4 The $M/M/1$ system

The basic mathematical model of a production system is the so-called $M/M/1$ queueing model. In this model jobs arrive according to a Poisson process and the processing times or service times of the jobs are independent and identically exponentially distributed. The production system is simplified to a single machine. The arrival rate is denoted by λ. The processing rate is denoted by μ, so the mean processing time is $1/\mu$. The jobs are processed in order of arrival (FCFS). We require that

$$\rho = \frac{\lambda}{\mu} < 1,$$

since, otherwise, the queue length will explode (see section 2.2). The quantity ρ is the fraction of time the machine is processing jobs. Although in reality a production system is never as mathematically simple as this, the model contains most of its essential characteristics. The analysis of this system will clearly show the sometimes devastating consequences of randomness both in the arrival process and in the service times.

4.1 The equilibrium distribution

In order to understand the behaviour of such a production system, it will be analysed as a stochastic process. Our main interest concerns the distribution of the number of jobs in the system at an arbitrary point in time. From this distribution we will see how the number of jobs in the system fluctuates and we will be able to compute important performance characteristics such as the mean number of jobs in the system and the fraction of jobs that will have a total throughput time less than, for instance, a week.

The assumptions we made about the system (i.e., Poisson arrivals, exponential processing times and FCFS servicing) make it possible to describe the state of the system at an arbitrary point in time by simply the number of jobs in the system. Without these assumptions, the state description would be very complicated and would have to contain not only the number of jobs in the system, but also, for example, the residual processing time of the job in service. The reason for this simplification is that, in the case of exponential interarrival times and exponential processing times, the distribution of the time until the next arrival or service completion is not affected by the time that elapsed since the last arrival and the last service completion. This is due to the *memoryless property* of the exponential distribution (see section 1.2.3) Further, the FCFS order of processing means that the past gives no information about the jobs waiting in the queue. Note that if, for instance, the processing order would be Shortest Processing Time First, then the jobs waiting in the queue will on average be longer than an arbitrary job.

Let us first have a look at a formal derivation of the equilibrium or limiting distribution, via the time-dependent behaviour. Let $p_k(t)$ denote the probability that at time t there are k jobs in the system. Then the evolution of the process in time can be described by the following set of equations:

$$p_0(t+h) = p_0(t)(1-\lambda h) + p_1(t)\mu h + o(h),$$

$$p_k(t+h) \;=\; p_{k-1}(t)\lambda h + p_k(t)(1-\lambda h-\mu h) + p_{k+1}(t)\mu h + o(h), \qquad k \geq 1.$$

Here, $o(h)$ is a shorthand notation for a function, $g(h)$ say, for which $g(h)/h$ tends to zero when h tends to zero. Letting h tend to 0 we get the set of differential equations

$$\begin{aligned} p_0'(t) &= -\lambda p_0(t) + \mu p_1(t), \\ p_k'(t) &= \lambda p_{k-1}(t) - (\lambda+\mu)p_k(t) + \mu p_{k+1}(t), q \quad k \geq 1, \end{aligned}$$

One may formally prove that if t tends to infinity, $p_k'(t)$ tends to 0 and $p_k(t)$ converges to p_k. From this we conclude that the limiting probabilities p_k satisfy the equations

$$0 = -\lambda p_0 + \mu p_1, \tag{1}$$

$$0 = \lambda p_{k-1} - (\lambda+\mu)p_k + \mu p_{k+1}, \qquad k = 1, 2 \ldots, \tag{2}$$

Clearly, the probabilities p_k also satisfy

$$\sum_{k=0}^{\infty} p_k = 1, \tag{3}$$

which is called the normalization equation.

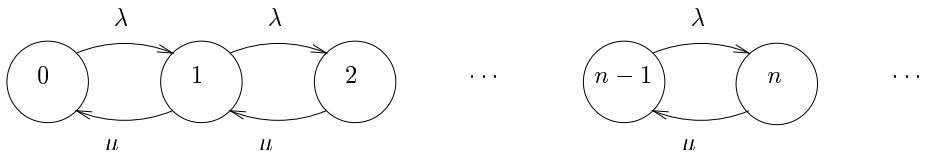

Figure 1: Flow diagram for the $M/M/1$ model

It is also possible to derive the equations (1) and (2) directly from a *flow diagram* as shown in figure 1. The arrows indicate possible transitions. The rate at which a transition occurs is λ for a transition from k to $k+1$ (an arrival) and μ for a transition from $k+1$ to k (a departure). The number of transitions per unit time from k to $k+1$, which is also called the *flow* from k to $k+1$, is equal to p_k, the fraction of time the system is in state k, times λ, the rate at which arrivals occur while the system is in state k. The equilibrium equations (1) and (2) follow by equating the flow out of state k and the flow into state k.

Remark 4.1 The use of flow diagrams is quite intuitive, but also quite risky. It is absolutely necessary that the states give a *complete description* of the state of the production process. As we discussed above, the number of jobs in the system is usually not the full state description. If e.g., the processing times are not exponential, then the state description has to contain the remaining processing time, and thus a simple flow diagram is not available.

A further simplification of the equations can be obtained by using the typical transition structure in the system. State changes are always from k to $k+1$ or from k to $k-1$.

Equating the mean number of transitions out of the set $\{0, 1, ..., k\}$ to the mean number of transitions into that set, we get

$$p_k\lambda = p_{k+1}\mu, \qquad k \geq 0.$$

Using the notation $\rho = \lambda/\mu$ for the fraction of time the machine is producing, we can rewrite this equation as

$$p_{k+1} = \rho p_k, \qquad k \geq 0.$$

From this, we obtain $p_k = \rho^k p_0$ and, by setting the sum of the probabilities equal to 1,

$$p_k = (1-\rho)\rho^k, \qquad k \geq 0.$$

From now on, we will call this the *equilibrium distribution* of the system.

4.2 Performance characteristics

From the equilibrium distribution we can compute the most important performance measures, such as the mean number of jobs in the system, denoted by $E(L)$, and the mean throughput time or system time, denoted by $E(S)$. The first one is easily obtained as

$$E(L) = \sum_{k=0}^{\infty} kp_k = \sum_{k=0}^{\infty} k(1-\rho)\rho^k = \frac{\rho}{1-\rho}.$$

As we see, if ρ, the load of the system or utilization, approaches 1 the mean number of jobs in the system goes to infinity. For example, if the load is 0.95 and the mean job size is 4 hours then the mean amount of work in the system is equal to 76 hours! This dramatic behaviour is caused by the variation in the arrival and service process and it is characteristic for almost every queueing system. Equally, or even more important, is the mean throughput time. From Little's formula we get

$$E(S) = E(L)/\lambda = \frac{\rho}{1-\rho}\frac{1}{\lambda} = \frac{1}{1-\rho}\frac{1}{\mu}.$$

So, we see that the behaviour of $E(S)$ is similar to the behaviour of $E(L)$, when ρ approaches 1. For $\rho = 0.95$, the mean throughput time is 20 times as big as the mean processing time.

4.3 The Mean Value Approach

There is another way to directly compute the mean number of jobs in the system $E(L)$ and the mean throughput time $E(S)$, without knowing the probabilities p_k. This approach uses three important properties. The first one is Little's formula, the second one is the PASTA property and the third one is the fact that if a job has an exponentially distributed production time, then the residual production time of the job in service on an arrival instant is again exponential.

Based on the PASTA property we know that the mean number of jobs in the system seen at an arrival instant of a job equals $E(L)$. Furthermore, by the third property, each of them (also the one in production) has a (residual) production time with mean $1/\mu$. Finally, the throughput time of a job also includes its own production time. Hence,

$$E(S) = E(L)\frac{1}{\mu} + \frac{1}{\mu} .$$

This relation is known as the *arrival relation*. Together with Little's formula,

$$E(L) = \lambda E(S),$$

we have two equations from which we get

$$E(S) = \lambda E(S)\frac{1}{\mu} + \frac{1}{\mu} = \rho E(S) + \frac{1}{\mu}.$$

Thus,

$$E(S) = \frac{1}{1-\rho}\frac{1}{\mu},$$

and

$$E(L) = \frac{\rho}{1-\rho}.$$

4.4 The distribution of the throughput time

The mean value approach is, although a very powerful tool, not able to lead us to the distribution of the throughput time. We can, however, compute this distribution from the equilibrium distribution. To do so, note that if an arriving job finds k jobs in the system, then the throughput time of this job consists of $k+1$ independent exponential production times (one of which may be a residual production time). Recall that the sum of $k+1$ independent and identically distributed production times is Erlang distributed with parameters $k+1$ and μ, so with density

$$f_{k+1}(t) = \mu\frac{(\mu t)^k}{k!}e^{-\mu t}.$$

By PASTA, the probability that an arriving job finds k jobs in the system is equal to p_k. So, we get for the overall density

$$f_S(t) = \sum_{k=0}^{\infty} p_k f_{k+1}(t) = \sum_{k=0}^{\infty}(1-\rho)\rho^k\mu\frac{(\mu t)^k}{k!}e^{-\mu t} = \mu(1-\rho)e^{-\mu(1-\rho)t}.$$

Hence, the throughput time is also exponentially distributed, but with parameter $\mu(1-\rho)$. For this production system, the probability that the actual throughput time of a job is larger than a times the mean throughput time is given by

$$P[S > aE(S)] = e^{-a}.$$

Hence, throughput times of 2, 3 and even 4 times the mean throughput time are not uncommon.

4.5 Arrival and departure distribution

Let a_k denote the probability that an arriving customer finds k customers in the system and d_k the probability that a departing customer leavse behind k customers, $k \geq 0$. The probabilities a_k are called the arrival distribution, the d_k are the departure distribution. By PASTA we know that a_k is equal to the equilibrium probability p_k, so

$$a_k = p_k = (1-\rho)\rho^k, \qquad k = 0, 1, 2, \ldots$$

To determine d_k we observe the following. Let $D_k(t)$ be the number of departures in $(0, t)$ leaving behind k customers and $A_k(t)$ the number of arrivals in $(0, t)$ finding k customers in the system. Since customers arrive and leave one by one (i.e., we have no batch arrivals or batch departures) it holds for any $t \geq 0$,

$$D_k(t) = A_k(t) \pm 1.$$

Hence,

$$\lambda d_k = \lim_{t\to\infty} D_k(t)/t = \lim_{t\to\infty} A_k(t)/t = \lambda a_k,$$

so the arrival and departure distribution are the same.

4.6 The output process

We now look at the output this production system. The output rate of the machine is of course the same as the input rate, so λ. To find the distribution of the time between two departures, let us consider an arbitrary departing customer. The probability that this customer leaves behind an empty system is equal to $d_0 = 1-\rho$. Then the time till the next departure is the sum of an exponential interarrival time with mean $1/\lambda$ and an exponential service time with mean $1/\mu$. If the system is nonempty upon departure, the time till the next departure is only a service time. Hence, the density of the time till the next departure is

$$f_D(t) = (1-\rho)\frac{\lambda\mu}{\lambda-\mu}\left(e^{-\mu t} - e^{-\lambda t}\right) + \rho\mu e^{-\mu t} = \lambda e^{-\lambda t},$$

from which we see that the interdeparture time is exponentially distributed with mean $1/\lambda$. In fact it can also be shown that the interdeparture times are *independent* (see, e.g., [1, 2]). So the output of the $M/M/1$ system is again a Poisson process.

References

[1] P.J. Burke, *The output of a queuing system.* Opns. Res., 4 (1956), pp. 699–704.

[2] D. Gross, C.M. Harris, *Fundamentals of queueing theory*, Wiley, Chichester, 1985.