

Chapter 2

Closed Queueing Networks

Queueing networks in general are networks of processing stations with intermediate storage areas called buffers. At each of the stations, a service is provided that takes up time. Workpieces approach the stations and are processed immediately if the server is idle. Otherwise, the workpieces line up in the buffer in front of a station and wait to receive service. Apart from that, queueing networks differ.

In Sect. 2.1, we present the assumptions of the closed queueing networks considered in this thesis and contrast these to other common assumptions. Subsequently, in Sect. 2.2, we introduce the characteristics of the closed queueing networks in focus.

2.1 Assumptions

Material flow. With regard to the material flow, queueing networks are mainly distinguished between open queueing networks (OQN) and closed queueing networks (CQN). Open systems are characterized by an arrival process which is independent of the departure process. In these systems, the first station is never starved and the last station is never blocked.¹

In closed queueing networks, workpieces circulate through the system. The arrival stream at the first station conforms with the departure process at the last station. In contrast to open queueing systems, the last station of a closed queueing network may become blocked if the buffer in front of the input station is full of workpieces. This holds true under the assumption that the buffer capacities within the system are finite. Furthermore, the first station may become starved if no carriers with workpieces are available in the buffer in front of the input station. Closed systems, therefore, are characterized by high dependencies not only between

¹See Dallery and Gershwin (1992).

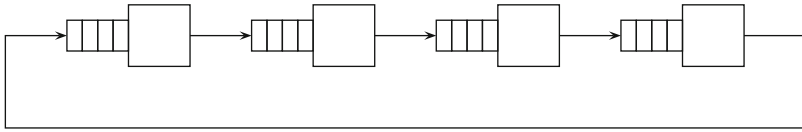


Fig. 2.1 Closed queueing network with four stations

Table 2.1 Notation

b_i	Buffer capacity at station i
c_i	Coefficient of variation of the processing time distribution of station i
d_i	Capacity of the station i , $d_i = b_i + s_i$
d^{min}	Capacity of the station with least buffer and server capacity, $d^{min} = \min_i \{b_i + 1\}$
M	Number of stations in the system
μ_i	Mean processing rate at station i
n	Number of workpieces/customers/jobs/pallets circling in the system
n^{NS}	Minimum number of workpieces for which no starving occurs in the complete networks
N	Maximum number of workpieces in the CQN
$P_i^B(n)$	Probability of blocking at station i if n workpieces circulate in the system
$P_i^S(n)$	Probability of starving at station i if n workpieces circulate in the system
$PR(n)$	Production rate with n workpieces in the system
s_i	Server capacity of station i
T_i	Random variable representing the processing time at station i

mid-stations, but also between the first and the last station. The closed-loop flow reflects the main assumption regarding the queueing networks examined in this thesis.

An example of a closed queueing network with four stations and linear flow of material is depicted in Fig. 2.1. Each station consists of a server (taller rectangle) and buffer space in front of the server (smaller rectangles).

Within the class of closed queueing networks, there are systems with a linear flow of material, flexible manufacturing systems,² systems with arbitrary routing,³ and closed assembly and disassembly systems.⁴ Here, the flow of material is assumed to be linear, i.e. the processing stations are connected in series. In this case, each workpiece receives service in the same order. We denote the stations by the index i in topological order, ranging from 1 to M , with M denoting the number of stations. In closed systems, the successor of station M constitutes station 1, and vice versa, the predecessor of station 1 represents station M . The notation used in this chapter is given in Table 2.1.

Work-in-process. The material consists of discrete parts. The number of items circulating within the closed queueing network is constant. This constitutes the central assumption in closed queueing networks. The circular flow of material is

²See Tempelmeier and Kuhn (1993).

³See Koenigsberg (1982).

⁴See Duenyas (1994).

often due to the requirement that workpieces must be led through the production system by carriers. In this case, a raw material is attached to an empty carrier when it passes in front of the first station. The carrier leads the workpiece through the system. Behind the last station, the finished product is released, and the empty carrier picks up raw material anew in the buffer between the last and the first station. No carriers are added or removed causing the number of carriers to stay constant.

Alternatively, the circulating items can be production-authorization cards, also called CONWIP (constant work-in-process) cards. In a CONWIP system, cards are attached to each processing unit. The purpose of these cards is to control the work-in-process of the system: After the completion of a unit at the last station, the CONWIP card is released from the finished workpiece. This free card authorizes a raw unit to enter the system in front of the first station.

We assume that, at the instant at which a finished product is removed from the carrier (or card), a new workpiece is loaded onto that carrier (or attached to the card) infinitely fast. In other words, the time to change a finished product into a raw product amounts to zero. As a result, not only are the carriers or cards constant, but so is the work-in-process.

The number of carriers or workpieces is denoted by n . n ranges from 1 to N , where N denotes the system capacity minus one.⁵ The number of carriers greatly influences the production rate. This issue is investigated in Sect. 2.2.

Processing time distribution. We assume that the processing time at station i , denoted by T_i , represents a random variable describing the time a server needs for the processing of a workpiece. T is assumed to be independent and identically distributed for all workpieces. The processing time per station may vary because of products taking different amounts of time or due to manual labor. Moreover, machine failures may contribute to the variability of the processing time if the machine failures are implicitly considered as part of the processing time per workpiece.⁶

The most frequently used processing time distribution is the exponential distribution. Systems under this assumption are mathematically easier to handle. However, the exponential distribution implies a very high variability, which is usually not present in real-life systems.⁷

In order to model the processing time more accurately, the variance should be regarded as well. We assume the processing time to be specified by the mean processing rate, denoted by μ , and the coefficient of variation, denoted by c (both of arbitrary values),⁸ or that it follows a phase-type distribution.⁹ Both settings also include the exponential distribution.

⁵For the range of the number of workpieces, see also page 9.

⁶See Gaver (1962) and “Machine failures and repairs” in this section.

⁷See Sect. 5.2.2 for further details.

⁸This is assumed in Chap. 4.

⁹In Chap. 5, both cases are considered.

If the processing times are stochastic, the so-called starving effect occurs. Starvation takes place if a station is operative but not supplied with material. It is measured by the starving probability that corresponds to the percentage of time in which starving occurs. It holds that the higher the starving probability, the lower the production rate.

Buffer capacity. Under the assumption of stochastic processing times, buffer space between the stations is very important because it mainly contributes to the productivity of the network. Buffer capacity ranges from infinitely large (unlimited buffer capacity), over a defined number (limited buffer capacity), to no buffer capacity. Unlimited buffer capacity is a theoretical construct that is easier to handle in performance-analysis procedures. Procedures for limited buffer-capacity systems are able to consider the complete range of buffer capacity from infinite to nonexistent. Within this thesis, limited buffer capacity is assumed. The capacity of the buffer in front of station i is denoted by b_i .

If the buffer space is limited, blocking may occur. A station is blocked if it is unable to work because the succeeding buffer is full and is, therefore, prevented from working on the next job. The higher the buffer capacity, the lower the frequency of blockages, which results in a higher production rate.

Although higher buffer capacity increases the production rate, it is not beneficial to install as many buffers as possible. In buffer optimization, the buffer capacity and the buffer distribution constitute important decision variables in the optimization of queueing networks and production systems in particular.¹⁰ In closed queueing systems, the number of workpieces represents a further decision variable in the optimization.

Blocking mechanism. Blocking may occur at different points during operation. The two most common blocking mechanisms are blocking-after-service and blocking-before-service.¹¹ This thesis is based on the blocking-after-service discipline.

Under the blocking-before-service (BBS) mechanism—also called type-2 blocking or service blocking—a machine can only start processing if a space is available in the downstream buffer. This means that a station is blocked if the succeeding buffer is full at the instant that the processing is supposed to start.

Blocking-after-service (BAS)—also called type-1 blocking or production blocking—occurs if, at the instant of completion of a part, the downstream buffer is full. The finished workpiece stays on the machine and prevents the machine from further production—thus the server is blocked. As soon as the downstream machine releases its current workpiece, it starts processing the next workpiece and makes buffer space available. This is the instant in time, in which blocking is resolved.

¹⁰See for instance Gershwin and Schor (2000) on a buffer optimization procedure.

¹¹See Dallery and Gershwin (1992, p. 12), for these and other blocking mechanisms.

Server. A network is distinguished by single and multiple servers. We assume that the processing unit consists of a single server. That means only one workpiece may be on the server. The three states of the processing unit are starved, blocked, and busy.

Machine failures and repairs. In some systems, machines are prone to failure. There are two major types of failures described in the relevant literature: operation-dependent failures and time-dependent failures.¹² In this thesis, machine failures are not considered explicitly, i. e. the servers are assumed to be completely reliable. However, operation-dependent failures may be implicitly included in the processing times by the Completion-Time Concept of Gaver (1962). This concept is explained in detail in Manitz (2005).¹³

In summary, our assumptions are as follows: The flow of material is assumed to be linear. At each of the $i = 1, \dots, M$ stations, a single server operates with stochastic processing times without failures. The service time distribution is described by the processing rate μ_i and the coefficient of variation c_i at each station i or by a phase-type distribution. In front of each station, a buffer with finite capacity b_i is located. Processing takes place according to the first-come first-served service discipline. After the process completion, the workpieces are led into the subsequent buffer if space is available. Otherwise, the server is blocked. The blocking mechanism is assumed to be blocking-after-service (BAS). Behind the last station, finished products are released and empty carriers pick up raw material in the buffer in front of the first station.

2.2 Characteristics

The defining characteristic of closed queueing networks is the constant number of workpieces. This mainly influences the performance measures of the system. The production-rate function subject to the number of workpieces is characteristic for CQN. We will, therefore, investigate the effect of the number of workpieces on the blocking and starving probabilities leading to the typical production-rate function according to the assumptions made above.

Range of the number of workpieces. The number of workpieces ranges from 1 to N . It is restricted to N due to the finite capacity of the system. If the number of workpieces amounted to the number of places in the system, each server would be blocked by the parts in the succeeding buffer, and the production rate would amount to zero. This effect is called deadlock.¹⁴ Therefore, n must not exceed one workpiece less than the sum of the buffer and server capacities, b_i and s_i , of all stations i .

¹²For further details, see Dallery and Gershwin (1992, pp. 14ff).

¹³See Manitz (2005, pp. 35ff).

¹⁴See Yüzükirmizi (2005, pp. 17f).

The maximum number of workpieces, N , is given by

$$N = \sum_{i=1}^M (b_i + s_i) - 1. \quad (2.1)$$

Blocking. Blocking is measured by the blocking probability which corresponds to the percentage of time in which station i is blocked. It is denoted by P_i^B . As long as the number of workpieces in the system, n , is less or equal to the minimum station capacity, d^{min} , with $d^{min} = \min_i \{b_i + s_i\}$, no blocking can occur at any station:

$$P_i^B(n) = 0 \quad \text{for } n \leq d^{min}, \forall i. \quad (2.2)$$

With one more workpiece, $n = d^{min} + 1$, blocking can occur at the station located upstream of the minimum-capacity station. We denote the index of the station with the minimum station capacity by j . If, at any instant in time, the number of workpieces at station j equals $n_j = d^{min}$ and the remaining workpiece is located at station $j - 1$, $n_{j-1} = 1$, then station $j - 1$ becomes blocked if it finishes its current workpiece faster than station j .¹⁵ Generally, station i may be blocked as soon as the number of workpieces in the system is greater than the station capacity of the succeeding station, $d_{i+1} = b_{i+1} + s_{i+1}$:

$$P_i^B(n) > 0 \quad \text{for } n > d_{i+1}, \forall i. \quad (2.3)$$

In the performance analysis, blocking must be taken into account as soon as the first station might become blocked, i. e. if $n > d^{min}$.

Starving. Starving is expressed by the starving probability, which represents the percentage of time in which a station is not supplied with material. This probability is denoted by P_i^S with regard to station i . With only one workpiece in the system, the starving probability is as high as possible. By adding more workpieces to the system, the probability of starvation decreases more and more.

Upon reaching a high enough quantity of workpieces, station i cannot be starving anymore, and the server of station i is occupied at any time. This occurs at a workpiece level such that, even if all buffer places and servers of all other stations are occupied, at least one workpiece still remains to be at station i :

$$P_i^S(n) = 0 \quad \text{for } n > \sum_{\substack{k=1, \\ k \neq i}}^M (b_k + s_k) \quad \forall i. \quad (2.4)$$

¹⁵See Onvural and Perros (1989b, p. 112).

Table 2.2 Exemplary configuration of a CQN

i	1	2	3	4	5
μ_i	1	1	1	1	1
c_i^2	0.5	0.6	0.5	0.6	0.7
b_i	4	4	4	4	4
s_i	1	1	1	1	1

Starving occurs if less than the afore-described number of workpieces resides at the station:

$$P_i^S(n) > 0 \quad \text{for } n \leq \sum_{\substack{k=1, \\ k \neq i}}^M (b_k + s_k) \quad \forall i. \quad (2.5)$$

There is no starving in the complete network if the number of workpieces is higher than n^{NS} , where n^{NS} denotes the system capacity of all stations except for the station with the minimum capacity:

$$n > n^{NS} = \sum_{\substack{i=1, \\ i \neq j}}^M (b_i + s_i), \quad \text{with } j = \arg \min_i \{b_i + s_i\}. \quad (2.6)$$

Production rate function. The production rate is denoted by PR . It constitutes the average number of finished parts per time unit. The production rate is subject to the number of workpieces and corresponds to the processing rate of station i , μ_i , multiplied by the percentage of time the station works at that processing rate, see Eq. (2.7).

$$PR(n) = \mu_i \cdot [1 - P_i^B(n) - P_i^S(n)] \quad \forall n, i. \quad (2.7)$$

The percentage of time the station works is called utilization, U_i , and represents the counter-probability of the event that a station cannot work because it is starving or blocked: $U_i = 1 - P_i^B - P_i^S$. The production rate is equal for all stations i . This corresponds to a law called conservation of flow.¹⁶

In the following, an exemplary configuration is investigated. The data of this example are given in Table 2.2. Figure 2.2 shows the course of blocking and starving probabilities for the given configuration.

The probability of starvation is very high for small n and decreases down to $P^S(n) = 0$ for $n > n^{NS}$. The probability of blocking equals zero for $n \leq d_i \quad \forall i$ and increases until N workpieces are reached. Note that in this example, the capacities are equal over all stations. Hence, for each station, the same limits of n hold regarding whether or not blocking or starving occurs.

¹⁶See Dallery and Gershwin (1992, p. 20).

Fig. 2.2 Blocking and starving probabilities

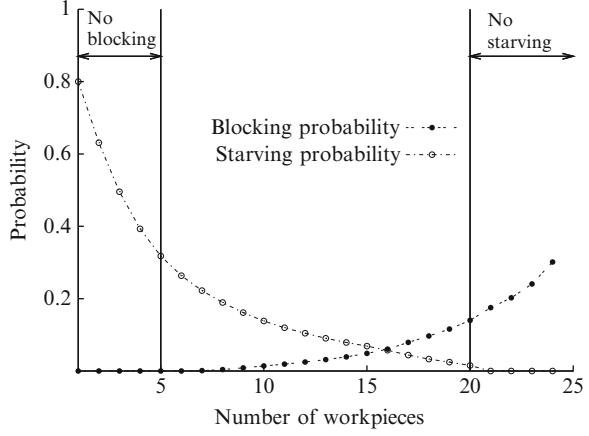
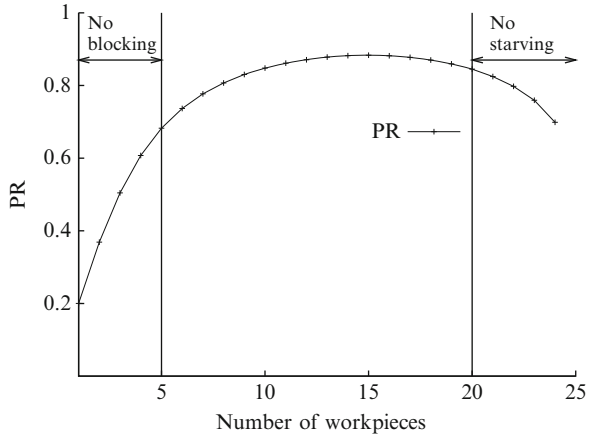


Fig. 2.3 Production rate



In summary, three ranges can be distinguished. For increasing n holds that

- For $n \leq d^{min}$, the blocking probability equals zero and the starving probability decreases
- For $n > d^{min}$ and $n < n^{NS}$, both blocking and starving effects occur. The starving probability decreases and the blocking probability increases
- For $n \geq n^{NS}$, the starving probability amounts to zero, and blocking further increases.

The corresponding production-rate function is depicted in Fig. 2.3. The sum of the blocking and starving probability is very high for low n due to the starving probability. It decreases with increasing n because the starving probability decreases. For a high number of n , it increases due to the blocking probability. In accordance with Eq. (2.7), the effect of the production rate is in reverse to the sum of the blocking and starving probabilities: The production-rate function first increases

for increasing n and then decreases. This concave function has the typical shape of the production-rate function of CQN.¹⁷

There is a maximal production rate for a medium number of workpieces. The number of workpieces for which the production rate is maximal, n^* , is found in the range of $d^{min} \leq n^* \leq n^{NS}$. This is true for the following reason: As long as the number of workpieces n is less than d^{min} ($n < d^{min}$), an additional workpiece must increase the production rate at least marginally because no blocking occurs and starvation decreases. As soon as the number of workpieces n exceeds n^{NS} ($n > n^{NS}$), no starvation exists and an additional workpiece will increase blocking and will, therefore, decrease the production rate.

Under optimization aspects, the establishment of a workpiece level n' for which holds $n^* < n'$ is not useful: The same or higher production rate can be achieved with less work-in-process, $PR(n^*) > PR(n')$. A value of n in the range of $1 \leq n \leq d^{min}$ is not of interest from the analytical standpoint because blocking is not taken into account. Hence, the most interesting range of the workpiece level equals $[d^{min} + 1, \dots, n^*]$.

¹⁷Compare Yao (1985) for statements on closed queueing networks with infinite capacities.

Performance Analysis of Closed Queueing Networks

Lagershausen, S.

2013, XXIII, 169 p. 49 illus., 5 illus. in color., Softcover

ISBN: 978-3-642-32213-6