4 The $M/M/1$ queue

In this chapter we will analyze the model with exponential interarrival times with mean $1/\lambda$, exponential service times with mean $1/\mu$ and a single server. Customers are served in order of arrival. We require that

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

since, otherwise, the queue length will explode. The quantity ρ is the fraction of time the server is working. In the following section we will first study the time-dependent behaviour of this system. After that, we consider the limiting behaviour.

4.1 Time-dependent behaviour

The exponential distribution allows for a very simple description of the state of the system at time t, namely the number of customers in the system (i.e. the customers waiting in the queue and the one being served). Neither we do have to remember when the last customer arrived nor we have to register when the last customer entered service. Since the exponential distribution is memoryless, this information does not yield a better prediction of the future.

Let $p_n(t)$ denote the probability that at time t there are n customers in the system, $n = 0, 1, \ldots$ Based on the memoryless property, we get, for $\Delta t \to 0$,

$$\begin{aligned} p_0(t + \Delta t) &= (1 - \lambda\Delta t)p_0(t) + \mu\Delta t p_1(t) + o(\Delta t), \\ p_n(t + \Delta t) &= \lambda\Delta t p_{n-1}(t) + (1 - (\lambda + \mu)\Delta t)p_n(t) + \mu\Delta t p_{n+1}(t) + o(\Delta t), \\ &\qquad n = 1, 2, \ldots \end{aligned}$$

Hence, by letting $\Delta t \to 0$, we obtain the following infinite set of differential equations for the probabilities $p_n(t)$.

$$\begin{aligned} p_0'(t) &= -\lambda p_0(t) + \mu p_1(t), \\ p_n'(t) &= \lambda p_{n-1}(t) - (\lambda + \mu)p_n(t) + \mu p_{n+1}(t), \qquad n = 1, 2, \ldots \end{aligned} \tag{1}$$

It is difficult to solve these differential equations. An explicit solution for the probabilities $p_n(t)$ can be found in [5] (see p. 77). The expression presented there is an infinite sum of modified Bessel functions. So already one of the simplest interesting queueing models leads to a difficult expression for the time-dependent behavior of its state probabilities. For more general systems we can only expect more complexity. Therefore, in the remainder we will focus on the *limiting or equilibrium behavior* of this system, which appears to be much easier to analyse.

4.2 Limiting behavior

One may show that as $t \to \infty$, then $p'_n(t) \to 0$ and $p_n(t) \to p_n$ (see e.g. [3]). Hence, from (1) it follows that the limiting or equilibrium probabilities p_n satisfy the equations

$$0 = -\lambda p_0 + \mu p_1, \tag{2}$$
$$0 = \lambda p_{n-1} - (\lambda + \mu) p_n + \mu p_{n+1}, \qquad n = 1, 2, \ldots \tag{3}$$

Clearly, the probabilities p_n also satisfy

$$\sum_{n=0}^{\infty} p_n = 1, \tag{4}$$

which is called the normalization equation. It is also possible to derive the equations (2) and (3) directly from a *flow diagram*, as shown in figure 1.

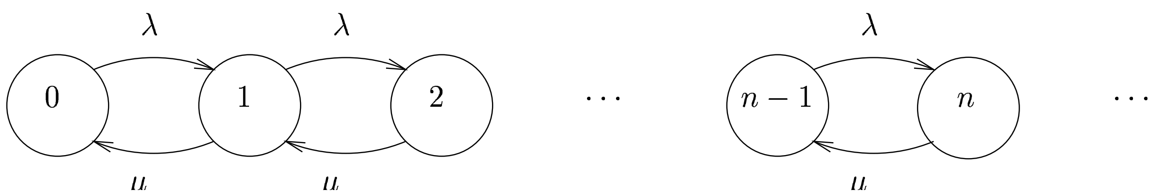

Figure 1: Flow diagram for the $M/M/1$ model

The arrows indicate possible transitions. The rate at which a transition occurs is λ for a transition from n to $n+1$ (an arrival) and μ for a transition from $n+1$ to n (a departure). The number of transitions per unit time from n to $n+1$, which is also called the *flow* from n to $n+1$, is equal to p_n, the fraction of time the system is in state n, times λ, the rate at arrivals occur while the system is in state n. The equilibrium equations (2) and (3) follow by equating the flow out of state n and the flow into state n.

For this simple model there are many ways to determine the solution of the equations (2)–(4). Below we discuss several approaches.

4.2.1 Direct approach

The equations (3) are a second order recurrence relation with constant coefficients. Its general solution is of the form

$$p_n = c_1 x_1^n + c_2 x_2^n, \qquad n = 0, 1, 2, \ldots \tag{5}$$

where x_1 and x_2 are roots of the quadratic equation

$$\lambda - (\lambda + \mu)x + \mu x^2 = 0.$$

This equation has two zeros, namely $x = 1$ and $x = \lambda/\mu = \rho$. So all solutions to (3) are of the form

$$p_n = c_1 + c_2 \rho^n, \qquad n = 0, 1, 2, \ldots$$

Equation (4), stating that the sum of all probabilities is equal to 1, of course directly implies that c_1 must be equal to 0. That c_1 must be equal to 0 also follows from (2) by substituting the solution (5) into (2).

The coefficient c_2 finally follows from the normalization equation (4), yielding that $c_2 = 1 - \rho$. So we can conclude that

$$p_n = (1-\rho)\rho^n, \qquad n = 0, 1, 2, \ldots \tag{6}$$

Apparently, the equilibrium distribution depends upon λ and μ only through their ratio ρ.

4.2.2 Recursion

One can use (2) to express p_1 in p_0 yielding

$$p_1 = \rho p_0.$$

Substitution of this relation into (3) for $n = 1$ gives

$$p_2 = \rho^2 p_0.$$

By substituting the relations above into (3) for $n = 2$ we obtain p_3, and so on. Hence we can recursively express all probabilities in terms of p_0, yielding

$$p_n = \rho^n p_0, \qquad n = 0, 1, 2, \ldots.$$

The probability p_0 finally follows from the normalization equation (4).

4.2.3 Generating function approach

The probability generating function of the random variable L, the number of customers in the system, is given by

$$P_L(z) = \sum_{n=0}^{\infty} p_n z^n, \tag{7}$$

which is properly defined for z with $|z| \le 1$. By multiplying the nth equilibrium equation with z^n and then summing the equations over all n, the equilibrium equations for p_n can be transformed into the following single equation for $P_L(z)$,

$$0 = \mu p_0(1 - z^{-1}) + (\lambda z + \mu z^{-1} - (\lambda + \mu))P_L(z).$$

The solution of this equation is

$$P_L(z) = \frac{p_0}{1-\rho z} = \frac{1-\rho}{1-\rho z} = \sum_{n=0}^{\infty}(1-\rho)\rho^n z^n, \tag{8}$$

where we used that $P(1) = 1$ to determine $p_0 = 1 - \rho$. Hence, by equating the coefficients of z^n in (7) and (8) we retrieve the solution (6).

4.2.4 Global balance principle

The global balance principle states that for *each set of states* A, the flow out of set A is equal to the flow into that set. In fact, the equilibrium equations (2)–(3) follow by applying this principle to a single state. But if we apply the balance principle to the set $A = \{0, 1, \ldots, n-1\}$ we get the very simple relation

$$\lambda p_{n-1} = \mu p_n, \qquad n = 1, 2, \ldots$$

Repeated application of this relation yields

$$p_n = \rho^n p_0, \qquad n = 0, 1, 2, \ldots$$

so that, after normalization, the solution (6) follows.

4.3 Mean performance measures and the mean value approach

From the equilibrium probabilities we can derive expressions for the mean number of customers in the system and the mean time spent in the system. For the first one we get

$$E(L) = \sum_{n=0}^{\infty} n p_n = \frac{\rho}{1-\rho},$$

and by applying Little's law,

$$E(S) = \frac{1/\mu}{1-\rho}. \tag{9}$$

If we look at the expressions for $E(L)$ and $E(S)$ we see that both quantities grow to infinity as ρ approaches unity. The dramatic behavior is caused by the variation in the arrival and service process. This type of behavior with respect to ρ is characteristic for almost every queueing system.

In fact, $E(L)$ and $E(S)$ can also be determined directly, i.e. without knowing the probabilities p_n, by combining Little's law and the PASTA property. Based on PASTA we know that the average number of customers in the system seen by an arriving customer equals $E(L)$ and each of them (also the one in service) has a (residual) service time with mean $1/\mu$. The customer further has to wait for its own service time. Hence

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

This relation is known as the *arrival relation*. Together with

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

we find expression (9). This approach is called the *mean value approach*.

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 $i = 1, 2$.

For type 1 customers the type 2 customers do not exist. Hence we immediately have

$$E(S_1) = \frac{1/\mu}{1-\rho_1}, \qquad E(L_1) = \frac{\rho_1}{1-\rho_1}. \tag{15}$$

Since the (residual) service times of all customers are exponentially distributed with the same mean, the total number of customers in the system does not depend on the order in which the customers are served. So this number is the same as in the system where all customers are served in order of arrival. Hence,

$$E(L_1) + E(L_2) = \frac{\rho_1 + \rho_2}{1 - \rho_1 - \rho_2}, \tag{16}$$

and thus, inserting (15),

$$E(L_2) = \frac{\rho_1 + \rho_2}{1 - \rho_1 - \rho_2} - \frac{\rho_1}{1-\rho_1} = \frac{\rho_2}{(1-\rho_1)(1-\rho_1-\rho_2)},$$

and applying Little's law,

$$E(S_2) = \frac{E(L_2)}{\lambda_2} = \frac{1/\mu}{(1-\rho_1)(1-\rho_1-\rho_2)}.$$

Example 4.2 For $\lambda_1 = 0.2$, $\lambda_2 = 0.6$ and $\mu = 1$, we find in case all customers are treated in order of arrival,

$$E(S) = \frac{1}{1-0.8} = 5,$$

and in case type 1 customers have absolute priority over type 2 jobs,

$$E(S_1) = \frac{1}{1-0.2} = 1.25, \qquad E(S_2) = \frac{1}{(1-0.2)(1-0.8)} = 6.25.$$

4.5.2 Non-preemptive priority

We now consider the situation that type 1 customers have nearly absolute priority over type 2 jobs. The difference with the previous rule is that type 1 customers are not allowed to interrupt the service of a type 2 customers. This priority rule is therefore called *non-preemptive*.

For the mean sojourn time of type 1 customers we find

$$E(S_1) = E(L_1)\frac{1}{\mu} + \frac{1}{\mu} + \rho_2\frac{1}{\mu}.$$

The last term reflects that when an arriving type 1 customer finds a type 2 customer in service, he has to wait until the service of this type 2 customer has been completed. According to PASTA the probability that he finds a type 2 customer in service is equal to the fraction of time the server spends on type 2 customers, which is ρ_2. Together with Little's law,

$$E(L_1) = \lambda_1 E(S_1),$$

we obtain

$$E(S_1) = \frac{(1+\rho_2)/\mu}{1-\rho_1}, \qquad E(L_1) = \frac{(1+\rho_2)\rho_1}{1-\rho_1}.$$

For type 2 customers it follows from (16) that

$$E(L_2) = \frac{(1-\rho_1(1-\rho_1-\rho_2))\rho_2}{(1-\rho_1)(1-\rho_1-\rho_2)},$$

and applying Little's law,

$$E(S_2) = \frac{(1-\rho_1(1-\rho_1-\rho_2))/\mu}{(1-\rho_1)(1-\rho_1-\rho_2)}.$$

Example 4.3 For $\lambda_1 = 0.2$, $\lambda_2 = 0.6$ and $\mu = 1$, we get

$$E(S_1) = \frac{1+0.6}{1-0.2} = 2, \qquad E(S_2) = \frac{1-0.2(1-0.8)}{(1-0.2)(1-0.8)} = 6.$$

4.6 Busy period

In a servers life we can distinguish *cycles*. A cycle is the time that elapses between two consecutive arrivals finding an empty system. Clearly, a cycle starts with a *busy period* BP during which the server is helping customers, followed by an *idle period* IP during which the system is empty.

Due to the memoryless property of the exponential distribution, an idle period IP is exponentially distributed with mean $1/\lambda$. In the following subsections we determine the mean and the distribution of a busy period BP.

4.6.1 Mean busy period

It is clear that the mean busy period divided by the mean cycle length is equal to the fraction of time the server is working, so

$$\frac{E(BP)}{E(BP)+E(IP)} = \frac{E(BP)}{E(BP)+1/\lambda} = \rho.$$

Hence,

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

4.6.2 Distribution of the busy period

Let the random variable C_n be the time till the system is empty again if there are now n customers present in the system. Clearly, C_1 is the length of a busy period, since a busy period starts when the first customer after an idle period arrives and it ends when the system is empty again. The random variables C_n satisfy the following recursion relation. Suppose there are $n(>0)$ customers in the system. Then the next event occurs after an exponential time with parameter $\lambda+\mu$: with probability $\lambda/(\lambda+\mu)$ a new customer arrives, and with probability $\mu/(\lambda+\mu)$ service is completed and a customer leaves the system. Hence, for $n = 1, 2, \ldots$,

$$C_n = X + \begin{cases} C_{n+1} & \text{with probability } \lambda/(\lambda+\mu), \\ C_{n-1} & \text{with probability } \mu/(\lambda+\mu), \end{cases} \tag{17}$$

where X is an exponential random variable with parameter $\lambda+\mu$. From this relation we get for the Laplace-Stieltjes transform $\widetilde{C}_n(s)$ of C_n that

$$\widetilde{C}_n(s) = \frac{\lambda+\mu}{\lambda+\mu+s}\left(\widetilde{C}_{n+1}(s)\frac{\lambda}{\lambda+\mu} + \widetilde{C}_{n-1}(s)\frac{\mu}{\lambda+\mu}\right),$$

and thus, after rewriting,

$$(\lambda+\mu+s)\widetilde{C}_n(s) = \lambda\widetilde{C}_{n+1}(s) + \mu\widetilde{C}_{n-1}(s), \qquad n = 1, 2, \ldots$$

For *fixed* s this equation is a very similar to (3). Its general solution is

$$\widetilde{C}_n(s) = c_1 x_1^n(s) + c_2 x_2^n(s), \qquad n = 0, 1, 2, \ldots$$

where $x_1(s)$ and $x_2(s)$ are the roots of the quadratic equation

$$(\lambda+\mu+s)x = \lambda x^2 + \mu,$$

satisfying $0 < x_1(s) \le 1 < x_2(s)$. Since $0 \le \widetilde{C}_n(s) \le 1$ it follows that $c_2 = 0$. The coefficient c_1 follows from the fact that $C_0 = 0$ and hence $\widetilde{C}_0(s) = 1$, yielding $c_1 = 1$. Hence we obtain

$$\widetilde{C}_n(s) = x_1^n(s),$$

and in particular, for the Laplace-Stieltjes transform $\widetilde{BP}(s)$ of the busy period BP, we find

$$\widetilde{BP}(s) = \widetilde{C}_1(s) = x_1(s) = \frac{1}{2\lambda}\left(\lambda+\mu+s-\sqrt{(\lambda+\mu+s)^2-4\lambda\mu}\right).$$

By inverting this transform (see e.g. [1]) we get for the density $f_{BP}(t)$ of BP,

$$f_{BP}(t) = \frac{1}{t\sqrt{\rho}}e^{-(\lambda+\mu)t}I_1(2t\sqrt{\lambda\mu}), \qquad t > 0,$$

where $I_1(\cdot)$ denotes the modified Bessel function of the first kind of order one, i.e.

$$I_1(x) = \sum_{k=0}^{\infty} \frac{(x/2)^{2k+1}}{k!(k+1)!} .$$

In table 2 we list for some values of ρ the probability $P(BP > t)$ for a number of t values. If you think of the situation that $1/\mu$ is one hour, then 10% of the busy periods lasts longer than 2 days (16 hours) and 5% percent even longer than 1 week, when $\rho = 0.9$. Since the mean busy period is 10 hours in this case, it is not unlikely that in a month time a busy period longer than a week occurs.

ρ		$P(BP > t)$						
	t	1	2	4	8	16	40	80
0.8		0.50	0.34	0.22	0.13	0.07	0.02	0.01
0.9		0.51	0.36	0.25	0.16	0.10	0.05	0.03
0.95		0.52	0.37	0.26	0.18	0.12	0.07	0.04

Table 2: Probabilities for the busy period duration for the $M/M/1$ with mean service time equal to 1

4.7 Arrival and departure distribution

Let L^a be the number of customers in the system just before the arrival of a customer and let L^d be the number of customers in the system just after a departure. By PASTA we know that

$$P(L^a = n) = p_n = (1-\rho)\rho^n, \qquad n = 0, 1, 2, \ldots$$

To determine $P(L^d = n)$ we observe the following. Let $D_n(t)$ be the number of departures in $(0, t)$ leaving behind n customers and $A_n(t)$ the number of arrivals in $(0, t)$ finding n 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 $t \geq 0$,

$$D_n(t) = A_n(t) \pm 1.$$

Hence,

$$\lambda P(L^d = n) = \lim_{t\to\infty} D_n(t)/t = \lim_{t\to\infty} A_n(t)/t = \lambda P(L^a = n),$$

so the arrival and departure distribution are exactly the same.

4.8 The output process

We now look at the output process of the $M/M/1$ system. The output rate 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 $P(L^d = 0) = 1 - \rho$. 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., [2, 4]). So the output of the $M/M/1$ system is again a Poisson process.

References

[1] M. Abramowitz, I.A. Stegun, *Handbook of mathematical functions*, Dover, 1965.

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

[3] J.W. Cohen, *The single server queue*, North-Holland, Amsterdam, 1982.

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

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