2 \times q_1 q_2 (1-S)] \\ & + \sum_{x \geq 1} P_{x0}[1 \times ((1-q_2) + q_2 S) + 2 \times q_2(1-S)] \\ & + \sum_{y \geq 1} P_{0y}[1 \times ((1-q_1) + q_1 S) + 2 \times q_1(1-S)] \\ & + \sum_{x \geq 1, y \geq 1} P_{xy}[1 \times S + 2 \times (1-S)],\end{aligned}$$

where we detail the cases with one-cell and with two-cell transmission.

In the case of output buffering, the throughput is

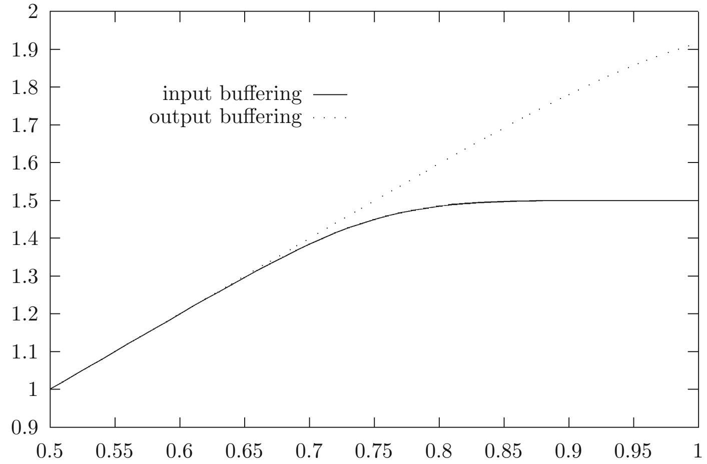


Fig. 11.13 Throughput

$$\delta = P_0^{(1)}\left(q_1(1-q_2)w_{11} + q_2(1-q_1)w_{21} + q_1q_2(1-w_{12}w_{22})\right) + \sum_{x\geq 1} P_x^{(1)}$$
$$+P_0^{(2)}\left(q_1(1-q_2)w_{12} + q_2(1-q_1)w_{22} + q_1q_2(1-w_{11}w_{21})\right) + \sum_{x\geq 1} P_x^{(2)}.$$

Figure 11.13 plots the throughput as a function of cell arrival probability, $q = q_1 = q_2$, for input and output buffering when the buffer length is limited to 5 and $w_{11} = w_{21} = 0.5$

In accordance with intuitive expectations, the throughput with input buffering is less than that with output buffering. As the arrival probability tends to 1, the throughput tends to 1.5 in the case of input buffering. A quick intuitive explanation of this property is as follows. If packets arrive at each time slot, the buffers are always busy, $\sum_{x\geq 1, y\geq 1} P_{xy}$ tends to 1, and δ tends to $1 \times S + 2 \times (1 - S)$ where $S = 1/2$.

11.3.5 Output Buffering in N × N Switch

Let us consider a single output of an $N \times N$ switch with output buffering and assume that cells arrive at the N input ports according to N independent identical Bernoulli processes. The probability that a packet arrives at a time slot is p. The arriving cells are directed to the output ports according to independent uniform distributions.

The number of cells directed to a tagged output port is binomially distributed:

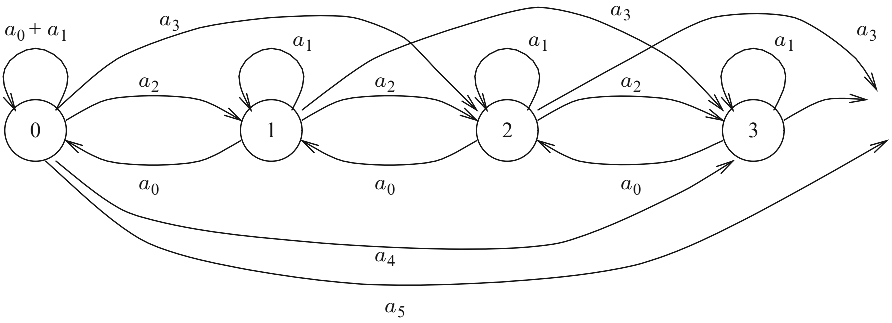


Fig. 11.14 Markov chain modeling a switch with output buffering

$$a_i = \mathbf{P}(i \quad \text{cells arrived in the time slot}) = \binom{N}{i} (p/N)^i (1 - p/N)^{N-i}.$$

Figure 11.14 shows the transition probability graph of a Markov chain describing the number of cells in a tagged output port. This Markov chain can also be described by an evolution equation of type II:

$$X_{n+1} = (X_n - 1 + Y_{n+1})^+,$$

and its state transition probability matrix is

$$\mathbf{\Pi} = \begin{bmatrix} a_0 + a_1 & a_2 & a_3 & a_4 & \cdots & a_{k+1} & \cdots \\ a_0 & a_1 & a_2 & a_3 & \cdots & a_k & \cdots \\ 0 & a_0 & a_1 & a_2 & \cdots & a_{k-1} & \cdots \\ 0 & 0 & a_0 & a_1 & \cdots & a_{k-2} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}. \tag{11.1}$$

The stationary probabilities satisfy the following linear equations:

$$p_0 = p_0 a_0 + p_0 a_1 + p_1 a_0, \tag{11.2}$$

$$p_k = p_k a_1 + p_{k+1} a_0 + \sum_{i=0}^{k-1} p_i a_{k+1-i} = \sum_{i=0}^{k+1} p_i a_{k+1-i}. \tag{11.3}$$

Let us introduce the following z-transform functions

$$P(z) = \sum_{k=0}^{\infty} p_k z^k, \quad A(z) = \sum_{k=0}^{\infty} a_k z^k.$$

From Eqs. (11.2) and (11.3) we have

$$\begin{aligned} P(z) &= p_0a_0 + p_0a_1 + p_1a_0 + \sum_{k=1}^{\infty}\sum_{i=0}^{k+1} p_i a_{k+1-i} z^k \\ &= p_0a_0 + \sum_{k=0}^{\infty}\sum_{i=0}^{k+1} p_i a_{k+1-i} z^k \\ &= p_0a_0 + \sum_{k=0}^{\infty}\sum_{i=1}^{k+1} p_i a_{k+1-i} z^k + \sum_{k=0}^{\infty} p_0 a_{k+1} z^k \\ &= p_0a_0 + \sum_{i=1}^{\infty} p_i \sum_{k=i-1}^{\infty} a_{k+1-i} z^k + z^{-1} \sum_{k=0}^{\infty} p_0 a_{k+1} z^{k+1} \\ &= p_0a_0 + z^{-1} \sum_{i=1}^{\infty} p_i z^i \sum_{l=0}^{\infty} a_l z^l + z^{-1} p_0 \sum_{m=1}^{\infty} a_m z^m \\ &= p_0a_0 + z^{-1}(P(z) - p_0)A(z) + p_0 z^{-1}(A(z) - a_0), \end{aligned}$$

whence

$$P(z) = \frac{(1 - z^{-1}) p_0 a_0}{1 - z^{-1}A(z)} = p_0 a_0 \frac{(z-1)}{z - A(z)}.$$

Considering that $\lim_{z\to 1} P(z) = 1$ and applying l'Hospital's rule we have

$$1 = p_0 a_0 \frac{1}{1 - A'(z)} \; |_{z=1},$$

where $A'(1)$, the mean number of cells arriving at the tagged output in a time slot, can be computed;

$$p_0 a_0 = 1 - A'(z) = 1 - p,$$

and using this

$$P(z) = \frac{(1-p)(z-1)}{z - A(z)} = \frac{(1-p)(1-z)}{A(z) - z}.$$

To check the obtained results we set $N = 2$. In this case the probability that i cells arrive at the tagged output post is

$$a_i = \binom{2}{i} \left(\frac{p}{2}\right)^i \left(1 - \frac{p}{2}\right)^{2-i},$$

whence

$$A(z) = \left(1 - \frac{p}{2} + z\frac{p}{2}\right)^2$$

and

$$P(z) = \frac{(1-p)(1-z)}{\left(1-\frac{p}{2}+z\frac{p}{2}\right)^2 - z}.$$

The probability that the buffer is idle, p_0, is easily obtained from the transform domain expression

$$p_0 = P(z)\mid_{z=0} = \frac{1-p}{\left(1-\frac{p}{2}\right)^2}.$$

This result can be checked easily by considering that for $N = 2$ the Markov chain has a birth-death structure with forward probability a_2 and backward probability a_0. From this Markov chain we have

$$p_0 = \frac{1}{1+\sum_{k=1}^{\infty}\left(\frac{a_2}{a_0}\right)^k} = \frac{1-p}{\left(1-\frac{p}{2}\right)^2}.$$

11.3.6 Throughput of N × N Switch with Input Buffering

The end of Sect. 11.3.4 shows the throughput computation of a 2 × 2 switch with input buffering. In this section we compute the throughput of larger switches, $N > 2$, with input buffering.

We assume that the cells are indistinguishable, the switch chooses one of the cells in conflict with independent uniform distribution, and the cells are directed to output ports in a uniformly distributed manner.

To obtain the maximal throughput of the system, we assume that cells are waiting in each buffer of the switch, i.e., none of the input buffers is idle.

We use the following notations:

- Let R_m^i be the number of cells that are at the head of a buffer at time m, are directed to output i, and are not forwarded due to collision.
- Let A_m^i be the number of cells that arrive at the head of a buffer at time m and are directed to output i.

R_m^i is a DTMC. It can be described by the evolution equation

$$R_m^i = (0, R_{m-1}^i + A_m^i - 1)^+,$$

where the sum on the right-hand side is reduced by 1 due to the cell that is transmitted to output i. A_m^i follows a binomial distribution

$$\mathbf{P}\left(A_m^i = k\right) = \binom{F_{m-1}}{k}(1/N)^k(1-1/N)^{F_{m-1}-k}, \qquad k = 0, 1, \cdots, F_{m-1}, \tag{11.4}$$

Table 11.1 Per output port throughput as a function of N

N	Throughput
1	1.0000
2	0.7500
3	0.6825
4	0.6553
5	0.6399
6	0.6302
7	0.6234
8	0.6184
∞	0.5858

where the number of new cells at the head of the buffer is

$$F_{m-1} = N - \sum_{i=1}^{N} R_{m-1}^{i}. \tag{11.5}$$

Equation (11.5) is based on the assumption that none of the input buffers is idle. Due to the same assumption the number of new cells arriving at the head of the buffers is equal to the number of cells successfully transmitted in a time slot. Consequently, the throughput output i is $\delta^i = \lim_{m\to\infty} E(A_m^i)$.

The parameters of the binomial distribution are F_{m-1} and $1/N$ since there are F_{m-1} new cells at the heads of the buffers and they choose their destination according to a uniform distribution.

With all elements of the evolution equation defined it is possible compute the stationary distribution of the Markov chain numerically. From the stationary distribution we also have

$$E(R^i) = \lim_{m\to\infty} E(R_m^i).$$

Taking the expectation of A^i based on Eq. (11.4) and both sides of Eq. (11.5) we get

$$E(A^i) = E(F)/N \text{ and } E(F) = N - \sum_{i=1}^{N} E(R^i) = E(F) = N - NE(R),$$

where we utilized the symmetry of the uniform output selection in the last step. Introducing $\delta = \delta^i \ E(A) = E(A^i) \ E(R) = E(R^i)$ we get

$$\delta = E(A) = 1 - E(R).$$

For any finite N the previously discussed numerical method results in the throughput of one output port. Table 11.1 presents the throughput as a function of N. For $N = 2$ we already computed the result in Sect. 11.3.4, but there we computed the throughput of the whole switch, not for one output port.

When $N \to \infty$, the same evolution equation holds, but A_m^i tends to be Poisson distributed with the parameter δ. At the limit we obtain a Markov chain with the same structure as that in Eq. (11.1). Following the same transform domain analysis and using $A(z) = e^{\delta(z-1)}$ we obtain $E(R)$ and from $\delta = 1 - E(R)$ the limiting throughput of the switch.

11.4 Conflict Resolution Methods of Random Access Protocols

One of the main functions of medium access control (MAC) is to share common resources between randomly arriving customer requests. Different random access protocols are developed for this purpose. In the case of random access protocols, several users try to communicate through a common transmission channel. The users do not know the activity of the others. In this kind of environment stable communication requires the application of a protocol that under a system-dependent load level ensures

- Stable communication (with finite mean delay),
- Transmission of all packets,
- Fairness (users obtain the same service).

These protocols work based on the information available about the state of the users. The following procedures are different members of the set of random access protocols, which differ (1) in the way users are informed about the status of the common channel and, indirectly, the activity of the other users and (2) in the design goals to adopt to the alternation of the traffic load.

11.4.1 ALOHA Protocol

The ALOHA protocol [2] is the simplest random access protocol. It was developed for simple radio communication between radio terminals and a central station. It uses two radio channels, one of which is used by the terminals for communication to the central station and the other for communication from the central station to all terminals (Fig. 11.15).

If more than one terminal sends a message to the central station, then the signals interfere and the central station cannot receive any correct messages. This case is referred to as a collision of messages. Collision can only happen in the first radio channel since the second radio channel is used only by the central station. Successfully received messages are acknowledged by the central station. The terminals are informed about the success of their message transmission by these acknowledgements. If no acknowledgement arrives within a given deadline, then the

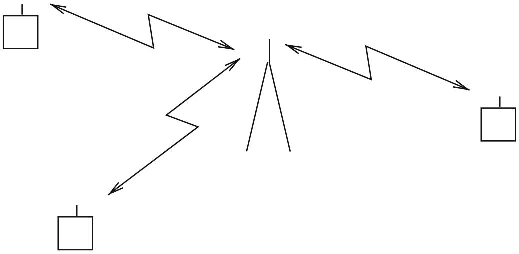


Fig. 11.15 Radio terminal system

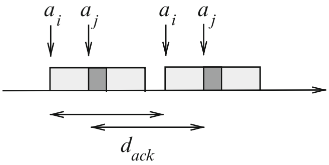


Fig. 11.16 Packet retransmission without random delay

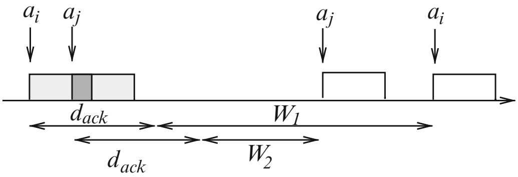


Fig. 11.17 Packet retransmission with random delay

terminal assumes that the transmission failed. The ALOHA protocol is designed to ensure communication in this system.

ALOHA functions as follows. As soon as a terminal has a new packet to transmit it starts sending it away without any attention to the activity of the other terminals. If no acknowledgement arrives within a given deadline, the terminal assumes that the message collided and it switches to message retransmission mode. Terminals in message retransmission mode are referred to as blocked terminals. In this mode the terminal retransmits the message until it receives an acknowledgement about successful transmission.

If after an unsuccessful transmission a terminal were to retransmit the message immediately, then there would be multiple collisions among the same set of terminals (Fig. 11.16).

To avoid these multiple collisions in the same set of terminals, the terminals wait for a random amount of time before retransmitting messages (Fig. 11.17). The cited figures distinguish between the period of collision (dark gray) and the additional period that is wasted due to the collision (light gray).

The quantitative behavior of a system is straightforward. If the terminals choose a large random delay, then the probability of consecutive collisions with the same set of terminals is low, but the time to successful transmission is high due to the long delay till retransmission. In the opposite case, if the delay is short, then the probability of subsequent collisions is higher, and this could cause a longer time to successful transmission due to the high number of repeated transmission attempts. The optimal behavior of the system is somewhere between these extremes.

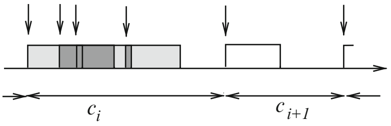


Fig. 11.18 Cycles of busy and idle periods

The modeling and analysis of ALOHA systems has been a well-studied area since 1970s. It is still an interesting research area because several random access protocols that were subsequently introduced contain elements of the basic ALOHA protocol (as is detailed in the following sections).

There exists a wide variety of performance studies. These studies differ in their assumptions about the behavior of users and systems. It is practically impossible to analyze the simplest ALOHA protocol in all its minute technical details. To reduce the complexity of the models, several simplifying assumptions are used. The obtained simplified models often closely approximate real system behavior.

In the following sections we introduce some of the simplest models of the basic ALOHA system and their analysis.

Continuous Time ALOHA System

We adopt the following modeling assumptions:

- The aggregate arrival process of new and retransmitted messages is a Poisson process with parameter λ.

  This model is not a correct model of the aggregate arrival process (in general) but there are several cases (e.g., the number of blocked terminals is negligible compared to the number of all terminals) when it properly approximates the real system behavior. This kind of model, where the arrivals of the new and the retransmitted messages are considered in an aggregate flow, is referred to as a zero-order model.
- The length of the messages is fixed and the time of a message transmission is T.

With the zero-order model we evaluate which portion of the new and repeated messages, which arrive according to a Poisson process with parameter λ, is transmitted successfully, and what is the related transmission delay and collision probability.

As shown in Fig. 11.18 we divide the time axes according to the busy and idle periods of the common channel.

The probability of a successful message transmission in a busy period equals the probability that after the beginning of a busy period the next message arrives later than T. Its probability is $\mathrm{e}^{-\lambda T}$.

In order to determine the long-term idle ratio, successful busy and unsuccessful busy periods, we determine the average length of these periods. The interarrival time in a Poisson process with parameter λ is exponentially distributed with parameter λ. The length of an idle period is the remaining time of an exponentially distributed interarrival time, which is exponential again with the same parameter. Thus the mean length of an idle period is $1/\lambda$.

The length of a successful busy period is T. The difficult question is the length of the unsuccessful period. An unsuccessful busy period is composed of $N - 1$ ($N \geq 2$) interarrival intervals shorter than T and a final interval of length T. The case where $N = 1$ is the successful busy period. Due to the memoryless property of Poisson processes we can compute the number of colliding messages during the unsuccessful busy period independently of the length of the interarrival times,

$$Pr(N = n) = (1 - \mathrm{e}^{-\lambda T})^{n-1}\mathrm{e}^{-\lambda T}.$$

The CDF of the length of an interarrival interval shorter than T, denoted as U, is

$$F_U(t) = Pr(U < t) = Pr(\tau < t|\tau < T) = \begin{cases} \dfrac{1 - \mathrm{e}^{\lambda t}}{1 - \mathrm{e}^{\lambda T}} & 0 < t < T, \\ 1 & T < t, \end{cases}$$

whence $E(U) = \dfrac{1 - \mathrm{e}^{-\lambda T} - \lambda T \mathrm{e}^{-\lambda T}}{\lambda(1 - \mathrm{e}^{-\lambda T})}$. Consequently, in a cycle composed of a busy and an idle period

- The mean length of the idle period is $E(I) = 1/\lambda$,
- The probability of a successful message transmission is $E(S) = \mathrm{e}^{-\lambda T} T$, and
- The mean length of an unsuccessful busy period is

$$E(L) = \sum_{n=2}^{\infty} Pr(N = n)\Big((n - 1)E(U) + T\Big) = \frac{1 - \mathrm{e}^{-\lambda T} - \lambda T \mathrm{e}^{-\lambda T}}{\lambda \mathrm{e}^{-\lambda T}}.$$

System utilization is characterized by the portion of time associated with successful message transmission:

$$\rho = \frac{E(S)}{E(I) + E(S) + E(L)} = \lambda T \mathrm{e}^{-2\lambda T}.$$

It can be seen that utilization depends only on the λT product. The maximum of utilization is obtained through the derivative of ρ as a function of λT. The maximum is found at $\lambda T = 0.5$ and is $\rho = 1/2e \sim 0.18394$. Figure 11.19 shows that utilization decreases significantly as the load increases above 0.5; consequently these systems should be operated with a load lower than 0.5.

The mean number of arriving messages in a Δ long interval is $\lambda\Delta$. In the same interval the mean time of successful message transmission is $\rho\Delta$. During this interval the mean number of successfully transmitted messages is $\rho\Delta/T$. The ratio between the number of successfully transmitted messages and the number of all message transmission attempts, which is the mean number of transmission attempts per message, is $E(R) = \lambda T/\rho = \mathrm{e}^{2\lambda T}$.

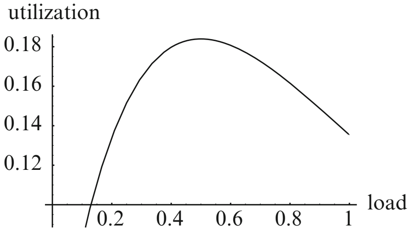


Fig. 11.19 Utilization ρ as a function of load λT

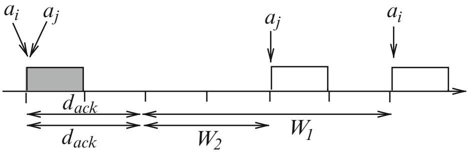


Fig. 11.20 Slotted ALOHA system

Having the mean number of transmission attempts per message we can compute the message transmission delay:

$$E(D) = E\left(\sum_{r=1}^{\infty} Pr(R=r)\left(rT + \sum_{i=1}^{r-1} d_{\text{ack}} + W_i\right)\right)$$
$$= E(R)T + (E(R)-1)(d_{\text{ack}} + E(W)),$$

where d_{ack} is the time a terminal waits for message acknowledgement and W is the random delay spent before message retransmission.

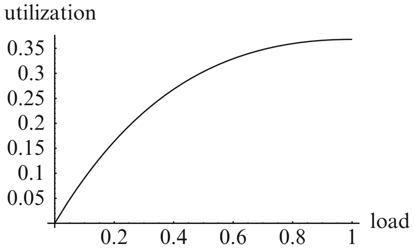
Discrete-Time (Slotted) ALOHA System

The main disadvantage of continuous-time ALOHA systems is that the wasted time when messages collide is large. This phenomenon can be seen in Figs. 11.16 and 11.18. The dark gray period denotes the overlapping intervals of colliding messages, while light gray periods are additional wasted time intervals that cannot be used for useful message transmission.

With a simple modification of the ALOHA system this additional wasted time interval can be avoided. If all terminals work in a synchronized manner and initiate message transmission only at the beginning of time slots, then the length of time the colliding messages occupy the common channel reduces to T. Naturally in this system the delay of a message retransmission should be an integer multiple of the time slot, T. This system is commonly referred to as a slotted ALOHA system (Fig. 11.20).

The zero-order model of a slotted ALOHA system assumes that the terminals generate a Poisson distributed number of new and repeated messages in a time slot, where the parameter of the Poisson distribution is λT. This model of message arrivals is similar to the zero-order model of continuous-time ALOHA systems

Fig. 11.21 Utilization of slotted ALOHA system



assuming that the messages are generated continuously according to a Poisson process, but the messages generated during a time slot are delayed till the beginning of the next time slot.

With these assumptions, utilization of the zero-order model of a slotted ALOHA system can be computed based on the analysis of a single time slot. Let N be the number of packets generated in a time slot. In this case,

$$\rho = Pr(\text{successful message transmission}) = Pr(N = 1) = \lambda T \mathrm{e}^{-\lambda T}.$$

Maximum utilization is obtained at $\lambda T = 1$, and it is $\rho = 1/e \sim 0.367879$. Compared to the continuous-time ALOHA system, the optimal throughput doubles and the aggregated load (new and repeated messages) can be increased to the capacity of the system ($\lambda T = 1$), as is plotted in Fig. 11.21.

The mean number of retransmission attempts, R, can be computed as the ratio between the successfully transmitted and all messages:

$$E(R) = \frac{E(N)}{E(\text{successfully transmitted messages})} = \frac{\lambda T}{\lambda T \mathrm{e}^{-\lambda T}} = \mathrm{e}^{\lambda T}.$$

Similar to the continuous-time case, the message transmission delay is

$$\begin{aligned} E(D) &= E\left(\sum_{r=1}^{\infty} Pr(R = r) \left(rT + \sum_{i=1}^{r-1} d_{\text{ack}} + W_i \right) \right) \\ &= E(R)T + (E(R) - 1)(d_{\text{ack}} + E(W)). \end{aligned}$$

The more complex models of the ALOHA system distinguish the states of the terminals (message generation, new message transmission attempt, waiting random delay, message retransmission attempt) and characterize the arrival of new and repeated messages according to those states [27].

In contrast to the terminology of ALOHA systems, the general terminology of random access protocols refers to stations instead of terminals and packets instead of messages.

Fig. 11.22 CSMA system

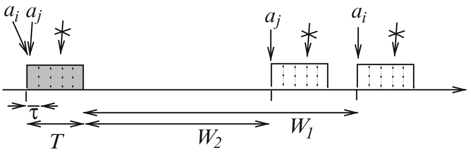
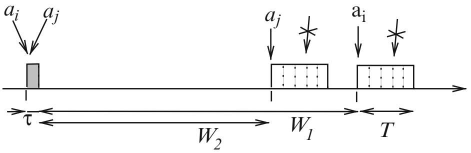


Fig. 11.23 CSMA/CD system



11.4.2 CSMA and CSMA/CD Protocols

The more advanced random access protocols aim to enhance channel utilization based on the information available for the stations by the given physical media.

In the introduction of the slotted ALOHA system we saw that the reduction of the time period while a collision makes the channel unavailable enhances the performance of the protocol. In the case of radio terminal systems where the terminals can be outside of each other's propagation range, it is hard to further reduce the ineffective time of the channel. In the case of wired systems, the stations can sense each other's signals, but not immediately; there is a propagation delay of the medium. This direct sensing of the stations can be used to enhance the performance of the multiple access protocol in the following two ways.

- If one station senses that another station is sending a packet when it has a packet to send, then the first station does not start sending the packet.
- If by accident the packets of two stations collide (because they are sent within the propagation delay), then the stations can recognize that the packets collide and finish the useless packet transmission immediately.

The first way is referred to as *carrier sense multiple access* (CSMA) and the second as *collision detection* (CD).

Figures 11.22 and 11.23 demonstrate the behavior of the CSMA and the CSMA/CD systems. In these systems the time is slotted and the time unit is the maximal propagation delay between the most remote stations, τ. Collision can happen only among packets transmitted within the same slot because in the next time slot all stations are aware of the busy state of the channel. A station can initiate packet transmission only if the channel is idle. In the case of CSMA without CD, colliding packets are transmitted completely. Thus a significant portion of the channel capacity is lost (Fig. 11.22). In the case of CSMA with CD, the collision is recognized within one time slot and packet transmission is finished immediately (Fig. 11.23).

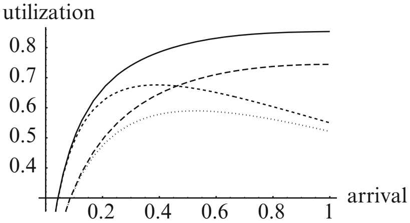


Fig. 11.24 Utilization of slotted CSMA and CSMA/CD systems

Performance of Slotted CSMA System

We analyze the zero-order model of a system by the analysis of the intervals between consecutive packet transmission attempts. The beginning of these intervals is indicated by the arrows below the time axes in Figs. 11.22 and 11.23. It can also happen that, in contrast to the figures, there is no idle period between two consecutive packet transmission attempts. According to the zero-order model of a system, we assume that after an idle time slot or the last time slot of a successful packet transmission there is a Poisson distributed number of (new and repeated) packet transmissions initiated with parameter $\tau\lambda$. That is, in contrast with the previously discussed zero-order models, the state of the channel affects the arrival process of the packets. Packets can arrive in the aforementioned time slots and not otherwise. The success of the packet transmission depends on the number of arriving packets, N:

$$Pr(\text{succesfull packet transmission}) = Pr(N = 1|N \geq 1) = \frac{\lambda\tau \mathrm{e}^{-\lambda\tau}}{1 - \mathrm{e}^{-\lambda\tau}}.$$

After a successful or colliding packet transmission the channel remains idle until the next packet arrives. Let I denote the number of idle time slots until the next packet arrives. Due to the memoryless property of the arrival process, I is geometrically distributed. $Pr(I = i) = \mathrm{e}^{-\lambda\tau i}(1 - \mathrm{e}^{-\lambda\tau})$. Consequently, in an interval between consecutive packet transmission attempts

- The mean length of the idle period is $E(I)\tau = \frac{\tau\ \mathrm{e}^{-\lambda\tau}}{1-\mathrm{e}^{-\lambda\tau}}$,
- The mean length of successful packet transmission is $E(S) = \frac{T\ \lambda\tau\mathrm{e}^{-\lambda\tau}}{1-\mathrm{e}^{-\lambda\tau}}$, and
- The mean length of unsuccessful packet transmission is $E(L) = \frac{T\ (1-(1+\lambda\tau)\mathrm{e}^{-\lambda\tau})}{1-\mathrm{e}^{-\lambda\tau}}$.

Finally, utilization is obtained as

$$\rho = \frac{E(S)}{E(I) + E(S) + E(L)} = \frac{T\ \lambda\tau\mathrm{e}^{-\lambda\tau}}{T(1 - \mathrm{e}^{-\lambda\tau}) + \tau\mathrm{e}^{-\lambda\tau}}.$$

Figure 11.24 plots utilization as a function of $\lambda\tau$ when $\tau/T = 0.2$ (dotted line) and when $\tau/T = 0.1$ (short dashed line). It can be seen that the probability of collision is lower and utilization is higher in the case of shorter propagation time ($\tau/T = 0.1$). In any case, utilization reaches an optimum and starts decreasing when the load is increasing. The optimal load level depends on the τ/T ratio.

Performance of Slotted CSMA/CD System

The zero-order model of the CSMA/CD system is very similar to that of the CSMA system. It differs only in the time of unsuccessful packet transmission, which is shorter due to collision detection: $E(L) = \frac{\tau\,(1-(1+\lambda\tau)\mathrm{e}^{-\lambda\tau})}{1-\mathrm{e}^{-\lambda\tau}}$. As a result, utilization is

$$\rho = \frac{E(S)}{E(I)+E(S)+E(L)} = \frac{T\ \lambda\tau\mathrm{e}^{-\lambda\tau}}{T\lambda\tau\mathrm{e}^{-\lambda\tau} + \tau(1-\lambda\tau\mathrm{e}^{-\lambda\tau})}.$$

Figure 11.24 plots the utilization of a CSMA/CD system as a function of load, $\lambda\tau$, together with that of the CSMA system. The propagation delay is $\tau/T = 0.2$ (long dashed line) and $\tau/T = 0.1$ (solid line). Also, in these cases the shorter propagation delay increases utilization. In contrast with the CSMA system, utilization is continuously increasing with the load due to the efficient utilization of the channel.

Slotted Persistent CSMA/CD System

Up to now we have not discussed the behavior of a station when it has a packet to transmit but the channel is busy. Indeed we implicitly assumed that these stations assumed that their packet collided and delayed the next packet retransmission attempt accordingly. This behavior is referred to as nonpersistent station behavior.

The stations sense the channel and know the history of the channel state from which they can compute when the packet under transmission finishes. Knowing this information, a station with a packet to transmit can reduce the packet retransmission time by attempting a packet transmission immediately when the channel becomes idle next. This behavior is referred to as persistent station behavior.

In the zero-order model of persistent CSMA systems we assume that in each τ long time slot stations generate a Poisson distributed number of new and repeated packets, and those stations that generate packets during a packet transmission attempt to transmit packets when the channel becomes idle next.

The analysis of this system is based on the analysis of successful (S), colliding (L), and idle (I) intervals because the system behavior is memoryless at the beginning of these periods. The mean length of these intervals is as follows: $E(S) = T$, $E(L) = \tau$, and $E(I) = \frac{1}{1-\mathrm{e}^{-\lambda\tau}}$.

To compute the utilization we also need to know how often these intervals occur. The following transition probability matrix defines the probability of the occurrence of various consecutive intervals:

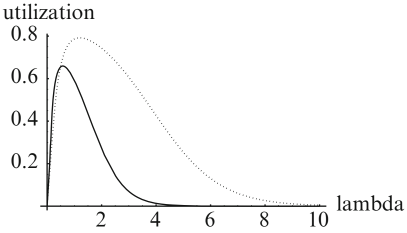


Fig. 11.25 Utilization of slotted persistent CSMA/CD system

$$\Pi = \begin{array}{|ccc|c} S & L & I & \\ \hline P(\lambda T, 1) & P(\lambda T, >1) & P(\lambda T, 0) & S \\ P(\lambda\tau, 1) & P(\lambda\tau, >1) & P(\lambda\tau, 0) & L \\ \dfrac{P(\lambda\tau, 1)}{P(\lambda\tau, >0)} & \dfrac{P(\lambda\tau, >1)}{P(\lambda\tau, >0)} & 0 & I \end{array}$$

where $P(a, i) = \mathrm{e}^{-a} a^i / i!$ and $P(a, > i) = \sum_{j=i+1}^{\infty} P(a, j)$. This is a DTMC whose stationary solution is obtained by the solution of the linear system of equations $\pi\Pi = \pi$, $\pi_S + \pi_L + \pi_I = 1$, where $\pi = \pi_S, \pi_L, \pi_I)$. Given the stationary probabilities π_S, π_L, and π_I, the utilization is

$$\begin{aligned} \rho &= \frac{\pi_S E(S)}{\pi_S E(S) + \pi_L E(L) + \pi_I E(I)} \\ &= \frac{\lambda\tau\, \lambda T}{1 - \lambda T \mathrm{e}^{-\lambda T} + \lambda\tau \mathrm{e}^{-\lambda T}\Big(1 - \lambda\tau + \lambda T(1 - \mathrm{e}^{\lambda\tau})\Big) + \lambda\tau\Big(\mathrm{e}^{\lambda\tau} + \lambda T - 1\Big)}. \end{aligned}$$

Figure 11.25 plots utilization as a function of load, λ, with propagation delay $\tau/T = 0.1$ (dotted line) and with $\tau/T = 0.2$ (solid line). As with the previous cases, shorter propagation delays result in a lower probability of collision and better utilization. Utilization decreases when the load is high. This is because the probability of collision after a successful packet transmission becomes very high due to persistent station behavior.

In summary, nonpersistent behavior is beneficial when the delay due to a collision at the end of a successful packet transmission is less than the normal message retransmission delay. At low load levels, persistent behavior decreases the delay (because the probability of collision is low), while at high load levels nonpersistent behavior performs better.

There is a continuous transition between persistent and nonpersistent behaviors. It is obtained when a station follows persistent behavior with probability p and nonpersistent behavior with probability $1 - p$. This behavior is referred to as p-persistent behavior. Obviously $p = 1$ results in persistent and $p = 0$ nonpersistent behavior. For a given load level we can optimize the system utilization by setting p to an optimal value.

11.4.3 IEEE 802.11 Protocol

One of the most commonly used ways to wirelessly access computer networks currently is the wireless fidelity (WF or wifi), which is defined in the IEEE 802.11 standard. The core of this rather complicated protocol is also an enhanced version of the slotted ALOHA protocol. The IEEE 802.11 protocol [1] is designed to meet the following requirements:

- The random delay for packet transmission is bounded.
- The protocol operates in a wide range of traffic load and adapts the actual level of traffic load.
- If large packets are transmitted, priority is given to the completion of the already started packets.

According to these requirements the ALOHA protocol is modified as follows: [11,32].

- At a collision a station draws a random number uniformly distributed between 1 and M_i and retransmits the collided packet after the expiration of that many slots.
- The upper limit of the uniformly distributed delay depends on the number of unsuccessful transmission trials. After the first collision this value is $M_1 = 8$ and it doubles after each consecutive collision $M_i = 8 * 2^{i-1}$ until a predefined upper limit, $M_{\max}$, is reached.
- Large packets are transmitted in several small segments. In the case of an ALOHA system these segments are transmitted one by one and each of them can collide with other segments and got delayed by the collision resolutoin procedure. Thus the packet transmission delay, which is determined by the largest delay of the segments, can be very high. IEEE 802.11 reduces the packet transmission delay by giving priority to the consecutive segments of a packet under transmission. Thus only the first segment of a packet participates in the contention and other segments are transmitted with high priority. The protocol implements this feature by the introduction of two different delays. The stations in contention consider the medium available if it is idle for a *distributed interframe space* (DIFS) period, while packet segments can be sent within a *short interframe space* (SIFS) period, which is shorter than a DIFS period.

The IEEE 802.11 protocol is built on a so-called basic access method, which is practically identical with the ALOHA protocol and combines with various reservation methods. One of these reservation methods is the aforementioned DIFS- and SIFS-based packet transmission. The mathematical description of these complex reservation methods is rather complex. In this section we present an analytical model of the basic access method, which is introduced in [11]. This model is also based on a simplifying assumption. It assumes that there are so many independently working stations in the system that a packet transmission trial will be unsuccessful with probability p in each time slot independent of the past history of the system. Furthermore, to compute the maximal throughput we assume that stations always have packets to transmit.

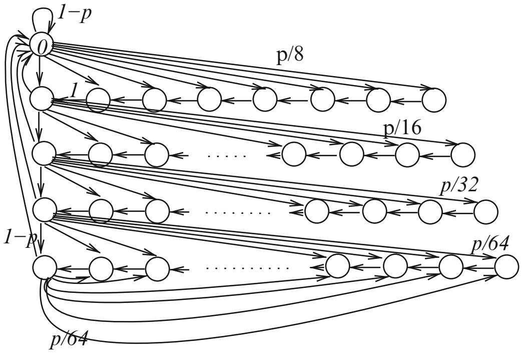


Fig. 11.26 Markov chain describing basic access method of IEEE 802.11 standard

With this assumption we can describe the behavior of a station with a DTMC. The state of the Markov chain describes the phase of the actual packet transmission attempt. Figure 11.26 shows the state-transition graph of this Markov chain. State 0 indicates that the station just finished transmitting a packet and is trying to transmit the next one in the next time slot. If it is unsuccessful, which will happen with probability p, then it draws a uniformly distributed random sample between 1 and $M_1 = 8$. Transitions to the right with probability 1 describe a situation where the station waits until the given delay expires. When the chain arrives at the leftmost state, it attempts to transmit the packet again and go back to state 0, or it moves to the next row, etc. In this Markov chain the retransmission delay is limited to $M_{\max} = 64$. The time between two consecutive visits to state 0 represents the packet transmission delay. The throughput of the station is p_0 and the mean packet transmission delay is $1/p_0$ if p_0 is the stationary probability of state 0 in this Markov chain.

11.5 Priority Service Systems

Priority systems appear in different fields [37,39,45,91]. Several aspects of telecommunications, data management, planning of computer networks, organization of health services, and automatization of production processes could be mentioned. For example, in mobile cellular networks the coverage area is partitioned into cells, and each cell can serve at most c simultaneous communications and use some channels from other cells. There are calls initiated by subscribers from the cell and handover calls from others. Handover calls already use the network resources and should be prioritized with respect to new calls. Different approaches are possible, e.g., a special channel or a priority queue of handover calls.

The problem may be formulated as follows. Customers of different types enter the service system, each of them belonging to a priority class, indexed by a subscript. We assume that customers with a lower index have higher priority in service; this way customers with a lower index can leave the queue earlier than customers with a higher index which were already in the queue at the arrival of customers with a lower index. There are two possible cases: either the entry of customers of higher priority does not interrupt the actual service with lower priority customers or immediately starts its service (in the first case we speak of relative, in the second case of absolute priority). In the second case we have again two choices, whether or not the work up to this moment will be taken into account in the future. With respect to the first possibility, one must complete the residual service, for the second one must complete the residual service, for the second one must complete the whole service later, when the higher proirity customers are served. Both of these possibilities occur in computer systems. For example, the results of computations are either regularly saved or not. In the first case results are not lost at a system error. Similar situations appear in other fields. When a disaster occurs, one must first to divert the danger and after that to deal with less urgent tasks. For example, a dentist must first see patients who are in pain; other patients can wait.

We will consider service systems with two Poisson arrival processes where the service time will have exponential and general distributions. In the exponential case we will follow the usual method – find the system of differential equations describing the functioning of system, solve it, and at $t \to \infty$ determine the equilibrium distribution. In the general case, we examine the virtual waiting time by means of the Laplace–Stieltjes transform; the approach is mainly based on the Pollaczek–Khinchin formula concerning waiting time (8.19) (e.g., [70]).

11.5.1 Priority System with Exponentially Distributed Service Time

Let us consider the following problem. We have m homogeneous servers, and two types of customers. Type i customers arrive to the system according to a Poisson process with parameter $\lambda_i (i = 1, 2)$. If upon the entry of a type 1 customer all servers are occupied, but some servers handle customers of the second type, then a server will change its service and the type 2 customer will be lost. Thus, customers of the second type may be lost not only if a type 2 customer arrives and all servers are occupied, but if customers of the first type show up as well. First type customers are refused only when there are customers of the same type.

The service times of type 1 and type 2 customers are exponentially distributed with parameters μ_1 and μ_2, respectively.

It is quite clear that the service of type 1 customers is denied if all servers were busy and there were no type 2 customers. Thus, the probability of loss of type 1 customers clearly equals

$$p_v = \frac{\frac{\rho_1^m}{m!}}{\sum\limits_{i=0}^{m} \frac{\rho_1^i}{i!}}, \qquad \rho_1 = \frac{\lambda_1}{\mu_1}.$$

Let $p_{ij}(t)$ be the probability of the event that at moment t there are i type 1 and j type 2 customers being served $(0 \le i + j \le m)$. Furthermore, let

$$p_{i\cdot}(t) = \sum_{j=0}^{m-i} p_{ij}(t) \qquad \text{and} \qquad p_{\cdot j}(t) = \sum_{i=0}^{m-j} p_{ij}(t).$$

The sum $\sum_{i+j=m} p_{ij}(t)$ is the probability of loss of a type 2 customer at moment t. The probability of loss of a type 2 customer during its service is

$$\sum_{i+j=m} p_{ij}(t) - p_{m0}(t).$$

11.5.2 Probabilities $p_{ij}(t)$

The differential equations determining $p_{ij}(t)$ are

$$p'_{00}(t) = -(\lambda_1 + \lambda_2)p_{00}(t) + \mu_1 p_{10}(t) + \mu_2 p_{01}(t); \tag{11.6}$$

if $1 \le i < m$, then

$$p'_{i0}(t) = -(\lambda_1+\lambda_2 + i\mu_1)p_{i0}(t)+\lambda_1 p_{i-1,0}(t) + (i + 1)\mu_1 p_{i+1,0}(t) + \mu_2 p_{i1}(t), \tag{11.7}$$

$$p'_{m0}(t) = -m\mu_1 p_{m0}(t) + \lambda_1[p_{m-1,0}(t) + p_{m-1,1}(t)]; \tag{11.8}$$

in the case of $1 \le j < m$,

$$p'_{0j}(t) = -(\lambda_1+\lambda_2+j\mu_2)p_{0j}(t)+\lambda_2 p_{0,j-1}(t)+\mu_1 p_{1j}(t)+(j + 1)\mu_2 p_{0,j+1}(t), \tag{11.9}$$

$$p'_{0m}(t) = -(\lambda_1 + m\mu_2)p_{0m}(t) + \lambda_2 p_{0,m-1}(t); \tag{11.10}$$

in the case of $i \ge 1$, $j \ge 1$, $i + j < m$,

$$\begin{aligned} p'_{ij}(t) = &-(\lambda_1 + \lambda_2 + i\mu_1 + j\mu_2)p_{ij}(t) + \lambda_1 p_{i-1,j}(t) \\ &+\lambda_2 p_{i,j-1}(t) + (i + 1)\mu_1 p_{i+1,j}(t) + \mu_2 p_{i,j+1}(t); \end{aligned} \tag{11.11}$$

in the case of $i > 0$, $j > 0$, $i + j = m$, $i \ne m$, $j \ne m$,

$$p'_{ij}(t) = -(\lambda_1 + i\mu_1 + j\mu_2)p_{ij}(t) + \lambda_1[p_{i-1,j}(t) + p_{i-1,j+1}(t)] + \lambda_2 p_{i-1,j}(t). \tag{11.12}$$

Summing up Eqs. (11.6), (11.9), and (11.12) by j from 0 to m we obtain

$$p'_{0\cdot}(t) = -\lambda_1 p_{0\cdot}(t) + \mu_1 p_{1\cdot}(t). \tag{11.13}$$

Summing up Eqs. (11.7), (11.11), and (11.12) by j from 0 to m, in the case $1 \leq i < m$,

$$p'_{i.}(t) = -(\lambda_1 + i\mu_1)p_{i.}(t) + \lambda_1 p_{i-1,.}(t) + (i+1)\mu_1 p_{i+1,.}(t). \tag{11.14}$$

Equation (11.8) may be rewritten in the form

$$p'_{m.}(t) = -m\mu_1 p_{m.}(t) + \lambda_1 p_{m-1,.}(t). \tag{11.15}$$

The summation of Eqs. (11.6)–(11.8) by i leads to

$$p'_{.0}(t) = -\lambda_2[p_{.0}(t) - p_{m0}(t)] + \lambda_1 p_{m-1,1}(t) + \mu_2 p_{.1}(t).$$

Summing up Eqs. (11.9), (11.11), and (11.12) by i at $1 \leq j < m$:

$$\begin{aligned} p'_{.j}(t) = {} & -(\lambda_2 + j\mu_2)[p_{.j}(t) - p_{m-j,j}(t)] + \lambda_2[p_{.,j-1}(t) - p_{m-j+1,j-1}(t)] \\ & + (j+1)\mu_2 p_{.,j+1}(t) - j\mu_2 p_{m-j,j}(t) - \lambda_1 p_{m-j,j}(t) + \lambda_1 p_{m-j-1,j+1}(t). \end{aligned}$$

Equation (11.10) may be written in the form

$$p'_{.m}(t) = -(\lambda_1 + m\mu_2)p_{.m}(t) + \lambda_2[p_{.m}(t) - p_{1,m-1}(t)].$$

From these equations one can see that for type 2 customers the situation is more complicated; type 1 customers play an essential role in the service process.

Let us consider the case $m = 1$. Then Eqs. (11.6)–(11.12) lead to the equations

$$\begin{aligned} p'_{00}(t) &= -(\lambda_1 + \lambda_2)p_{00}(t) + \mu_1 p_{10}(t) + \mu_2 p_{01}(t), \\ p'_{10}(t) &= -\mu_1 p_{10}(t) + \lambda_1[p_{00}(t) + p_{01}(t)], \\ p'_{01}(t) &= -(\lambda_1 + \mu_2)p_{01}(t) + \lambda_2 p_{00}(t). \end{aligned}$$

This system may be solved easily; the initial conditions are

$$p_{00}(0) = 1, \qquad p_{10}(0) = 0, \qquad p_{01}(0) = 0.$$

We have

$$\begin{aligned} p_{10}(t) = {} & \frac{\lambda_1}{\lambda_1 + \mu_1}\left(1 - \mathrm{e}^{-(\lambda_1+\mu_1)t}\right), \\ p_{01}(t) = {} & \frac{\lambda_2\mu_1}{(\lambda_1 + \mu_1)(\lambda_1 + \lambda_2 + \mu_2)} + \frac{\lambda_1\lambda_2}{(\lambda_1 + \mu_1)(\lambda_2 + \mu_2 - \mu_1)}\mathrm{e}^{-(\mu_1+\mu_2)t} \\ & - \left(\frac{\lambda_2\mu_1}{(\lambda_1 + \mu_1)(\lambda_1 + \lambda_2 + \mu_2)} + \frac{\lambda_1\lambda_2}{(\lambda_1 + \mu_1)(\lambda_2 + \mu_2 - \mu_1)}\right)\mathrm{e}^{-(\lambda_1+\lambda_2+\mu_2)t}, \end{aligned}$$

$$p_{00}(t) = \frac{\mu_1(\lambda_1+\mu_2)}{(\lambda_1+\mu_1)(\lambda_1+\lambda_2+\mu_2)} + \frac{\lambda_1(\mu_2-\mu_1)}{(\lambda_1+\mu_1)(\lambda_2+\mu_2-\mu_1)}\mathrm{e}^{-(\lambda_1+\mu_1)t}$$
$$+\frac{\lambda_2}{\lambda_1+\mu_1}\left(\frac{\mu_1}{\lambda_1+\lambda_2+\mu_2}+\frac{\lambda_1}{\lambda_2+\mu_2-\mu_1}\right)\mathrm{e}^{-(\lambda_1+\lambda_2+\mu_2)t}.$$

Consequently,

$$p_{0.}(t) = p_{00}(t) + p_{01}(t) = \frac{\mu_1}{\lambda_1+\mu_1} + \frac{\lambda_1}{\lambda_1+\mu_1}\mathrm{e}^{-(\lambda_1+\mu_1)t},$$
$$p_{1.}(t) = p_{10}(t) = \frac{\lambda_1}{\lambda_1+\mu_1}\left(1-\mathrm{e}^{-(\lambda_1+\mu_1)t}\right),$$
$$p_{.0}(t) = \frac{\lambda_1(\lambda_1+\lambda_2+\mu_2)+\mu_1(\lambda_1+\mu_2)}{(\lambda_1+\mu_1)(\lambda_1+\lambda_2+\mu_2)} - \frac{\lambda_1\lambda_2}{(\lambda_1+\mu_1)(\lambda_2+\mu_2-\mu_1)}\mathrm{e}^{-(\lambda_1+\mu_1)t}$$
$$+\frac{\lambda_2}{\lambda_1+\mu_1}\left(\frac{\mu_1}{\lambda_1+\lambda_2+\mu_2}+\frac{\lambda_1}{\lambda_2+\mu_2-\mu_1}\right)\mathrm{e}^{-(\lambda_1+\lambda_2+\mu_2)t},$$
$$p_{.1}(t) = p_{01}(t) = \frac{\lambda_2\mu_1}{(\lambda_1+\mu_1)(\lambda_1+\lambda_2+\mu_2)} + \frac{\lambda_1\lambda_2}{(\lambda_1+\mu_1)(\lambda_2+\mu_2-\mu_1)}\mathrm{e}^{-(\lambda_1+\mu_1)t}$$
$$-\frac{\lambda_2}{\lambda_1+\mu_1}\left(\frac{\mu_1}{\lambda_1+\lambda_2+\mu_2}+\frac{\lambda_1}{\lambda_2+\mu_2-\mu_1}\right)\mathrm{e}^{-(\lambda_1+\lambda_2+\mu_2)t}.$$

The stationary probabilities at $t \to \infty$ are

$$\begin{aligned}
p_{00} &= \mu_1(\lambda_1+\mu_2)/\left((\lambda_1+\mu_1)(\lambda_1+\lambda_2+\mu_2)\right),\\
p_{01} &= p_{.1} = \lambda_2\mu_1/\left((\lambda_1+\mu_1)(\lambda_1+\lambda_2+\mu_2)\right),\\
p_{10} &= p_{1.} = \lambda_1/(\lambda_1+\mu_1),\\
p_{0.} &= \mu_1/(\lambda_1+\mu_1),\\
p_{.0} &= \frac{\lambda_1(\lambda_1+\lambda_2+\mu_2)+\mu_1(\lambda_1+\mu_2)}{(\lambda_1+\mu_1)(\lambda_1+\lambda_2+\mu_2)}.
\end{aligned}$$

It is interesting to note the probability of the event that a service in process at a given moment will not be interrupted. This happens if during the service no type 1 customers enter, namely,

$$\int_0^\infty \mathrm{e}^{-\lambda_1 x}\mu_2\mathrm{e}^{-\mu_2 x}\,\mathrm{d}x = \frac{\mu_2}{\lambda_1+\mu_2},$$

then service will be interrupted with probability $\lambda_1/(\lambda_1+\mu_2)$.

11.5.3 Priority System with General Service Time

Now we come to priority systems with generally distributed service time. We will consider three cases:

1. If a type 1 customer enters, then the service of a type 2 customer is interrupted and is continued after all type 1 customers have been served. The performed work is taken into account, and the service time is decreased with the work done.
2. The service is realized as above, but when a type 2 customer is served, the performed work will not be taken into account; the service decreases with time spent.
3. When a type 1 customer enters, the actual service is interrupted, and the customer is lost.

In all three cases we assume the entering customers constitute Poisson processes with parameters λ_1 and λ_2, the service times are arbitrarily distributed random variables with distribution functions $B_1(x)$ and $B_2(x)$, respectively. The Laplace–Stieltjes transforms of the service time is

$$b_i^{\sim}(s) = \int\limits_0^{\infty} \mathrm{e}^{-sx}\,\mathrm{d}B_i(x), \qquad i = 1, 2. \tag{11.16}$$

Let us denote the mean values of service times by τ_1 and τ_2, let $V_i(t)$ be the waiting time of a type i customer on the condition that it entered at moment t, and let $\hat{V}_i(t)$ be the time till completion of service. Let

$$F_i(x) = \lim_{t\to\infty} \mathbf{P}\left(V_i(t) < x\right), \qquad i = 1, 2, \ldots,$$

$$\hat{F}_i(x) = \lim_{t\to\infty} \mathbf{P}(\hat{V}_i(t) < x), \qquad i = 1, 2, \ldots,$$

be the distribution of the waiting time and the time till service completion and let their Laplace–Stieltjes transforms according to Eq. (11.16) be $f_i^{\sim}(s)$ and $\hat{f}_i^{\sim}(s)$.

Type 1 customers are served independently of type 2 customers, so by Eq. (8.19) (if the condition $\lambda_1\tau_1 < 1$ is fulfilled)

$$f_1^{\sim}(s) = \frac{1 - \lambda_1\tau_1}{1 - \lambda_1 \frac{1 - b_1^{\sim}(s)}{s}}.$$

The time interval till completion consists of two parts: the waiting time and the service time. They are independent random variables, so for $\hat{V}_1(t)$ we obtain

$$\hat{f}_1^{\sim}(s) = \frac{(1 - \lambda_1\tau_1) b_1^{\sim}(s)}{1 - \frac{\lambda_1}{s}(1 - b_1^{\sim}(s))}.$$

At the service of type 2 customers the service of type 1 customers may be interpreted as the breakdown of a server. Let L be a random variable denoting the time from the beginning of service of a type 2 customer till the beginning of service of the next one and

$$b_L^{\sim}(s) = \int_0^{\infty} \mathrm{e}^{-sx} \, \mathrm{d}\mathbf{P}(L < x).$$

At a fixed moment we have two possibilities: a type 1 customer is absent in the system with probability $1 - \lambda_1\tau_1$ and present with probability $\lambda_1\tau_1$, and in its presence according to the service discipline it is being served. If there are no type 1 customers, then by Eq. (8.19) the Laplace–Stieltjes transform of the remaining service time is

$$\frac{1 - \lambda_2 \mathbf{E}(L)}{1 - \frac{\lambda_2}{s}(1 - b_L^{\sim}(s))}.$$

In the presence of a type 1 customer, we must first finish the serving existing and entering type 1 customers, then serve type 2 customers (taking into account type 1 customers that enter in the meantime). The Laplace–Stieltjes transform of the service time for existing and entering type 1 customers is

$$\frac{1 - b_0^{\sim}(s)}{s \frac{\tau_1}{1-\lambda_1\tau_1}},$$

where $b_0^{\sim}(s)$ is the solution of the functional equation

$$b_0^{\sim}(s) = b_1^{\sim}(s + \lambda_1 - \lambda_1 b_0^{\sim}(s)),$$

i.e., the Laplace–Stieltjes transform of a busy period for type 1 customers. After having served the type 1 customers we come to the previous situation. Thus, the Laplace–Stieltjes transform of the time period till the service of type 2 customers entering at a given moment is

$$\begin{aligned}
&(1 - \lambda_1\tau_1)\frac{1 - \lambda_2 \mathbf{E}(L)}{1 - \frac{\lambda_2}{s}(1 - b_L^{\sim}(s))} + \lambda_1\tau_1 \frac{1 - b_0^{\sim}(s)}{s\frac{\tau_1}{1-\lambda_1\tau_1}} \cdot \frac{1 - \lambda_2 \mathbf{E}(L)}{1 - \frac{\lambda_2}{s}(1 - b_L^{\sim}(s))} \\
&= \frac{(1 - \lambda_1\tau_1)(1 - \lambda_2\mathbf{E}(L))[s + \lambda_1(1 - b_0^{\sim}(s))]}{s - \lambda_2(1 - b_L^{\sim}(s))} = f_2^{\sim}(s). \qquad (11.17)
\end{aligned}$$

In this expression $b_L^{\sim}(s)$ is still unknown, but we will find it for our three models.

1. From the point of view of a type 2 customer we can interpret the system behavior such that the entry of a type 1 customer is a failure and the end of the busy period generated by this type 1 customer as maintenance. Based on this interpretation our model can be considered a system with server breakdowns. Thus,

$$b_L^{\sim}(s) = b_2^{\sim}(s + \lambda_1 - \lambda_1 b_0^{\sim}(s)).$$

2. Let us consider the sequences of independent random variables $\{U_n\}$, $\{H_n\}$ and $\{A_n\}$, which have the following meaning:

   U_i: service time of a type 2 customer [with Laplace–Stieltjes transform $b_2^{\sim}(s)$];
   H_i: length of busy period for type 1 customers [the corresponding Laplace–Stieltjes transform is $b_0^{\sim}(s)$].
   A_i: interarrival time for type 1 customers (exponentially distributed random variable with parameter λ_1).

   If $U_1 \leq A_1$, then $L = U_1$ (during the service of type 2 customers no type 1 customers enter, so the type 2 customer leaves after U_1 time from the beginning of service).

   If $A_1 < U_1, U_2 \leq A_2$, then $L = H_1 + A_1 + U_2$ (during the service of a type 2 customer after A_1 time a type 1 customer enters, and for its and the entering customers' service we try time H_1, then the service of a type 2 customer is realized for U_2 without interruption). Similarly, if $A_1 < U_1, A_2 < U_2, \ldots, A_n < U_n, U_{n+1} \leq A_{n+1}$, then $L = A_1 + H_1 + A_2 + H_2 + \ldots + A_n + H_n + U_{n+1}$.

   By the formula of total probability,

$$\mathbf{P}(L < x) = \sum_{n=0}^{\infty} \mathbf{P}(A_i < U_i, 1 \leq i \leq n; U_{n+1} \leq A_{n+1}; A_1 + H_1 + A_2 + H_2 + \ldots + A_n + H_n + U_{n+1} < x).$$

Since

$$\int_0^\infty \mathrm{e}^{-sx}\, \mathrm{d}_x P\{A_i < x,\ A_i < U_i\} = \lambda_1 \int_0^\infty \mathrm{e}^{-sx}(1 - B_2(x))\mathrm{e}^{-\lambda_1 x}\, \mathrm{d}x = \frac{\lambda_1}{s + \lambda_1}[1 - b_2^{\sim}(s + \lambda_1)]$$

and

$$\int_0^\infty \mathrm{e}^{-sx}\, \mathrm{d}_x P\{U_i < x,\ U_i \leq A_i\} = \int_0^\infty \mathrm{e}^{-(s+\lambda_1)x}\, \mathrm{d}B_2(x) = b_2^{\sim}(s + \lambda_1),$$

we obtain

$$b_L^{\sim}(s) = \sum_{n=0}^{\infty} \left\{ \frac{\lambda_1}{s + \lambda_1}[1 - b_2^{\sim}(s + \lambda_1)] b_0^{\sim}(s) \right\}^n b_2^{\sim}(s + \lambda_1) = \frac{(s + \lambda_1) b_2^{\sim}(s + \lambda_1)}{s + \lambda_1 - \lambda_1[1 - b_2^{\sim}(s + \lambda_1)] b_0^{\sim}(s)}.$$

3. Using the random variables U_i, H_i, A_i we have

$$L = \begin{cases} U_1, & \text{ha } U_1 \leq A_1, \\ A_1 + H_1, & \text{ha } U_1 > A_1. \end{cases}$$

Consequently,

$$b_L^{\sim}(s) = b_2^{\sim}(s + \lambda_1) + \frac{\lambda_1}{s + \lambda_1}[1 - b_2^{\sim}(s + \lambda_1)]b_0^{\sim}(s).$$

Now let us find the functions $\hat{f}_2^{\sim}(s)$. In the first two cases the time from the moment t till the end of service is $U_2(t) + L$. They are independent random variables, so in both cases

$$\hat{f}_2^{\sim}(s) = f_2^{\sim}(s)b_L^{\sim}(s).$$

In the third case we can lose the type 2 customer; this happens if during the service of a type 2 customer a type 1 customer appears, and the probability of loosing the type 2 customer is

$$\mathbf{P}(A_1 < U_1) = 1 - b_2^{\sim}(\lambda_1).$$

Obviously,

$$\hat{U}_2(t) = U_2(t) + \min(A_2, U_1).$$

Since

$$\int\limits_{x=0}^{\infty} \mathrm{e}^{-sx}\, \mathrm{d}\mathbf{P}(\min(A_1, U_1) = x) = b_2^{\sim}(s + \lambda_1) + \frac{\lambda_1}{s + \lambda_1}[1 - b_2^{\sim}(s + \lambda_1)],$$

then

$$\hat{f}_2^{\sim}(s) = f_2^{\sim}(s)\left\{b_2^{\sim}(s + \lambda_1) + \frac{\lambda_1}{s + \lambda_1}[1 - b_2^{\sim}(s + \lambda_1)]\right\}.$$

These formulas are true if the process has an equilibrium distribution. On the basis of Eq. (11.17), this means that the inequalities $\lambda_1\tau_1 < 1$ and $\lambda_2\mathbf{E}(L) < 1$ hold.

- In the first model $\mathbf{E}(L) = \tau_2/(1-\lambda_1\tau_1)$, from which the condition of equilibrium is $\lambda_2\tau_2 < 1 - \lambda_1\tau_1$.
- In the second model $\mathbf{E}(L) = [1 - b_2^{\sim}(\lambda_1)]/\lambda_1(1 - \lambda_1\tau_1)b_2^{\sim}(\lambda_1)$, from which the condition of equilibrium is $\lambda_2[1 - b_2^{\sim}(\lambda_1)] < \lambda_1(1 - \lambda_1\tau_1)b_2^{\sim}(\lambda_1)$.
- In the third model $\mathbf{E}(L) = [1 - b_2^{\sim}(\lambda_1)]/\lambda_1(1 - \lambda_1\tau_1)$, from which the condition of equilibrium is $\lambda_2[1 - b_2^{\sim}(\lambda_1)] < \lambda_1(1 - \lambda_1\tau_1)$.

11.6 Systems with Several Servers and Queues

11.6.1 Multichannel Systems with Waiting and Refusals

Let $(X_n, Y_n),\ n = 1, 2, \ldots$ be a sequence of i.i.d. random vector variables, where $X_1, X_2, \ldots$ are the interarrival periods of successive customers (the nth one enters at the moment $t_n = X_1 + \cdots + X_n,\ n = 1, 2, \ldots$), and Y_n is the service time of nth customer.

Let us consider a $G/G/m$ system. We introduce the *waiting time vector* of the nth customer:

$$W_n = (W_{n,1}, \ldots, W_{n,m}),\ n = 1, 2, \ldots,$$

where $W_{n,i}$ means the random time interval the nth customer (entering at t_n) has to wait till i servers become free from all earlier (with numbers $1, \ldots, n-1$) customers.

If the initial random vector variable W_0 (on the same probability space) is given, then the sequence $W_n,\ n \geq 0$, is uniquely determined and a recurrence relation is valid for W_n, i.e., $\{W_n,\ \ n\ \geq\ 0\}$ is a recurrent process. For the arbitrary $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$ let

$$x^+ = (x_1^+, \ldots, x_m^+), \quad \text{where} \quad s^+ = \max(s, 0),\ s \in \mathbb{R},$$

$$R(x) = (x_{i_1}, \ldots, x_{i_m}), \quad x_{i_1} \leq x_{i_2} \leq \cdots \leq x_{i_m},$$

i.e., the function $R(x)$ arranges the components of vector x in increasing order. We introduce the vectors

$$e = (\overset{(1)}{1},\ \overset{(2)}{0}, \ldots,\ \overset{(m)}{0}), \qquad i = (\overset{(1)}{1}, \ldots,\ \overset{(m)}{1}).$$

Theorem 11.4. *For the sequence $W_n,\ n \geq 0$, the recurrence relation*

$$W_{n+1} = R\left([(W_n + Y_n e) - X_n i]^+\right), \quad n \geq 0. \tag{11.18}$$

holds.

Proof. Using the definition of included quantities, this is trivial. □

In the investigation of queueing systems the existence of a limit distribution for the basic characteristics is an important question. Using results from the theory of recurrence processes one can prove a theorem valid in the more general case where instead of total independence we assume stationarity in a narrower sense and the ergodicity of the process $\{(X_n, Y_n),\ n \geq 1\}$ (see [14]).

Theorem 11.5. *Let* $\{(X_n, Y_n),\ n \geq 1\}$ *be a sequence of i.i.d. random variables and* $\mathbf{E}(Y_1 - mX_1) < 0,\ \ W_0 = 0$*; then there exists a stationary, in a narrow sense, process* $\{W^{(n)},\ n \geq 0\}$ *satisfying Eq. (11.19), and the distribution function of* W_n *monotonically converges to the distribution function of* $W^{(0)}$.

***G/G/m/0* Systems with Refusals** Since we are considering a system with refusals, one can speak of waiting only in a virtual sense. Thus, the component $W_{n,i}$ of W_n means the possible waiting time of customers entering at moment t_n till i servers become free from all earlier (with numbers $1, \ldots, n-1$) customers (if $W_{n,1} > 0$, then the nth one will not be serviced). We can write the recurrence relation

$$W_{n+1} = R\left(\left[(W_n + Y_n e \mathcal{I}_{\{W_{n,1}=0\}}) - X_n i\right]^+\right).$$

The sufficient condition similar to the previous theorem is

$$\mathbf{P}(Y_1 \leq mX_1) > 0, \ \ \mathbf{E}(Y_1) < \infty.$$

If $(X_n, Y_n),\ -\infty < n < \infty$ is not a sequence of independent random variables with the same distribution but a stationary (stationary in a narrower sense), ergodic sequence, even in this case we can give a sufficient condition for the existence of a limit distribution, namely,

$$\mathbf{P}(Y_0 \leq X_0 + \ldots + X_{m-1}, Y_{-1} \leq X_{-1} + X_0 + \ldots + X_{m-2}, \ldots, Y_{-m+1} \\ \leq X_{-m+1} + X_{-m+2} + \ldots + X_0) > 0,$$

$$\mathbf{E}(Y_1) < \infty.$$

If we consider instead of the virtual waiting time the queue length L_n at the arrival moment of the nth customer, then it also has a nondegenerate limiting distribution.

***G/G/∞* system** Now we have an infinite number of servers, so one cannot speak of queueing or losses. In this case the basic characteristic is the queue length: $L_k,\ k \geq 1$, denotes the number of customers at the arrival moment of the kth customer [at an arbitrary moment t the number of customers present $L(t)$ is left continuous]. Actually, it is the number of occupied servers. At the beginning the system is empty, i.e., $L_1 = 0$.

For the sake of simplicity let $X_n,\ n \geq 1$, denote the interarrival time of the nth and $(n+1)$st customers, $Y_n,\ n \geq 1$, the service time of the nth customer. Then

$$L_{k+1} = \mathcal{I}_{\{Y_k > X_k\}} + \mathcal{I}_{\{Y_{k-1} > X_{k-1} + X_k\}} + \cdots + \mathcal{I}_{\{Y_1 > X_1 + \cdots + X_k\}}, \quad k \geq 1.$$

Theorem 11.6. *If* $\{(X_n, Y_n),\ -\infty < n < \infty\}$ *is a sequence of i.i.d. random variables and* $0 < \mathbf{E}(Y_1) < \infty$ *is fulfilled, then*

$$L = \sum_{k\geq 1} \mathcal{I}_{\{Y_{-k} > X_{-k} + \cdots + X_{-1}\}}$$

defines a finite random variable with probability 1, the random variables

$$L_{-n} = \sum_{k=1}^{n} \mathcal{I}_{\{Y_{-k} > X_{-k} + \cdots + X_{-1}\}}, \quad n = 1, 2, \ldots$$

and L_n have the same distribution, and this distribution monotonically converges to the distribution of L.

Proof. For the proof it is enough to show that L is finite with probability 1. We need the following lemma; from it with probability 1 follows the finiteness of L. □

Lemma 11.7. *Let $U_1, U_2, \ldots$ be a sequence of i.i.d. random variables for which $\mathbf{P}(U_1 \geq 0) = 1$ and $h = \mathbf{E}\left(\mathrm{e}^{-U_1}\right) < 1$, i.e., the distribution of U_i is not concentrated at the point 0. Let V be an arbitrary random variable (not necessarily independent of U_i) for which $\mathbf{E}\left(V^+\right) < \infty$. Furthermore, let $\kappa = \frac{1}{2}\log\frac{1}{h}$, $G_V(x) = 1 - \mathbf{P}(V < x)$, $x \in \mathbb{R}$. Then for arbitrary $n, N \geq 1$*

$$\mathbf{P}(U_1 + \cdots + U_n < V) < \mathrm{e}^{-n\kappa} + G_V(n\kappa), \tag{11.19}$$

$$\sum_{n\geq N} \mathbf{P}(U_1 + \cdots + U_n < V) < \frac{1}{1 - \mathrm{e}^{-\kappa}} \mathrm{e}^{-N\kappa} + \frac{1}{\kappa} \mathbf{E}\left(V \mathcal{I}_{\{N\kappa \leq V\}}\right) \tag{11.20}$$

is true.

Proof. For arbitrary $x > 0$

$$\begin{aligned}
&\mathbf{P}(U_1 + \cdots + U_n < V) \\
&\quad = \mathbf{P}(U_1 + \cdots + U_n < V,\ V \leq nx) + \mathbf{P}(U_1 + \cdots + U_n < V,\ nx < V) \\
&\quad \leq \mathbf{P}(U_1 + \cdots + U_n < nx) + \mathbf{P}(nx \leq V).
\end{aligned}$$

Using the Markov inequality we obtain

$$\begin{aligned}
\mathbf{P}(U_1 + \cdots + U_n < nx) &\leq \mathbf{E}(\exp\{nx - (U_1 + \cdots + U_n)\}) \\
&= \mathrm{e}^{nx} \prod_{i=1}^{n} \mathbf{E}\left(\mathrm{e}^{-U_i}\right) \\
&= \mathrm{e}^{n(x + \log h)},
\end{aligned}$$

where at $x = \kappa$ Eq. (11.19) follows.

Proof of Eq. (11.20): From inequality (11.19)

$$\sum_{n\geq N} \mathbf{P}\left(U_1 + \cdots + U_n < V\right) \leq \sum_{n\geq N} \{\mathrm{e}^{-n\kappa} + G_V(n\kappa)\}$$

$$= \frac{1}{1-\mathrm{e}^{-\kappa}} \mathrm{e}^{-N\kappa} + \sum_{j=0}^{\infty} \mathbf{P}\left((N+j)\kappa \leq V\right).$$

Since

$$\mathbf{E}\left(V\mathcal{I}_{\{N\kappa\leq V\}}\right) \geq \sum_{j=0}^{\infty}(N+j)\kappa\mathbf{P}\left((N+j)\kappa \;\leq\; V \;<\; (N+j+1)\kappa\right)$$

$$= (N-1)\kappa \sum_{j=0}^{\infty} \mathbf{P}\left((N+j)\kappa \;\leq V \;<\; (N+j+1)\kappa\right)$$

$$+ \sum_{j=0}^{\infty}(j+1)\kappa\mathbf{P}\left((N+j)\kappa \;\leq\; V \;<\; (N+j+1)\kappa\right)$$

$$= (N-1)\kappa\mathbf{P}\left(N\kappa \leq V\right) + \kappa \sum_{j=0}^{\infty} \mathbf{P}\left((N+j)\kappa \leq V\right),$$

then

$$\sum_{j=0}^{\infty} \mathbf{P}\left((N+j)\kappa \leq V\right) \leq \frac{1}{\kappa}\mathbf{E}\left(V\mathcal{I}_{\{N\kappa\leq V\}}\right) - (N-1)\mathbf{P}\left(N\kappa \leq V\right)$$

$$\leq \frac{1}{\kappa}\mathbf{E}\left(V\mathcal{I}_{\{N\kappa\leq V\}}\right).$$

Using Eq. (11.19) we obtain Eq. (11.20). □

11.6.2 Closed Queueing Network Model of Computers

The queueing network in Fig. 11.27 may be considered the simplest mathematical model for computers.

In a system there are continuously n customers (tasks) and they can move along the routes indicated in the figure. In front of each service unit there is a waiting buffer of corresponding capacity (for at most $n-1$ customers). On the units the service is realized by the FCFS rule; the service times are independent and on the ith unit have distribution function $F_i(x),\ 0 \leq i \leq M$. After having completed a service on the 0th unit, the customer moves to the ith unit with probability $p_i,\ 0 \leq i \leq M\ \ (p_i \geq 0,\ p_0 + \cdots + p_M = 1)$, which does not depend on the state of the system or the service time. If the service of a customer is completed on the ith

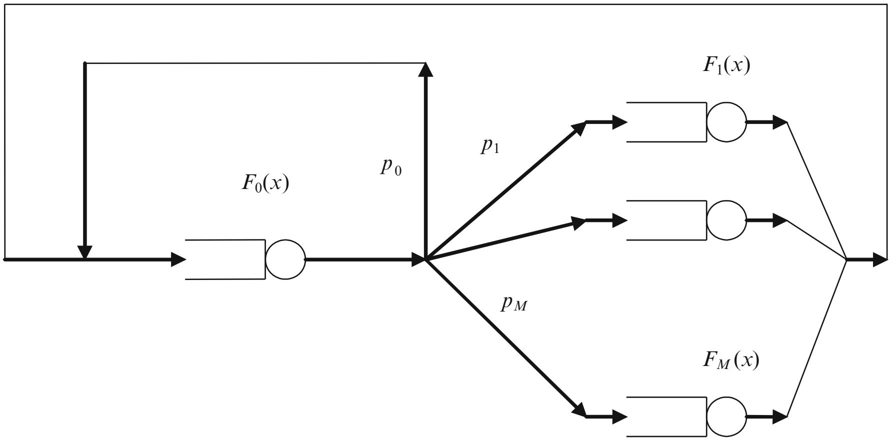


Fig. 11.27 Closed queueing network model

$(1 \le i \le M)$ unit, then the customer goes to a unit with index 0 with probability 1. The unit 0 plays a special role in the network and is called a *central unit*.

It may seem too strict a restriction that the number of customers in a computer is fixed, but this model gives accurate results for several important performance parameters. For example, when we are interested in the maximum performance of a computer, we can assume that the load of the computer is maximal, that is, after completion of a task, a new one enters the system immediately.

If we consider the successive moments when all customers stay at the central unit and the service of a customer has just started, these moments are regeneration points for the network. If the service times have finite pth $(p \ge 1)$ order moments, then one can show the pth moment of a regenerative cycle is also finite [86, 87]. It follows that if the mean values of services are finite, then the different characteristics for the system have limiting distributions.

11.7 Retrial Systems

11.7.1 Continuous-Time Retrial System

If in the case of a phone call the subscriber is occupied, one usually repeats the attempt while the conversation is realized. So the system has two types of requests: primary calls and calls generated by occupied lines. Models constructed for systems with losses do not describe this situation, and they do not take into account repetitions. These problems appeared in Erlang's time, but due to a lack of corresponding theoretical results, these repetitions were considered new arrivals.

Retrial queues constitute a special field of queueing systems; their distinguishing feature is that in the case of a busy server, entering customers leave the service area (go to the orbit) and, after a certain (generally random) time, reinitiate their service.

Let us consider some examples of the retrial phenomenon. The first example is connected with the functioning of call centers used by companies to communicate with customers. When a call arrives, it is sent to a call distribution switch. If all agents are busy, then the call center may announce an estimated waiting time. Some customers might decide to wait for a free agent, while some will interrupt the connection immediately or after some time. A portion of these customers will return after some random time.

Random access protocols provide a motivation for the design of communication protocols with retransmission control. If two or more stations transmit packets at the same time, then a collision takes place. Then all packets are destroyed and should be retransmitted. To avoid collisions in the next period, this transmission is realized with a certain random delay for each station. This fact motivates the investigation of the retrial feature of computer networks.

Two textbooks have been published in this field. The book by Falin and Templeton [30] gives analytical solutions in terms of generating functions and Laplace-Stieltjes transforms, and the one by Artalejo and Gómez-Corral [6] focuses on the application of algorithmic methods studying the M/G/1 and M/M/c retrial queues and using matrix-analytic techniques to solve some retrial queues with QBD, GI/M/1, and M/G/1 structures.

We will consider a model connected with the landing process of airplanes in the case of continuous time. The model was introduced in [59], and the results for waiting time are contained in [61].

Let us consider the landing process of airplanes. An airplane appears at an airport ideally positioned for landing. If it is not possible (due to insufficient distance or a waiting queue), it starts circling. The next request for service is possible when it returns to the starting geometrical point on the condition that there are no other airplanes ahead of it.

Similar problems appear at the transmission of optical signals. Signals entering the node must be sent in the order of arrival, but they cannot be stored. They go to delay lines and upon their return can reinitiate their transmission. If all previous signals have already been sent, then the signal is transmitted; in the opposite case they pass through the delay queue again, and the process is repeated.

Let us formulate the queueing problem. We investigate a service system where the service may start at the moment of arrival (if the system is available) or at moments differing from it by multiples of a given cycle time T (in the case of busy server or queue). Service of a customer can be started if all customers who had entered the system earlier have already left (i.e., the FIFO rule works). In such a system the service process is not continuous; during the "busy period" there are idle intervals; these idle intervals are necessary to reach the starting point; for them there is no real service.

Let the service of the nth customer begin at moment t_n, and let us consider the number of customers present at the moment just before service begins. Then the number of customers present is determined by the recursive formula

$$N_{n+1} = \begin{cases} \Delta_n, & \text{if } N_n = 0, \\ N_n - 1 + \Delta_n, & \text{if } N_n > 0, \end{cases}$$

where Δ_n is the number of customers appearing for $[t_n, t_{n+1})$. We show that these values constitute a Markov chain.

Let us consider the time intervals during which we record the entering customers. Let $\{Z_i\}$ and $\{Y_i\}$ $(i = 1, 2, \ldots)$ be two independent sequences of independent random variables. Z_i means the interarrival time between the ith and $i + 1$th customers (it has an exponential distribution with the parameter λ), Y_i is the service time of the ith customer (in our case it has an exponential distribution with parameter μ).

Let us assume that at the beginning of service there is one customer in the system. If $Z_i \geq Y_i$, then the time till the beginning of service of the next customer is Z_i (the service of the existing customer will be completed, and the server arrives at a free state and the next customer appears later). If $Z_i < Y_i$, then during the service of the first customer a second one appears, and after this moment there will be intervals with length T while we pass the moment of service of the first customer (from the viewpoint of entering customers we are interested in the time from the entry of the second customer till the beginning of its service). The length of this interval is the function of random variables Z_i and Y_i, i.e., a certain $f_1(Z_i, Y_i)$.

If at the beginning of service of the first customer the second one is already present, then the time till the starting moment of its service is determined in the following way. Divide the service time of the first customer into intervals of length T (the last period generally is not full). Since the starting moments for both customers differ by multiples of T from the moments of arrivals, on each interval of length T there is one point where the service of the second customer may start. In reality, this happens at the first moment after the service of the first customer has completed, so the required time period is determined by the service time of the first customer and the interarrival time. Consequently, it will be a certain function of Y_i and Z_i, i.e., $f_2(Y_i, Z_i)$.

Thus, the time intervals for which we consider the number of entering customers are only functions of random variables Y_i and Z_i, consequently they are independent. Taking into account the fact that entering customers form a Poisson process, the quantities Δ_i of these customers are independent random variables, and N_n is a Markov chain.

To describe the functioning of the system we use the embedded Markov chain technique. Our result is formulated in the following theorem.

Theorem 11.8. *Let us consider a service system in which the entering customers form a Poisson process with parameter λ, and the service time is exponentially distributed with parameter μ. The service of a customer may be started at the moment of arrival (in the case of a free system) or at moments differing from it by the multiples of a cycle time T (in the case of a busy server or queue); the service* *discipline is FIFO. Let us define a Markov chain whose states correspond to the number of customers in the system at moments $t_k - 0$ (t_k is the starting moment of service of the kth customer). The matrix of transition probabilities of this Markov chain has the form*

$$\begin{bmatrix} a_0 & a_1 & a_2 & a_3 & \dots \\ a_0 & a_1 & a_2 & a_3 & \dots \\ 0 & b_0 & b_1 & b_2 & \dots \\ 0 & 0 & b_0 & b_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{11.21}$$

and its elements are determined by the generating functions

$$A(z) = \sum_{i=0}^{\infty} a_i z^i = \frac{\mu}{\lambda + \mu} + \frac{\lambda}{\lambda + \mu} z \frac{(1 - \mathrm{e}^{-\mu T})\mathrm{e}^{-\lambda(1-z)T}}{1 - \mathrm{e}^{-[\lambda(1-z)+\mu]T}}, \tag{11.22}$$

$$\begin{aligned} B(z) &= \sum_{i=0}^{\infty} b_i z^i \\ &= \frac{1}{(1 - \mathrm{e}^{-\lambda T})(1 - \mathrm{e}^{-[\lambda(1-z)+\mu]T})} \\ &\times \left[\frac{1}{2 - z} \left(1 - \mathrm{e}^{-\lambda(2-z)T}\right) \left(1 - \mathrm{e}^{-[\lambda(1-z)+\mu]T}\right) \right. \\ &\left. - \frac{\lambda}{\lambda(2 - z) + \mu} \left(1 - \mathrm{e}^{-[\lambda(2-z)+\mu]T}\right) \left(1 - \mathrm{e}^{-\lambda(1-z)T}\right) \right]. \end{aligned} \tag{11.23}$$

The generating function of the ergodic distribution of this chain is

$$P(z) = p_0 \frac{B(z)(\lambda z + \mu) - zA(z)(\lambda + \mu)}{\mu[B(z) - z]}, \tag{11.24}$$

where

$$p_0 = 1 - \frac{\lambda}{\lambda + \mu} \frac{1 - \mathrm{e}^{-(\lambda+\mu)T}}{\mathrm{e}^{-\lambda T}(1 - \mathrm{e}^{-\mu T})}. \tag{11.25}$$

The ergodicity condition is

$$\frac{\lambda}{\mu} < \frac{\mathrm{e}^{-\lambda T}(1 - \mathrm{e}^{-\mu T})}{1 - \mathrm{e}^{-\lambda T}}. \tag{11.26}$$

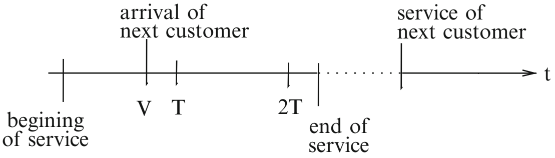


Fig. 11.28 One customer at the beginning of service

Proof. Our original system, where during the busy period there are possible idle intervals, too, is replaced by another one. In it the service process is continuous, and the service time of a customer consists of two parts: the first part is the real service, the second part holds from the end of service till the second one gets to the corresponding position.

For a description of the operation we use an embedded Markov chain; its states are the number of customers in the system at moments $t_k - 0$, i.e., we consider it at moments just before starting service. Let us find the transition probabilities for this chain. We have to distinguish two cases: at the starting moment of service the next customer is present or not. First we will consider the second possibility (Fig. 11.28), which happens for the states 0 and 1. Suppose that the service time of the first customer is U, the second customer enters at V time after the beginning of service. The probability of event $\{U - V < t\}$ is

$$\begin{aligned} P(t) &= \mathbf{P}(U - V < t) \\ &= \int_0^t \int_0^U \lambda \mathrm{e}^{-\lambda V} \mu \mathrm{e}^{-\mu U} \mathrm{d}V \mathrm{d}U + \int_t^\infty \int_{U-t}^U \lambda \mathrm{e}^{-\lambda V} \mu \mathrm{e}^{-\mu U} \mathrm{d}V \mathrm{d}U \\ &= \frac{\lambda}{\lambda + \mu} \left(1 - \mathrm{e}^{-\mu t}\right). \end{aligned} \tag{11.27}$$

The time from the entry of the second customer till the beginning of its service equals

$$(I(U - V) + 1)\, T,$$

where $I(x)$ denotes an integer part of number x/T. This expression is valid for all points excluding multiples of T, but the probability of an event for this time to equal a multiple of T is equal to zero. To determine the transition probabilities, we need the number of customers entering during this period. According to Eq. (11.27) the time from the entry till the beginning of service is equal to iT with probability

$$\frac{\lambda}{\lambda + \mu} \left(\mathrm{e}^{-\mu(i-1)T} - \mathrm{e}^{-\mu i T}\right),$$

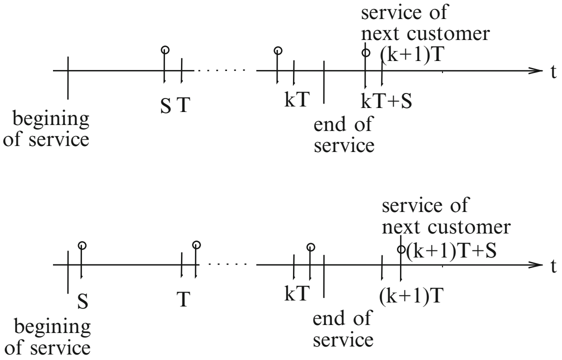


Fig. 11.29 More than one customer at the beginning of service

and the generating function of entering customers equals

$$\frac{\lambda}{\lambda+\mu}\sum_{k=0}^{\infty}\sum_{i=1}^{\infty}\left(\mathrm{e}^{-\mu(i-1)T}-\mathrm{e}^{-\mu iT}\right)\frac{(\lambda iTz)^k}{k!}\mathrm{e}^{-\lambda iT}$$

$$=\frac{\lambda}{\lambda+\mu}\sum_{i=1}^{\infty}\left(\mathrm{e}^{-\mu(i-1)T}-\mathrm{e}^{-\mu iT}\right)\mathrm{e}^{-\lambda iT(1-z)}=\frac{\lambda}{\lambda+\mu}\frac{\mathrm{e}^{-\lambda(1-z)T}(1-\mathrm{e}^{-\mu T})}{1-\mathrm{e}^{-[\lambda(1-z)+\mu]T}}.$$

This formula is valid if for U at least one customer enters the system, so the desired generating function is

$$A(z)=\frac{\mu}{\lambda+\mu}+\frac{\lambda}{\lambda+\mu}z\frac{(1-\mathrm{e}^{-\mu T})\mathrm{e}^{-\lambda T(1-z)}}{1-\mathrm{e}^{-[\lambda(1-z)+\mu]T}},$$

where $\frac{\mu}{\lambda+\mu}=\int_0^{\infty}\mathrm{e}^{-\lambda x}\mu\mathrm{e}^{-\mu x}\mathrm{d}x$ is the probability that for the service time no customer appears.

Now we find the transition probabilities for all other states. In this case at the beginning of service the next customer is already present (Fig. 11.29). Let $R=U-I(U)T$ and let S be the *mod T* interarrival time. One can easily see that S has a truncated exponential distribution with distribution function $\frac{1-\mathrm{e}^{-\lambda S}}{1-\mathrm{e}^{-\lambda T}}$. The time between the starting moments of two successive customers is

$$I\,(U)\,T+S\quad\text{if}\quad R\le S\qquad\text{and}\qquad(I\,(U)+1)\,T+S\quad\text{if}\quad R>S.$$

k customers enter in the two cases with probabilities

$$\frac{(\lambda\,\{I\,(U)\,T+S\})^k}{k}!\exp\left(-\lambda\,\{I\,(U)\,T+S\}\right)\tag{11.28}$$

and

$$\frac{\left(\lambda\left\{\left[I\left(U\right)+1\right]T+S\right\}\right)^{k}}{k!}\exp\left(-\lambda\left\{\left[I\left(U\right)+1\right]T+S\right\}\right). \qquad (11.29)$$

Let us fix S and divide the service time of the customer into intervals of length T. S divides each such interval into two parts (the first has length S, the second $T-S$), and the corresponding probability for the first part is Eq. (11.28), for the second part Eq. (11.29). Let $I(U)=i$. The generating function of the number of entering customers, denoted by N, assuming that the interarrival time *mod T* is equal to S is as follows

$$\begin{aligned}\mathbf{E}\left(z^{N}|S\right) &= \sum_{k=0}^{\infty}\sum_{i=0}^{\infty}\left(\int\limits_{iT}^{iT+S}\frac{[\lambda(iT+S)z]^{k}}{k}!\mathrm{e}^{-\lambda(iT+S)}\mu\mathrm{e}^{-\mu U}\mathrm{d}U\right.\\ &\quad\left.+\int\limits_{iT+S}^{(i+1)T}\frac{\left[\lambda\left((i+1)T+S\right)z\right]^{k}}{k!}\mathrm{e}^{-\lambda((i+1)T+S)}\mu\mathrm{e}^{-\mu U}\mathrm{d}U\right)\\ &= \frac{1}{1-\mathrm{e}^{-[\lambda(1-z)+\mu]T}}\left(\mathrm{e}^{-\lambda(1-z)S}-\mathrm{e}^{-[\lambda(1-z)+\mu]S}\right.\\ &\quad\left.+\mathrm{e}^{-\lambda(1-z)T}\mathrm{e}^{-[\lambda(1-z)+\mu]S}-\mathrm{e}^{-\lambda(1-z)S}\mathrm{e}^{-[\lambda(1-z)+\mu]T}\right),\end{aligned}$$

Multiplying this expression by $\frac{\lambda\mathrm{e}^{-\lambda S}}{1-\mathrm{e}^{-\lambda T}}$ and integrating by S from 0 to T we obtain the generating function of transition probabilities

$$\begin{aligned}B(z) &= \sum_{i=0}^{\infty}b_{i}z^{i}\\ &= \frac{1}{(1-\mathrm{e}^{-\lambda T})(1-\mathrm{e}^{-[\lambda(1-z)+\mu]T})}\\ &\quad\times\left[\frac{1}{2-z}\left(1-\mathrm{e}^{-\lambda(2-z)T}\right)\left(1-\mathrm{e}^{-[\lambda(1-z)+\mu]T}\right)\right.\\ &\quad\left.-\frac{\lambda}{\lambda(2-z)+\mu}\left(1-\mathrm{e}^{-[\lambda(2-z)+\mu]T}\right)\left(1-\mathrm{e}^{-\lambda(1-z)T}\right)\right].\end{aligned}$$

Consider a Markov chain describing the functioning of the system; the matrix of transition probabilities has the form Eq. (11.21). Let us denote the ergodic distribution by p_i $(i=0,1,2,\ldots)$ and introduce the generating function $P(z)=\sum\limits_{i=0}^{\infty}p_{i}z^{i}$. Then

$$p_j = p_0 a_j + p_1 a_j + \sum_{i=2}^{j+1} p_i b_{j-i+1},$$

whence

$$\begin{aligned}\sum_{j=0}^{\infty} p_j z^j &= p_0 A(z) + p_1 A(z) + \sum_{j=0}^{\infty}\sum_{i=2}^{j+1} p_i b_{j-i+1} z^j \\ &= \frac{1}{z} P(z) B(z) - \frac{1}{z} p_0 B(z) + p_0 A(z) + p_1 A(z) - p_1 B(z),\end{aligned}$$

i.e.,

$$P(z) = \frac{p_0\,[zA(z) - B(z)] + p_1 z\,[A(z) - B(z)]}{z - B(z)}.$$

This expression contains two unknown probabilities – p_0 and p_1 – but

$$p_0 = p_0 a_0 + p_1 a_0,$$

i.e.,

$$p_1 = \frac{1 - a_0}{a_0} p_0 = \frac{\lambda}{\mu} p_0.$$

p_0 can be found from the condition $P(1) = 1$,

$$p_0 = \frac{1 - B'(1)}{1 + A'(1) - B'(1) + \frac{\lambda}{\mu}[A' - B'(1)]}.$$

The embedded chain is irreducible, so $p_0 > 0$. Using

$$\begin{aligned}A'(1) &= \frac{\lambda}{\lambda + \mu}\left(1 + \frac{\lambda T}{1 - \mathrm{e}^{-\mu T}}\right), \\ B'(1) &= 1 - \frac{\lambda T \mathrm{e}^{-\lambda T}}{1 - \mathrm{e}^{-\lambda T}} + \frac{\lambda}{\lambda + \mu} \lambda T \frac{1 - \mathrm{e}^{-(\lambda+\mu)T}}{(1 - \mathrm{e}^{-\lambda T})(1 - \mathrm{e}^{-\mu T})},\end{aligned}$$

we obtain

$$\left(1 + \frac{\lambda}{\mu}\right) A'(1) - \frac{\lambda}{\mu} B'(1) = \frac{\lambda}{\lambda + \mu} \lambda T \frac{1 - \mathrm{e}^{-(\lambda+\mu)T}}{(1 - \mathrm{e}^{-\lambda T})(1 - \mathrm{e}^{-\mu T})} > 0,$$

so the condition $1 - B'(1) > 0$ must be fulfilled. This leads to the inequality

$$\frac{\lambda T \mathrm{e}^{-\lambda T}}{1 - \mathrm{e}^{-\lambda T}} - \frac{\lambda}{\lambda + \mu} \lambda T \frac{1 - \mathrm{e}^{-(\lambda+\mu)T}}{(1 - \mathrm{e}^{-\lambda T})(1 - \mathrm{e}^{-\mu T})} > 0,$$

i.e.,

$$\frac{\lambda}{\lambda+\mu} < \frac{\mathrm{e}^{-\lambda T}(1-\mathrm{e}^{-\mu T})}{1-\mathrm{e}^{-(\lambda+\mu)T}}.$$

This is equivalent to Eq. (11.26). Substituting the corresponding values we obtain

$$p_0 = 1 - \frac{\lambda}{\lambda+\mu}\,\frac{1-\mathrm{e}^{-(\lambda+\mu)T}}{\mathrm{e}^{-\lambda T}(1-\mathrm{e}^{-\mu T})}.$$

The theorem is proved. □

During the busy period there are idle intervals, which are necessary to get to the starting position, and they alternate between 0 and T. It is clear that if T decreases, then their influence will become increasingly attenuated. In the limit case, the service process becomes continuous, and after having served a customer, we immediately change to the next one.

Theorem 11.9. *The limiting distribution for the system described above as $T \to 0$ is*

$$P^*(z) = \frac{1-\rho}{1-\rho z} \qquad \left(\rho = \frac{\lambda}{\mu}\right),$$

i.e., it is geometrical with parameter ρ.

Proof. We find p_0, $A(z)$ and $B(z)$ as $T \to 0$, and the limiting values are denoted by p_0^*, $A^*(z)$, and $B^*(z)$. On the basis of Eqs. (11.25), (11.22), and (11.23),

$$p_0^* = \lim_{T\to 0} p_0 = \lim_{T\to 0}\left(1 - \frac{\lambda}{\lambda+\mu}\,\frac{1-\mathrm{e}^{-(\lambda+\mu)T}}{\mathrm{e}^{-\lambda T}(1-\mathrm{e}^{-\mu T})}\right) = 1 - \frac{\lambda}{\mu} = 1-\rho,$$

$$\begin{aligned} A^*(z) = \lim_{T\to 0} A(z) &= \lim_{T\to 0}\left(\frac{\mu}{\lambda+\mu} + \frac{\lambda z}{\lambda+\mu}\,\frac{\mathrm{e}^{-\lambda(1-z)T}(1-\mathrm{e}^{-\mu T})}{1-\mathrm{e}^{-[\lambda(1-z)+\mu]T}}\right) \\ &= \frac{\mu}{\lambda(1-z)+\mu}, \end{aligned}$$

$$\begin{aligned} B^*(z) = \lim_{T\to 0} B(z) = {} & \frac{1}{(1-\mathrm{e}^{-\lambda T})[1-\mathrm{e}^{-[\lambda(1-z)+\mu]T}]} \\ & \times \left\{ \frac{1}{2-z}\left(1-\mathrm{e}^{-\lambda(2-z)T}\right)\left(1-\mathrm{e}^{-[\lambda(1-z)+\mu]T}\right) \right. \\ & \left. - \frac{\lambda}{\lambda(2-z)+\mu}\left(1-\mathrm{e}^{-[\lambda(2-z)+\mu]T}\right)\left(1-\mathrm{e}^{-\lambda(1-z)T}\right) \right\} \\ = {} & \frac{\mu}{\lambda(1-z)+\mu}. \end{aligned}$$

Using these values

$$P^*(z) = (1-\rho)\frac{(\lambda z+\mu)\frac{\mu}{\lambda(1-z)+\mu} - z(\lambda+\mu)\frac{\mu}{\lambda(1-z)+\mu}}{\mu\left(\frac{\mu}{\lambda(1-z)+\mu} - z\right)} = \frac{1-\rho}{1-\rho z}.$$

The preceding formula is the generating function for an M/M/1 system and coincides with the previous results. □

11.7.2 Waiting Time for Continuous Retrial System

Let us consider the previously described system. Using Koba's results [57] we determine the distribution of the waiting time. Let t_n denote the moment of arrival of the nth customer; then its service may be started at the moment $t_n + T \cdot X_n$, where X_n is a nonnegative integer. Let $Z_n = t_{n+1} - t_n$, and let Y_n be the service time of the nth customer. If $X_n = i$, then between X_n and X_{n+1} the following relation holds. If

$$(k-1)T < iT + Y_n - Z_n \le kT \qquad (k \ge 1),$$

then $X_{n+1} = k$. In this case X_n is a homogeneous Markov chain with transition probabilities p_{ik}, where

$$p_{ik} = \mathbf{P}\left((k-i-1)T < Y_n - Z_n \le (k-i)T\right)$$

if $k \ge 1$, and

$$p_{i0} = \mathbf{P}\left(Y_n - Z_n \le -iT\right).$$

Introduce the notations

$$f_j = \mathbf{P}\left((j-1)T < Y_n - Z_n \le jT\right), \tag{11.30}$$

$$p_{ik} = f_{k-i} \quad \text{ha} \quad k \ge 1, \quad p_{i0} = \sum_{j=-\infty}^{-i} f_j = \hat{f}_i. \tag{11.31}$$

The ergodic distribution of the Markov chain satisfies the system of equations

$$p_j = \sum_{i=0}^{\infty} p_i p_{ij} \quad (j \ge 0), \qquad \sum_{j=0}^{\infty} p_j = 1.$$

Theorem 11.10. *Let us consider the system described in Theorem 11.8. Define a Markov chain whose states correspond to the waiting times of customers at moments of arrivals. The matrix of transition probabilities has the form*

$$\begin{bmatrix} \sum\limits_{j=-\infty}^{0} f_j & f_1 & f_2 & f_3 & f_4 & \dots \\ \sum\limits_{j=-\infty}^{-1} f_j & f_0 & f_1 & f_2 & f_3 & \dots \\ \sum\limits_{j=-\infty}^{-2} f_j & f_{-1} & f_0 & f_1 & f_2 & \dots \\ \sum\limits_{j=-\infty}^{-3} f_j & f_{-2} & f_{-1} & f_0 & f_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{11.32}$$

and its elements are determined by formulas (11.30)–(11.31). Then the generating function of the waiting time is

$$P(z) = \left[1 - \frac{\lambda}{\mu}\frac{1-\mathrm{e}^{-\lambda T}}{\mathrm{e}^{-\lambda T}(1-\mathrm{e}^{-\mu T})}\right] \times \frac{\dfrac{\mu}{\lambda+\mu} - \dfrac{\mu(1-\mathrm{e}^{-\lambda T})}{\lambda+\mu}\dfrac{z}{z-\mathrm{e}^{-\lambda T}}}{1 - \dfrac{\lambda(1-\mathrm{e}^{-\mu T})}{\lambda+\mu}\dfrac{z}{1-z\mathrm{e}^{-\mu T}} - \dfrac{\mu(1-\mathrm{e}^{-\lambda T})}{\lambda+\mu}\dfrac{z}{z-\mathrm{e}^{-\lambda T}}}, \tag{11.33}$$

and the stability condition is

$$\frac{\lambda}{\mu} < \frac{\mathrm{e}^{-\lambda T}(1-\mathrm{e}^{-\mu T})}{1-\mathrm{e}^{-\lambda T}}. \tag{11.34}$$

Proof.

$$\mathbf{P}(Z < x) = 1 - \mathrm{e}^{-\lambda x}, \qquad \mathbf{P}(Y < x) = 1 - \mathrm{e}^{-\mu x}.$$

The distribution function of $Y - Z$ is

$$F(x) = \begin{cases} \frac{\mu}{\lambda+\mu}\mathrm{e}^{\lambda x} & \text{if } x \le 0, \\ 1 - \frac{\lambda}{\lambda+\mu}\mathrm{e}^{-\mu x} & \text{if } x > 0. \end{cases}$$

We find the transition probabilities. In the case $j > 0$,

$$f_j = 1 - \frac{\lambda}{\lambda+\mu}\mathrm{e}^{-\mu(j-1)T} - 1 + \frac{\lambda}{\lambda+\mu}\mathrm{e}^{-\mu j T} = \frac{\lambda}{\lambda+\mu}(1-\mathrm{e}^{-\mu T})\mathrm{e}^{-\mu(j-1)T},$$

for the negative values ($j \ge 0$)

$$f_{-j} = \frac{\mu}{\lambda+\mu}\mathrm{e}^{-\lambda j T} - \frac{\mu}{\lambda+\mu}\mathrm{e}^{-\lambda(j+1)T} = \frac{\mu}{\lambda+\mu}(1-\mathrm{e}^{-\lambda T})\mathrm{e}^{-\lambda j T},$$

$$p_{i0} = \hat{f}_i = \sum_{j=-\infty}^{-i} f_j = \sum_{j=i}^{\infty} \frac{\mu}{\lambda+\mu}(1-\mathrm{e}^{-\lambda T})\mathrm{e}^{-\lambda jT} = \frac{\mu}{\lambda+\mu}\mathrm{e}^{-\lambda iT}.$$

Using the matrix of transition probabilities (11.32) we get the system of equations

$$\begin{aligned}
p_0 &= p_0\hat{f}_0 + p_1\hat{f}_1 + p_2\hat{f}_2 + p_3\hat{f}_3 + \dots\\
p_1 &= p_0 f_1 + p_1 f_0 + p_2 f_{-1} + p_3 f_{-2} + \dots\\
p_2 &= p_0 f_2 + p_1 f_1 + p_2 f_0 + p_3 f_{-1} + \dots\\
&\ \vdots
\end{aligned}$$

Multiplying the jth equation by z^j, summing up by j from 0 to infinity for the generating function $P(z) = \sum_{j=0}^{\infty} p_j z^j$ we obtain

$$P(z) = P(z)F_+(z) + \sum_{j=1}^{\infty} p_j z^j \sum_{i=0}^{j-1} f_{-i} z^{-i} + \sum_{j=0}^{\infty} p_j \hat{f}_j,$$

where

$$\begin{aligned}
F_+(z) &= \sum_{i=1}^{\infty} f_i z^i = \frac{\lambda z}{\lambda+\mu}(1-\mathrm{e}^{-\mu T}) \sum_{i=1}^{\infty} \mathrm{e}^{-\mu(i-1)T} z^{i-1}\\
&= \frac{\lambda(1-\mathrm{e}^{-\mu T})}{\lambda+\mu} \frac{z}{1-z\mathrm{e}^{-\mu T}},
\end{aligned}$$

$$\sum_{i=0}^{j-1} f_{-i} z^{-i} = \frac{\mu(1-\mathrm{e}^{-\lambda T})}{\lambda+\mu} \sum_{i=0}^{j-1} \mathrm{e}^{-\lambda iT} z^{-i} = \frac{\mu(1-\mathrm{e}^{-\lambda T})}{\lambda+\mu} \frac{1-\left(\frac{\mathrm{e}^{-\lambda T}}{z}\right)^j}{1-\frac{\mathrm{e}^{-\lambda T}}{z}},$$

$$\sum_{i=0}^{\infty} p_i \hat{f}_i = \sum_{i=0}^{\infty} p_i \frac{\mu}{\lambda+\mu}\mathrm{e}^{-\lambda iT} = \frac{\mu}{\lambda+\mu} P\left(\mathrm{e}^{-\lambda T}\right).$$

Using the preceding equations

$$\begin{aligned}
P(z) &= P(z)F_+(z) + \sum_{j=1}^{\infty} p_j z^j \frac{\mu(1-\mathrm{e}^{-\lambda T})}{\lambda+\mu} \frac{1-\left(\frac{\mathrm{e}^{-\lambda T}}{z}\right)^j}{1-\frac{\mathrm{e}^{-\lambda T}}{z}} + \frac{\mu}{\lambda+\mu} P\left(\mathrm{e}^{-\lambda T}\right)\\
&= P(z)F_+(z) + \frac{\mu(1-\mathrm{e}^{-\lambda T})}{\lambda+\mu} \frac{z}{z-\mathrm{e}^{-\lambda T}} \left[P(z) - P\left(\mathrm{e}^{-\lambda T}\right)\right]
\end{aligned}$$

$$+\frac{\mu}{\lambda+\mu} P\left(\mathrm{e}^{-\lambda T}\right)$$

or

$$\begin{aligned} P(z) & \left[1-F_{+}(z)-\frac{\mu(1-\mathrm{e}^{-\lambda T})}{\lambda+\mu} \frac{z}{z-\mathrm{e}^{-\lambda T}}\right] \\ & = P\left(\mathrm{e}^{-\lambda T}\right)\left[\frac{\mu}{\lambda+\mu}-\frac{\mu(1-\mathrm{e}^{-\lambda T})}{\lambda+\mu} \frac{z}{z-\mathrm{e}^{-\lambda T}}\right]. \end{aligned}$$

$P(\mathrm{e}^{-\lambda T})$ may be computed from the condition $P(1)=1$,

$$P\left(\mathrm{e}^{-\lambda T}\right)=1-\frac{\lambda}{\mu} \frac{1-\mathrm{e}^{-\lambda T}}{\mathrm{e}^{-\lambda T}(1-\mathrm{e}^{-\mu T})}.$$

So for the generating function we get Eq. (11.33). From it we get the probability of the event that the waiting time is equal to zero:

$$p_0=\left[1-\frac{\lambda}{\mu} \frac{1-\mathrm{e}^{-\lambda T}}{\mathrm{e}^{-\lambda T}(1-\mathrm{e}^{-\mu T})}\right] \frac{\mu}{\lambda+\mu}.$$

In order to have $p_0>0$, the inequality

$$\frac{\lambda}{\mu} \frac{1-\mathrm{e}^{-\lambda T}}{\mathrm{e}^{-\lambda T}(1-\mathrm{e}^{-\mu T})}<1$$

must be fulfilled. It leads to condition (11.34) and coincides with the stability condition for the number of customers. □

11.8 Exercises

Exercise 11.1. A transmission link with capacity $C = 5\,\mathrm{MB/s}$ serves two kinds of CBR connections. Type i connections arrive according to a Poisson process at a rate λ_i and occupy c_i bandwidth of the link for an exponentially distributed amount of time with the parameter μ_i $(i = 1, 2)$, where $c_1 = 1\,\mathrm{MB}$ and $c_2 = 2\,\mathrm{MB}$.

1. Describe the system behavior with a CTMC and compute the loss probability of type 1 customers if $\lambda_2 = 0$.
2. Describe the system behavior with a CTMC when both λ_1 and λ_2 are positive, and compute the loss probability of types 1 and 2 connections and the overall loss probability of connections.
3. Which loss probability is higher, that of type 1 or that of type 2 connections? Why?

4. Compute the link utilization factor when both arrival intensities are positive.
5. Compute the link utilization of type 1 and type 2 connections.

Exercise 11.2. There is a transmission link with a capacity of $C = 13$ MB/s that serves adaptive connections. The connections arrive according to a Poisson process at a rate λ, and their length is exponentially distributed with the parameter μ. The minimal and maximal bandwidths of the adaptive connections are $c_{\text{min}} = 2$ MB/s and $c_{\text{max}} = 3$ MB/s, respectively. Compute the average bandwidth of an adaptive connection in equilibrium.

Exercise 11.3. There is a transmission link with a capacity of $C = 13$ MB/s that serves elastic connections. The connections arrive according to a Poisson process at a rate λ, and during an elastic connection an exponentially distributed amount of data is transmitted with the parameter γ. The minimal and maximal bandwidths of the elastic connections are $c_{\text{min}} = 2$ MB/s and $c_{\text{max}} = 3$ MB/s, respectively. Compute the average bandwidth of an elastic connection in equilibrium. Compute the average time of an elastic connection in equilibrium.

Exercise 11.4. A transmission link with a capacity of $C = 3$ MB/s serves two kinds of elastic connections. Type 1 connections arrive according to a Poisson process at a rate $\lambda_1 = 0.5$ 1/s and transmit an exponentially distributed amount of data with the parameter $\gamma_1 = 4$ 1/MB. The minimal and maximal bandwidths of type 1 connections are $\check{c}_1 = 1$ MB/s and $\hat{c}_1 = 1$ MB/s, respectively. Type 2 connections are characterized by $\lambda_2 = 0.1$ 1/s, $\gamma_1 = 2$ 1/MB, $\check{c}_2 = 1$ MB/s, and $\hat{c}_2 = 2$ MB/s.

(a) Describe the system behavior with a CTMC.
(b) Compute the mean number of type 1 and type 2 connections.
(c) Compute the mean channel utilization.
(d) Compute the loss probability of type 1 and type 2 connections.
(e) Compute the average bandwidth of type 2 connections.

Exercise 11.5. A transmission link with a capacity of $C = 3$ MB/s serves two kinds of connections, elastic and adaptive. Type 1 elastic connections arrive according to a Poisson process at a rate λ_1 [1/s] and transmit an exponentially distributed amount of data with parameter γ_1 [1/MB]. The minimal and maximal bandwidths of type 1 connections are $\check{c}_1 = 0.75$ MB/s and $\hat{c}_1 = 1.5$ MB/s, respectively. Type 2 adaptive connections arrive according to a Poisson process at a rate λ_2 [1/s] and stay in the system for an exponentially distributed amount of time with the parameter μ_2 [1/s]. The minimal and maximal bandwidths of type 2 connections are $\check{c}_2 = 1$ MB/s and $\hat{c}_2 = 2$ MB/s, respectively.

(a) Describe the system behavior with a CTMC.
(b) Compute the loss probability of type 1 and type 2 connections.
(c) Compute the average bandwidth of type 1 and type 2 connections.
(d) Compute the mean number of type 1 and type 2 connections on the link.

Exercise 11.6. Compute the mean value of the waiting time in a cyclic waiting system.

Exercise 11.7. Let us consider our cyclic waiting system in the case of discrete time. Divide the cycle time T into n equal parts and suppose that for an interval T/n a new customer enters with probability r (there is no entry with probability $1 - r$), and the service in process for such an interval is continued with probability q and completed with probability $1 - q$ (i.e., the service time has a geometrical distribution). The service may be started at the moment of arrival or at moments differing from it by multiples of T.

(a) Show that the number of customers in the system at moments $t_k - 0$ constitute a Markov chain, and find its transition probabilities.

(b) Find the generating function of the number of customers in a system in equilibrium and the stability condition.