RJ10512 (ALM1309-008) September 18, 2013
Computer Science

IBM Research Report

Parameter Tolerance in Queueing Models

Ying Tat Leung
IBM Research Division
Almaden Research Center
650 Harry Road
San Jose, CA 95120-6099
USA

Manjunath Kamath, Juan Ma
School of Industrial Engineering and Management
Oklahoma State University
322 Engineering North
Stillwater, OK 74078
USA

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Parameter Tolerance in Queueing Models

Ying Tat Leung[1] Manjunath Kamath[2] Juan Ma[2]

[1]IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120
U.S.A.
ytl@us.ibm.com
1-408-927-2122
(Corresponding author)

[2]School of Industrial Engineering & Management
Oklahoma State University
322 Engineering North
Stillwater, OK 74078
U.S.A.
m.kamath@okstate.edu
juan.ma@okstate.edu

July 2013

Abstract

Analytical models based on queueing theory have been widely used in analyzing the dynamic behavior of manufacturing and service systems. Such models allow the user to easily experiment with different system designs or configurations for a given set of input parameters. However, an input parameter of the model could be inaccurate, due to, for example, estimation difficulties. In some cases a parameter may not be known (e.g., the customer order arrival rate for a new product) and we can only provide a good guess. In order to determine if the analysis results (and hence the system design) are robust to estimation errors, sensitivity analysis can be performed, using an analytical derivation or a numerical estimation.

In this paper we propose an alternative to the traditional approach of sensitivity analysis. We select a subset of the model parameters as the uncertain set and specify a tolerance range of a system performance measure, such as within 10% of the nominal value resulting from the baseline estimates of the parameters. We then calculate a feasible region of the uncertain parameters for which the performance measure will be within the tolerance limits. This is more convenient in cases where the system performance measure is required to be within a target range, or when it is an interface parameter to other models. We illustrate this approach by analyzing the basic exponential queues and then apply it to a more realistic model – a queueing model of a typical order fulfillment process in a distribution center.

Keywords*: queueing; parameter tolerance; sensitivity analysis; exponential queues; distribution center operation*

1. Introduction

In planning the capacity of a manufacturing or service operation, queueing models have long been recognized as a useful tool for decision support (see e.g. Buzacott and Shanthikumar 1993, Gans et al. 2003, Suri et al. 1995). These models can capture critical dynamic behavior of the system such as the number of parts or customers waiting in line for processing, and are practical in terms of data and computational requirements. As operations are increasingly outsourced to third-party providers, such models are correspondingly more useful. Operation-oriented performance measures estimated using these models, e.g., the average waiting/response time, will take an additional role as an external measure reported to and monitored by the outsourcing client. In some cases, the attainment or its failure has a direct impact on the financial rewards of a third-party provider. For example, a third-party logistics provider may provide a warehousing and customer order fulfillment service to its client who requires an incoming customer order for its goods to be shipped within 24 hours of order receipt on the average. At the end of each month, the logistics provider has to report statistics on the customer order handling times for all orders received that month, and may have to pay a financial penalty to its client if the order handling requirement is not met. The customer order handling time is the system time in a queueing model, making such models indispensable in planning the operation when new outsourcing client contracts are signed. Other similar situations arise in service businesses, such as customer service centers which can be walk-in facilities, or more commonly nowadays, telephone call centers. There, a common operation-oriented performance measure is how long an incoming customer has to wait before he is served by an agent, whether in person or on the phone. Typically, key performance measures of an operation and their target values (like those mentioned above) are specified in the service level agreement (SLA) of an outsourcing relationship.

Given an estimated business volume provided by the customer and the SLA specification, the service provider can plan its capacity in terms of the number of people and/or machines needed, and in more detail, the work schedule of these people and machines. One important aspect in planning the capacity of the service provider is analyzing the conditions under which the planned capacity becomes inadequate to deliver the performance required by the SLA. There are a number of sources of uncertainty that lead to an inadequate capacity. In this paper we focus on the following issues of estimation. First, the "expected" business volume, i.e., the arrival rate in a queueing model, provided by the customer is their best guess and may not be very accurate. For example, in information technology (IT) outsourcing it is not uncommon to have a customer being unaware of certain existing systems (hardware and software) that need to be supported. These systems will help generate a higher volume of support requests than the estimate. Similarly, the estimated amount of work per request, represented by the service time in a queueing model, as provided by the customer or estimated by the service provider itself, may not be accurate. In IT outsourcing, service requests are classified into types (e.g., desktops, then operating system) and the service provider may have historical statistics on serving requests of this type. However, each customer has slightly different configurations and requirements and the requests of a particular customer may not have the same characterization as that of the historical data.

A related issue not studied in this paper is that even though the estimated total business volume (and hence the arrival rate) may be fairly accurate over a long time

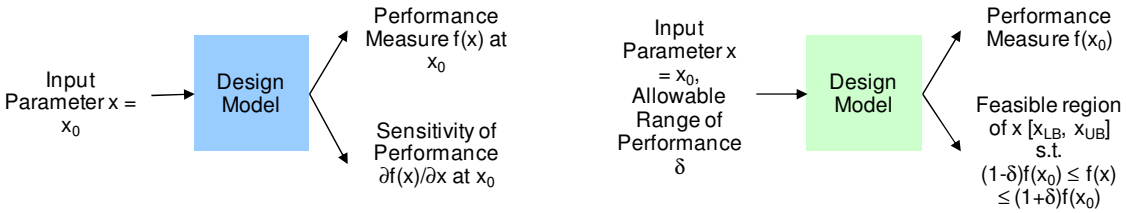
period, e.g. a month, the arrival rate may fluctuate significantly over time, such that the peak periods have much higher arrival rates than the long term average. For example, this phenomenon in telephone call centers is well documented, see, e.g. Gans et al. (2003). Characterizations of such variations over time may or may not be provided by the customer. Even if they are, estimates of arrival rates will be subject to the same, and most often higher, possibility of errors as the overall long term average.

In this paper, we assume that a queuing model is used to plan the capacity of a manufacturing or service operation, and study the conditions under which a specified target performance level will not be met with the planned capacity. Specifically, we ask the following question: For a given set of system parameters which include the estimated business volume (estimated arrival rate), the planned capacity (planned service rate), and a specified SLA, how much more business volume or reduction in planned capacity we can tolerate before the SLA is breached? In other words, what is the feasible region of a system parameter such as arrival rate or service rate such that a selected system performance measure such as average waiting time is within the SLA specification?

To gain some insights, we study the above question in two steps. First, in Section 3 we select a basic, analytically tractable situation where the ubiquitous M/M/1 and M/M/c queues are analyzed. Then, in Section 4, we study a more realistic example of a customer order fulfillment operation at a distribution center. We develop a simple but reasonably accurate queueing model for this operation and use it to answer our question above. These clearly represent basic steps in a subject not thoroughly explored in the literature which is reviewed in Section 2. Ultimately we would like to see such analysis as standard features in queueing model based capacity planning tools. Some additional concluding remarks are given in Section 5.

2. Related Concepts and Literature

A closely related concept that can be used to partially answer our research question is sensitivity analysis of performance measures. This typically gives the derivative or a derivative-like quantity of the performance with respect to a chosen system parameter. Of course, due to the nonlinearity of practically all queueing systems the feasible region cannot be directly deduced from the derivative information. Nevertheless the latter yields useful insights such as what parameter has the largest impact at the design point and hence represents a high risk area. Intuitively speaking, traditional sensitivity analysis is a *forward* calculation to obtain the difference in a performance measure given a change in a parameter, while the present study is a *backward* calculation of the allowable change in a parameter given a tolerance region of performance. Figure 2.1 summarizes the difference between sensitivity analysis and the present study denoted as parameter tolerance analysis.



Traditional Sensitivity Analysis **Parameter Tolerance Analysis**

Figure 2.1: Sensitivity Analysis vs. Parameter Tolerance Analysis

Kleijnen (1997) reviews different types of sensitivity analyses and thus develops a general framework to study them systematically. In that framework our present study falls under uncertainty analysis to quantify the effect of uncertain model inputs. Kleijnen commented that "uncertainty analysis has hardly been applied to stochastic models such as queueing models…" To a large extent this remains to be true even today, as seen by the sparse existence of such papers in the literature. Several works in sensitivity analysis of queueing models appeared before Kleijnen's paper, but few did after that.

Gordon and Dowdy (1980) analyze the effect of errors in relative utilization in a closed product-form queueing network on performance measures such as throughput, absolute utilization, and mean queue lengths. A key insight obtained is that for single-class load-independent networks the resulting errors in throughput and utilization are about the same percentage as that in the relative utilization estimates, but this is not true in networks with load-dependent servers or multiple customer classes. Sensitivity of more general performance functions in the form of an arbitrary function of the state of a BCMP network (open or closed) are obtained in Liu and Nain (1991). Similar to Gordon and Dowdy (1980), Tay and Suri (1985) contains a sensitivity analysis for closed queueing networks under the operational analysis framework rather than the classical stochastic product-form solution framework. Specifically they calculate bounds on performance measures such as throughput, utilization, queue lengths given errors in input parameters such as visit ratios and service time ratios.

Opdahl (1995) analyzes the performance sensitivity of a combined software-hardware model of a computer system, modeled as a queueing network under the operational analysis framework. In addition to providing feedback to designers for improving the system performance, the author proposes that "sensitivity analysis is useful for pointing out where model refinement and parameter capture effort should be focused..."

A more recent paper by Whitt (2006) studies the sensitivity of the performance of an M/M/c + M (multi-server exponential queue with abandonment) with respect to the arrival rate, service rate, and abandonment rate. Motivated by call center operations, different heavy traffic approximations are utilized to calculate the sensitivity results which are compared with results from finite difference methods.

More complex queueing models do not have analytical solutions and we have to resort to simulation to estimate the performance function. Efficient algorithms have been developed to compute the sensitivity, in the form of a gradient, alongside the performance function itself. A review of such techniques is contained in Fu (2006).

We also note that there is a second type of sensitivity in queueing models – the sensitivity of the performance with respect to some of the structural assumptions (rather than parameter values). For example, Suri (1983) studies the impact of some of the basic service time assumptions used in the aforementioned operational analysis framework; Davis et al. (1995) identifies scenarios in an Erlang loss model where the performance is sensitive to the service time distribution beyond its mean.

3. Parameter Tolerance Analysis for Exponential Queueing Models

In this section we utilize two basic exponential queues, namely, M/M/1 and M/M/c to motivate and develop our approach. Similar to the practical situations discussed in Section 1 but at a vastly simplified level, assume that we are planning our service capacity to serve a client who is sending their transactions to our server over a period of time under contract. The client informs us of their business volume in terms of a (long-run) transaction arrival rate and as well specifies a target average system time or target average waiting time of a transaction as part of the SLA. We can then calculate the required transaction service rate in order to meet the target average system or waiting time. (This is in fact the minimum required service rate to meet the SLA.) We call the system at this design point the *nominal system*.

To facilitate our discussion, we define the following notations:

λ	transaction arrival rate.
μ	transaction service rate.
T	average time a transaction spends in the system.
W	average waiting time of a transaction in queue.
$\lambda_0, \mu_0, \mathrm{T}_0, W_0$	corresponding parameters to the above in the nominal system.
x	half-width of the tolerance level promised in SLA.
p_λ	λ/λ_0
p_μ	μ/μ_0

3.1 The M/M/1 Case

For the input parameter tolerance of an M/M/1 queue, the problem can be described as follows. Given a nominal system specification, what is the feasible region of arrival rate λ and service rate μ such that the resulting average time in system (or average waiting time in queue) lies in the interval $(1 \pm x)\ \mathrm{T}_0$ (or $(1 \pm x)\ W_0$). To solve this problem, say for the case of the average system time, it is sufficient to solve the following inequality system:

$$\begin{cases} \lambda < \mu \\ (1-x)T_0 \le 1/(\mu - \lambda) \le (1+x)T_0 \\ \lambda, \mu > 0 \end{cases} \qquad \text{Equation 3.1}$$

The first inequality is to ensure stability of the queueing system and the second is to ensure that the average system time is within the tolerance region. If we are interested in characterizing the feasible region of λ and μ in terms of percentages of λ_0 and μ_0 respectively, we can simply replace λ with $p_\lambda * \lambda_0$ and μ with $p_\mu * \mu_0$ in the above inequality system. Note that p_λ and p_μ are positive scalars.

Equation 3.1 can be solved numerically and the results plotted using available commercial software such as Mathematica® and MATLAB®. In this paper, we specifically used Mathematica® version 8.0 (Wolfram Research 2010) as an inequality solver (by utilizing the built-in tool *Reduce*) and as a results visualization tool (by utilizing the built-in tool *Plot* for two-dimensional graphs and *Plot3D* for three-dimensional graphs respectively). For instance, give the nominal system specification ($\lambda_0 = 1, \mu_0 = 1.25, \mathrm{T}_0 = 4, W_0 = 3.2$), the feasible region of λ and μ is shown in Fig. 3.1.

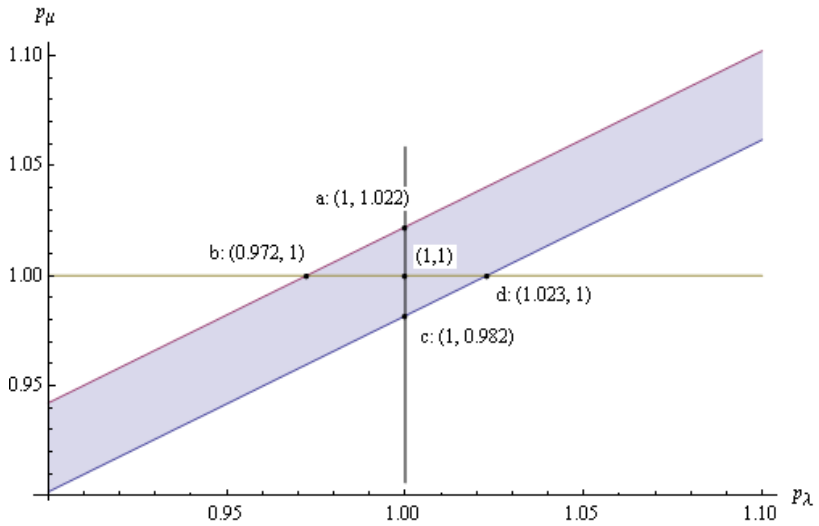


Figure 3.1: 10% Tolerance Region for the Nominal Averge System Time (T_0) in the M(1)/M(1.25)/1 System

The horizontal and vertical lines in Fig 3.1 show the tolerance in λ or μ if the other parameter is held at the nominal value. The entire shaded region between the two parallel lines shows the range of λ and μ for which the average system time is within the 10% tolerance zone of the nominal value, $T_0 = 4$. This is a partial feasible region of arrival rate and service rate satisfying the system of inequalities in Equation 3.1. Specifically, the values of points b and d on a horizontal axis are the lower and upper bounds respectively of λ, given that service rate μ is fixed to be μ_0. Similarly, the values of points a and c on a vertical axis are upper and lower bounds respectively of μ, given that arrival rate λ is fixed to be λ_0. As expected from the nonlinearity of queues, points a (b) and c (d) are not symmetrical with respect to the nominal point. Further, λ has a slightly larger tolerance range (in terms of percentages) than μ when the other parameter is held constant. This is good news since transaction arrival rates are usually more difficult to estimate than service rates.

In the following, we will show that the coordinates of points a, b, c and d are a function of nominal system utilization rate and half-width value of the tolerance zone, namely, x.

Coordinates of b and d can be obtained by solving

$$(1-x)T_0 \le 1/(\mu_0 - \lambda) \le (1+x)T_0 \qquad \text{Equation 3.2}$$

which yields

$$\Rightarrow \frac{\rho_0 - x}{(1-x)\rho_0} \le p_\lambda = \frac{\lambda}{\lambda_0} \le \frac{\rho_0 + x}{(1+x)\rho_0}$$

$$\Rightarrow \begin{cases} b: \left(\frac{\rho_0 - x}{(1-x)\rho_0}, 1\right) \\ d: \left(\frac{\rho_0 + x}{(1+x)\rho_0}, 1\right) \end{cases} \qquad \text{Equation 3.3}$$

Similarly, to get the coordinates of a and c we solve

$$(1-x)T_0 \le 1/(\mu-\lambda_0) \le (1+x)T_0 \qquad \text{Equation 3.4}$$

which yields

$$\Rightarrow \frac{1+\rho_0 x}{1+x} \le p_\mu = \frac{\mu}{\mu_0} \le \frac{1-\rho_0 x}{1-x}$$

$$\Rightarrow \begin{cases} a: (1, \dfrac{1-\rho_0 x}{1-x}) \\ c: (1, \dfrac{1+\rho_0 x}{1+x}) \end{cases} \qquad \text{Equation 3.5}$$

Numerical results for different system utilizations are given in Table 3.1

Table 3.1: 10% Tolerance Region for the Averge System Time (T) in an M/M/1 queue

ρ_0	p_λ: (b, d)	p_μ: (c, a)
	$x = 0.1$	$x = 0.1$
0.7	(0.9524, 1.0390)	(0.9727, 1.0333)
0.8	(0.9722, 1.0227)	(0.9818, 1.0222)
0.9	(0.9876, 1.0101)	(0.9909, 1.0111)

If average waiting time in queue is the promised target, the following inequality system can be solved to get the feasible region:

$$\begin{cases} \lambda < \mu \\ (1-x)W_0 \le \lambda/\mu(\mu-\lambda) \le (1+x)W_0 \\ \lambda, \mu > 0 \end{cases} \qquad \text{Equation 3.6}$$

Again replacing λ with $p_\lambda * \lambda_0$ and μ with $p_\mu * \mu_0$ will give the feasible region in terms of percentage of the nominal system parameter values. Similar to the case of time in system, we can solve equation 3.6 numerically using Mathematica or the equivalent. In Figure 3.2, the entire shaded region between the two lines (not straight nor parallel) shows the range of λ and μ for which the average waiting time is within the 10% tolerance zone of the nominal value, $W_0 = 3.2$. Figure 3.2 shows that the tolerance region for average waiting time in queue becomes smaller as both λ and μ decrease. The plot in Figure 3.3, in terms of the percentage of the nominal system parameter values, shows an enlarged view of the region of interest in the neighborhood of (λ_0, μ_0).

Similar to Fig. 3.1, Fig. 3.3 shows that the tolerance range of λ is larger than that of μ, when both parameters are not near zero. Comparing the shaded regions in Figures 3.1 and 3.3, we can also see that the tolerance ranges of λ and μ are smaller (in terms of percentages) in the case of W than the case of T. In other words, specifying SLA in terms of system time is less risky for a service provider, assuming that the specified tolerance percentages of the final performance measure remain the same. The latter assumption is not unreasonable, as the criteria of $(1 \pm x)$ deviation of the nominal system time or waiting time do not appear to be radically different.

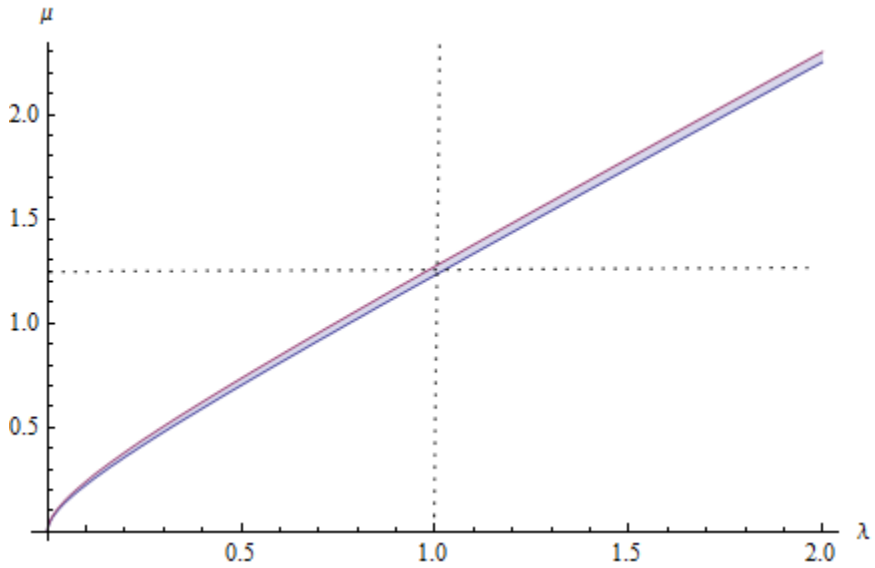


Figure 3.2: 10% Tolerance Region for the Nominal Averge Waiting Time (W_0) in the M(1)/M(1.25)/1 System

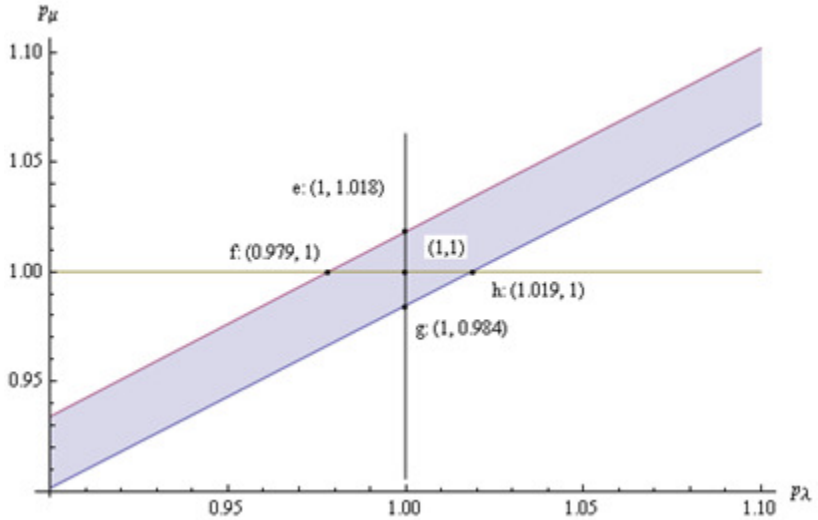


Figure 3.3: 10% Tolerance Region (enlarged) for the Nominal Averge Waiting Time (W_0) in the M(1)/M(1.25)/1 System

To get the coordinates of f and h, given

$$(1-x)W_0 \le \lambda/\mu_0(\mu_0-\lambda) \le (1+x)W_0 \qquad \text{Equation 3.7}$$

we get $\lambda_0[\frac{1}{\rho_0} - \frac{1-\rho_0}{\rho_0-\rho_0^2 x}] \le \lambda \le \lambda_0[\frac{1}{\rho_0} - \frac{1-\rho_0}{\rho_0+\rho_0^2 x}]$

$$\Rightarrow \frac{1-x}{1-\rho_0 x} \le p_\lambda = \frac{\lambda}{\lambda_0} \le \frac{1+x}{1+\rho_0 x}$$

$$\Rightarrow \begin{cases} f: (\frac{1-x}{1-\rho_0 x}, 1) \\ h: (\frac{1+x}{1+\rho_0 x}, 1) \end{cases}$$ Equation 3.8

To get the coordinates of e and g, given

$$(1-x)W_0 \leq \lambda_0/\mu(\mu-\lambda_0) \leq (1+x)W_0$$ Equation 3.9

we get

$$\Rightarrow \frac{\lambda_0}{2}[1+\sqrt{1+\frac{4(1-\rho_0)}{(1+x)\rho_0^2}}] \leq \mu \leq \frac{\lambda_0}{2}[1+\sqrt{1+\frac{4(1-\rho_0)}{(1-x)\rho_0^2}}]$$

$$\Rightarrow \frac{\rho_0}{2}[1+\sqrt{1+\frac{4(1-\rho_0)}{(1+x)\rho_0^2}}] \leq p_\mu = \frac{\mu}{\mu_0} \leq \frac{\rho_0}{2}[1+\sqrt{1+\frac{4(1-\rho_0)}{(1-x)\rho_0^2}}]$$

$$\Rightarrow \begin{cases} e: (1, \frac{\rho_0}{2}[1+\sqrt{1+\frac{4(1-\rho_0)}{(1+x)\rho_0^2}}]) \\ g: (1, \frac{\rho_0}{2}[1+\sqrt{1+\frac{4(1-\rho_0)}{(1-x)\rho_0^2}}]) \end{cases}$$ Equation 3.10

Numerical results for different system utilizations are given in Table 3.2

Table 3.2: 10% Tolerance Region for the Averge Waiting Time (W) in an M/M/1 queue

ρ_0	$p_\lambda: (f, h)$	$p_\mu: (e, g)$
	$x = 0.1$	$x = 0.1$
0.7	(0.9677, 1.0280)	(0.9787, 1.0252)
0.8	(0.9783, 1.0185)	(0.9846, 1.0182)
0.9	(0.9890, 1.0092)	(0.9917, 1.0100)

Finally, we solve equations 3.1 and 3.6 for a range of nominal average times in system and average waiting times in queue, and plot the 10% tolerance region in Figures 3.4 and 3.5 using Mathematica®. A slice of Fig. 3.4 (3.5) at a fixed T (W) will yield a figure similar to Fig. 3.1 (3.2). An interesting observation is that as the nominal values of T or W become smaller, the 10% tolerance region becomes wider because change in the average service time dominates, while for larger values of T or W, change in the average waiting time dominates as shown by the narrower tolerance region. A smaller T or W implies a lower utilization which usually means a higher operating cost per transaction. But in addition to greater customer satisfaction from less waiting, we also have a lower risk of not meeting SLA.

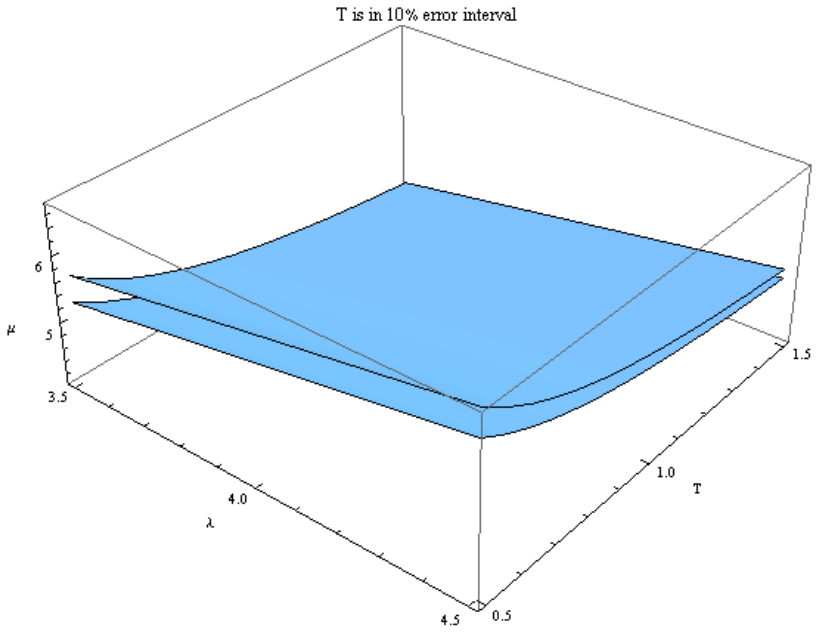


Figure 3.4: 10% Tolerance Region for Average Time in System

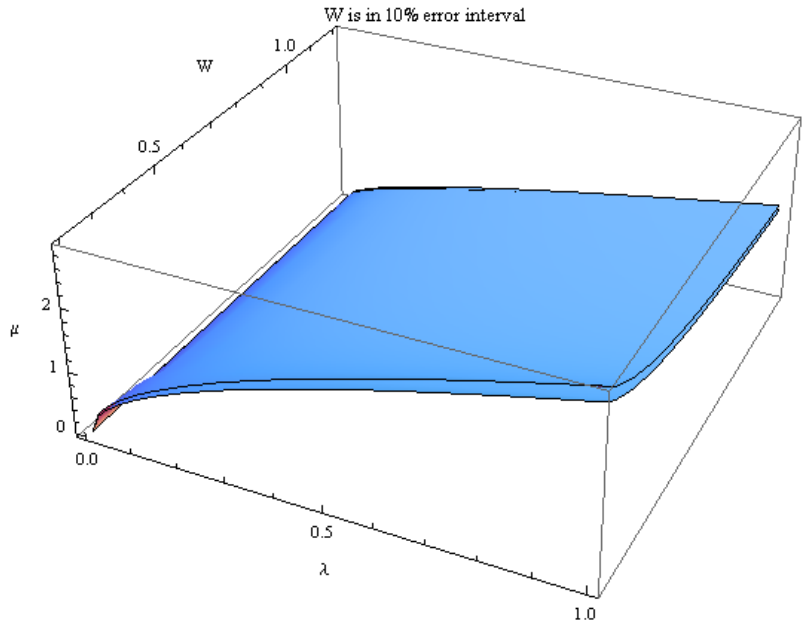


Figure 3.5: 10% Tolerance Region for Average Waiting Time in Queue

3.2 The M/M/c Case

In this section we consider a classic multi-server queueing model, namely, the M/M/c system. We use the notation defined previously, with the addition of *c*, the number of parallel servers or service channels. For simplicity of exposition, we use an

approximate expression for the average waiting time in queue for M/M/c, which is based on an approximation proposed for a GI/G/c queueing system by Sakasegawa (1977).

$$W = \frac{\rho^{\sqrt{2(c+1)}}}{\lambda(1-\rho)} \qquad \text{Equation 3.11}$$

where $\rho = \lambda/c\mu$.

In a manner similar to the M/M/1 case, to get the feasible region of the arrival rate and the service rate given that average time in system varies within an interval of $(1 \pm x)$, we need to solve the following inequality system:

$$\begin{cases} \lambda < c\mu \\ (1-x)T_0 \leq \dfrac{1}{\mu} + \dfrac{\rho^{\sqrt{2(c+1)}}}{\lambda(1-\rho)} \leq (1+x)T_0 \\ \lambda, \mu > 0 \end{cases} \qquad \text{Equation 3.12}$$

The feasible region obtained by solving the above inequality system is shown in Figure 3.6 for two different utilization levels, 70% (left column) and 90% (right column), and five different values of *c*, 1 (special case of single-server), 2, 7, 17, and the special infinite server case. This allows us to see how the feasible region changes as a function of both system load and the number of parallel channels.

As *c* increases, the feasible region changes from a narrow band between two steep parallel lines to a combination of an initial broader horizontal band transitioning to a narrow band between two almost linear boundary lines. Furthermore, as *c* increases, the horizontal band becomes well defined and longer, while the narrow band tends to become less steep. In the limiting case of the infinite-server queue, the entire region is a uniform, horizontal band. As *c* or the number of parallel servers increases, the growth in the initial broader horizontal band of the feasible region can be explained by the increasing dominance of the service time component of the time in system measure. In the limiting case, the feasible region is an $(1 \pm x)$ interval around the nominal value of the mean service time.

Comparing the plots in the left and right columns allows us to see the effect of the utilization level. By comparing the plots for the same *c,* one immediate observation is that the feasible region becomes tighter as the utilization increases. Similar to the single server case, the cost per transaction is lower with a system at a higher utilization but the risk of not meeting SLA is higher. Also, the growth of the initial horizontal band as *c* increases is slower at the higher utilization level. This can be explained by the greater influence of the waiting time component of the time in system measure as utilization increases.

In all the plots, we have kept the nominal service rate constant (=1.0). As *c* increases, the arrival rate will have to change to yield the desired utilization level (0.7 or 0.9). As the service time component becomes more dominant, the feasible region becomes more horizontal and more centered around the nominal service rate. This means that the system can tolerate larger deviations in the arrival rate and can still remain within $\pm x$ interval around the nominal value of the average time in system. The feasible region becomes tighter either as *c* decreases (fewer service channels) or utilization level increases.

To get the feasible region of the arrival rate and service rate given that average waiting time in queue varies within an $(1 \pm x)$ interval, one needs to solve the following inequality system:

$$\begin{cases} \lambda < c\mu \\ (1-x)W_0 \leq \dfrac{\rho^{\sqrt{2(c+1)}}}{\lambda(1-\rho)} \leq (1+x)W_0 \\ \lambda, \mu > 0 \end{cases} \qquad \text{Equation 3.13}$$

As before, the feasible region obtained by solving the above inequality system is shown in Figure 3.7 for two different utilization levels, 70% (left column) and 90% (right column), and four different values of *c*, 1 (special case of single-server), 2, 7, and 17. This allows us to see how the feasible region changes as a function of both system load and the number of parallel channels. In all the plots, we have kept the nominal service rate constant (=1.0). As *c* increases, the arrival rate will have to change to yield the desired utilization level (0.7 or 0.9).

As *c* increases, the shape of the feasible region remains essentially the same - a narrow band between two boundary lines. Furthermore, as *c* increases, the narrow band tends to become less steep or more horizontal. In the limiting case of the infinite-server queue, there is no waiting time and hence, the above exercise of finding a feasible region becomes meaningless.

Comparing the plots in the left and right columns allows us to see the effect of the utilization level. By comparing the plots for the same *c,* one immediate observation is that the feasible region becomes tighter as the utilization increases similar to the time in system case. This can be explained by the greater sensitivity of the waiting time to changes in either arrival or service rate as utilization increases.

From the graphs, for a fixed $c > 1$, we see that our comments earlier on the single server case on higher utilization resulting in higher risk, a larger tolerance in λ than that in μ, specification of tolerance in terms of T resulting in a lower risk than that of W, all apply. In addition, as the business volume scales up and the service provider employs more people or machine (i.e., increases *c*) to handle the volume, we see the following.

(1) The slope of the tolerance region is less steep. This means that when λ changes or we discover an error in λ, we may not have to change the service rate μ so much to compensate. In particular, a horizontal band means a fixed percentage change in μ can handle a relatively large range of λ and the horizontal band gets larger with increasing c.

(2) The area of the tolerance region around the nominal design point increases as *c* increases. This means that the system can tolerate a wider range of situations.

These are secondary, risk oriented advantages of economy of scale. (A primary advantage of economy of scale is the fact that we need less than 10x the number of servers to handle 10x the arrival rate to maintain the same system or waiting time, for a fixed service rate.)

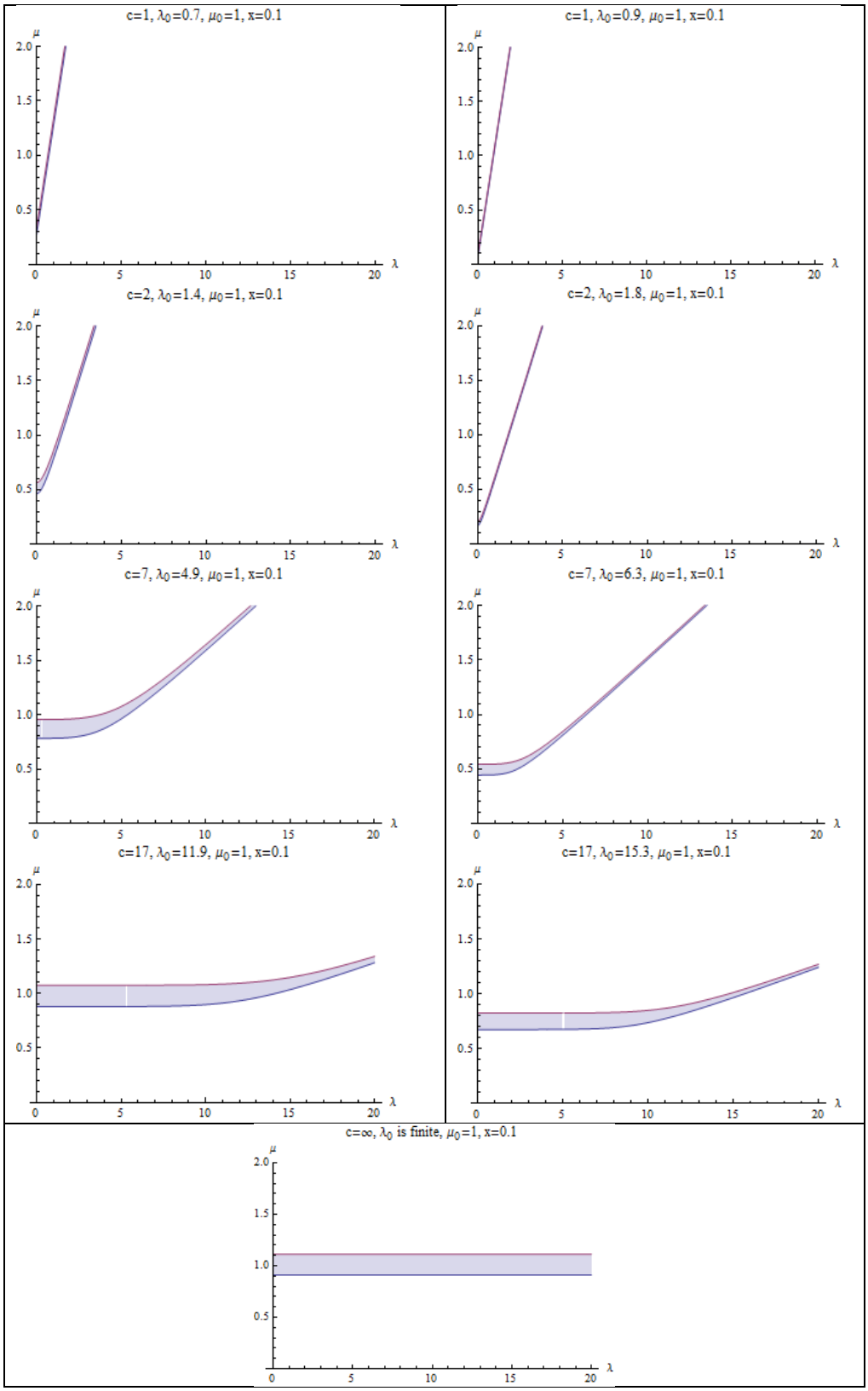


Figure 3.6: 10% Tolerance Region for Average Time in System for an M/M/c queue

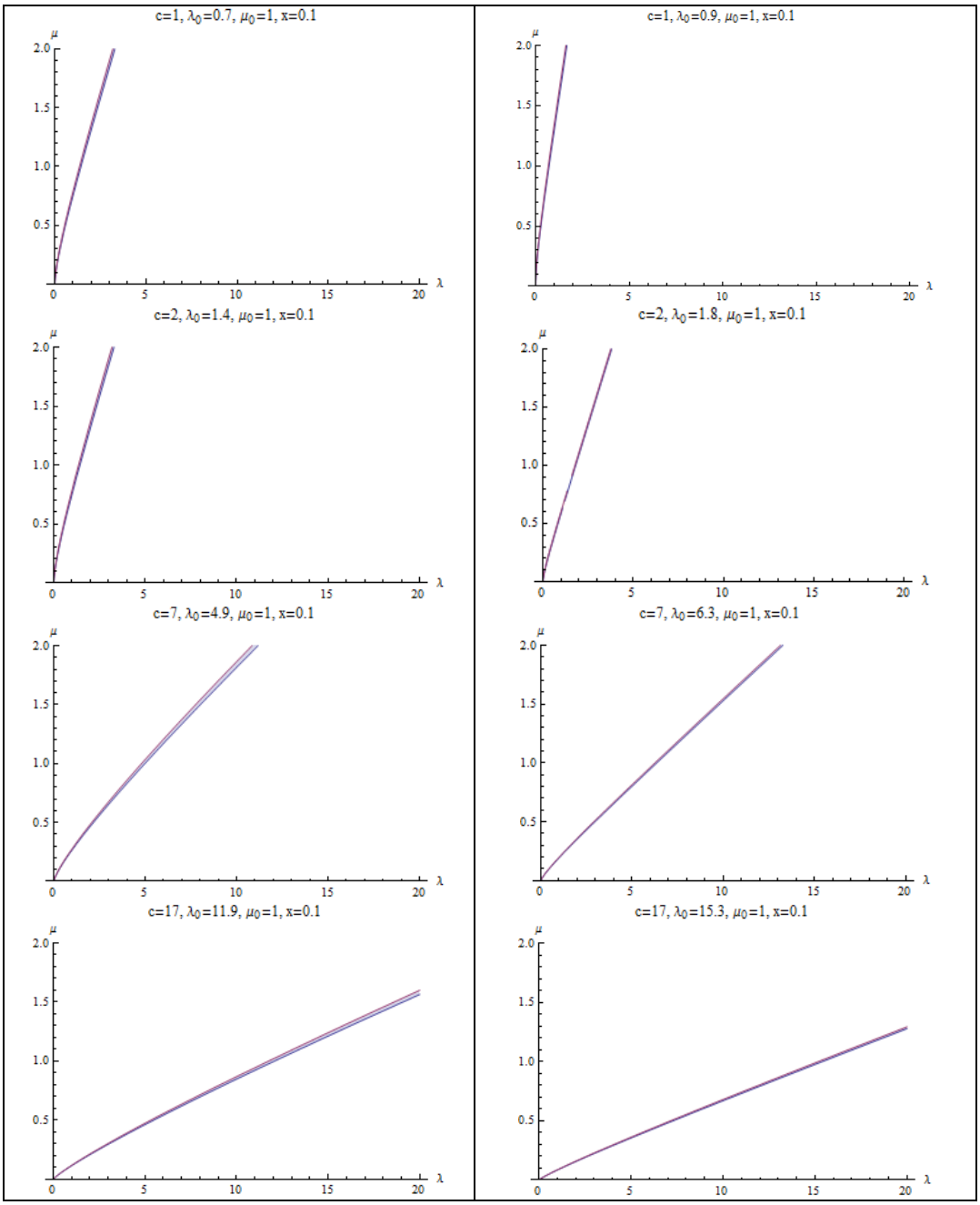


Figure 3.7: 10% Tolerance Region for Average Waiting Time for an M/M/c queue

4. Parameter Tolerance Analysis of a Distribution Center Operation Model

In this section we study a more realistic example of a customer order fulfillment operation at a distribution center (DC). This example was motivated by the work of Le-Duc & de Koster (2002, 2004), who modeled the order fulfillment operations in a distribution center shown in Figure 4.1. They assumed that orders arrived according to a Poisson process, each order has one order line and that k orders are batched for picking. The DC uses a random assignment policy for storing items in the storage racks and it was assumed that a picker travels at constant speed. Under these assumptions, Le-Duc and de Koster (2004) showed how to calculate the first and second moments of the pick time for a storage layout configuration with a central aisle shown in Figure 4.1. Le-Duc and de Koster (2002) showed that the order picking process can be modeled by a $M/G^k/1$ queue – a queue with batch service. They used the approach suggested by Tijms (1994) wherein the mean waiting time in a $M/G^k/1$ queue is approximated by a convex combination of the mean waiting times in a batch-service queue with deterministic service times and a batch-service queue with exponential processing times as follows:

$$W_{M/Gk/1} = (1 - c_S^2)\, W_{M/Dk/1} + c_S^2\, W_{M/Mk/1} ,$$

where c_S is the coefficient of variation of the service time distribution.

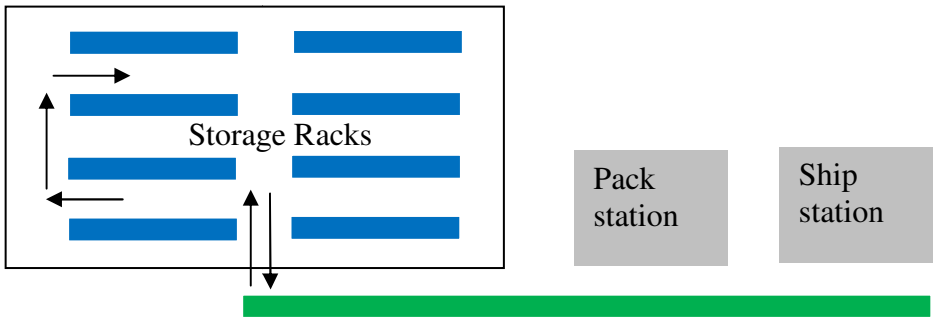


Figure 4.1: Distribution Center Operations

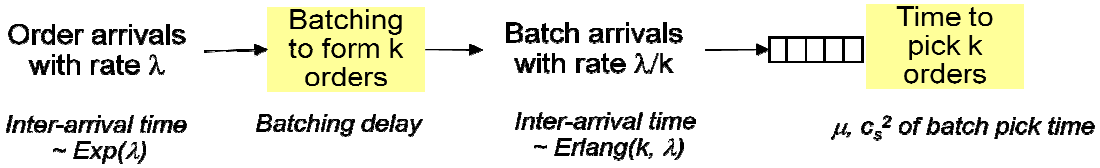


Figure 4.2: Modeling the Order Picking Process

Le-Duc and de Koster (2002) obtain $W_{M/Dk/1}$ and $W_{M/Mk/1}$ by solving the respective queueing models exactly as described in Gross & Harris (1998). We use an alternative approach to model the order picking process that allows us to still use the more realistic example while simplifying the exposition of the tolerance analysis procedure. The conceptual model shown in Figure 4.2 indicates that there are two main components of the average time to pick an order. The first component involves a batching delay and the

second is waiting for the order picker and the pick time. This is shown in Equation 4.1. We use a simple approximation for waiting time in a GI/G/1 queue for the waiting time in the order picker queue. As before, to facilitate our discussion, we define the following notations:

λ	order arrival rate.
c_a^2	Squared coefficient of variation (SCV) of the inter-arrival time of batches of orders.
μ	order picker service rate (for a batch of *k* orders).
c_s^2	SCV of the order picking time.
k	order picking batch size.
T	average time an order spends in the system.
$\lambda_0, \mu_0, \mathrm{T}_0$	corresponding system parameters in the nominal system.
x	half-width of the tolerance level promised in SLA.

Average time an order spends in the DC $= W_{\text{batch}} + (W_{\mathrm{GI}(\frac{\lambda}{k})/\mathrm{G}(\mu)/1} + S)$

$$E[T] \cong \frac{1}{\lambda}\frac{k-1}{2} + \frac{c_a^2 + c_s^2}{2} W_{\mathrm{M}(\frac{\lambda}{k})/\mathrm{M}(\mu)/1} + \frac{1}{\mu} \qquad \text{Equation 4.1}$$

$$\Rightarrow E[T] \cong \frac{1}{\lambda}\frac{k-1}{2} + \frac{c_a^2 + c_s^2}{2} \frac{\lambda}{(k\mu - \lambda)\mu} + \frac{1}{\mu}$$

4.1 Feasible Region in (λ, μ) for the Order Picking Operation

To get the feasible region of the order arrival rate and order picker service rate such that average system time T is within $(1 \pm x)\ \mathrm{T}_0$ where T_0 is the nominal average system time, it suffices to solve the following inequality system:

$$\begin{cases} \lambda < k\mu \\ (1-x)T_0 \le \dfrac{1}{\lambda}\dfrac{k-1}{2} + \dfrac{1/k + c_s^2}{2}\dfrac{\lambda}{(k\mu - \lambda)\mu} + \dfrac{1}{\mu} \le (1+x)T_0 \\ \lambda, \mu > 0 \end{cases} \qquad \text{Equation 4.2}$$

As the order arrival process is Poisson, the batch arrival process is Erlang-k, where k is the batch size. Hence, the SCV of the inter-arrival time to the order picker queue $c_a^2 = 1/k$. Figures 4.3 and 4.4 show plots of feasible region of (λ, μ) for the following two example configurations. In each plot, the nominal point has been identified by the dashed lines.

Case 1 (70% utilization) - $k = 4, \lambda_0 = 0.4, \mu_0 = \frac{1}{7}, c_a^2 = 0.25, c_s^2 = 0.2, \rho = 0.7$ and $T_0 = 14.425$.

Case 2 (90% utilization) - $k = 4, \lambda_0 = 0.4, \mu_0 = \frac{1}{9}, c_a^2 = 0.25, c_s^2 = 0.2, \rho = 0.9$ and $T_0 = 30.975$.

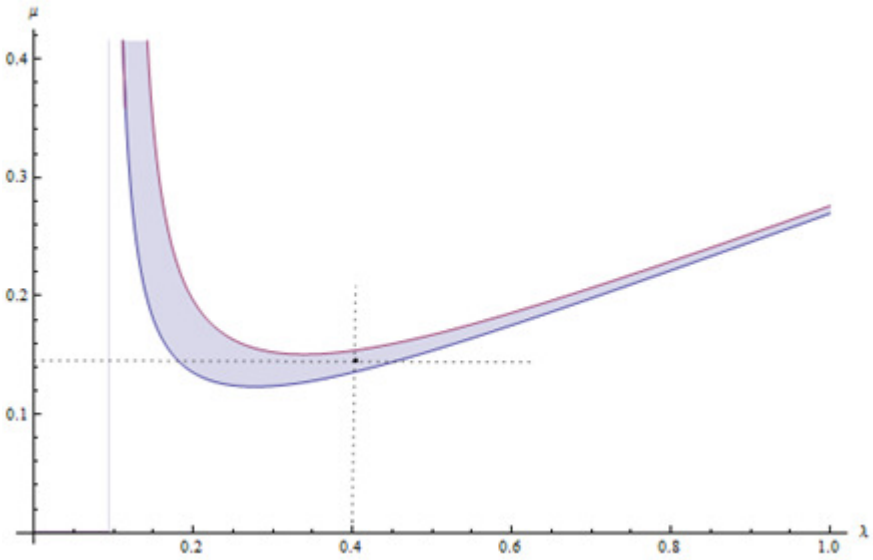


Figure 4.3: 10% Tolerance Region for Average Time in System for an Order in DC (70% uitilization)

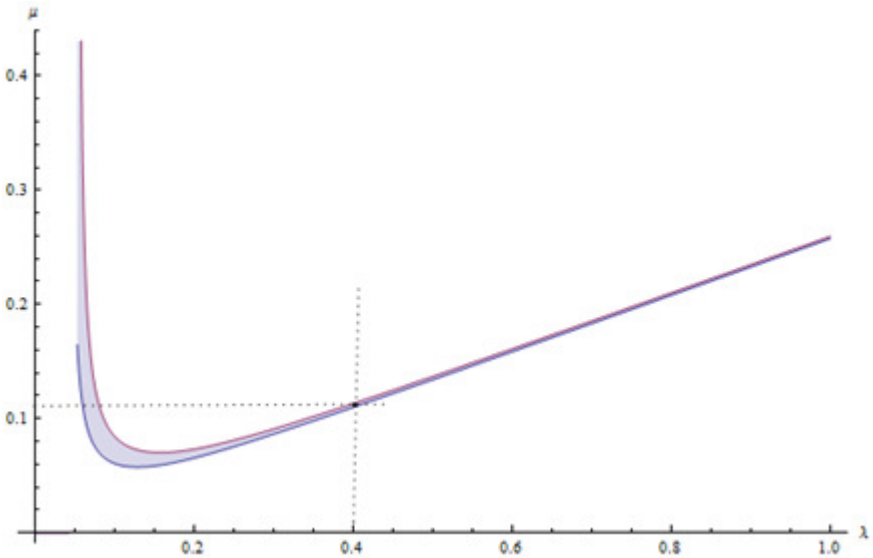


Figure 4.4: 10% Tolerance Region for Average Time in System for an Order in DC (90% uitilization)

Based on the two plots, we can make the following observations. When λ is small, the batching delay component dominates the average time in system measure, so μ has to be large to keep the waiting time and pick time smaller. When μ becomes very small, it is not possible to keep the system time within tolerance no matter how small λ is. By comparing the plots in figures 4.3 and 4.4, one immediate observation is that the feasible region becomes tighter as the utilization increases similar to the time in system case for the exponential queues in the previous section, resulting in a higher risk of not meeting the SLA. In a small neighborhood of the nominal design point, we see that the tolerance is again not symmetrical in two ways:

(1) Not symmetrical in μ (or λ) – the range of μ(or λ) is different depending on whether λ(or μ) is smaller or larger than the nominal point. In particular, the range of μ is smaller when λ is larger than the nominal point than that when λ is

smaller than the nominal point. This difference is rather small at low utilizations but increases when the utilization is higher. Therefore, at higher utilizations (which will be the norm in practice) we have to be more careful in estimating the order arrival rate. This is consistent with the more general observation above that the entire feasible region becomes smaller as utilization increases.

(2) Not symmetrical between μ and λ – the tolerance range for λ is larger for a given μ than that for μ for a given λ. Again this is advantageous in practice since order arrival rates are usually harder to estimate than service rates.

4.2 Feasible Region in (μ, c_s^2) for the Order Picking Operation

To get the feasible region of (μ, c_s^2), we solve the inequality system shown in Equation 4.2 by taking μ and c_s^2 as unknowns. This exercise allows us to develop some insight into the role played by the variability in the service (picking) process. Figures 4.5 and 4.6 show plots of feasible region of (μ, c_s^2), for the following two example configurations. In each plot, the nominal point has been identified by the dashed lines.

Case 1 (70% utilization) - $k = 4, \lambda_0 = 0.4, \mu_0 = \frac{1}{7}, c_a^2 = 0.25, c_s^2 = 0.2, \rho = 0.7$ and $T_0 = 14.425$.

Case 2 (90% utilization) - $k = 4, \lambda_0 = 0.4, \mu_0 = \frac{1}{9}, c_a^2 = 0.25, c_s^2 = 0.2, \rho = 0.9$ and $T_0 = 30.975$.

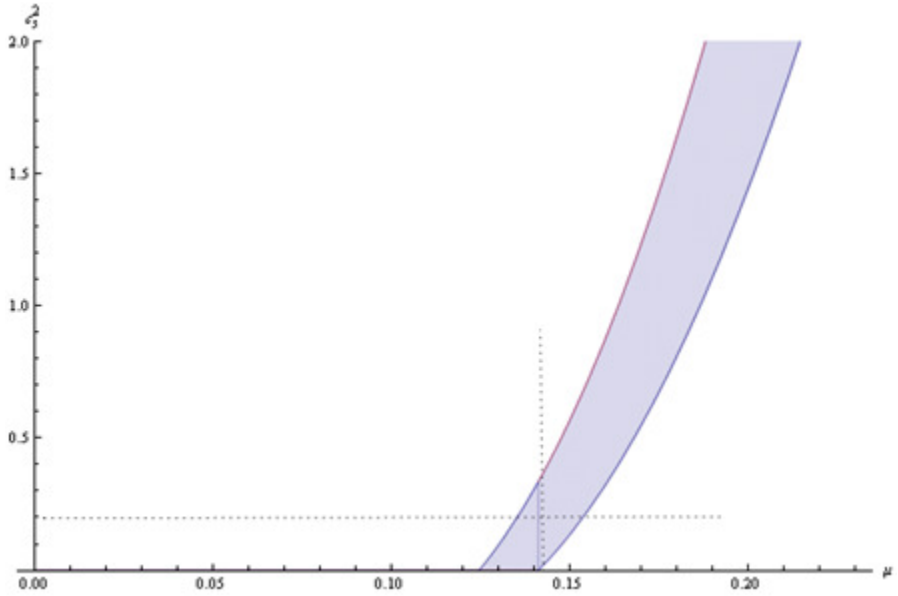


Figure 4.5: 10% Tolerance Region for Average Time in System for an Order in DC (70% uitilization)

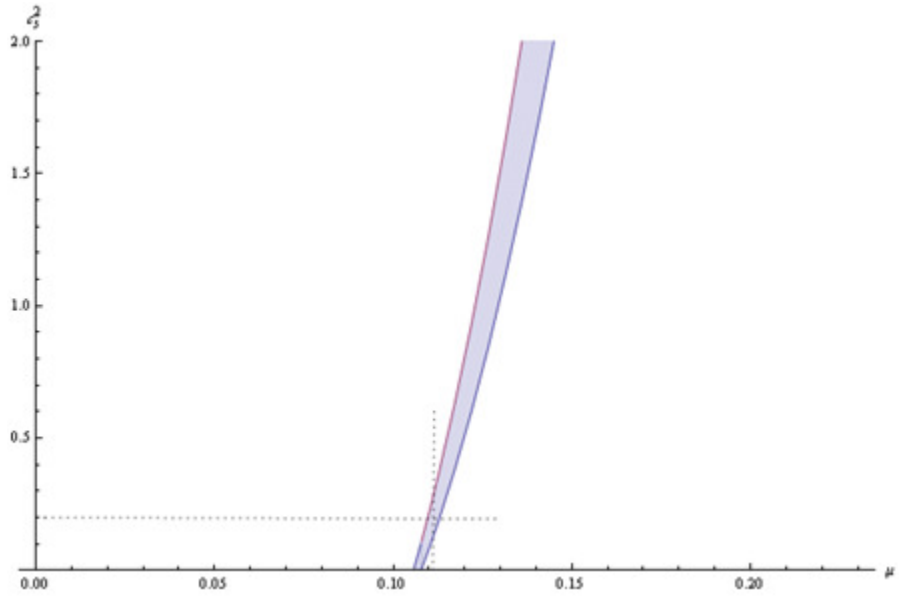


Figure 4.6: 10% Tolerance Region for Average Time in System for an Order in DC (90% uitilization)

Based on the two plots in figures 4.5 and 4.6, we can make the following observations. The feasible region becomes much tighter as the utilization increases similar to previous cases. As the service (picking) rate is increased, the tolerance region for c_s^2 becomes wider as indicated by the length of the vertical line within the feasible region for a particular μ. In both plots, the batching delay component remains fixed as μ and k are held constant. So the effect of the variability in the picking time is felt only through the waiting time for a batch of orders for the picker. As a higher service rate (μ) decreases both the waiting time and picking (service) time components, the system can tolerate higher levels of variability and still stay within the SLA.

5. Concluding Remarks

We introduce a form of sensitivity of the performance of a queueing system by finding the feasible region of selected model parameters that would result in an acceptable range of a given performance measure. Through basic exponential queues and a more complex model of a customer order fulfillment operation we learn some interesting properties of these feasible regions. Such an analysis provides complementary information to traditional sensitivity analysis, which usually takes the form of gradient estimation. In contrast to the latter, we call the type of analysis performed in this paper tolerance analysis. In practice, tolerance analysis is useful in analyzing the robustness of a system design, providing some concrete information for managing the risk of not conforming to performance targets. For example, the size of the feasible region of the most important parameters will give a sense of how likely the system will go out of performance specification. The size of the feasible region can hence be used to rank different system designs in terms of performance risk.

While we believe that tolerance analysis will give important information for operational risk management, many challenges remain. We use some simple examples in this first study, which we can compute manually due to the analytically tractability of the models and the small number of model parameters. For models with a large number

of parameters, visualizing the feasible region is not so straightforward. We can always resort to using a tabular form, but such a form does not convey insights as readily as a graphical form.

For models that are not analytically solvable, finding a feasible region may take more effort. Many queueing models do at least have a numerical solution. For these models, a straightforward way to find the feasible region of a system parameter is to do a search using the model. Since queueing models are often monotonic in a number of parameters (see, e.g., Shanthikumar and Yao 1989), we can use an efficient search technique such as a binary search in these cases. For example, to find the feasible region of arrival rate in a model that is monotonic with respect to this parameter, we first calculate the target average waiting time using the given parameters, then set the average waiting time to the maximum allowed based on the SLA and then search for the arrival rate that corresponds to the maximum average waiting time. The literature on monotonicity properties of queueing models will be useful to identify whether a specific model has the appropriate property.

For models that are not solvable even numerically, simulation is the only practical alternative. We can still use a search procedure to find a feasible region, but the total computational effort required may become prohibitive. Akin to the development of gradient estimation in simulations over two decades ago (e.g., Fu 2006), finding feasible regions in a simulation model may become a fruitful area for future research.

References

[1] Buzacott, J.A. and Shanthikumar, J.G. (1993), "Stochastic Models of Manufacturing Systems," Prentice Hall, Englewood Cliffs, NJ.

[2] Davis, J.L., Massey, W.A., and Whitt, W. (1995), "Sensitivity to the service-time distribution in the nonstationary Erlang loss model," Management Science, Vol. 41, No. 6, 1107-1116.

[3] Fu, M.C. (2006), "Stochastic gradient estimation," Chapter 19, Handbook on Operations Research and Management Science: Simulation, S.G. Henderson and B.L. Nelson, editors, Elsevier, 575-616.

[4] Gans, N., Koole, G. and Mandelbaum, A. (2003), 'Telephone call centers: tutorial, review, and research prospects', Manufacturing and Service Operations Management, Vol. 5, No. 2, 79-141.

[5] Gordon, K.D. and Dowdy, L.W. (1980), "The impact of certain parameter estimation errors in queueing network models," Proceedings of the 1980 international symposium on Computer performance modelling, measurement and evaluation, 3-9.

[6] Gross, D. and Harris, C. M. (1998), "Fundamentals of Queueing Theory," 3rd Ed., Wiley, New York.

[7] Kleijnen, J.P.C. (1997), “Sensitivity analysis & related analyses: A review of some statistical techniques,” J. Statist. Comput. Simul., Vol. 57, 111-142.

[8] Le-Duc, T. and De Koster, M.B.M. (2002), “Determining the Optimal Order Picking Batch Size in Single Aisle Warehouses”, ERIM Report Series Reference No. ERS-2002-64-LIS. Available at SSRN: http://ssrn.com/abstract=1097857

[9] Le-Duc, T. and De Koster, M.B.M. (2007), "Travel time estimation and order batching in a 2-block warehouse," European Journal of Operational Research, Elsevier, vol. 176(1), pages 374-388, January.

[10] Liu, Z. and Nain, P. (1991), “Sensitivity results in open, closed, and mixed product-form queueing networks,” Performance Evaluation Vol. 13 No. 4, 237-251.

[11] Opdahl, A.L. (1995), “Sensitivity analysis of combined software and hardware performance models: open queueing networks,” Performance Evaluation, Vol. 22, 75-92.

[12] Sakasegawa, H. (1977), “An approximation formula Lq ≃α·ρ^β/(1-ρ)”, Annals of the Institute of Statistical Mathematics, Vol. 29, No. 1, 67-75.

[13] Shanthikumar, J.G. and Yao, D.D. (1989), “Stochastic monotonicity in general queueing networks,” Journal of Applied Probability, Vol. 26, 413-417.

[14] Suri, R. (1983), “Robustness of queueing network formulas,” Journal of the ACM, Vol. 30, No. 3, 564-594.

[15] Suri, R., Diehl, G.W., de Treville, S., Tomsicek, M. (1995), “From CAN-Q to MPX: Evolution of queueing software for manufacturing,” Interfaces, Vol. 25, No. 5, 128-150.

[16] Tay, Y.C. and Suri, R. (1985), “Error bounds for performance prediction in queueing networks,” ACM Transactions on Computer Systems, Vol. 3, No. 3, 227-254.

[17] Tijms, H.C. (1994), “Stochastic models: an algorithmic approach,” New York: John Wiley & Sons.

[18] Whitt, W. (2006), “Sensitivity of performance in the Erlang-A queueing model to changes in the model parameters,” Operations Research, Vol. 54, No. 2, 247-260.

[19] Wolfram Research (2010), Inc., Mathematica, Version 8.0, Champaign, IL.