

Chapter 2

Markovian Systems

The common characteristic of all *markovian systems* is that all interesting distributions, namely the distribution of the interarrival times and the distribution of the service times are exponential distributions and thus exhibit the markov (memoryless) property. From this property we have two important conclusions:

- The state of the system can be summarized in a single variable, namely the number of customers in the system. (If the service time distribution is not memoryless, this is not longer true, since not only the number of customers in the system is needed, but also the remaining service time of the customer in service.)
- Markovian systems can be directly mapped to a *continuous time markov chain* (CTMC) which can then be solved.

In this chapter we will often proceed as follows: deriving a CTMC and solve it by inspection or simple numerical techniques.

2.1 The M/M/1-Queue

The M/M/1-Queue has iid interarrival times, which are exponentially distributed with parameter λ and also iid service times with exponential distribution with parameter μ . The system has only a single server and uses the FIFO service discipline. The waiting line is of infinite size. This section is mainly based on [9, chapter 3].

It is easy to find the underlying markov chain. As the system state we use the number of customers in the system. The M/M/1 system is a pure birth-/death system, where at any point in time at most one event occurs, with an event either being the arrival of a new customer or the completion of a customer's service. What makes the M/M/1 system really simple is that the arrival rate and the service rate are not state-dependent. The state-transition-rate diagram of the underlying CTMC is shown in figure 2.1.

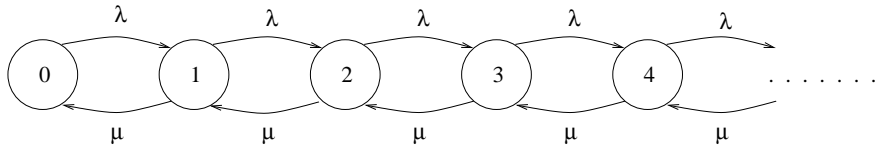


Figure 2.1: CTMC for the M/M/1 queue

2.1.1 Steady-State Probabilities

We denote the steady state probability that the system is in state k ($k \in \mathbb{N}$) by p_k , which is defined by

$$p_k := \lim_{t \rightarrow \infty} P_k(t)$$

where $P_k(t)$ denotes the (time-dependent) probability that there are k customers in the system at time t . Please note that the steady state probability p_k does not depend on t . We focus on a fixed state k and look at the *flows* into the state and out of the state. The state k can be reached from state $k - 1$ and from state $k + 1$ with the respective rates $\lambda P_{k-1}(t)$ (the system is with probability $P_{k-1}(t)$ in the state $k - 1$ at time t and goes with the rate λ from the predecessor state $k - 1$ to state k) and $\mu P_{k+1}(t)$ (the same from state $k + 1$). The total flow into the state k is then simply $\lambda P_{k-1}(t) + \mu P_{k+1}(t)$. The state k is left with the rate $\lambda P_k(t)$ to the state $k + 1$ and with the rate $\mu P_k(t)$ to the state $k - 1$ (for $k = 0$ there is only a flow coming from or going to state 1). The total flow out of that state is then given by $\lambda P_k(t) + \mu P_k(t)$. The total rate of change of the flow into state k is then given by the difference of the flow into that state and the flow out of that state:

$$\frac{dP_k(t)}{dt} = (\lambda P_{k-1}(t) + \mu P_{k+1}(t)) - (\lambda P_k(t) + \mu P_k(t)),$$

, however, in the limit ($t \rightarrow \infty$) we require

$$\frac{dP_k(t)}{dt} = 0$$

so we arrive at the following steady-state flow equations:

$$\begin{aligned} 0 &= \mu p_1 - \lambda p_0 \\ 0 &= \lambda p_0 + \mu p_2 - \lambda p_1 - \mu p_1 \\ 0 &= \dots \\ 0 &= \lambda p_{k-1} + \mu p_{k+1} - \lambda p_k - \mu p_k \\ 0 &= \dots \end{aligned}$$

These equations can be recursively solved in dependence of p_0 :

$$p_k = \left(\frac{\lambda}{\mu}\right)^k p_0$$

Furthermore, since the p_k are probabilities, the *normalization condition*

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

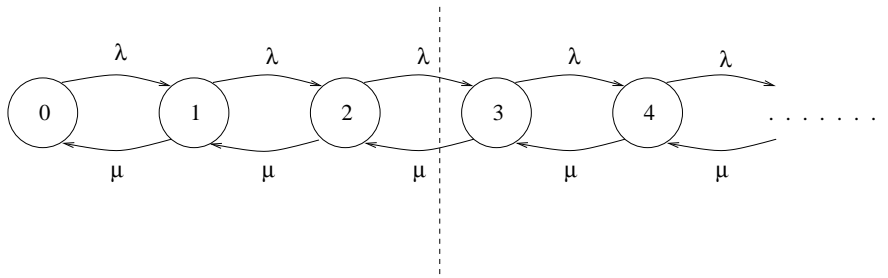


Figure 2.2: CTMC for the M/M/1 queue

says that

$$1 = p_0 + \sum_{k=1}^{\infty} p_k = p_0 + \sum_{k=1}^{\infty} p_0 \left(\frac{\lambda}{\mu}\right)^k = p_0 \left(\sum_{k=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^k\right) = p_0 \frac{1}{1 - \frac{\lambda}{\mu}}$$

which gives

$$p_0 = 1 - \frac{\lambda}{\mu} =: 1 - \rho \tag{2.1}$$

To summarize the results, the **steady state probabilities** of the M/M/1 markov chain are given by

$$p_0 = 1 - \frac{\lambda}{\mu} \tag{2.2}$$

$$p_k = \left(\frac{\lambda}{\mu}\right)^k p_0 \tag{2.3}$$

Obviously, in order for p_0 to exist it is required that $\lambda < \mu$, otherwise the series will diverge. This is the *stability condition* for the M/M/1 system. It makes also sense intuitively: when more customers arrive than the system can serve, the queue size goes to infinity.

A second derivation making use of the flow approach is the following: in the steady state we can draw a line into the CTMC as in figure 2.2 and we argue, that in the steady state the following principle holds: the flow from the left side to the right side equals the flow from the right side to the left side. Transforming this into flow equations yields:

$$\begin{aligned} \lambda p_0 &= \mu p_1 \\ \lambda p_1 &= \mu p_2 \\ \dots &= \dots \\ \lambda p_{k-1} &= \mu p_k \\ \dots &= \dots \end{aligned}$$

This approach can be solved using the same techniques as above.

The just outlined method of deriving a CTMC and solving the flow equations for the steady state probabilities can be used for most markovian systems.

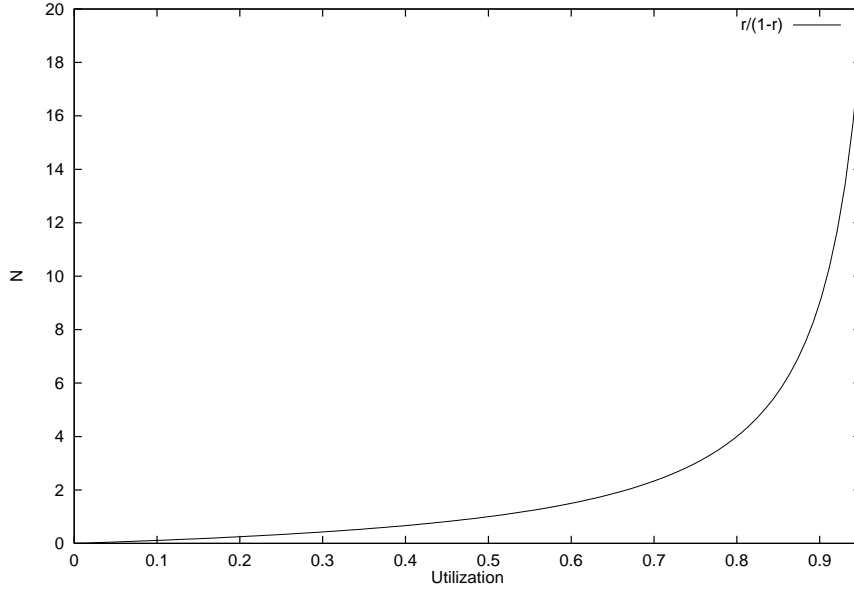


Figure 2.3: Mean Number of Customers vs. Utilization

2.1.2 Some Performance Measures

Utilization

The utilization gives the fraction of time that the server is busy. In the M/M/1 case this is simply the complementary event to the case where the system is empty. The utilization can be seen as the steady state probability that the system is not empty at any time in the steady state, thus

$$\text{Utilization} := 1 - p_0 = \rho \quad (2.4)$$

Mean number of customers in the system

The mean number of customers in the system is given by

$$\bar{N} = E[N] = \sum_{k=0}^{\infty} k p_k = p_0 \left(\sum_{k=0}^{\infty} k \rho^k \right) = (1 - \rho) \frac{\rho}{(1 - \rho)^2} = \frac{\rho}{1 - \rho} \quad (2.5)$$

where we have used the summation

$$\sum_{k=0}^{\infty} k x^k = \frac{x}{(1 - x)^2}$$

for $|x| < 1$

The mean number of customers in the system for varying utilizations is shown in figure 2.3. As can be seen \bar{N} grows to infinity as $\rho \rightarrow 1$, thus for higher utilizations the system tends to get unstable. This trend is especially observable for utilizations of 70 % or more.

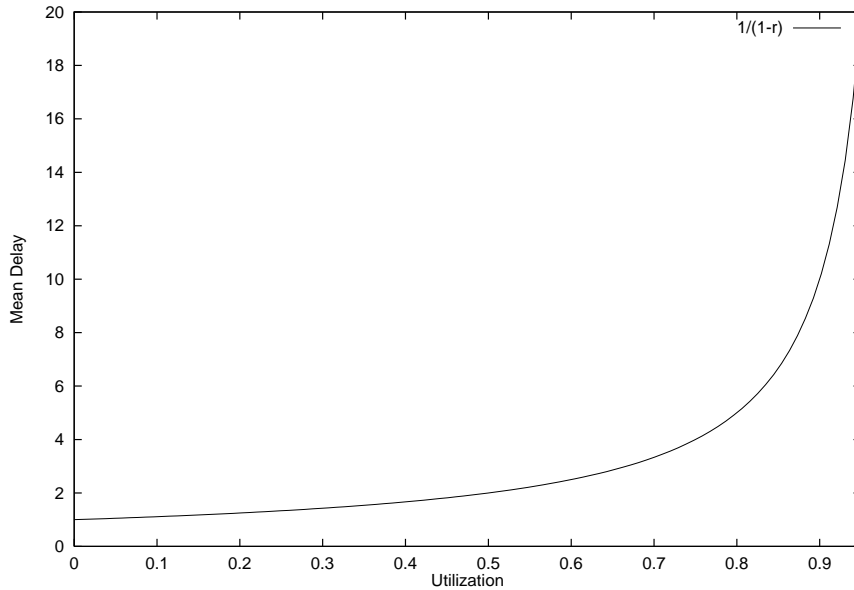


Figure 2.4: Mean Delay vs. Utilization

Mean Response Time

The mean response time T is the mean time a customer spends in the system, i.e. waiting in the queue and being serviced. We simply apply Little's law to find

$$\bar{T} = \frac{\bar{N}}{\lambda} = \frac{1/\mu}{1-\rho} = \frac{1}{\mu-\lambda} \quad (2.6)$$

For the case of $\mu = 1$ the mean response time (mean delay) of a customer is shown in figure 2.4 (for $\mu = 1$). This curve shows a behaviour similar to the one for the mean number of customers in the system.

Tail Probabilities

In applications often the following question arises: we assume that we have an M/M/1 system, however, we need to restrict the number of customers in the system to a finite quantity. If a customer arrives at a full system, it is lost. We want to determine the size of the waiting line that is required to lose customers only with a small probability. As an example consider e.g. a router for which the buffer space is finite and packets should be lost with probability 10^{-6} . In principle this is a M/M/1/N queue, however, we use an M/M/1 queue (with infinite waiting room) as an approximation. We are now interested in the probability that the system has k or more customers (the probability $\Pr[N > k]$ is called a *tail probability*) and thus would lose a customer in reality. We have

$$\Pr[N > k] = 1 - \Pr[N \leq k] = 1 - \sum_{\nu=0}^k p_{\nu} = 1 - p_0 \frac{1 - \rho^{k+1}}{1 - \rho} = \rho^{k+1} \quad (2.7)$$

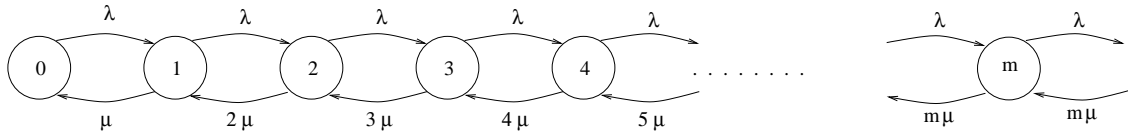


Figure 2.5: CTMC for the M/M/m queue

2.2 The M/M/m-Queue

The M/M/m-Queue ($m > 1$) has the same interarrival time and service time distributions as the M/M/1 queue, however, there are m servers in the system and the waiting line is infinitely long. As in the M/M/1 case a complete description of the system state is given by the number of customers in the system (due to the memoryless property). The state-transition-rate diagram of the underlying CTMC is shown in figure 2.5. The M/M/m system is also a pure birth-death system.

2.2.1 Steady-State Probabilities

Using the above sketched technique of evaluating the flow equations together with the well-known geometric summation yields the following steady state probabilities:

$$p_0 = \left[\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \left(\frac{(m\rho)^m}{m!} \right) \left(\frac{1}{1-\rho} \right) \right]^{-1} \quad (2.8)$$

$$p_k = \begin{cases} p_0 \frac{(m\rho)^k}{k!} & : k \leq m \\ p_0 \frac{\rho^k m^m}{m!} & : k \geq m \end{cases} \quad (2.9)$$

with $\rho = \frac{\lambda}{\mu}$ and clearly assuming that $\rho < 1$.

2.2.2 Some Performance Measures

Mean number of customers in the system

The mean number of customers in the system is given by

$$\bar{N} = E[N] = \sum_{k=0}^{\infty} k p_k = m\rho + \rho \frac{(m\rho)^m}{m!} \frac{p_0}{(1-\rho)^2} \quad (2.10)$$

The mean response time again can be evaluated simply using Little's formula.

For the case of $M=10$ we show the mean number of customers in the system for varying ρ in figure 2.6.

Queueing Probability

We want to evaluate the probability that an arriving customer must enter the waiting line because there is currently no server available. This is often used in telephony and denotes the probability that a newly arriving call at a central office will get no trunk, given that the interarrival times and service times (call durations) are exponentially distributed (in "real life" it is not so easy to justify

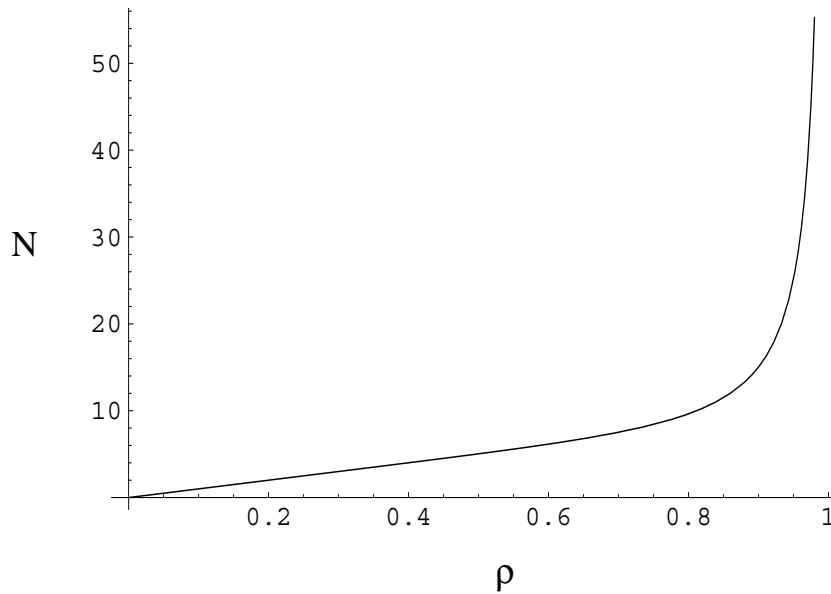


Figure 2.6: Mean Number of Customers in the system for the M/M/10-Queue

this assumption). This probability can be calculated as follows:

$$\Pr[\text{Queueing}] = \sum_{k=m}^{\infty} p_k = \sum_{k=m}^{\infty} p_0 \frac{(m\rho)^k}{m!} \frac{1}{m^{k-m}} = \frac{\left(\frac{(m\rho)^m}{m!}\right) \left(\frac{1}{1-\rho}\right)}{\left[\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \left(\frac{(m\rho)^m}{m!}\right) \left(\frac{1}{1-\rho}\right)\right]} \quad (2.11)$$

and is often denoted as *Erlangs C Formula*, abbreviated with $C(m, \rho)$

2.3 The M/M/1/K-Queue

The M/M/1/K-Queue has exponential interarrival time and service time distributions, each with the respective parameters λ and μ . The customers are served in FIFO-Order, there is a single server but the system can only hold up to K customers. If a new customer arrives and there are already K customers in the system the new customer is considered lost, i.e. it drops from the system and never comes back. This is often referred to as *blocking*. This behaviour is necessary, since otherwise (e.g. when the customer is waiting outside until there is a free place) the arrival process will be no longer markovian. As in the M/M/1 case a complete description of the system state is given by the number of customers in the system (due to the memoryless property). The state-transition-rate diagram of the underlying CTMC is shown in figure 2.7. The M/M/1/K system is also a pure birth-death system. This system is better suited to approximate “real systems” (like e.g. routers) since buffer space is always finite.

2.3.1 Steady-State Probabilities

One can again using the technique based on evaluation of the flow equations to arrive at the steady state probabilities p_k . However, since the number of customers in the system is limited, the arrival process is state dependent: if there are fewer than K customers in the system the arrival rate is λ ,

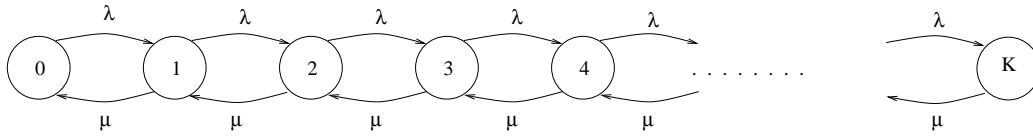


Figure 2.7: CTMC for the M/M/1/K queue

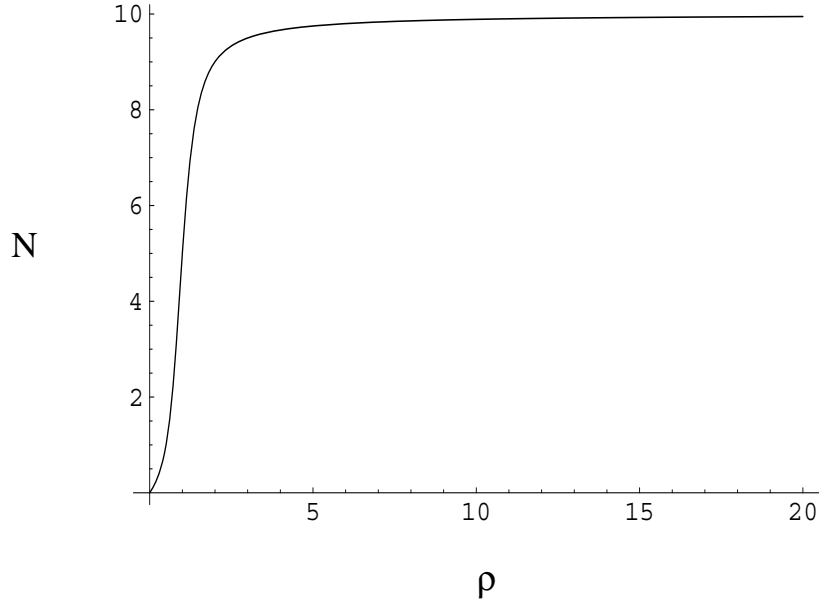


Figure 2.8: Mean number of Customers in the system for M/M/1/10-queue

otherwise the arrival rate is 0. It is then straightforward to see that the steady state probabilities are given by:

$$p_0 = \frac{1 - \rho}{1 - \rho^{K+1}} \quad (2.12)$$

$$p_k = p_0 \rho^k \quad (2.13)$$

where $1 \leq k \leq K$ and again $\rho = \frac{\lambda}{\mu}$ holds. It is interesting to note that the system is stable even for $\rho > 1$

2.3.2 Some Performance Measures

Mean number of customers in the system

The mean number of customers in the system is given by

$$\bar{N} = E[N] = \sum_{k=0}^K k p_k = \dots = \begin{cases} \frac{\rho}{1-\rho} - \frac{K+1}{1-\rho^{K+1}} \rho^{K+1} & : \rho \neq 1 \\ \frac{K}{2} & : \rho = 1 \end{cases} \quad (2.14)$$

The mean number of customers in the system is shown in figure 2.8 for varying ρ and for $K = 10$. Please note that for this queue ρ can be greater than one while the queueing system remains stable.

The mean response time again can be evaluated simply using Little's formula.

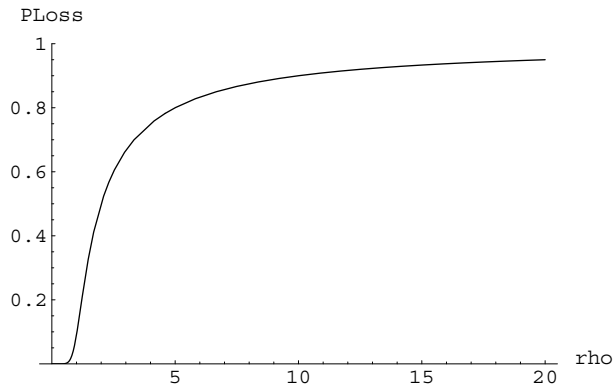


Figure 2.9: Loss Probability for M/M/1/20

Loss Probability

The loss probability is simply the probability that an arriving customer finds the system full, i.e. the loss probability is given as p_K with

$$p_{Loss} := p_K = \begin{cases} \frac{\rho^K - \rho^{K+1}}{1 - \rho^{K+1}} & : \rho \neq 1 \\ \frac{1}{K+1} & : \rho = 1 \end{cases} \quad (2.15)$$

For the case of 10 servers the loss probability for varying ρ is shown in figure 2.9

In section 2.1 we have considered the problem of dimensioning a router's buffer such that customers are lost only with a small probability and used the M/M/1 queue as an approximation, where an M/M/1/K queue with unknown K may be more appropriate. However, it is not possible to solve equation 2.15 algebraically for K when p_{Loss} is given (at least if no special functions like LambertW [1] are used).

2.4 A comparison of different Queueing Systems

In this section we want to compare three different systems in terms of mean response time (mean delay) vs. offered load: a single M/M/1 server with the service rate $m\mu$, a M/M/m system and a system where m queues of M/M/1 type with service rate μ are in parallel, such that every customer enters each system with the same probability.

The answer to this question can give some hints on proper decisions in scenarios like the following: given a computer with a processor of type X and given a set of users with long-running number cruncher programs. These users are all angry because they need to wait so long for their results. So the management decides that the computer should be upgraded. There are three possible options:

- buy $n - 1$ additional processors of type X and plug these into the single machine, thus yielding a multiprocessor computer
- buy a new processor of type Y, which is n times stronger than processor X and replacing it, and let all users work on that machine
- provide each user with a separate machine carrying a processor of type X, without allowing other users to work on this machine

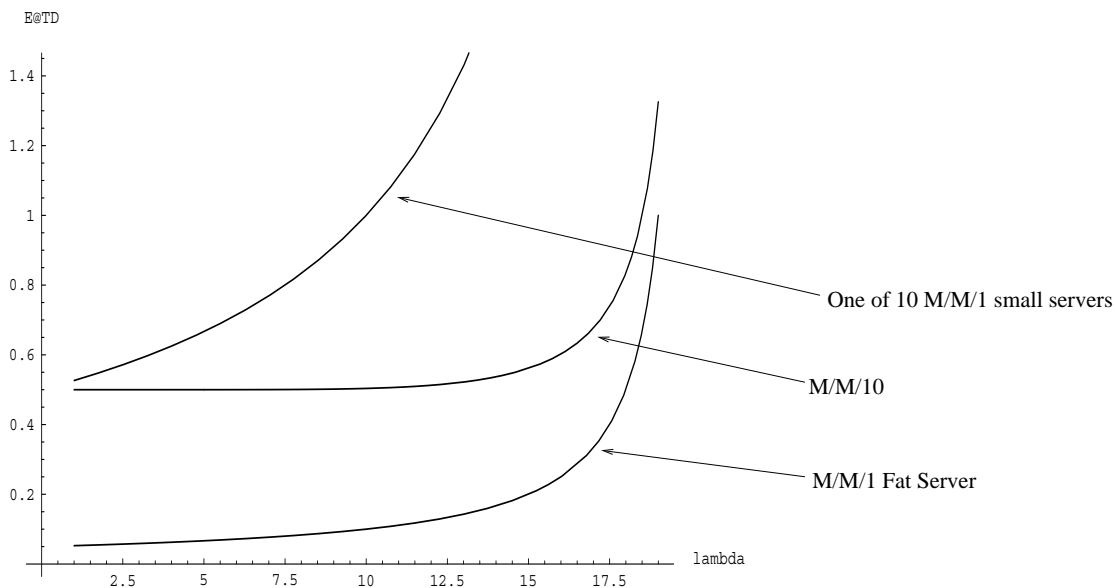


Figure 2.10: Mean Response Times for three different systems

We show that the second solution yields the best results (smallest mean delays), followed by the first solution, while the last one is the worst solution. The first system corresponds to an $M/M/m$ system, where each server has the service rate μ and the arrival rate to the system is λ . The second system corresponds to an $M/M/1$ system with arrival rate λ and service rate $m \cdot \mu$. And, from the view of a single user, the last system corresponds to an $M/M/1$ system with arrival rate λ/m and service rate μ . The mean response times for $m = 10$ and $\mu = 2$ are for varying λ shown in figure 2.10.

An intuitive explanation for the behaviour of the systems is the following: in the case of 10 parallel $M/M/1$ queues there is always a nonzero probability that some servers have many customers in their queues while other servers are idle. In contrast to that, in the $M/M/m$ case this cannot happen. In addition to that, the fat single server is especially for lighter loads better than the $M/M/10$ system, since if there are only $k < 10$ customers in the system the $M/M/10$ system has a smaller overall service rate $k \cdot \mu$, while in the fat server all customers are served with the full service rate of $10 \cdot \mu = 20$