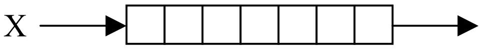
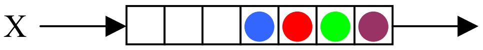
Dr. Jozo Dujmović

An Elementary Analysis of the Single Server Model

A queue is a memory element that contains a line of customers that wait for service. It is graphically presented as follows:

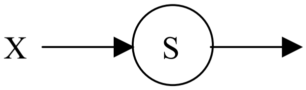


Each box represents a memory space that can contain one customer request. The customers arrive randomly, and the mean arrival rate is X customers per second. The average number of customers in the queue is denoted Q and customers are sometimes represented as dots stored in queue boxes:

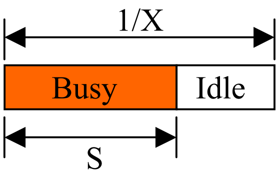


In this example the queue length is Q=4 customers.

A server is a delay element where customers spend an average service time S. Graphical symbol of a server is a circle:



If X customers arrive per second than the average interarrival time is 1/X. On the average, each 1/X seconds there is an arrival and this period consists of two parts: during S seconds the server is busy, and during the time 1/X-S the server is idle, as follows:



From this picture it is obvious that the condition for normal work is that S is less than or equal to 1/X:

$$S \le 1/X$$
$$SX \le 1$$
$$X \le 1/S = X_{\max}$$

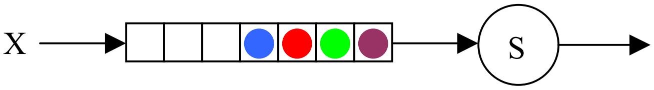
We can define the server utilization U as the fraction of time that the server is busy:

$$U = \frac{S}{1/X} = SX$$
$$0 \le U \le 1$$

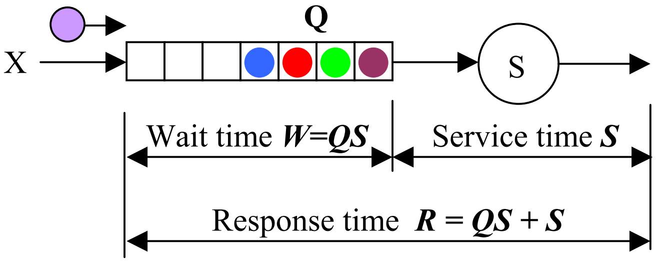
A typical working pattern of a server can be illustrated as follows:

Busy	Idle	Busy	Idle	Busy	Idle	Busy	Idle

A single server system is a system that has a single queue and a server as follows:



Let us analyze this system at the moment of arrival of a new customer. This is the moment when a new customer enters the queue, as shown in the following example:

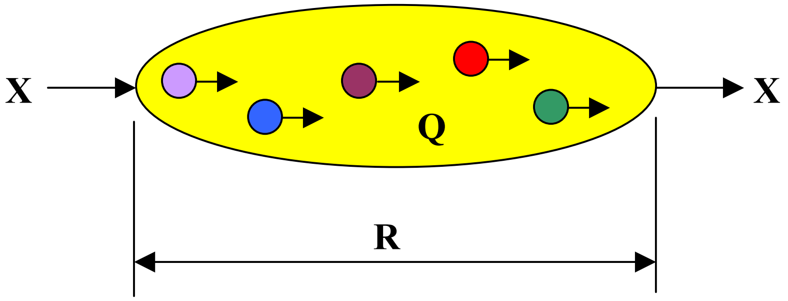


The arriving customer at the moment of arrival encounters (on the average) Q customers already in the system. In this example, at the moment of arrival the queue contains 4 customers, but the customer at the head of queue is being served and a part of his/her service may be completed before the arrival of the new customer. Therefore, the average number of customers encountered in the system, Q, is generally a real number. For example, if Q is 3.5 this denotes that (at the moment of arrival of a new customer) 3 customers have to be completely served and the customer in service already received 50% of his/her service. Therefore, the arriving customer will have to wait in queue the time $W=QS$ before receiving service. Consequently, the average value of the total time a customer spends in the system is

$$R = QS + S = S(Q+1)$$

This time is called the *response time* and it is the most important of all performance indicators. This formula is not particularly useful since generally we don't know what is the average number of customers in the system, Q.

To complete our analysis we need to know what is the relationship between Q, R, and X. This relationship can be illustrated as follows:



The average departure rate is called *throughput* and it must be equal to the average arrival rate, X. Indeed, if the departure rate were less than the arrival rate this would mean a permanent increase of the number of customers in the system, and consequently the average number of customers in the system, Q, could not be constant. Similarly, if the average departure rate were greater than the average arrival rate, the system would eventually become empty, yielding $Q=0$. Therefore, the average arrival rate (measured during a long observation period) must be equal to the average departure rate.

The customers spend in the system the average time R, and Q must be the average number of customers that enter the system during the time R, and this is RX. Therefore,

$$Q = RX$$

This simple formula is called Little's formula and it is the most important result in queuing theory. It can be shown that this is a powerful formula, since it holds not only for the system as a whole, but also for each subsystem or component of the analyzed system.

We have now shown three important formulas necessary for the analysis of a single server model:

$$
\begin{aligned}
&U = SX \\
&Q = RX \\
&R = QS + S = S(Q+1)
\end{aligned}
$$

The utilization formula *U=SX* can be interpreted as Little's formula applied only to the server: the utilization of server is the average number of customers at the server.

We can now use these formulas to compute the average response time (R), the average queue length (Q), the average wait time (W), the average server utilization (U), and the average number of customers that are waiting in the queue before service (Q_W), as follows:

$$
\begin{aligned}
&R = QS + S = RXS + S = RU + S \\
&R(1-U) = S \\
&R = \frac{S}{1-U} = \frac{S}{1-SX} = \frac{1}{1/S - X} = \frac{1}{X_{\max} - X} \\
&Q = RX = \frac{SX}{1-SX} = \frac{U}{1-U} \\
&Q - QU = U \\
&U = \frac{Q}{Q+1} = \frac{QS}{R} \\
&W = R - S \\
&Q_W = WX = RX - XS = Q - U = QU \\
&Q = Q_W + U
\end{aligned}
$$

The presented short derivation of the above formulas was completely based on elementary intuitive reasoning. Of course, intuitive reasoning is not a proof of correctness. If you have a sharp eye, and you should have it, you have noticed one imprecision in the above presentation: at the beginning we said that Q is the average number of customers in the system encountered by an arriving customer. In other words, Q is the mean value of the number of customers in the system computed only for moments of arrival of new customers, and this is computed for an infinite number of customers. Then later (for the Little formula) we said that Q is just the mean number of customers in the system. In other words, Q is now the average number of customers in the system computed over an infinite interval of time by an external observer. If external observers see something different from arriving customers, our derivation would be incorrect.

A more precise mathematical analysis is necessary to show under what conditions these formulas are correct. Such a mathematical analysis shows that some of these formulas hold only in the case when both interarrival times and service times are exponentially distributed. Indeed, we avoided to prove two important results: the Little formula, and the fact that only for exponential interarrival times the mean queue length encountered by an arriving customer is equal to the mean queue length observed by an external observer. Nevertheless, the above models are robust, useful, and illustrate the type of elementary reasoning that is necessary to understand basic dynamic phenomena in queuing systems.

Limits of intuitive reasoning

Everybody agrees that program testing can prove the presence of programming errors, but cannot prove their absence. Similarly, intuitive reasoning seems to be better instrument for detecting errors in mathematical models, than for proving their correctness. In queuing theory intuitive reasoning is sometimes similar to optical illusions. Following are three examples of wrong conclusions

Wrong conclusion #1. Customers arrive for service completely independently and randomly. Consequently, and if we take from the arriving customers samples of the number of customers in the system encountered at the moment of arrival, we can compute in this way the average number of customers in the system. True of false?

False. This approach is acceptable only for exponential interarrival *and service times*.

Wrong conclusion #2. The arrival rate X is the average number of customers that independently and randomly arrive for service in a time unit. The average service rate 1/S is the average number of customers that a server can serve in a time unit. If X=1/S then the system will work fine and the server will be permanently busy. True of false?

Mostly false. The system will not "work fine" since the response time will be $R = 1/(1/S - X) = +\infty$. Not surprisingly, it is true that the server will be permanently busy: $U = SX = 1$.

Wrong conclusion #3. System administrator reports to management: "Our servers are 95% busy and we offer an excellent service to our customers. The quality of service is practically the same as last month when the servers were 90% busy." True or false?

False. If U=0.95, then the response time is R=S/(1-U)=20S. So, to get S seconds of service the customer must spend 19S seconds waiting in clogged queues. This is unacceptable, except in cases where S is extremely small. If U=0.9 then R=10S, what is generally also unacceptable. When utilization increases from 90% to 95% the response time doubles, from 10S to 20S and the quality of service is certainly not the same.