2 Queueing models and some fundamental relations

Queueing models have been proved to be very useful in many practical applications in areas such as, e.g., production systems, inventory systems and communication systems. These applications concern in particular *design problems*, where we need to answer questions like: Is the capacity sufficient?, What should be the layout? or How do we have to divide work among several capacities? In many applications the variability in the arrival and service processes are essential to the performance of the system. Queueing models help us to understand and quantify the effect of variability.

In this chapter we describe the basic queueing model and we discuss some important fundamental relations for this model. These results can be found in every standard textbook on this topic, see, e.g., [1, 3, 5].

2.1 Queueing models and Kendall's notation

The basic queueing model is shown in figure 1.

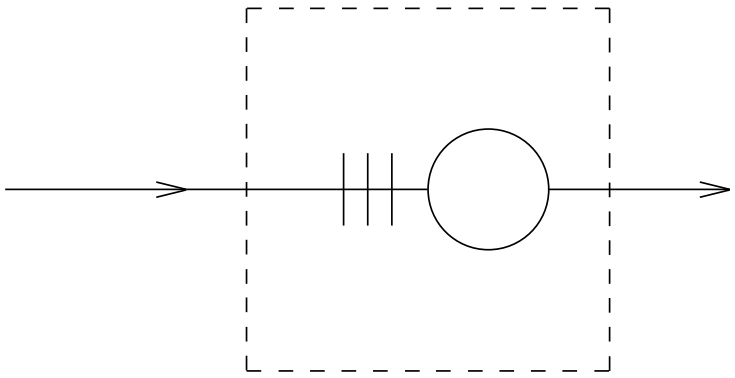

Figure 1: Basic queueing model

Among others, a queueing model is characterized by:

- The arrival process of customers.
  Usually we assume that the interarrival times are independent and have a common distribution. In many practical situations customers arrive according to a Poisson stream (i.e., exponential interarrival times). Customers may arrive one by one, or in batches. An example of batch arrivals is the customs office at the border where travel documents of bus passengers have to be checked.

- The behaviour of customers.
  Customers may be patient and willing to wait (for a long time). Or customers may be impatient and leave after a while. For example, in call centers, customers will hang up when they have to wait too long before an operator is available, and they possibly try again after a while.

- The service times.
  Usually we assume that the service times are independent and identically distributed, and that they are independent of the interarrival times. For example, the service times can be deterministic or exponentially distributed. It can also occur that service times are dependent of the queue length. For example, the processing rates of the machines in a production system can be increased once the number of jobs waiting to be processed becomes too large.

- The service discipline.
  Customers can be served one by one or in batches. We have many possibilities for the order in which they enter service. We mention:

  - first come first served, i.e., in order of arrival;
  - random order;
  - last come first served (e.g., in a computer stack or a shunt buffer in a production line);
  - priorities (e.g., rush orders first, shortest processing time first);
  - processor sharing (in computers that equally divide their processing power over all jobs in the system).

- The service capacity.
  There may be a single server or a group of servers helping the customers.

- The waiting room.
  There can be limitations with respect to the number of customers in the system. For example, in a data communication network, only finitely many cells can be buffered in a switch. The determination of good buffer sizes is an important issue in the design of these networks.

Kendall introduced a shorthand notation to characterize a range of these queueing models. It is a three-part code $a/b/c$. The first letter specifies the interarrival time distribution and the second one the service time distribution. For example, for a general distribution the letter G is used, M for the exponential distribution (M stands for Memoryless) and D for deterministic times. The third and last letter specifies the number of servers. Some examples are $M/M/1$, $M/M/c$, $M/G/1$, $G/M/1$ and $M/D/1$. The notation can be extended with an extra letter to cover other queueing models. For example, a system with exponential interarrival and service times, one server and having waiting room only for N customers (including the one in service) is abbreviated by the four letter code $M/M/1/N$.

In the basic model, customers arrive one by one and they are always allowed to enter the system, there is always room, there are no priority rules and customers are served in order of arrival. It will be explicitly indicated (e.g., by additional letters) when one of these assumptions does not hold.

2.2 Occupation rate

In a single-server system $G/G/1$ with arrival rate λ and mean service time $E(B)$ the amount of work arriving per unit time equals $\lambda E(B)$. The server can handle 1 unit work per unit time. To avoid that the queue eventually grows to infinity, we have to require that $\lambda E(B) < 1$. Without going into details, we note that the mean queue length also explodes when $\lambda E(B) = 1$, except in the $D/D/1$ system, i.e., the system with no randomness at all.

It is common to use the notation

$$\rho = \lambda E(B).$$

If $\rho < 1$, then ρ is called the *occupation rate* or *server utilization*, because it is the fraction of time the server is working.

In a multi-server system $G/G/c$ we have to require that $\lambda E(B) < c$. Here the occupation rate per server is $\rho = \lambda E(B)/c$.

2.3 Performance measures

Relevant performance measures in the analysis of queueing models are:

- The distribution of the waiting time and the sojourn time of a customer. The sojourn time is the waiting time plus the service time.
- The distribution of the number of customers in the system (including or excluding the one or those in service).

In particular, we are interested in mean performance measures, such as the mean waiting time and the mean sojourn time.

2.4 Little's law

Little's law gives a very important relation between $E(L)$, the mean number of customers in the system, $E(S)$, the mean sojourn time and λ, the average number of customers entering the system per unit time. Little's law states that

$$E(L) = \lambda E(S). \tag{1}$$

Here it is assumed that the capacity of the system is sufficient to deal with the customers (i.e., the number of customers in the system does not grow to infinity).

Intuitively, this result can be understood as follows. Suppose that all customers pay 1 dollar per unit time while in the system. This money can be earned in two ways. The first possibility is to let pay all customers "continuously" in time. Then the average reward earned by the system equals $E(L)$ dollar per unit time. The second possibility is to let customers pay 1 dollar per unit time for their residence in the system when they leave. In equilibrium, the average number of customers leaving the system per unit time is equal

to the average number of customers entering the system. So the system earns an average reward of $\lambda E(S)$ dollar per unit time. Obviously, the system earns the same in both cases. For a rigorous proof, see [2, 4].

To demonstrate the use of Little's law we consider the basic queueing model in figure 1 with one server. For this model we can derive relations between several performance measures by applying Little's law to suitably defined (sub)systems. Application of Little's law to the system consisting of queue plus server yields relation (1). Applying Little's law to the queue (excluding the server) yields a relation between the queue length L^q and the waiting time W, namely

$$E(L^q) = \lambda E(W).$$

Finally, when we apply Little's law to the server only, we obtain (cf. section 2.2)

$$\rho = \lambda E(B),$$

where ρ is the mean number of customers at the server (which is the same as the fraction of time the server is working) and $E(B)$ the mean service time.

2.5 PASTA property

For queueing systems with Poisson arrivals, so for $M/\cdot/\cdot$ systems, the very special property holds that arriving customers find on average the same situation in the queueing system as an outside observer looking at the system at an arbitrary point in time. More precisely, the fraction of customers finding on arrival the system in some state A is exactly the same as the fraction of time the system is in state A. This property is only true for Poisson arrivals.

In general this property is not true. For instance, in a $D/D/1$ system which is empty at time 0, and with arrivals at $1, 3, 5, \ldots$ and service times 1, every arriving customer finds an empty system, whereas the fraction of time the system is empty is $1/2$.

This property of Poisson arrivals is called PASTA property, which is the acronym for Poisson Arrivals See Time Averages. Intuitively, this property can be explained by the fact that Poisson arrivals occur completely random in time. A rigorous proof of the PASTA property can be found in [6, 7].

In the following chapters we will show that in many queueing models it is possible to determine mean performance measures, such as $E(S)$ and $E(L)$, directly (i.e., not from the distribution of these measures) by using the PASTA property and Little's law. This powerful approach is called the *mean value approach*.

References

[1] L. Kleinrock, *Queueing Systems, Vol. I: Theory*. Wiley, New York, 1975.

[2] J.D. Little, *A proof of the queueing formula $L = \lambda W$*, Opns. Res., 9 (1961), pp. 383–387.

[3] S.M. Ross, *Introduction to probability models*, 6th ed., Academic Press, London, 1997.

[4] S. Stidham, *A last word on* $L = \lambda W$, Opns. Res., 22 (1974), pp. 417–421.

[5] H.C. Tijms, *Stochastic models: an algorithmic approach*, John Wiley & Sons, Chichester, 1994.

[6] R.W. Wolff, *Poisson arrivals see time averages*, Opns. Res., 30 (1982), pp. 223–231.

[7] R.W. Wolff, *Stochastic modeling and the theory of queues*, Prentice-Hall, London, 1989.