

## **Configuring to Meet a Performance SLA – Part 5: Multiple Servers with General Service Times**

April 2009

An SLA often contains a performance specification in the form of “99% of all transactions will complete in less than 50 msec.” Whether or not we can meet this specification is the SLA question.

In Part 4 of this series on meeting a performance SLA specification, we used the Gamma distribution to provide a powerful method for determining the probability that the response time of a server with general service times will be less than a specified amount. Alternatively, we showed how to calculate the response time that could be achieved with a certain probability. We reduced these rather complex calculations to a set of charts and spreadsheets.

However, most systems comprise a set of servers acting in tandem through which an event must pass before a response to that event is generated. Each of these servers could have a different distribution of service times and could be carrying different loads. How do we calculate the performance SLA parameters for such a complex system?

In this article, we show how to use the results of the previous articles to solve this problem.

### **The Response Time of Servers with General Service-Time Distributions**

In Part 3, we solved the SLA problem for servers with exponential service times. In Part 4, we did so for servers with general service times. We review Part 4 here.

The solution for general service times depends upon the Gamma distribution. With its roots in the early days of telephony, the Gamma distribution has been shown to be a surprisingly good approximation to the distribution of response times in queuing systems.

The Gamma distribution provides its approximation to the response-time distribution based simply on the knowledge of the response time’s mean and variance. The problem then reduces to the determination of this mean and variance.

In the general case, there is no easy way to do this. The response-time mean and variance must be measured using an operational system with a known average service time and running under a specified load. In Part 4, we charted this relationship, repeated here as Figures 1a and 1b. Figure 1b is an explosion of the higher probabilities shown in Figure 1a.

These figures show the probability that the response time will be less than some specified maximum time,  $T_m$ , where

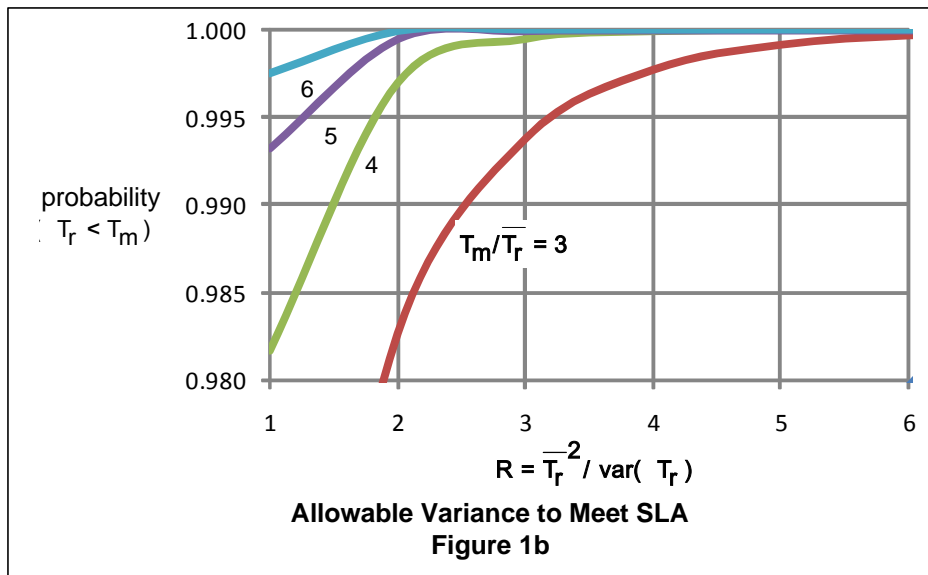
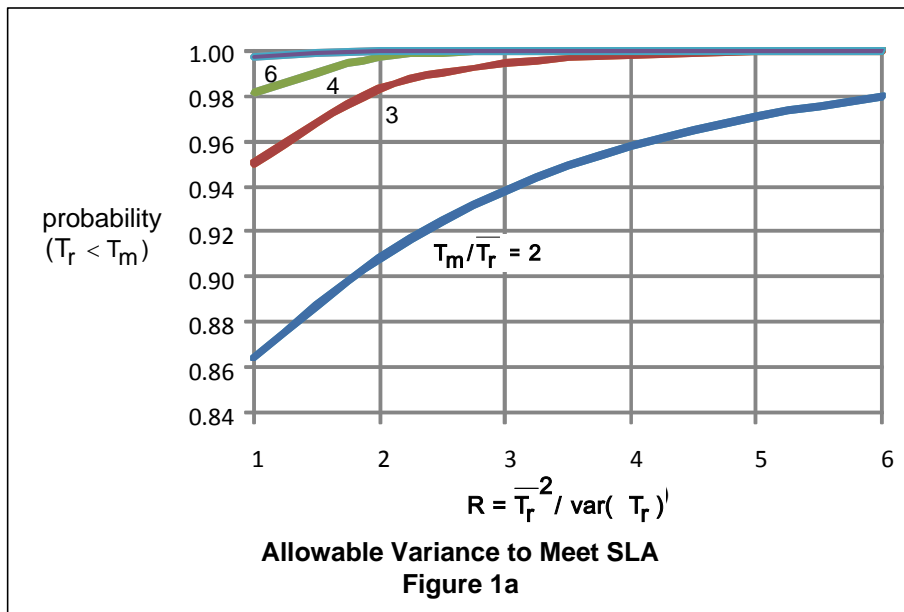
$T_m$  = the maximum response time specified by the SLA.

The Gamma distribution requires only one parameter,  $R$ , which is the ratio of the square of the average response time to the variance of the response time:

$$R = \frac{\bar{T}_r^2}{\text{var}(T_r)} \quad (1)$$

where

$T_r$  = response time  
 $\bar{T}_r$  = average response time  
 $\text{var}(T_r)$  = variance of the response time



Each curve in these figures represents a ratio of the maximum response time to the average response time,  $T_m/\bar{T}_r$ . For instance, the curve for  $T_m/\bar{T}_r = 3$  represents those probability values for the response time being less than three average response times as a function of  $R$ .

As an example, let us measure the response-time distribution of some server. It has an average service time of 5 msec., and we run it at a load of 50%. We find that the average response time at 50% load is 10 msec., and its variance is 40 (a standard deviation of 6.3 msec.). Thus,  $R$  is  $10^2/40 = 2.5$ . From Figure 1b, we see that the response time for this server will be less than three average response times (30 msec.) 99% of the time. Thus, we can state that 99% of all response times will be less than 30 msec. when the server is running at a 50% load.

However, having to make a physical measurement of the response times of a server under load may be difficult or impractical. Fortunately, there are several common cases in which the mean and variance can be calculated if the service-time distribution, average service time, and server load are known. These service-time distributions include the exponential distribution, the uniform distribution, and the constant distribution. For such distributions, the mean and variance of the response times as a function of mean service time and server load were given in Part 4 and are as follows:

- for exponential service times:

$$\begin{aligned}\bar{T}_r &= \frac{T_s}{(1-L)} \\ \text{var}(T_r) &= \frac{T_s^2}{(1-L)^2}\end{aligned}\tag{2}$$

- for uniform service times:

$$\begin{aligned}\bar{T}_r &= \frac{T_s}{(1-L)}\left(1 - \frac{L}{3}\right) \\ \text{var}(T_r) &= \frac{T_s^2}{(1-L)^2}\left(\frac{1}{3} + \frac{L^2}{9}\right)\end{aligned}\tag{3}$$

- for constant service times:

$$\begin{aligned}\bar{T}_r &= \frac{T_s}{(1-L)}\left(1 - \frac{L}{2}\right) \\ \text{var}(T_r) &= \frac{T_s^2}{(1-L)^2}\left(\frac{L}{3} - \frac{L^2}{12}\right)\end{aligned}\tag{4}$$

where

$T_s$  is the average service time for the server  
 $L$  is the load on the server (its utilization)

Spreadsheets for the probability of response times and Equations (2), (3), and (4) may be found at [http://www.availabilitydigest.com/public\\_articles/0403/composite\\_performance\\_sla.xls](http://www.availabilitydigest.com/public_articles/0403/composite_performance_sla.xls).

## Response-Time Distributions for Multiple Servers

Often, systems comprise a series of servers working in tandem, each having different characteristics – different service-time distributions, different capacities, and different loads. An event (such as a transaction) passes through each of these servers before generating a response.

Though the techniques summarized above can predict the response-time distribution for each of the servers, what is really desired is the response-time distribution for the set of servers taken as a whole.

This problem is solved based on three rules:

1. The mean of the sum of independent variables is the sum of their means. If  $z = x + y$ , then

$$\bar{z} = \bar{x} + \bar{y}$$

where the overbar denotes “mean.”

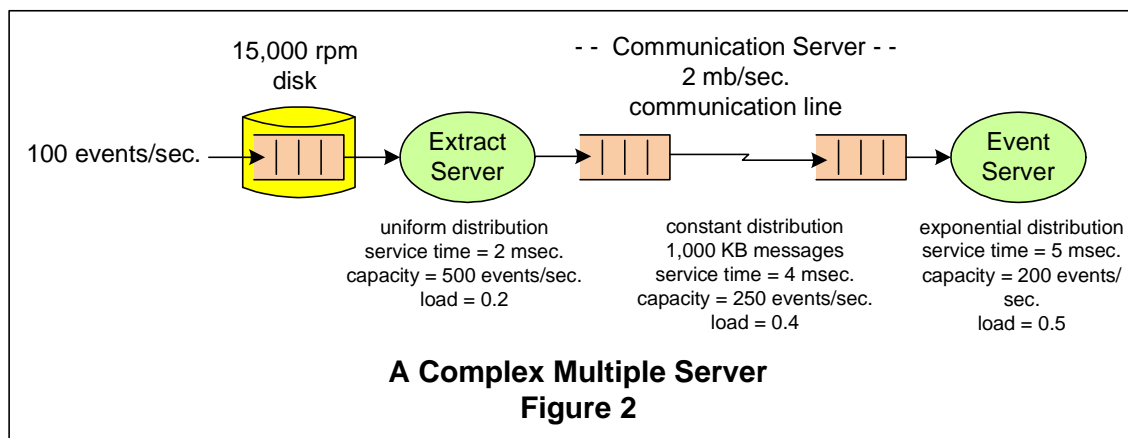
2. The variance of the sum of independent variables is the sum of their variances.<sup>1</sup> Using the above definition of  $z$ ,

$$\text{var}(z) = \text{var}(x+y) = \text{var}(x) + \text{var}(y)$$

3. If two or more variables follow a Gamma distribution, the sum of these variables also follows a Gamma distribution.<sup>2</sup>

Therefore, if we can determine the mean and variance of each of the servers in the tandem string, we simply need to add those means and variances to find the value of  $R$  for the system. We can then use the above techniques to answer the SLA question.

As an example, consider the system of Figure 2, in which incoming events contained in fixed-length messages are stored temporarily on disk, where they are queued until they are serviced. An Extract Server process reads the next event from the disk and sends it over a communication channel to an Event Server that will process it.



<sup>1</sup> W. H. Highleyman, Chapter 4, Basic Performance Concepts, pg 97, *Performance Analysis of Transaction Processing Systems*, Prentice-Hall; 1989.

<sup>2</sup> James Martin, Chapter 26, Probability and Queuing Theory, pg. 389, *Design of Real-Time Computer Systems*, Prentice-Hall; 1967.

In this system there are three servers:

1. The *Extract Server* reads the next event from the disk. Since events are stored randomly on the disk, the server must wait for the disk to rotate until that event is under the read head. With equal probability, this delay time can be anywhere from zero to  $n$  milliseconds, where  $n$  is the time that it takes for the disk to make one revolution. Thus, the service time for this server is uniformly distributed with a mean service time of  $n/2$ .
2. The *Communication Server* sends the event over the communication line. Since the messages are fixed length, the service time of the Communication Server is constant (ignoring the effect of retransmissions due to transmission errors).
3. The *Event Server* processes the event. We assume that its service time is exponential.

Figure 2 provides parameters for these servers:

- The event rate is 100 events per second.
- The disk rotates at 15,000 rpm. Thus, its rotational latency is 4 msec. The processing time of the Extract Server doing the extraction is measured in microseconds and can be ignored. Therefore, the mean service time of the Extract Server is 2 milliseconds. Its capacity is 500 events per second. At 100 events per second, it is carrying a load of  $100/500 = 0.2$ .
- The event message size is 1,000 kilobytes (8,000 kilobits). The communication line transmits at 2 megabits per second. Therefore, the service time of the Communication Server is 4 msec.; and it can handle 250 event messages per second. At 100 events per second, it is carrying a load of  $100/250 = 0.4$ .
- The Event Server has an exponentially-distributed service time with a mean service time of 5 milliseconds. It has a capacity of 200 events per second and is 50% loaded at an event rate of 100 events per second.

Using the spreadsheet referenced above, the following results are obtained for the average response time and the response time variance:

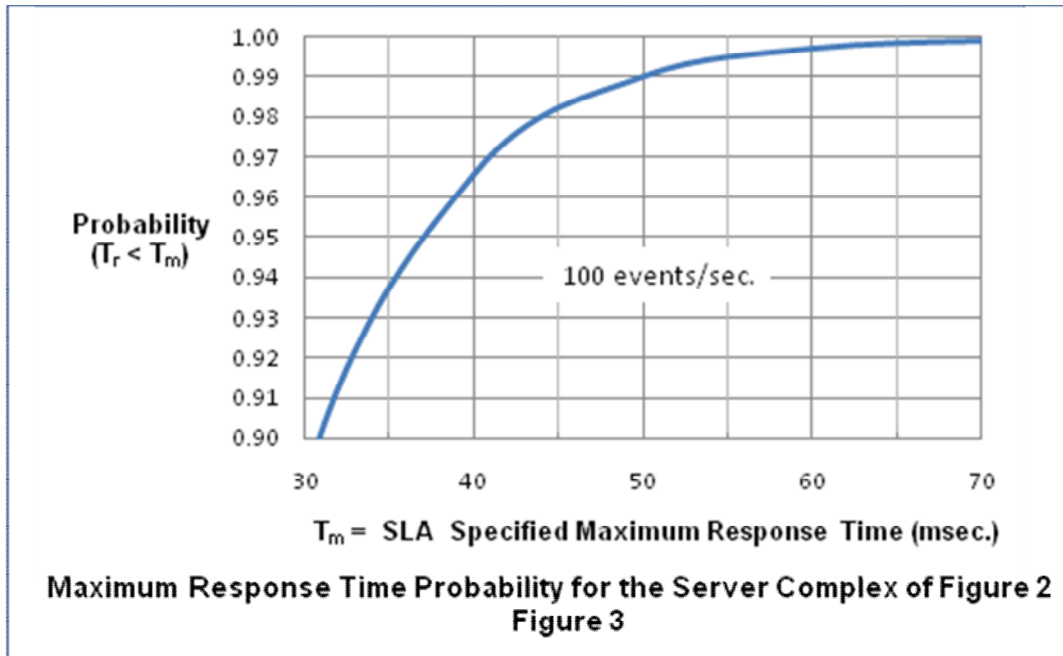
Server	Service Time Distribution	Mean Service Time (msec.)	Load	Mean Response Time (msec.)	Response Time Variance
Extract	uniform	2	0.2	2	.000002
Communications	constant	4	0.4	5	.000005
Event	exponential	5	0.5	10	.000100
			<b>Total</b>	<b>17</b>	<b>.000107</b>

**Response-Time Parameters for the Server Complex of Figure 2**  
Table 1

Summing these gives us a system mean response time of 17 msec. and a variance of .000107. Thus, the Gamma parameter  $R$  for this system is  $(.017)^2/.000107 = 2.70$ .

From Figure 1b or from the spreadsheet, we see, for example, that 99% of all response times will be less than 50 msec. ( $T_m/\bar{T}_r = 2.92$  from the spreadsheet). 95% will be less than 37 msec. ( $T_m/\bar{T}_r = 2.16$ ).

The maximum response time/probability curve for this example is given in Figure 3. From this curve, we can answer the question of what is the probability that a response will occur in less than a specified time.



If after going through this exercise, it is found that the SLA specification is not met, the only recourse (short of renegotiating the SLA) is to use faster servers – moving to a memory-resident input queue with disk backup, to a higher bandwidth communication line, or to a faster event server, for instance. The values given in Table 1 can be useful for deciding where to make the improvements in the system. The above calculation can then be repeated to verify the results.

## Summary

The Gamma distribution of a variable that is the sum of independent variables is itself a Gamma distribution. For complex systems of tandem servers with differing characteristics, this fact allows us to easily answer the SLA question, “What is the probability that the response time will be less than a specified amount?” If the SLA specification cannot be met, the details of the analysis serve as a valuable roadmap to the server modifications required to meet the SLA.

If the distribution of a server in the tandem configuration is unknown, an assumption of an exponential distribution is always conservative.

The charts of Figures 1a and 1b can be used to obtain approximate results. More accurate results can then be obtained from the referenced spreadsheet.