Abstract. The paper examines the transient behaviour of some queueing systems by the use of simulation techniques. After summarizing experimental results obtained for M/M/1/\infty system, a new simple rule to compute the length of the transient period is given. The paper differs from the existing literature because it investigates the length of the transient behaviour of the M/M/k/0 queueing system for emergency medical services: it is confirmed that, in this case also, a decaying exponential function can well approximate the behaviour of L(t), the expected number of ambulances which are busy at time t.

Keywords: queueing transient behaviour, emergency services, allocation.

Introduction.

In the analysis of an emergency medical service, as well as in many other framework in which a queueing system occurs, it is important to know, at least approximately, when the steady state is reached (or, equivalently, when the transient period comes to an end). The earlier works on the transient behaviour of queues in literature were published in the late 1950s and in the early 1960s. One of the most important achieved results states that, independently of the system’s initial state, for many queueing systems, a queue converges to its own steady state equilibrium at a rate that can be represented by an exponential term of the form $e^{-t/\tau}$ where $\tau$ is called the ‘relaxation time’, a time constant characteristic of the queueing system. This result was confirmed numerically by Odoni and Roth in 1983 ([5]) for M/M/1/\infty systems: in the first part of this paper we report it and give a new simple rule, empirically found ([3]), by which we can easily compute the length of transient behaviour in such systems (section 1).

In [5] much of the state of the art on the transient behaviour is also described. In [6], some results for the problems of correctly starting and stopping simulations to compute the length of the transient period, until the required accuracy of results is reached, are summarised, but with reference to systems with infinite capacity and only one server.

More recently, many authors studied the behaviour of the transient period also in M/M/k queueing systems, where $k$ is the number of servers available in a fixed region, but focusing on infinite capacity systems. Roth, for example, in [7] estimates the steady state expected number of customers by the ‘relaxation time heuristic’ technique.

On the other hand, for the M/M/\infty system, i.e. when servers are always available for a new call, the transient effects disappear (see [7]).

The main purpose of this paper is to investigate the transient behaviour in queueing systems related to emergency services, i.e. in M/M/k/0 models. The real life case which originated the problem is described in details in [1]; here it is enough to say that in such services no queue is allowed, due to the nature itself of the problem and to the fact that, when available, servers are always
dispatched immediately to any calls; so, if all ambulances are busy, demands remain unfulfilled (otherwise, they can be satisfied by extraordinary carriers, such as helicopters).

\( Q(t) \), i.e. the expected number of customers in queue at time \( t \), and \( L(t) \), i.e. the expected number of ambulances busy at time \( t \), are respectively investigated. It is easily seen that \( Q(t) \) and \( L(t) \) can vary a lot depending on the time \( t \), if we look at the start-up time period, while, after it, they become respectively very close to \( Q(\infty) \), the steady state expected queue length, and to \( L(\infty) \), the expected number of ambulances in steady state. We aim to state how long it takes for \( L(t) \) to reach \( L(\infty) \) or, equivalently, to establish how long ambulances work in a transient setting.

We first review methodology and results from the theory about the M/M/1/0 problem; then, we illustrate assumptions on the simulation model, which we use to study the behaviour of the health emergency system with more than one server (section 2). In the third and fourth sections we respectively describe results empirically obtained when two and three servers are available. Finally, observations and conclusions are traced.

1. A simple rule for M/M/1/\( \infty \).

The transient behaviour of M/M/1/\( \infty \) system was investigated by Odoni and Roth in [5]. They estimated \( Q(t) \) by a technique, which numerically solves the Kolmogorov differential equations of a system, which describe the rate of transition in a discrete-state, continuous-time Markov process. They solved it using the International Math-Science Library subroutine DVERK. They used this empirical approach to get information on the transient behaviour for two reasons: difficulties to obtain general and closed-form solutions (even for the simplest kinds of queueing systems) to use in practice and the fact that they want to obtain explicit expressions on which to draw conclusions. Authors found that the time to equilibrium is bounded from above by the decaying exponential function

\[
Q(t) = Q(\infty) \ast [1 - e^{-\frac{t}{\tau}}]
\]  

where \( Q(\infty) \) is the value toward which the function approaches as \( t \) increases and \( \tau \) is the relaxation time depending on system parameters. ‘\( \tau \) does not depend on the initial state’ ([5]) and is directly proportional to the service rate. \( \tau \) characterises the shape of the negative exponential function. It can be approximated by

\[
\tau = \frac{C_\sigma^2 + C_i^2}{2,8\mu_s(1 - \rho)^2}
\]

where \( C_\sigma^2 \) and \( C_i^2 \) are the coefficients of variation for the interarrival and service times, respectively, and \( 1/\mu_s \) is the mean service time (see [5]). Remember that for M/M/1/\( \infty \) system \( C_\sigma^2 = C_i^2 = 1 \).

For most practical purposes (see [5] and [7]) when \( Q(t) \) is within 2\% of \( Q(\infty) \), we can regard it as sufficiently closed to the steady state value and, thus, the gap can be ignored.

Remember here that, when servers are infinite, (1) becomes

\[
Q(t) = \rho \ast [1 - e^{-\frac{t}{1/\mu}}]
\]

i.e. the value, toward which the function approaches, is the utilization ratio and the time constant is \( 1/\mu \) (see [7]).

In any case, function (1) allows us to estimate the length of the transient period. We can also compute an upper bound to the instant to reach the steady state, using the formula \( Ub = 4\tau \) by Odoni and Roth in [5].
In an unpublished work, [3], the transient behaviour of this queueing system was tested, for different values of the utilization ratio $\rho$, through simulation, using a correlated sampling. The results are shown in table 1. The table reports from left to right the following variables: the utilization ratio $\rho$, the (theoretical) expected number of customers in the system in steady state $Q(\infty)$ (i.e. $\rho/(1-\rho)$), the arrival number $#C$ of the customer when the transient period expires (it is assumed that the steady state has been reached when the difference between $Q(\infty)$ and $Q(t)$ be less than 2.5%), $R$, which is the ratio $#C$ over $Q(\infty)$ and, finally, the same arrival number $#C_{OR}$ which is computed using the relaxation time and the $Ub$ of [5]. The initial number of customer is supposed to be zero.

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$Q(\infty)$</th>
<th>$#C$</th>
<th>$R$</th>
<th>$#C_{OR}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.053</td>
<td>3</td>
<td>56.60</td>
<td>5</td>
</tr>
<tr>
<td>0.10</td>
<td>0.11</td>
<td>4</td>
<td>36.36</td>
<td>6</td>
</tr>
<tr>
<td>0.20</td>
<td>0.25</td>
<td>6</td>
<td>24.00</td>
<td>9</td>
</tr>
<tr>
<td>0.30</td>
<td>0.43</td>
<td>8</td>
<td>18.60</td>
<td>14</td>
</tr>
<tr>
<td>0.40</td>
<td>0.67</td>
<td>12</td>
<td>17.91</td>
<td>21</td>
</tr>
<tr>
<td>0.50</td>
<td>1.00</td>
<td>20</td>
<td>20.00</td>
<td>33</td>
</tr>
<tr>
<td>0.60</td>
<td>1.50</td>
<td>30</td>
<td>20.00</td>
<td>56</td>
</tr>
<tr>
<td>0.70</td>
<td>2.55</td>
<td>68</td>
<td>26.67</td>
<td>107</td>
</tr>
<tr>
<td>0.80</td>
<td>4.00</td>
<td>154</td>
<td>38.50</td>
<td>256</td>
</tr>
<tr>
<td>0.90</td>
<td>9.00</td>
<td>330</td>
<td>36.67</td>
<td>1085</td>
</tr>
</tbody>
</table>

Table 1. The number of customers in the transient period.

The most interesting column for our purpose is the third one: it tells us that, after this number of arrivals, the transient period finishes and the system reaches the steady state.

The value of $#C_{OR}$ is computed dividing the relaxation time $\tau$ for the mean service rate. Comparing $#C$ with $#C_{OR}$ it is clear that the upper bound formula by Odoni and Roth overestimates the time necessary to reach the steady state.

By the fourth column of the table we can derive a simple general rule to easily compute how long the transient behaviour is. The rule, which is valid for a broad range of values, i.e. $0.2 \leq \rho \leq 0.7$, gives with a good approximation the arrival number of the first customer served which can be considered in steady state:

$$#C \approx 20 \times Q(\infty).$$

We can obviously compute also how long transient period $T$ is through the formula

$$T = #C / \lambda.$$

In [3], also, some cases in which initial customer number is not zero are investigated. The main results obtained by this analysis are:

- the behaviour of $Q(t)$ is strongly influenced by the initial state of the queueing system during the start-up period;
- (however) after a start-up period, $Q(t)$ approaches $Q(\infty)$ through a decaying exponential function, which is characterized by the time constant $\tau$, characteristic of the queueing system;
- the only parameter which influences $Q(\infty)$ is $\rho$. 
2. The assumptions in our simulation approach.

In this paper the number of servers is fixed according to the original problem described in [1]. Before discussing the problem of our interest, i.e. the three-server case, we want to review what happens when there is only one server and no queue is allowed, i.e. in the M/M/1/0 system. This last case has been fully investigated in [4]. There, the transient behaviour is theoretically found: results show that for the simplest case, using the first-order linear differential equation with constant coefficients, we are able to determine the state’s probabilities and also the behaviour of the queueing system for small values of times. The most interesting formula, which illustrates the transient solution depending on time, is

\[ p_1(t) = \frac{\lambda}{\lambda + \mu} + \left[ -\frac{\lambda}{\lambda + \mu} + p_1(0) \right] e^{-(\lambda+\mu)t} \]

where \( p_1(t) \) represents the probability that at an arbitrary time \( t \) there is one customer in the system.

In our problem, at time \( t = 0 \) the system can reasonably assumed to be empty: thus, the previous formula becomes

\[ p_1(t) = \frac{\lambda}{\lambda + \mu} \left[ 1 - e^{-(\lambda+\mu)t} \right] \]

i.e., a negative exponential function, which goes asymptotically to \( \lambda/(\lambda+\mu) \) as \( t \) increases.

We are interested to investigate what happens with more than one server and to establish if (2) is still valid. It is well known that ‘the problem becomes much more complicated when…multiple servers are considered’ ([4]). For this reason we investigated the transient behaviour of the M/M/2/0 and of the M/M/3/0 queueing systems by simulation.

In this paper we assume, both in the case of two or three servers, Manhattan metric, servers located at fixed positions, uniform density function for the demand and negative exponential law for inter-arrival and service times. Notice that the overwhelming quantity of the service time for a call is spent for the on-scene time: thus, the travel time will be negligible to determine how much time is needed in a service (see [2]). According to the effective organization of the emergency service, a duty will be eight hours long.

Without loss of generality, we assume also \( \lambda = 1 \) and we choose some values of \( \mu \geq 1 \). However, the limit distribution always exists!

Remember from [1] that the utilization ratio, which is more representative of the real case, is \( \rho = 0.5 \) and ([5]) that “for peak periods (\( \rho = 0.9 \) or \( \rho = 0.95 \)) it is inappropriate to use steady state expressions to estimate queue lengths”.

The program we created for simulation to analyse the transient behaviour of the health emergency systems is written in Visual Basic; it is really easy to use, even if it requires a lot of time to run (about 12 minutes in the case of three servers and 2000 iterations) on a Pentium 4.

Random numbers are used to simulate locations, time and duration of the accidents.

We need 10,000 iterations to obtain from simulation data with a precision of two points of percentage.

To improve the statistical efficiency of the simulation and to reduce the variance, correlated sampling is used, which induces a negative correlation between results and in different simulation runs.

The expected number of ambulances in steady state, \( L(\infty) \), can be found directly (see [4]). It is necessary to write the probabilities of the number of users in the system as a function of \( P_0 \). Now we are able to directly calculate this value by the formula:
Transient behaviour in some queueing systems related to an emergency service

\[ L(\infty) = \sum_{n=0}^{k} n \cdot P_n \]

where \( k \) is the number of servers (so, \( k = 2 \) or \( k = 3 \), in our case), while \( P_n \) represents the probability of \( n \) items in the system. As aforesaid, for most practical purposes, a value of \( L(t) \) which is within 2% of \( L(\infty) \) is considered sufficiently close to the steady state value (see [5], [6] or [7]).

The program returns us the average number of ambulances which are busy every 5 minutes in the duty. To take advantage of this information, we assume that this number, as evaluated in \( t \), is constant throughout the interval \((t-2,5; t+2,5)\), so that \( L(t) \) will result piecewise constant.

We can see that this curve can be well interpolated by a decaying exponential function similar to the ones in (2) and (1), so we can write

\[ L(t) = L(\infty) \cdot [1 - e^{-t/\tau}] \]

(3).

3. The M/M/2/0 case.

The M/M/2/0 case is studied here by simulation, as aforesaid. We tested that a negative exponential function, similar to the one in (3), remains valid: the asymptotic value \( L(\infty) \) can be easily computed by the Erlang’s loss formula (see [4] or [5]) and, consequently, the only value to be estimated is \( \tau \), which we can’t compute by the Odoni’s previous formula because it can’t be used in a no-queue system.

We assume that there are two servers, which are respectively located on the two points \( 1/3 \) and \( 2/3 \) on the interval \([0, 1]\) on the x-axis, while the boundary between their competence districts is the equidistance point, say, it is located on \( x = 1/2 \). The results obtained by simulation are summarized in the following table 2, in which we have, from left to right: the utilization ratio \( \rho \), the asymptotic value \( L(\infty) \), the inverse of the relaxation time, \( 1/\tau \), the instant in which the transient period ends, \( T_{ss} \) (expressed in unit and minutes), the upper bound of Odoni and Roth \( Ub \) and, finally, the number of calls after which the next one will be served in steady state \( S(t) \).

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>( L(\infty) )</th>
<th>( 1/\tau )</th>
<th>( T_{ss} )</th>
<th>( T_{ss}(\text{min}) )</th>
<th>( Ub )</th>
<th>( S(t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>0.243902</td>
<td>0.22</td>
<td>17.8</td>
<td>89</td>
<td>91</td>
<td>5.9</td>
</tr>
<tr>
<td>0.33</td>
<td>0.32</td>
<td>0.21</td>
<td>18.7</td>
<td>93.5</td>
<td>95.2</td>
<td>4.7</td>
</tr>
<tr>
<td>0.40</td>
<td>0.378378</td>
<td>0.2</td>
<td>19.6</td>
<td>98</td>
<td>100</td>
<td>4.1</td>
</tr>
<tr>
<td>0.50</td>
<td>0.461538</td>
<td>0.165</td>
<td>23.8</td>
<td>119</td>
<td>121.2</td>
<td>4</td>
</tr>
<tr>
<td>0.60</td>
<td>0.539326</td>
<td>0.15</td>
<td>26.1</td>
<td>130.5</td>
<td>133.3</td>
<td>3.6</td>
</tr>
<tr>
<td>0.75</td>
<td>0.646154</td>
<td>0.125</td>
<td>31.4</td>
<td>157</td>
<td>160</td>
<td>3.5</td>
</tr>
<tr>
<td>0.90</td>
<td>0.741866</td>
<td>0.11</td>
<td>35.7</td>
<td>178.5</td>
<td>181.8</td>
<td>3.3</td>
</tr>
<tr>
<td>1.00</td>
<td>0.8</td>
<td>0.105</td>
<td>37.3</td>
<td>186.5</td>
<td>190.5</td>
<td>3.1</td>
</tr>
</tbody>
</table>

Tab. 2. Table summing up the transient behaviour with two servers.

The most interesting values of \( \rho \) are the ones between 0.33 and 0.75: lower values would correspond to the cases in which the service time were too fast (not realistic) and, on the other side, values greater than 0.75 would imply too slow service times.

In order to estimate \( \tau \), we use the idea that the area under the negative exponential function be equal to the one obtained by the piecewise constant curve we mentioned beforehand.

Using the negative exponential function, we can now compute when the transient period ends.
We report the value of $T_{ss}$ expressed in two different ways: this will turn out to be useful for what follows. In the first way, one time unit equals 5 minutes and, in the second case, time is directly expressed in minutes. Looking at the first row of table 2 we can say, for example, that after 89 minutes the transient period comes to an end and the system reaches the steady state.

We must underline that the rule $Ub = 4\tau$ in [5] and [7] arose in a context of M/M/1 and M/E$\alpha$/1 systems, while we test the same formula here in systems with two (and in the next section three) servers. Results show that values of the upper bound are really close to the values of $T_{ss}$ in what they differ less than 2.2% in the two-server case.

We want to focus the attention on some observations.

First of all, the transient period is greater as the service time is smaller and, what appears most important, the servers work in transient conditions for a significant part of their time: at the best, for 1 hour and a half and, in the worst case, for 2 hours and a half. Notice that this amount of time is not negligible in a duty of 8 hours. Notice also that, apart from the first row, (after rounding!) values of $S(t)$ are always 4 or 5: so, we can say that after 4 or 5 emergency services the next call will be served under steady state conditions.

Figure 1 represents the case in which $\rho = 0.50$, $T_{ss}$ is on the x-axis and $L(t)$ is on the y-axis: the normalized piecewise constant curve obtained by the simulation and the negative exponential function

$$L(t) = 0.461538 \times (1 - e^{-0.165t})$$

are represented. The dashed line represents the time after which the transient period can be considered finished: it is easy to see that, after 120 minutes, $L(t)$ is “sufficiently close” to its steady state value.

![Negative exponential function and piecewise constant curve with $\rho = 0.5$ and two servers.](image)

As aforesaid, as $\rho \to 1$, i.e. as the system approaches saturation, the system requires a greater amount of time to reach equilibrium. Figure 2 represents the whole range of negative exponential functions, everyone normalized to one ($L(\infty) = 1$) in such a way to have a correct comparison, varying the service rate. It is easy to see that the transient period can be considered finished in a range that goes from 18 to 37 units of time, i.e. in a period that goes from 90 to 190 minutes.
4. The M/M/3/0 case.

The study of the M/M/3/0 case, as in the previous section, is by simulation. The three servers are respectively located on the three points 1/6, 1/2 and 5/6 (see [1]) on the interval [0, 1] on the x-axis. The boundaries are the equidistance points located on $x = 1/3$, to define competence districts between the first and the second server, $x = 2/3$ between the second and the third one and $x = 1/2$ between the first and the third one (when the central server is busy). The use of the equidistance point is justified in [1] and it is a reasonable boundary, which does not depend on the value of $\rho$.

Table 3 shows values of the most important parameters, which resulted from simulation, under the same assumptions and the same notations of the previous section.

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$L(\infty)$</th>
<th>$1/\tau$</th>
<th>$T_{ss}$</th>
<th>$T_{ss(min)}$</th>
<th>$U_b$</th>
<th>$S(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>0.249493</td>
<td>0.245</td>
<td>16</td>
<td>80</td>
<td>81.6</td>
<td>5.3</td>
</tr>
<tr>
<td>0.33</td>
<td>0.331858</td>
<td>0.215</td>
<td>18.2</td>
<td>91</td>
<td>93</td>
<td>4.6</td>
</tr>
<tr>
<td>0.40</td>
<td>0.397138</td>
<td>0.19</td>
<td>20.6</td>
<td>103</td>
<td>105.3</td>
<td>4.3</td>
</tr>
<tr>
<td>0.50</td>
<td>0.493671</td>
<td>0.14</td>
<td>28</td>
<td>140</td>
<td>142.9</td>
<td>4.7</td>
</tr>
<tr>
<td>0.60</td>
<td>0.588106</td>
<td>0.135</td>
<td>29</td>
<td>145</td>
<td>148.1</td>
<td>4</td>
</tr>
<tr>
<td>0.75</td>
<td>0.724907</td>
<td>0.11</td>
<td>35.6</td>
<td>178</td>
<td>181.8</td>
<td>4</td>
</tr>
<tr>
<td>0.90</td>
<td>0.854935</td>
<td>0.1</td>
<td>39.2</td>
<td>196</td>
<td>200</td>
<td>3.6</td>
</tr>
<tr>
<td>1.00</td>
<td>0.9375</td>
<td>0.085</td>
<td>46</td>
<td>230</td>
<td>235.3</td>
<td>3.8</td>
</tr>
</tbody>
</table>

*Tab. 3. Table summing up the transient behaviour with three servers.*

Points in figure 3 represent the instant of time after which the transient period can be considered finished in minutes (y-axis) varying $\rho$ (x-axis) in the two-server and three-server case (see the fifth column of table 2 and 3). We try to interpolate the simulated data by a straight line. Notice that the range of values of time which is necessary to reach the steady state is larger in the three-server case than in the previous one because it goes from 1 hour and 20 minutes, when the service time is really fast, to about 4 hours, in the case of the worst service (in this last case it would say that,
for one half of the time of the duty, servers work in transient behaviour). So, we can now emphasize how relevant is the transient period in the emergency health service. For small value of $\rho$, the transient period ends in fewer minutes than in the two-server case while, as $\rho$ grows up, it increases in a more than proportional way. In other words, we can say that, if the average service time increases, it takes a longer time for people to reach the steady state. We can also observe that the there is a monotone behaviour between the transient period and the utilization ratio.

In the case of Padua, where the utilization ratio is equal to 0.50, it takes 2 hours and 20 minutes to reach the steady state and, so, for a significant period of time servers don’t work in steady state.

![Fig.3 Time to reach steady state varying \rho.](image)

Figure 4 of the following page represents the piecewise constant curve based on data obtained by simulation when $\rho = 0.5$ and the negative exponential function, which is

$$f(t) = 0.493671 \times (1 - e^{-0.14t}).$$

Some other observations can be drawn from the analysis of table 3.

As aforesaid, we can easily see that Odoni’s upper bound is a very good bound also in the three-server case: the gap between values found by simulation and values deriving from his formula is less than 2.5%.

On the other side, for comments about $S(t)$ we can repeat what we said in the two-server case because, in practise, the same things happen.

Figure 5 represents the whole range of negative exponential functions varying the service rate (as figure 2): it is easy to compare these functions and to determine when the system reaches the steady state.
5. Results and conclusions.

This paper investigates the transient behaviour of queueing systems. The first result concerns the M/M/1/∞ system: we give a new very simple rule by which, knowing only the utilization ratio, we obtain the number of the first customer served in steady state or, equivalently, the time after which the steady state is reached.

But the main purpose of this paper deals with the M/M/k/0 systems of emergency medical services by cooperating servers. Simulation, which was implemented with \( k = 2 \) or \( k = 3 \), allows us to test the theoretical transient behaviour, which is well known to be too difficult to compute directly for more than one server.
Important observations we derived in both cases are:

- the negative exponential function (3) interpolates in a very good way the behaviour of queueing systems with no allowed queue;
- servers work in transient conditions for a significant part of their time and, for this reason, we can't investigate real problems only under the steady state conditions, but we must keep in mind also the transient behaviour;
- the quantity of time necessary to work under steady state conditions is directly proportional to the utilization ratio of the system: systems which work more efficiently reach earlier the steady state;
- the rule to compute the upper bound found by Odoni and Roth for M/M/1/∞ systems is a very good upper bound even if servers are more than one and no queue is allowed.

But the most important idea is in what follows. From results, which we obtained by simulation, we can generalize the behaviour of the transient period by a very simple rule. Keeping in mind the number of calls after which the next one will be served in steady state $S(t)$ of tables 2 and 3, we see that always 4 or 5 accidents are needed before the transient period comes to an end both for two and three-server cases. So, we can compute $S(t)$ by this simple rule, which depends only by the utilization ratio and by the number of servers $k$:

$$S(t) = \begin{cases} 
2+k & \text{for } 0.33 \leq \rho \leq 0.50 \\
1+k & \text{for } 0.60 \leq \rho \leq 1 
\end{cases}$$

Obviously, if $S(t)$ is multiplied for the mean service time we have the time after which the transient period comes to an end:

$$T = S(t) \ast (1/\mu).$$

References


