the Availability Digest

www.availabilitydigest.com

Configuring to Meet a Performance SLA – Part 4: Single Server with General Service Time

March 2009

Many applications carry with them a performance Service Level Agreement (SLA) that specifies the response times they must achieve. After all, if an application's response time is so slow that the application is not useful, the application is, in effect, down.

The performance requirement is often expressed as a probability that the system's transaction-response time will be less than a given interval. For instance, "When handling 50 transactions per second, 98% of all transactions must complete within 500 milliseconds."

In Part 1 of this series, we derived the basic average response-time expression for a single-server system. In Part 2, we extended that result to a multiserver system in which multiple servers work off a common work queue. In Part 3, we showed how to size a system to meet a performance SLA if service time is exponentially distributed.

If service time is not exponentially distributed, the solution is more complex and involves the Gamma distribution. In this part, we extend our analysis of Part 3 to servers with general service-time distributions.

First, we review the results of the first three parts of this series.

Reviewing the Average Response Time for a Single Server

In Part 1, we showed that the average response time for a single-server system was given by the Pollaczek-Khintchine equation:

$$T_r = \frac{T_s}{1-L}[1-(1-k)L] \qquad (1)$$

where

T_r is the average transaction-response time.
T_s is the average service time of the server.
L is the load on the server.
k is the distribution coefficient of the server's service time.

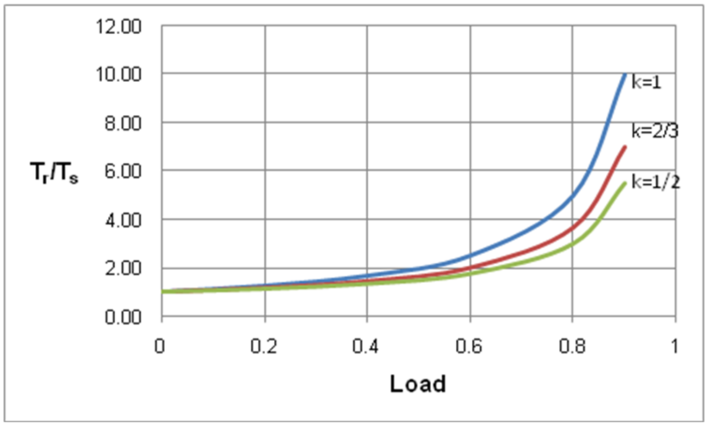
The distribution coefficient *k* depends upon the probability distribution of the server's service time. For instance, *k* = 1 applies to server distribution times that are random (that is, exponentially distributed), which is the common assumption.[1] In this case, Equation (1) reduces to

[1] More specifically, *k* is 1/2 the ratio of the service time's second moment to the square of its mean.

$$T_r = \frac{T_s}{(1-L)} \qquad (2)$$

For many of you, Equation (2) is the well-recognized expression relating transaction-response time to server load.

The response times for exponential service-time distributions (k = 1), uniform service-time distributions (k = 2/3), and constant service-time distributions (k =1/2) are shown in Figure 1. This chart shows the response time, T_r, relative to the service time, T_s, (that is, T_r/T_s) as a function of server load.

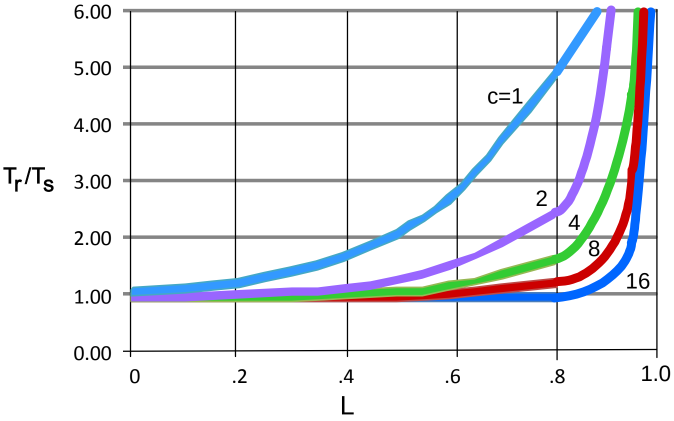


Single Server Response Time
Figure 1

Reviewing the Average Response Time for a Multiserver System

The calculation of response time for a multiserver system is more complex, as shown in Part 2. In a multiserver system, several like servers process transactions from a common work queue. Examples of such servers are web farms and transaction-processing monitors, such as Tuxedo and NonStop Pathway, which distribute transactions to a pool of servers.

Using c to reflect the number of servers in the multiserver system, Figure 2 shows transaction-response time normalized to service time (T_r /T_s) as a function of load for multiserver systems using 1, 2, 4, 8, and 16 servers. The improvement in response time as servers are added is readily apparent.

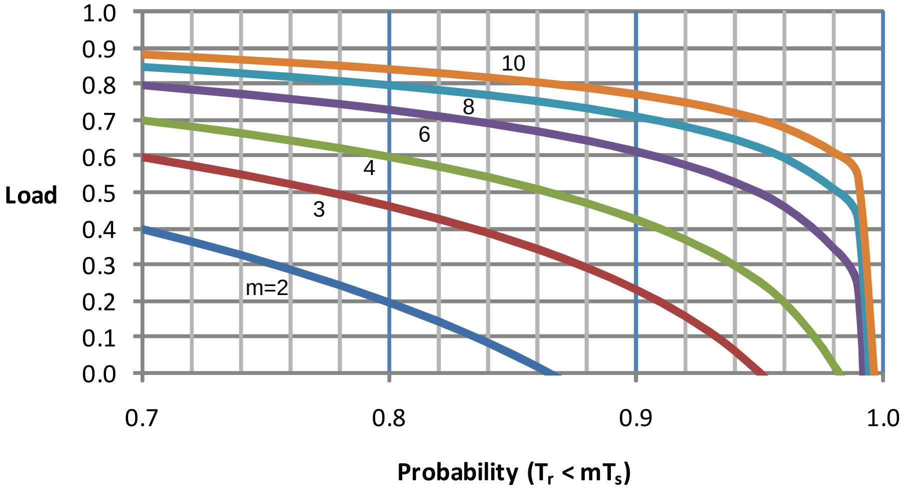


Multiserver Transaction Time
Figure 2

Reviewing the SLA Requirements for Exponential Service Times

In Part 3, our attention was turned to determining what average response time is needed in order to meet an SLA requirement such as “At 50 transactions per second, 98% of all transactions will complete within 500 msec.” We did this first for the case of exponentially-distributed service times.

The result is shown in Figure 3 for various values of *m*, where $m = T_m / T_s$ is the ratio of the specified response time, T_m, to the service time, T_s. For instance, assume that the server has a service time of 10 milliseconds. If 95% of all transactions are to complete in less than 60 msec. (*m* = 6), we can load the server up to 50%.



Allowable Load To Meet Performance SLA
Figure 3

SLA Requirements for General Service Times

If the service time of the server is not exponentially distributed, the analysis becomes more complex. What we need to know is the distribution of the response times so that we can determine the high-end tail that we are to avoid with the specified probability.

Fortunately, there is a distribution that has been shown to be a reasonable approximation of the distribution of response times for a server with an arbitrary service-time distribution so long as the arrivals at the server are random. This is the Gamma distribution.[2]

The Gamma Distribution

The Gamma distribution comes with a compelling history. In the early days of telephony, the Erlang distribution, developed by A. K. Erlang, was used to determine the number of telephone calls that might be waiting for telephone operators. This work has been expanded to determine the distribution of waiting times in queuing systems in general. The resulting probability distribution is the Gamma function. It has been shown to give surprisingly close results to actual response times measured in practice or by simulation.

[2] James Martin, Chapter 31, Queuing Calculations, pp. 438-444, *Systems Analysis for Data Transmission*, Prentice-Hall; 1972.

The Gamma function provides the probability that a variable will have a value less than a specified value – just what we need. It depends upon only one parameter represented by R, which is the ratio of the square of the distribution's mean to its variance.

Since we are interested in the distribution of response time, for our purposes *R* is given by

$$R = \frac{\left(\overline{T_r}\right)^2}{\mathrm{var}(T_r)} \tag{3}$$

where

T_r is the response time
$\overline{T_r}$ is the mean (average) response time
var(T_r) is the variance of T_r.

Thus, if we can determine the mean and the variance of the response time of a server, we can make a statement about the probability that the response time will be less than a certain value. As it turns out, the variance of the response-time distribution is a function of load. More about this later.

The mathematical representation of the Gamma function is not very palatable (see page 439 of the above footnote referencing Martin's book). Fortunately, Excel comes to the rescue with its GAMMADIST function. For our purposes, we use

$$\text{Probability } (T_r < T_m) = (\text{GAMMADIST}(zR, R, 1, \text{TRUE}) \tag{4}$$

where *z* is the ratio of the maximum specified response time to the mean response time:

$$z = \frac{\text{maximum response time}}{\text{mean response time}} = \frac{T_m}{\overline{T_r}} \tag{5}$$

and T_m is the maximum specified response time.[3]

For *R* = 1, the Gamma distribution becomes the exponential distribution; and our analysis of Part 3 holds. For *R* = ∞ (i.e., variance = 0), the Gamma distribution becomes a constant distribution.

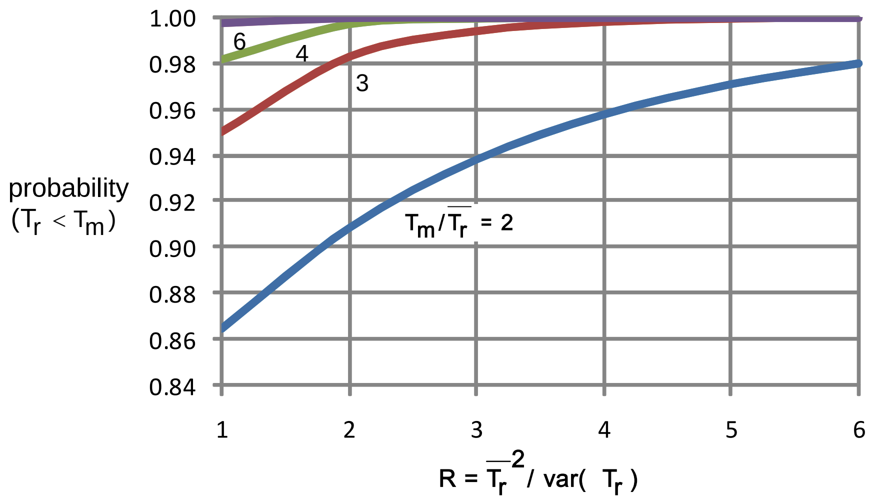
This distribution is shown in Figures 4a and 4b. Figure 4b explodes the higher probabilities.

For instance, if we wanted 99% of all responses to be less than three times the average response time (T_m/T_r =3), we find from Figure 4b that this can be met if $R \geq 2.5$. From Equation (3), this means that the variance of the response time must be less than the 0.4 of the mean response time squared. If the mean response time is 10 msec., its variance must be less than 0.4 x 100 = 40. (The standard deviation of the response time is the square root of the variance, or 6.3 msec. in this case.)

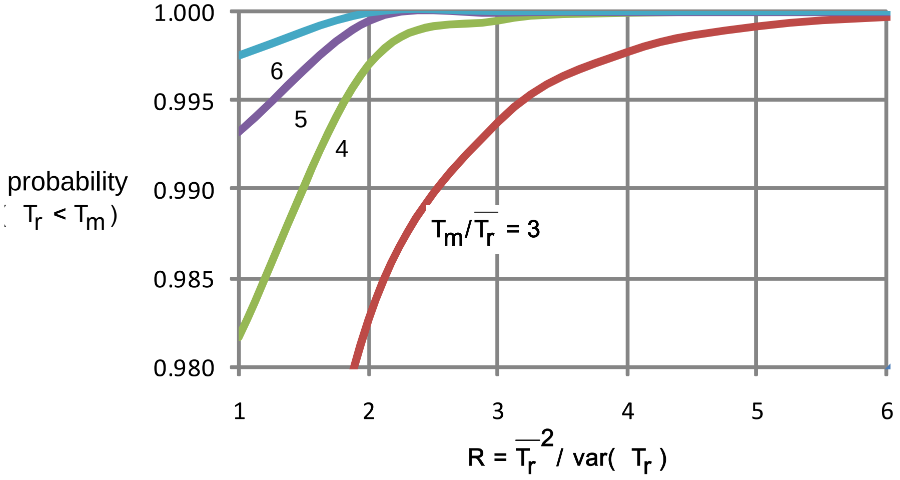
Calculating the SLA Parameters

This is all well and good, but it doesn't really tell us what we want to know. We are missing two parameters. We need to know how much we can *load* the server such that the response time will be less than some multiple of the *server's service time* with a stated probability. We need to convert *variance* to *load* and *mean response time* to *mean service time.*

[3] The third parameter, "1," returns the standard Gamma distribution. The parameter "TRUE" returns the cumulative distribution. "FALSE" returns the probability density function. Excel also provides an inverse Gamma function, which provides *zR* = GAMMAINV(*P*, *R*, 1).



Allowable Variance to Meet SLA
Figure 4a



Allowable Variance to Meet SLA
Figure 4b

One way to do this is to perform response-time measurements on the real server, either under actual conditions or under simulated conditions. We know the average service time of the server. By running the server at some fixed known load, the mean and variance of the response time can be measured. Knowing this, *R* can be calculated; and Figures 4a or 4b can be used to obtain the probability that response times will be less than some maximum. We then know that the response time will be less than that maximum for the service time of the server and the load imposed on the server.

Let us use the above example and turn it upside down. Let us assume that we have a server with a response time of 5 msec. We load it to 50% and measure the response times. We find that the average response time is 10 msec., and its variance is 40. Therefore, the applicable value of *R* is, from Equation (3), $(10)^2/40$, or 2.5. Using *R* = 2.5 in Figure 4b, we find that 99% of all responses will be less than three times the average response time of 10 msec. Since the server's service time is 5 milliseconds, we can make the SLA statement that 99% of all responses will complete in less than six service times – that is, in less than 60 msec. – when the server is 50% loaded.

Calculating Response Variance for Known Service-Time Distributions

An actual measurement of the server can be avoided if we know the relationship between server load, response time, and response-time variance. These relations do, in fact, exist for the cases of exponential service times, uniform service times, and constant service times.[4] Let

T_s	= mean service time
L	= load on the server
$\overline{T}_r$	= mean response time
var (T_r)	= variance of the response time

Using Equation (1) to calculate response time based on server time and server load, the relations are:

- for exponential service times:

$$\overline{T}_r = \frac{T_s}{(1-L)} \qquad \text{var}(T_r) = \frac{T_s^2}{(1-L)^2} \tag{6}$$

- for uniform service times:

$$\overline{T}_r = \frac{T_s}{(1-L)}\left(1-\frac{L}{3}\right) \qquad \text{var}(T_r) = \frac{T_s^2}{(1-L)^2}\left(\frac{1}{3}+\frac{L^2}{9}\right) \tag{7}$$

- for constant service times:

$$\overline{T}_r = \frac{T_s}{(1-L)}\left(1-\frac{L}{2}\right) \qquad \text{var}(T_r) = \frac{T_s^2}{(1-L)^2}\left(\frac{L}{3}-\frac{L^2}{12}\right) \tag{8}$$

Thus, for these cases, an actual measurement is not needed. The server response time and server load of interest can be used to calculate the mean response time and its variance. Knowing these, *R* is known; and the charts of Figures 4a and 4b can be used as described above to determine the probability that the response time will be less than some multiple of the service time at a specified load.

Note from Equations (6), *R* = 1 for exponential distributions independent of load, as mentioned previously.

[4] W. H. Highleyman, Chapter 4, Basic Performance Concepts, pp. 117-118, *Performance Analysis of Transaction Processing Systems*, Prentice-Hall; 1989.

Summary

Using the Gamma distribution, one can reasonably approximate the distribution of response times for a given server load if the mean response time and its variance are known. These parameters can be determined by physical measurement, or they can be calculated for certain service-time distributions as shown in Equations (6) through (8). Using the Gamma distribution, one can then determine the probability that response times will be less than some multiple of the mean service time. Alternatively, we can determine from the inverse Gamma function what response time will not be exceeded with a given probability. The results are what we need to know in order to satisfy an SLA.

An Excel spreadsheet that is useful in making many of these response-time calculations can be found in the Excel workbook at
http://www.availabilitydigest.com/public_articles/0403/composite_performance_sla.xls.
This workbook includes the spreadsheets covering the first three parts of this series as well.

If you have no reasonable way to determine the variance of the transaction-response time, a very conservative approach is to assume that the server has an exponential service time; and use a value of *R* =1 (or use the results of Part 3). This is typically the worst case.

In Part 5, our final part in this series, we introduce a very powerful feature of the Gamma distribution to demonstrate how to use these results to calculate the SLA for a complex series of servers.