The mean number of customers in the queue, $E(L^q)$, can be obtained from $E(L)$ by subtracting the mean number of customers in service, so

$$E(L^q) = E(L) - \rho = \frac{\rho^2}{1-\rho}.$$

The mean waiting time, $E(W)$, follows from $E(S)$ by subtracting the mean service time (or from $E(L^q)$ by applying Little's law). This yields

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

4.4 Distribution of the sojourn time and the waiting time

It is also possible to derive the distribution of the sojourn time. Denote by L^a the number of customers in the system just before the arrival of a customer and let B_k be the service time of the kth customer. Of course, the customer in service has a residual service time instead of an ordinary service time. But these are the same, since the exponential service time distribution is memoryless. So the random variables B_k are independent and exponentially distributed with mean $1/\mu$. Then we have

$$S = \sum_{k=1}^{L^a+1} B_k. \tag{10}$$

By conditioning on L^a and using that L^a and B_k are independent it follows that

$$P(S > t) = P(\sum_{k=1}^{L^a+1} B_k > t) = \sum_{n=0}^{\infty} P(\sum_{k=1}^{n+1} B_k > t) P(L^a = n). \tag{11}$$

The problem is to find the probability that an arriving customer finds n customers in the system. PASTA states that the fraction of customers finding on arrival n customers in the system is equal to the fraction of time there are n customers in the system, so

$$P(L^a = n) = p_n = (1-\rho)\rho^n. \tag{12}$$

Substituting (12) in (11) and using that $\sum_{k=1}^{n+1} B_k$ is Erlang-$(n+1)$ distributed, yields

$$\begin{aligned} P(S > t) &= \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{(\mu t)^k}{k!} e^{-\mu t} (1-\rho)\rho^n \\ &= \sum_{k=0}^{\infty} \sum_{n=k}^{\infty} \frac{(\mu t)^k}{k!} e^{-\mu t} (1-\rho)\rho^n \\ &= \sum_{k=0}^{\infty} \frac{(\mu\rho t)^k}{k!} e^{-\mu t} \\ &= e^{-\mu(1-\rho)t}, \qquad t \geq 0. \end{aligned} \tag{13}$$

Hence, S is exponentially distributed with parameter $\mu(1-\rho)$. This result can also be obtained via the use of transforms. From (10) it follows, by conditioning on L^a, that

$$\begin{aligned}\widetilde{S}(s) &= E(e^{-sS}) \\ &= \sum_{n=0}^{\infty} P(L^a = n) E(e^{-s(B_1+\ldots+B_{n+1})}) \\ &= \sum_{n=0}^{\infty} (1-\rho)\rho^n E(e^{-sB_1}) \cdots E(e^{-sB_{n+1}}).\end{aligned}$$

Since B_k is exponentially distributed with parameter μ, we have

$$E(e^{-sB_k}) = \frac{\mu}{\mu+s}\,,$$

so

$$\widetilde{S}(s) = \sum_{n=0}^{\infty} (1-\rho)\rho^n \left(\frac{\mu}{\mu+s}\right)^{n+1} = \frac{\mu(1-\rho)}{\mu(1-\rho)+s}\,,$$

from which we can conclude that S is an exponential random variable with parameter $\mu(1-\rho)$. So, for this system, the probability that the actual sojourn time of a customer is larger than a times the mean sojourn time is given by

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

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

To find the distribution of the waiting time W, note that $S = W + B$, where the random variable B is the service time. Since W and B are independent, it follows that

$$\widetilde{S}(s) = \widetilde{W}(s) \cdot \widetilde{B}(s) = \widetilde{W}(s) \cdot \frac{\mu}{\mu+s}\,.$$

and thus,

$$\widetilde{W}(s) = \frac{(1-\rho)(\mu+s)}{\mu(1-\rho)+s} = (1-\rho)\cdot 1 + \rho \cdot \frac{\mu(1-\rho)}{\mu(1-\rho)+s}\,.$$

From the transform of W we conclude that W is with probability $(1-\rho)$ equal to zero, and with probability ρ equal to an exponential random variable with parameter $\mu(1-\rho)$. Hence

$$P(W > t) = \rho e^{-\mu(1-\rho)t}, \qquad t \geq 0. \tag{14}$$

The distribution of W can, of course, also be obtained along the same lines as (13). Note that

$$P(W > t | W > 0) = \frac{P(W > t)}{P(W > 0)} = e^{-\mu(1-\rho)t},$$

so the *conditional waiting time* $W|W > 0$ is exponentially distributed with parameter $\mu(1-\rho)$.

In table 1 we list for increasing values of ρ the mean waiting time and some waiting time probabilities. From these results we see that randomness in the arrival and service process leads to (long) waiting times and the waiting times explode as the server utilization tends to one.

ρ	$E(W)$	$P(W > t)$		
	t	5	10	20
0.5	1	0.04	0.00	0.00
0.8	4	0.29	0.11	0.02
0.9	9	0.55	0.33	0.12
0.95	19	0.74	0.58	0.35

Table 1: Performance characteristics for the $M/M/1$ with mean service time 1

Remark 4.1 *(PASTA property)*
For the present model we can also derive relation (12) directly from the flow diagram 1. Namely, the average number of customers per unit time finding on arrival n customers in the system is equal to λp_n. Dividing this number by the average number of customers arriving per unit time gives the desired fraction, so

$$P(L^a = n) = \frac{\lambda p_n}{\lambda} = p_n.$$

4.5 Priorities

In this section we consider an $M/M/1$ system serving different types of customers. To keep it simple we suppose that there are two types only, type 1 and 2 say, but the analysis can easily be extended the situation with more types of customers (as we will see later). Type 1 and type 2 customers arrive according to independent Poisson processes with rate λ_1, and λ_2 respectively. The service times of all customers are exponentially distributed with the same mean $1/\mu$. We assume that

$$\rho_1 + \rho_2 < 1,$$

where $\rho_i = \lambda_i/\mu$, i.e. the occupation rate due to type i customers. Type 1 customers are treated with priority over type 2 jobs. In the following subsections we will consider two priority rules, preemptive-resume priority and non-preemptive priority.

4.5.1 Preemptive-resume priority

In the preemptive resume priority rule, type 1 customers have absolute priority over type 2 jobs. Absolute priority means that when a type 2 customer is in service and a type 1 customer arrives, the type 2 service is interrupted and the server proceeds with the type 1 customer. Once there are no more type 1 customers in the system, the server resumes the service of the type 2 customer at the point where it was interrupted.

Let the random variable L_i denote the number of type i customers in the system and S_i the sojourn time of a type i customer. Below we will determine $E(L_i)$ and $E(S_i)$ for $i = 1, 2$.