Modeling user perceived unavailability due to long response times

Magnos Martinello, Mohamed Kaâniche, Karama Kanoun, Carlos Aguilar Melchor

Modeling User Perceived Unavailability due to Long Response Times

Magnos Martinello*, Mohamed Kaâniche, Karama Kanoun and Carlos Aguilar Melchor

LAAS-CNRS
7, Av. du Colonel Roche 31077 Toulouse - France
(magnos,kaaniche,kanoun,caguilar)@laas.fr

Abstract

In this paper, we introduce a simple analytical modeling approach for computing service unavailability due to long response time, for infinite and finite single-server systems as well as for multi-server systems. Closed-form equations of system unavailability based on the conditional response time distributions are derived and sensitivity analyses are carried out to analyze the impact of long response time on service unavailability. The evaluation provides practical quantitative results that can help distributed system developers in design decisions.

1 Introduction

Distributed systems widely sustain our day-to-day life providing service for a large number of enterprises, producing business opportunities and offering new services to customers. Such systems should ideally remain operational supporting correct service despite the occurrence of undesirable events.

Service unavailability may result from several causes such as i) failures in service hosts, in the communication infrastructure [1] or in the user site, or ii) heavy loads leading to overloaded servers. From the user viewpoint, the service is perceived as degraded or even unavailable if the response time is too long compared to what he or she is expecting. A long response time may discourage some users who will visit other service providers. For instance, if a request takes 30 seconds to complete, users may consider the request failed.

The goal of this paper is to provide a modeling approach for computing service unavailability due to long response time, relying on Markov reward models and queueing theory. We analyze the long response time effects on service unavailability without distinguishing the various causes leading to long response time. We introduce a flexible mathematical abstraction that is general enough to capture the essence of unavailability behavior due to long response time. We analyze this measure at steady state, starting with a single-server queueing system, then multi-server queueing systems are considered. The aim of our work is twofold:

- Evaluate the effects of long response times on service unavailability.
- Provide practical quantitative results that can help in design decisions.

The derivation of the response-time distribution is widely recognized to be not trivial [8]. Some interesting approaches have been proposed to define measures combining performance and dependability issues [7, 6, 3]. [7, 6] introduced a model for hard and soft real-time systems, while [3] has considered transactional systems in which failures may be due to frequent violation of response time constraints.

The modeling approach presented this paper builds on and extends the work introduced in [3]. The latter work uses i) a Markov model to evaluate *system availability*, and ii) a tagged job approach to compute the response time distribution. We take a step further by providing closed-form equations for response-time distribution and for service unavailability due to long response time in single and multi-server systems. The closed-form equations are derived using the well-known gamma function. We present several sensitivity analyses to illustrate how the designers can use the models to guide the system design.

The paper is organized as follows. Section 2 defines the availability measure based on response time. In section 3, we introduce the modeling approach using single server queueing systems. This is followed by sensitivity analysis results illustrating the measure behavior. Section 4 provides a modeling approach using multi-server queueing systems with sensitivity analysis. Section 5 concludes the paper.

*M. Martinello has a researcher fellowship from CAPES-Brazil.

2 Availability measure definition

Steady *service availability* may be defined as the long-term fraction of time of actually delivered service. We assume that the service states as perceived by the users are partitioned into two sets: i) a set of states in which the service is perceived as available and ii) the complementary set in which the service is perceived as unavailable.

Let Ω denote the set of all service states, and p_i: be the probability that the service is in state i at steady-state.

In order to define the service availability based on the response time, let us introduce

- $R(i)$: the random variable denoting the response time given that the system is in i at steady-state;
- d: the maximum acceptable response time (i.e., if the response time is longer, the service is considered as unavailable), this metric can indicate network delays or an overloaded server; it is also referred to as the maximum response time requirement;
- ϕ: the quality of service requirement (or the accepted quality of service) representing the minimum fraction of requests that satisfy the maximum response time requirement;
- $P[R(i) \leq d]$: the conditional response-time distribution (i.e., the probability that the response time of a request is lower than or equal to d, given that the system is in state i at steady-state).

Using the definitions above, the service is said to be available if the following condition is satisfied

$$P[R(i) \leq d] > \phi \tag{1}$$

Let us denote by K the states in which the service is available (i.e., equation (1) is satisfied for all states i, $i = 0$ to K). Thus, the service availability A and the service unavailability UA are given by:

$$A = \sum_{i=0}^{K} p_i \Leftrightarrow UA = 1 - \sum_{i=0}^{K} p_i \tag{2}$$

The evaluation of the unavailability measure based on response time is carried out according to the following steps. First, one needs to specify the service model describing in particular the distribution of request arrivals and processing times as well as the servers capacity. Based on this specification, $P[R(i) \leq d]$ and p_i can be obtained for a given d. For a given ϕ, using $P[R(i) \leq d] > \phi$, K is derived, then the availability is computed by equation (2). It is worth to mention that the two parameters, d and ϕ characterize the quality of service and should be specified a priori. For example, one can specify ϕ to be equal to $\phi = 0.9$ and $d = 5$ seconds. This means that the service response time should be less than 5 seconds for at least 90% of all requests.

In the following sections, we will i) build analytical models for single-server systems and for multi-server systems, and ii) derive closed-from equations for the conditional response time probability and service unavailability.

3 Single server queueing systems

In this section, we assume that the server is modeled as a single queueing system with exponential arrival and service times. The modeling approach using single server queueing systems is carried out in two steps. First, we model the availability measure based on the response time distribution at steady-state and then some numerical sensitivity analysis results are presented.

3.1 Modeling

3.1.1 Conditional response time distribution

In this simplest example, we assume that there is only a single process serving the incoming requests at a constant rate μ requests/sec. The system is assumed to be in state i when there are i requests in the system ($i - 1$ waiting for service in the queue and one being served). By definition, if a request arrives given that there are already i requests in the system, then the total time spent in the system by the request, denoted as $R(i)$, is given by a sum of $i + 1$ random variables. Since the random variables are independent and identically distributed with mean $1/\mu$, it can be shown that $R(i)$ is described by an Erlang distribution [2] as follows

$$P[R(i) \leq d] = 1 - \sum_{j=0}^{i} \frac{(\mu d)^j}{j!} e^{-\mu d} \tag{3}$$

Let us consider the incomplete gamma function[1] defined as $\Gamma(i+1, \mu d) = \int_{-\mu d}^{\infty} e^{-t} t^i dt$.

Using the fact that

$$\left[1 + z + \frac{z2}{2!} + \cdots + \frac{z^j}{j!}\right] e^{-z} = \frac{\Gamma(j+1, z)}{\Gamma(j+1)}$$

$P[R(i) \leq d]$ can be expressed as follows

$$P[R(i) \leq d] = 1 - \frac{\Gamma(i+1, \mu d)}{\Gamma(i+1)} \tag{4}$$

[1] Note that $\Gamma(i+1)$ is defined by the following integral $\Gamma(i+1) = \int_0^{\infty} e^{-t} t^i dt$. If i is a positive integer, then $\Gamma(i+1) = i!$. It is also important to note that i is not a complex number.

3.1.2 Service unavailability

Consider a system accessible to a very large population. The arrival process is characterized by requests arriving at an average arrival rate of λ requests/sec. This assumption is known as the single class or homogeneous workload.

We first assume that all requests arriving are queued for service. This assumption is known as infinite buffer. Then we will consider the finite buffer case. All the analyses presented assume that the system being analyzed is in operational equilibrium.

3.1.3 Infinite buffer

Requests arrive at the system at a rate of λ requests/sec, queue for service, get served at rate μ requests/sec and depart. Such a system is a traditional M/M/1 queue system [2], in which the probability (p_i) that there are i requests at steady-state is well-known

$$p_i = (1-\rho)\rho^i \tag{5}$$

where $\rho = \frac{\lambda}{\mu}$ refers to the load.

Therefore, equation (2) becomes $A = \sum_{i=0}^{K}(1-\rho)\rho^i$.

Using the fact that $\sum_{i=0}^{K}\rho^i = \frac{1-\rho^{K+1}}{1-\rho}$, we obtain

$$A = 1-\rho^{K+1} \Leftrightarrow UA = \rho^{K+1} \tag{6}$$

3.1.4 Finite buffer

For a system supporting at most b requests including the request being processed (finite buffer) denoted by M/M/1/b queue system, we have

$$A = \frac{1-\rho^{K+1}}{1-\rho^{b+1}} \Leftrightarrow UA = 1-\left[\frac{1-\rho^{K+1}}{1-\rho^{b+1}}\right] \tag{7}$$

where $K < b$.

Table 1 summarizes the equations for user perceived unavailability due to long response time in single server queueing systems. Recall that the computation of UA requires to calculate K which corresponds to the maximum value of i satisfying equation (1).

3.2 Sensitivity analysis

In this section, some numerical results are presented in order to illustrate the behavior of UA using the equations derived in Table 1.

Conditional response time probability $$P[R(i) \leq d] = 1 - \frac{\Gamma(i+1,\mu d)}{\Gamma(i+1)}$$
Unavailability due to long response time for an M/M/1 queue system $$UA = \rho^{K+1}$$
Unavailability due to long response time for an M/M/1/b queue system $$UA = 1 - \left[\frac{1-\rho^{K+1}}{1-\rho^{b+1}}\right]$$

Table 1. Closed-form equations for single server queueing systems

3.2.1 Variation of response time

An important consideration for design purposes is to define when the service is "too slow". In fact, one has to specify the threshold d for the acceptable response time. Considering the example of web servers, practical experiences have suggested that *ten seconds* is well above the normal response time for all the sites studied in [5]. The latter divides timing problems affecting sites availability into "medium" (ten seconds) and "severe" (thirty seconds) problems.

Figure 1 shows the conditional response time distribution ($P[R(i) \leq d]$), given in Table 1, as a function of the number of requests, considering different values for μd. In fact, the evaluation of this distribution allows us to determine the K states for which ($P[R(i) \leq d] > \phi$).

As it can be seen, the response time probability is directly affected by the product μd. It is noteworthy that μd corresponds to the average number of requests processed by the server during a period of time d. Another observation is that K increases with μd. For example, setting the quality of service parameter ϕ to 0.9, $K = 7$ for $\mu d = 12.5$ and $K = 63$ for $\mu d = 75$. In fact, the greater is K, the higher is the probability that the response time is lower than d. Figure 1 also shows that lower values for the quality of service parameter ϕ (e.g., $\phi = 0.8$) clearly lead to greater values for K. In other words[2], the greater is K, the more requests arriving at the server are likely to be satisfied within the acceptable response time.

Such analyses are useful for design decisions, since the

[2]It can be noticed that for a given ϕ, the longer is the response time requirement, the greater is K.

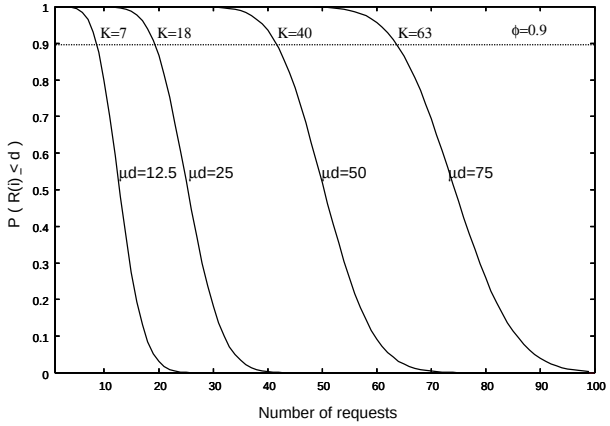

Figure 1. $P[R(i) \leq d]$ **for single server systems**

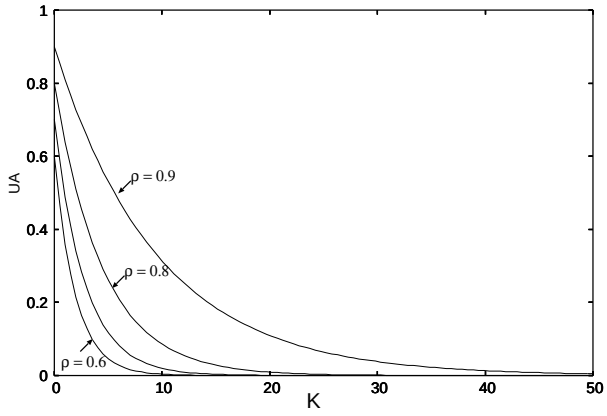

Figure 2. UA **as a function of** K **(M/M/1 system)**

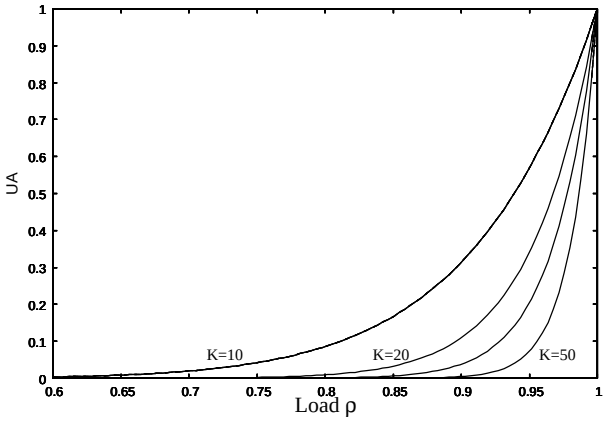

Figure 3. UA **as a function of** ρ **(M/M/1 system)**

expected level of degradation of response time probability as a function of the number of requests queued in the system can be evaluated.

Once K is evaluated, one can compute the service unavailability using equation (6). Table 2 shows how UA varies as μd increases for different loads ρ and for $\phi = 0.9$. We clearly observe that UA decays faster as ρ decreases.

μd	K	$\rho = 0.9$	$\rho = 0.8$	$\rho = 0.7$	$\rho = 0.6$
12.5	7	4.3e-01	1.6e-01	5.7e-02	1.6e-02
25	18	1.3e-01	1.4e-02	1.1e-03	6.0e-05
50	40	1.3e-02	1.0e-04	4.4e-07	8.0e-10
75	63	1.2e-03	6.2e-07	1.2e-10	6.3e-15

Table 2. UA **as** μd **increases (**$\phi = 0.9$**)**

3.2.2 Effects of K and ρ on UA

UA provides a useful indicator to analyze the impact of response time on service unavailability. By definition, parameter K represents the set of states for which the response time is acceptable for a given quality of service requirement. Figure 2 shows UA as a function of K for different loads ρ. The system is assumed to be composed by one server with infinite buffer (M/M/1). Note that by definition, $\rho = 1$ implies $UA = 1$ and $\lim_{K \to \infty} UA = 0$.

From the figure, we can see that UA is very sensitive to the load ρ. UA decays slowly for heavy loads ρ. In contrast, for "light" loads $\rho < 0.6$, the unavailability due to long response time is negligible. On the other hand, the greater is K, the lower is UA. Such behaviour is better illustrated on Figure 3 that plots UA as a function of the load ($\rho > 0.6$) for different values of K. In particular, for systems in which $K >> 50$, there is a small probability that the service is perceived as unavailable due to long response time. Thus, according to the figure, it is likely that service unavailability due to long response time will be very low.

3.2.3 Finite buffer effects on UA

All the evaluations presented along the previous section can also be applied to a system with a finite buffer (i.e., considering an M/M/1/b queue).

Figure 4 shows a comparison between a system with a finite buffer (M/M/1/b) and a system with an infinite buffer (M/M/1). The results for M/M/1/b (dotted lines) are obtained using equation (7). As expected, the greater is b, the lower is the difference between the models. For $\rho = 0.9$ M/M/1/40 is very close to M/M/1. The difference is significant for M/M/1/20. However, for lower loads ($\rho = 0.6$ and 0.7), the curves for M/M/1/20 and M/M/1/40 are the same

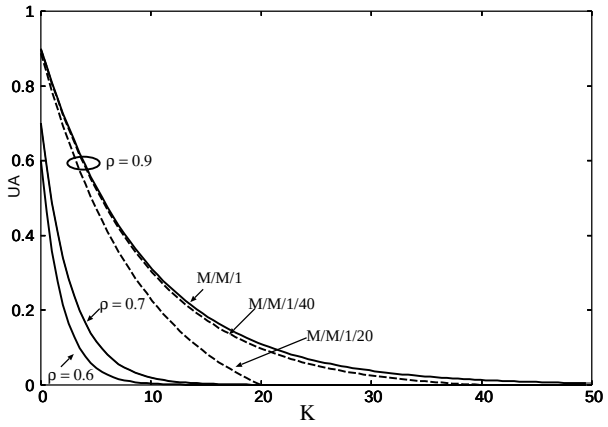

Figure 4. The effect of finite buffer size b on UA

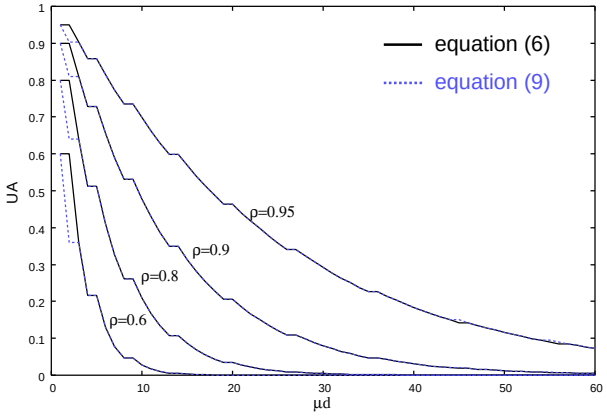

Figure 5. UA as a function of μd

as for M/M/1.

For a given request, the greater is its position in the buffer, the lower is the probability that it is served within the maximum response time requirement. In finite queueing systems, the requests arriving when the buffer capacity b is full are rejected, and therefore, they are not considered as leading to service unavailability due to long response time. This explains the fact that UA for M/M/1/b is lower than UA for M/M/1.

3.2.4 Approximation for UA

The evaluation of UA requires the calculation of parameter K which represents the set of states after which all arriving requests probably perceive the service as unavailable. K is not known a priori. It is computed in an intermediate step. We have investigated a more direct approach for computing UA based on an approximation of K, in order to obtain an analytical equation of UA as a function of only well known parameters, such as service rate (μ) and required maximum response time (d), without needing to compute K in an intermediate step based on the conditional response time distribution.

By analyzing the properties of the finite series $f(n) = \sum_{j=0}^{n} \frac{(\mu d)^j}{j!}$, we found the following approximation[3] for K

$$K \approx \left\lceil \mu d - \alpha\sqrt{\mu d} \right\rceil \quad (8)$$

where α is a constant that can be set to support a given quality of service (e.g., $\alpha = 1.35$ for $\phi = 0.9$).

Accordingly, UA can be obtained as follows

$$UA \approx \rho^{(\lceil \mu d - \alpha\sqrt{\mu d} \rceil)+1} \quad (9)$$

Thus, UA can be evaluated directly as a function of μ and d only, for a given ϕ. Figure 5 shows a comparison between the unavailability computed by equation (6) (where K is an integer value obtained from equation (4)) and the approximation given by equation (9) ($\alpha = 1.35$). UA is plotted as a function of μd, where dotted lines represent the approximation. As it can be seen, the approximation is very accurate. It differs from the exact value only for $\mu d < 5$. In other words, for any $\mu d > 5$ (which is the case of most server systems capable of handling much more than 5 requests per second), there is no difference between the exact and the approximated value of UA.

Equation (9) and the results of figure 5 show that UA approaches 0 as μd approaches infinity, (i.e., $\lim_{\mu d \to \infty} UA = 0$). It means that for a given $d > 0$ with a server having infinite service rate $\mu = \infty$, there is no service unavailability due to long response time. At the same time, for a given service rate $\mu > 0$, there is no service unavailability for a user with an infinite patience $d = \infty$.

4 Multi-server queueing systems

Let us consider a multi-server queueing system consisting of a queueing buffer of finite or infinite size, with multiple identical servers. Such an elementary queueing system is also referred to as a multi-server system. In section 4.1 we present closed-form equations for the condition response time distribution and the service unavailability. Sensitivity analysis results are presented in section 4.2.

[3] In fact, K is obtained through a sub-linear approximation to the inflexion point of f(n) around μd.

4.1 Modeling

4.1.1 Conditional response time distribution

Let us suppose that the multi-server system is composed of c identical servers, where each server is capable of handling μ requests/sec. Let $R_c(i)$ be the random variable denoting the response time in steady-state of an arriving request at a system with c servers and i requests. If a request arrives when there are already i requests in the system, two different cases can be distinguished to model the corresponding conditional response time distribution:

- If $i < c$, the new arrival is processed immediately by one of the free servers. Thus, $R_c(i)$ is an exponential random variable with parameter μ.
- If $i \geq c$, the new arrival must wait for $i - c + 1$ service completions before receiving service (If $i = c$, the new request must wait for one service completion. If $i = c + 1$, two service completions are required, etc.). In this case, $R_c(i)$ is the sum of an Erlang random variable X corresponding to the request waiting time and an exponential random variable Y denoting the service time. Therefore, by convolution $P[R_c(i) = X + Y \leq d] = \int_0^d F(d-y)g(y)dy$, where $F(x, i-c+1) = 1 - \sum_{j=0}^{i-c} \frac{(\mu c x)^j}{j!} e^{-\mu c x}$ and $g(y) = \mu e^{-\mu y}$.

After a set of transformations (see [4] for all details), we obtain equation (10)

$$P[R_c(i) \leq d] = \begin{cases} 1 - e^{-\mu d} & \text{, if } i < c \\ 1 - (\frac{c}{c-1})^{i-c+1} e^{-\mu d} \left[1 - \frac{\Gamma(i-c+1,(c-1)\mu d)}{\Gamma(i-c+1)}\right] & \\ \quad - \frac{\Gamma(i-c+1,c\mu d)}{\Gamma(i-c+1)} & \text{, if } i \geq c \end{cases} \quad (10)$$

4.1.2 Service unavailability

Let us take the same system consisting of c identical servers, where each server is capable of handling μ requests/sec. We need to compute the probability that the system with c servers has i requests at steady-state denoted $p_i(c)$. Assuming that the sequence of interarrival times is described by independent and identical exponential random variables of rate λ (a traditional M/M/c in which $p_i(c)$ is well-known [2] with $\rho = \frac{\lambda}{c\mu}$), we obtain for service unavailability the following closed-form equation (see [4] for all details)

$$UA = \begin{cases} 1 - p_0 \frac{\Gamma(K+1,c\rho)e^{c\rho}}{\Gamma(K+1)} & \text{, if } K < c \\ 1 - p_0 \frac{\Gamma(c,c\rho)e^{c\rho}}{\Gamma(c)} & \\ \quad + \frac{(c\rho)^c}{c!} \frac{(1-\rho^{K-c+1})}{1-\rho} & \text{, if } K \geq c \end{cases} \quad (11)$$

To summarize, for $i = 0, 1, 2, ...$ requests, $P[R_c(i) \leq d]$ can be computed using equation (10). Based on this distribution, K is obtained as the maximum value of i satisfying $P[R_c(i) \leq d] > \phi$. Then, we can proceed to calculate UA for multi-servers in an infinite and finite queueing systems using equation (11).

4.2 Sensitivity analysis

In this section, we study the effect of the number of servers on UA using the modeling approach developed in the previous section. The analysis is divided in two parts. First, in 4.2.1 the response time distribution is studied in order to quantify how the response time is affected by the number of servers. A simple example of system is presented illustrating some possible configurations. Then, we evaluate UA itself taking into account the response time variation, the load effects in 4.2.2 and the number of servers in 4.2.3.

4.2.1 Variation of response time distribution

Figure 6 shows the response time distribution ($P[R_c(i) \leq d]$) as a function of the number of requests computed by equation (10). This function is evaluated varying the number of servers c and the product μd. As it can be seen, as c or μd increases the response time probability is improved. This is illustrated by the increase of K. Clearly, the greater is K, the lower is UA. The effect of K on UA for multi-servers is similar to the case of single-server (discussed in section 3.2.2).

These results can be used for supporting design decisions. For instance, let us define by μc the aggregated service rate provided by c servers. We consider a set of system configurations designed to support the same aggregated service rate of $\mu c = 150$ requests/sec. Table 3 identifies four possible system configurations using only multi-servers, i) to iv), and one configuration, v), with a single server.

The values of K are $K = [116, 126, 130, 131, 133]$ corresponding to configurations i), ii), iii), iv) and v) respectively. This result shows that a configuration with only 1 server provides the greatest K. Clearly, the response time is longer when the aggregated service rate is split among the servers. This fact explains why K decreases for configurations that employ various servers with low service rates (e.g., $K = 116$ for $c = 12$ and $\mu = 12.5$, compared to $K = 131$ for $c = 2$ and $\mu = 75$).

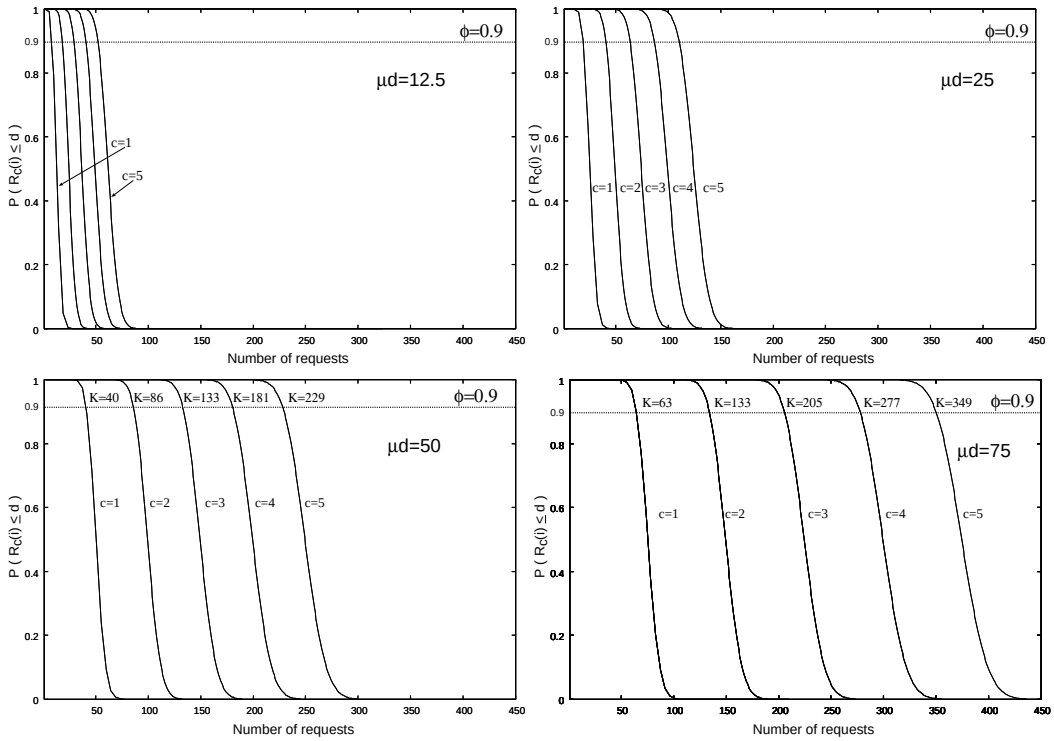

Figure 6. $P[R_c(i) \leq d]$ **variation for multi-server queuing systems**

Configuration	c	μ
i)	12	12.5
ii)	6	25
iii)	3	50
iv)	2	75
v)	1	150

Table 3. Configurations for $\mu c = 150$ **req/sec.**

	UA in days:hours:minutes per year		
Configuration	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.95$
i)	0	00:00:05	00:32:28
ii)	0	00:00:01	00:15:17
iii)	0	0	00:11:07
iv)	0	0	00:10:16
v)	0	0	00:09:04

Table 4. Impact of ρ **on** UA **(**$\mu c = 150$**).**

4.2.2 Load effects on UA

Table 4 shows the impact of c and μd on UA for different loads ρ (setting d to 1 to simplify the analysis) assuming the same aggregated service rate of $\mu c = 150$ requests/sec. The service unavailability is given in days:hours:minutes per year for various loads $\rho = [0.8, 0.9, 0.95]$. UA is computed using equation (11) based on the value of K obtained from equation (10). Note that for $\rho \leq 0.8$, there is a very small unavailability due to long response time.

According to Table 4, it can be seen that the lowest UA is obtained for a single powerful server (configuration v). Nevertheless, for all configurations, UA is less than 5 min and 30 sec per year for $\rho \leq 0.9$, which is relatively low. UA is significantly much higher for $\rho = 0.95$.

4.2.3 Impact of the number of servers c on UA

Table 5 shows the impact of c for three values of service rates $\mu = [25, 50, 75]$, when the load is set to $\rho = 0.9$. It is important to note that increasing c is efficient for reducing UA especially when the load is not heavy $\rho < 0.9$. For instance, if $\mu = 25$ and $\rho = 0.9$, then UA is 49 days per year for $c = 1$ compared to ($UA \approx 16$ minutes per year) for $c = 3$. It becomes more efficient as μ increases (e.g., for $\mu = 50$, $UA \approx 4$ days per year for $c = 1$ compared to $UA \approx 1$ minute per year for $c = 2$).

c	$\mu = 25$	$\mu = 50$	$\mu = 75$
1	49:07:22	04:20:29	00:10:16
2	06:07:24	00:01:09	0
3	00:15:51	0	0
4	00:01:38	0	0
5	00:00:07	0	0

Table 5. UA in days:hours:minutes per year for $\rho = 0.9$.

5 Conclusion

In this paper we have provided an analytic modeling approach for computing service unavailability due to long response times using queueing systems theory. Closed-form equations of the response time distribution and the service unavailability were developed in order to illustrate fundamental availability issues for single and multi-server systems.

The models presented in this paper are aimed to allow the system developers to draw some interesting and practical conclusions concerning the impact of various parameters on the user perceived unavailability. For example, the results have shown that for "light" loads (i.e., $\rho \leq 0.7$), the unavailability due to long response time is negligible. From the designer perspective, the evaluations suggest that systems with low service rate subject to a heavy load $\rho \geq 0.9$ tend to exhibit the highest unavailability due to long response time. The effect of heavy loads on UA is less substantial as the aggregate service rate increases. Also, the obtained results have suggested that the difference on UA among the configurations becomes negligible as the aggregate service rate increases. Finally, increasing the number of servers (c) has been efficient for reducing UA especially for low loads $\rho < 0.9$.

It has been shown that it is possible to provide a service satisfying a response time requirement using only servers with low service rate, although this is not the optimal configuration. In fact, we should employ either a powerful single server or various servers preventing as much as possible the overloaded periods ($\rho \geq 0.9$). For multi-servers systems, the response time is longer as the aggregated service rate is shared among the servers. This fact explains why configurations that employ various servers with low service rates are not the optimal configuration.

All the analyzes presented have focused on the unavailability due to long response time, assuming that all the servers are available. We emphasize the fact that if we take into account the failures of one or more servers, the impact of long response time on service unavailability should be more significant. Although the optimal configuration consists of a powerful single server, it represents a single point of failure under the availability viewpoint. Therefore, an alternative configuration employing more than a single server should provide a better tradeoff supporting degradable service under the presence of failures.

Finally, it is recognized that the service unavailability in the context of widely distributed server systems might be due to problems with the host (e.g., the remote host is too busy handling other requests), problems with the underlying network (e.g., a proper route to the site does not exist) or problems in the user host. In this paper, our attention was devoted to the service unavailability due to long response time concentrated at the server side. In order to analyze the impact of the response time on the end-to-end service unavailability as perceived by users, it is necessary to include other components affecting the time spent by a user request, e.g. the network delay (latency and transmission time), etc.

References

[1] M. Dahlin, B. Chandra, L. Gao, and A. Nayate. End-to-end wan service availability. *ACM/IEEE Transactions on Networking*, 11(2), 2003.

[2] L. Kleinrock. *Queueing Systems*, volume I - Theory. Wiley, 1975.

[3] V. Mainkar. Availability Analysis of Transaction Processing Systems based on User-Perceived Performance. *Proceedings of 16th Symposium on Reliable Distributed Systems*, pages 10–17, 1997.

[4] M. Martinello. Availability Modeling and Evaluation of Web-based Services : A Pragmatic Approach. *PhD Thesis LAAS-CNRS*, (05552), 2005.

[5] M. Merzbacher and D. A. Patterson. Measuring End-User Availability of the Web:Practical Experience. *Dependable Computing and Network*, 2002.

[6] J.K. Muppala, S.P. Woolet, and K.S. Trivedi. Real-time systems performance in the presence of failures. *IEEE Computer*, pages 37–47, 1991.

[7] K.G. Shin and C. M. Krishna. New performance measures for design and evaluation of real-time multiprocessors. *Computer System Science and Engineering*, 4(1):179–192, 1986.

[8] K. S. Trivedi, S. Ramani, and R. Fricks. Recent advances in modeling response-time distributions in real-time systems. *Proceeding of the IEEE*, 91(7):1023–1037, 2003.