Parameter Tolerance in Capacity Planning Models

Ying Tat Leung
IBM Corporation
Bay Area Lab
Foster City, California 94404
U.S.A.
ytl@us.ibm.com

Manjunath Kamath
Oklahoma State University
School of Industrial Engineering
& Management
Stillwater, Oklahoma 74078
U.S.A.
m.kamath@okstate.edu

Juan Ma
Oklahoma State University
School of Industrial Engineering
& Management
Stillwater, Oklahoma 74078
U.S.A.
juan.ma@okstate.edu

Abstract

In capacity planning for a service operation, analytical models based on queueing theory allow the user to quickly estimate the capacity required and to easily experiment with different system designs or configurations, for a given set of input parameters. An input parameter of the model could be inaccurate or may not be known beyond a good guess. In order to determine if the analysis results (and hence the system design) are robust to parameter estimation errors, sensitivity analysis can be performed. We study an alternative approach that involves specifying a tolerance range of a system performance measure and calculating a feasible region of the uncertain parameters for which the performance measure will be within the tolerance range. We illustrate this approach using basic exponential queueing models as well as a model of an order fulfillment operation in a distribution center.

1. Introduction

In planning the capacity of a business 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, Gupta 2013, Mahdavi Pajouh and Kamath 2010, and 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, its attainment or failure has a direct impact on the financial rewards of a third-party provider. For example, a third-party logistics provider may provide a final assembly and customer order fulfilment service to its client who requires an incoming 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 order fulfillment times for all orders received that month, and may have to pay a financial penalty to its client if the fulfillment requirement is not met. The customer order fulfillment time is the system time in a queueing model, making such models indispensable in planning the operation when new outsourcing client contracts are signed.

Similar situations arise in other 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/she 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. Data centers, where arriving customers are machine requests, have similar SLA structures (e.g., Wustenhoff 2002).

Given an estimated business volume provided by the client and the SLA specification, the operation 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 operation 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 inadequate capacity. In this paper we focus on the following two issues in capacity estimation. First, the projected business volume, i.e., the arrival rate to the service or manufacturing system, provided by the client is their best guess and may not be very accurate.

URI: http://hdl.handle.net/10125/41343
ISBN: 978-0-9981331-0-2

For example, in information technology (IT) outsourcing it is not uncommon to have a client being unaware of certain existing systems that need to be supported. These systems will help generate a higher volume of support requests than the estimate. In call centers, arrival rates are known to be uncertain and its impact on performance has been studied using a simulation model (Robbins et al. 2006). Second, the estimated amount of work per customer arrival, represented by the service time in a queueing model, as provided by the client or estimated by the operation provider, may not be accurate.

In this paper, we assume that a queuing model is used to plan the capacity of a service operation, and 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 capacity can we tolerate before the SLA is breached? Or, what is the feasible region of the customer arrival rate and service rate such that a selected system performance measure is within the SLA specification? Although concepts discussed in this paper apply largely to both service and manufacturing operations, they are more important to service businesses since it is arguably more difficult to manage uncertainty in services for the lack of inventory as a buffering tool. Our work has been motivated by the needs of a service business and we will present our case in this context throughout the rest of the paper.

To illustrate our proposed approach and to gain some insights on its usefulness, we study the above question in the following manner. First, in Section 3 we select a basic situation where a single workstation modeled by the ubiquitous M/M/1 and M/M/c queues is analyzed. These models serve as convenient illustrations of our proposed approach. Then, in Section 4, we study a customer order fulfillment operation at a distribution center, where we show that our approach is feasible in a more complex example of a capacity planning model. 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 a standard feature in queueing model based capacity planning tools. Additional concluding remarks are given in Section 5.

2. Related Concepts in the 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, 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. Fig. 1 contrasts the two approaches. Each approach serves a slightly different purpose. In the context of planning for capacity of a service operation, especially under an outsourcing SLA, the proposed concept of tolerance analysis has some advantages. It is a direct reflection of typical terms in an SLA; it gives the entire feasible region in one step, providing a more comprehensive view; one can look up examples of extreme values in the feasible region to obtain more tangible insights; plots of feasible regions in the parameter space are friendly to, and therefore more likely to be considered by, a practitioner.

Kleijnen (1997) reviews different types of sensitivity analyses and 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…” This remains to be true even today. 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 on performance measures in a closed product-form queueing network such as throughput, absolute utilization and mean queue lengths. Sensitivity of more general performance functions in the form of an arbitrary function of the state of a 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, obtaining bounds on performance measures given errors in input parameters.

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 improving 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.

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 performance 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 this in a queueing network using operational analysis. Other papers analyze the sensitivity of the performance results when the actual service time distribution function of a queueing system is not what was assumed (typically exponential), e.g., Davis et al. (1995).

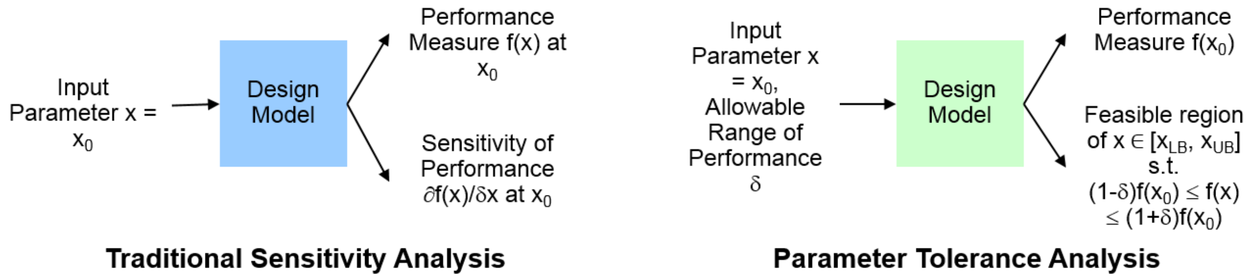
Input Parameter x = x0
Design Model
Performance Measure f(x) at x0
Sensitivity of Performance ∂f(x)/δx at x0
Traditional Sensitivity Analysis
Input Parameter x = x0, Allowable Range of Performance δ
Design Model
Performance Measure f(x0)
Feasible region of x ∈ [xLB, xUB] s.t. (1-δ)f(x0) ≤ f(x) ≤ (1+δ)f(x0)
Parameter Tolerance Analysis

Fig. 1. Comparison of Sensitivity Analysis & Parameter Tolerance Analysis

3. Parameter Tolerance Analysis for a Single Workstation

Similar to the practical situations discussed in Section 1, but at a simplified level, assume that we are planning the capacity of a service operation, consisting of a single workstation, to serve a client who is sending their transactions to our workstation 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 a target average system time for a transaction as part of the SLA. We then calculate the required transaction service rate in order to meet the target average system time. (This is in fact the minimum required service rate.) We call the system at this design point the nominal system. We define the following notations:

λ (μ) transaction arrival (service) rate;

T average time a transaction spends in the system;

λ_0, μ_0, T_0 the above quantities in the nominal system;

x half-width of the tolerance range; (SLA specification is typically one-sided – see explanation below)

p_λ (p_μ) $= \lambda/\lambda_0$ (μ/μ_0).

3.1. The M/M/1 Case

For a workstation with a single server modeled as an M/M/1 queue, our problem is that, given a nominal system specification, what the feasible region is for the values of arrival rate λ and service rate μ, such that the resulting average time in system lies within $(1\pm x)T_0$. We need to solve the following inequality system:

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

The first inequality is to ensure stability of the queueing system, the second the average system time (of an M/M/1 queue, e.g. Buzacott & Shanthikumar 1993) within the tolerance region. We include a lower bound for the average system time for completeness and for its potential usefulness in analyzing a priority type arrangement. It can be removed easily if one so desires. To characterize the feasible region of λ and μ in terms of percentages of λ_0 and μ_0 respectively, we 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.

Eq. (1) can then be solved analytically in terms of p_λ and p_μ, and the result for a specific numerical instance can be plotted using available commercial software. In this paper, we used Mathematica® version 8.0 (Wolfram Research 2010) as a results visualization tool (by utilizing the built-in tools Plot and Plot3D for 2- and 3-dimensional graphs respectively). For instance, given the nominal system specification (λ_0=1, μ_0=1.25, T_0=4), the feasible region of p_λ and p_μ is shown in Fig. 2.

In Fig. 2, the shaded region between the two parallel lines shows the range of p_λ and p_μ for which the average system time is within the 10% tolerance zone of the nominal value of T_0=4. This is a partial feasible region of arrival rate and service rate satisfying Eq. (1). The points b and d are respectively the lower and upper bounds of p_λ, given that service rate μ is fixed at the nominal value μ_0. Similarly, the points a and c are respectively the upper and lower bounds of p_μ, given a fixed arrival rate $\lambda = \lambda_0$. As expected from the nonlinearity of queues, points a (b) and c (d) are not symmetrical with respect to the nominal point (1, 1). 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.

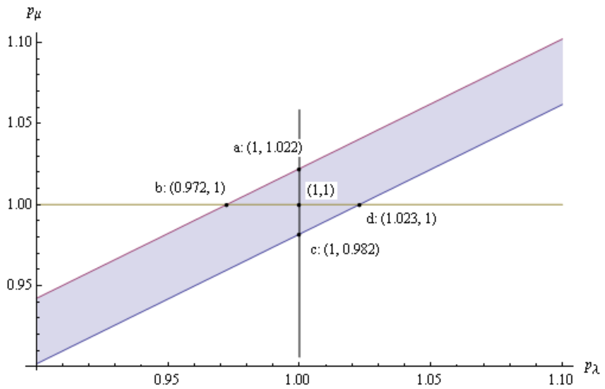

Fig. 2. The 10% tolerance region for the nominal average system time (T_0) in the M(1)/M(1.25)/1 system

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.

Coordinates of b and d can be obtained by solving

$$(1-x)T_0 \leq 1/(\mu_0 - \lambda) \leq (1+x)T_0 \qquad (2)$$

We can convert the inequality system into expressions of p_λ, x, and ρ_0 by plugging in the terms p_λ=λ/λ_0 and ρ_0=λ_0/μ_0. We obtain the coordinates as follows.

$$\begin{cases} b: ((\rho_0 - x)/[(1-x)\rho_0], 1) \\ d: ((\rho_0 + x)/[(1+x)\rho_0], 1) \end{cases} \qquad (3)$$

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

$$(1-x)T_0 \leq 1/(\mu - \lambda_0) \leq (1+x)T_0 \qquad (4)$$

to obtain the coordinates as follows.

$$\begin{cases} a: (1, (1-\rho_0 x)/(1-x)) \\ c: (1, (1+\rho_0 x)/(1+x)) \end{cases} \qquad (5)$$

Numerical results for the ranges of p_λ and p_μ with different system utilizations are given in Table I.

Finally, we solve Eq. (1) for a range of nominal average times in system and plot the 10% tolerance region in Fig. 3. A slice of Fig. 3 at a fixed T will yield a figure similar to Fig. 2. An interesting observation is that as the nominal values of T become smaller, the 10% tolerance region becomes wider because average service time dominates T, while for larger values of T, the average waiting time dominates T. A smaller T 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.

Table I. 10% Tolerance Region for the Average System Time (T) in an M/M/1 queue

$\boldsymbol{\rho_0}$	$\mathbf{p_\lambda}$**: (b, d)**	$\mathbf{p_\mu}$**: (a, c)**
	$\boldsymbol{x = 0.1}$	$\boldsymbol{x = 0.1}$
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)

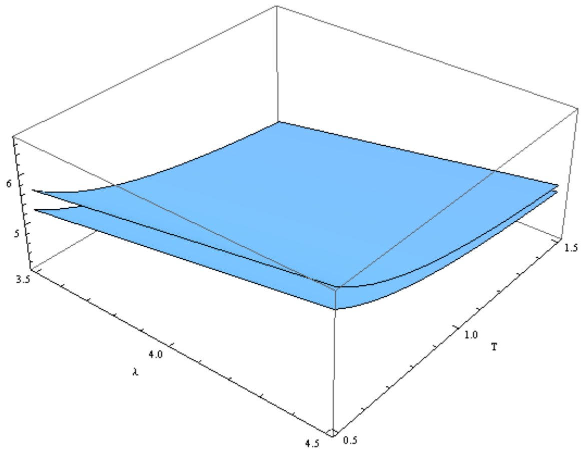

Fig. 3. The 10% tolerance region for average time in system

3.2. The M/M/c Case

For a multi-server workstation modeled as an M/M/c queue, we use an approximate expression for the average waiting time in queue, rather than the exact solution since the approximation gives a much simpler expression yet is adequate to serve our purpose. The expression is based on a well-known approximation proposed for a GI/G/c queueing system by Sakasegawa (1977). Let $\rho=\lambda/(c\mu)$, where c is the number of parallel servers. Then

$$W = \rho^{\sqrt{2(c+1)}}/[\lambda(1-\rho)] \quad (6)$$

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 $(1\pm x)T_0$, we need to solve the following inequality system:

$$\begin{cases} \lambda < c\mu \\ (1-x)T_0 \le 1/\mu + \rho^{\sqrt{2(c+1)}}/[\lambda(1-\rho)] \le (1+x)T_0 \\ \lambda, \mu > 0 \end{cases}$$
(7)

The feasible region obtained is shown in Fig. 4 for two different utilization levels, 70% (left column) and 90% (right column), and five different values of c: 1, 2, 7, 17, and the special infinite-server case.

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 trailed by a narrow band between two almost linear boundary lines. Furthermore, the horizontal band becomes longer, while the narrow band tends less steep. In the limiting case of the infinite-server queue, the feasible region is a uniform, horizontal band. As c increases, the growth in the initial broader horizontal band of the feasible region can be intuitively explained by the increasing dominance of the service time component of the time in system measure. In the limiting case, the feasible region is simply an $(1\pm x)$ interval around the nominal value of the mean service time.

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 the $(1\pm x)$ interval around the nominal average time in system. The feasible region becomes tighter as c decreases or utilization increases.

Comparing the plots in the left and right columns shows the effect of utilization with the same c. For a fixed $c > 1$, we see that our comments earlier on the single server case on higher utilization resulting in lower cost, but lower customer satisfaction and higher risk, and a larger tolerance in λ than that in μ apply. In addition, as the business volume scales up and the service provider employs more people or machines to handle the volume, we see the following.

1. The slope of the tolerance region is less steep and the horizontal section gets larger. 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.
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 is that we need less than 10x the number of servers to handle 10x the arrival rate to maintain the same system time, for a fixed service rate.)

The graphs shown in Figs. 2-4 are of course derived from known theoretical results in queueing theory. Our intention is to use them as feasibility tests to see if the proposed tolerance analysis can produce any useful insights for a practitioner who may not be well versed in queueing theory.

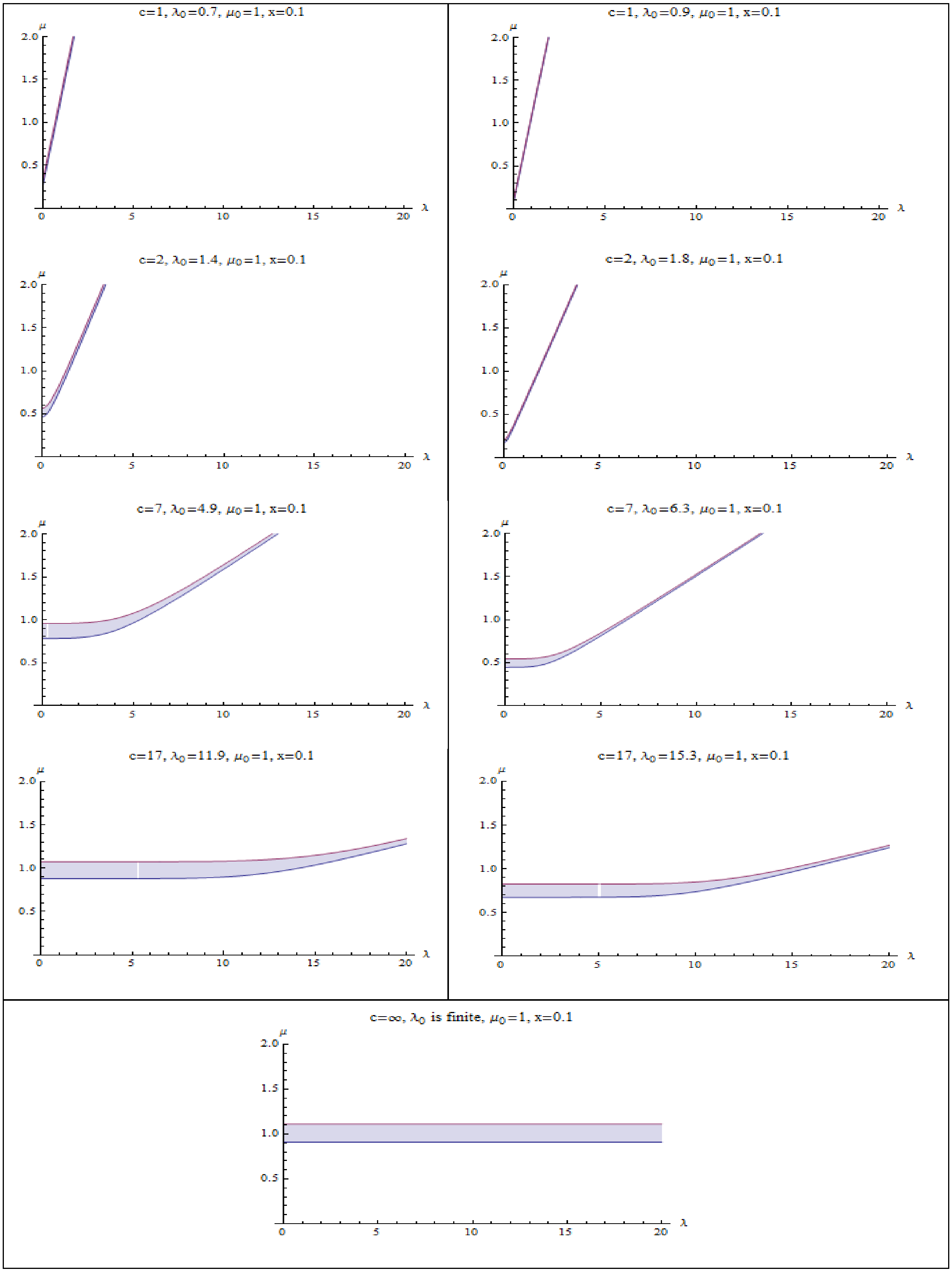

Fig. 4. 10% tolerance region for average time in system in an M/M/c queue

4. Parameter Tolerance Analysis of a Distribution Center Operation

In this section we study a more complex example motivated by the work of Le-Duc & de Koster (2002, 2004), who modeled an order fulfillment operation in a distribution center (DC; see Fig. 5). They assumed that customer orders arrived according to a Poisson process, each order having 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 pickers are assumed to travel at a 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 and that the order picking process can be modeled by an M/G$_k$/1 queue – a queue with batch service. To solve the latter, they used the approach suggested by Tijms (1994) using a convex combination of a batch-service queue with deterministic service times and one with exponential processing times. We use an alternative approach to model the order picking process that is simpler, as shown conceptually in Fig. 6. 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 Eq. (8).

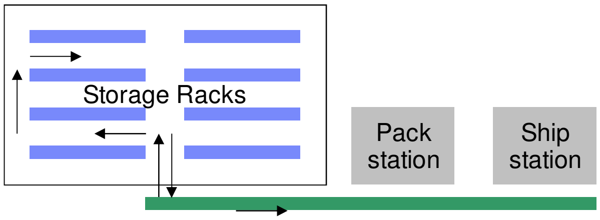

Fig. 5. Customer Order Fulfillment at a Distribution Center

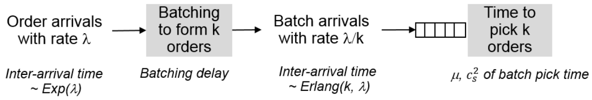
Order arrivals with rate λ
Batching to form k orders
Batch arrivals with rate λ/k
Time to pick k orders
Inter-arrival time ~ Exp(λ)
Batching delay
Inter-arrival time ~ Erlang(k, λ)
μ, c_s^2 of batch pick time

Fig. 6. Modeling the Order Picking Process

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.
λ_0, μ_0, T_0 respective quantities in the nominal system.
x half-width of the tolerance region (obtained from SLA specifications).

Then, the average time an order spends in the DC is:

$$E[T] = W_{\text{batch}} + (W_{\text{GI}(\lambda/k)/\text{G}(\mu)/1} + S) \quad (8)$$

$$\cong (k-1)/(2\lambda) + [(c_a^2 + c_s^2)/2]W_{\text{M/M/1}} + 1/\mu \quad (9)$$

$$\cong \frac{k-1}{2\lambda} + \frac{[(c_a^2+c_s^2\,)\lambda]}{[2(k\mu-\lambda)\mu]} + \frac{1}{\mu} \quad (10)$$

To obtain Eq. (9), we calculate each of the expected waiting times as follows. For the waiting time in the order picker queue, we use a well known approximation for GI/G/1 queues (Whitt 1993). For the batching delay, we observe that the expected waiting time for an arriving job to a batch seeing *j* jobs already in the batching queue is (*k*-*j*-1)/λ. Modeling the batching queue as a continuous time Markov chain, we can obtain the probability of an arriving job seeing *j* jobs to be 1/*k*. Hence, the expected batching delay is: $\frac{1}{k}\sum_{j=0}^{k-1}\frac{k-j-1}{\lambda} = \frac{k-1}{2\lambda}$.

4.1. Feasible Region in (*λ*, *μ*) for the Order Picking Process

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

$$\begin{cases} \lambda/k < \mu \\ (1-x)T_0 \le T \le (1+x)T_0 \\ \lambda, \mu > 0 \end{cases} \quad (11)$$

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 $c_a^2 = 1/k$. Figs. 7 and 8 show the feasible region of (λ, μ) for the following two example configurations. The nominal point is identified by the intersection of the dashed lines.

Case 1 (70% utilization):
$k = 4, \lambda_0 = 0.4, \mu_0 = 1/7, c_a^2 = 0.25, c_s^2 = 0.2, \rho = 0.7$ & $T_0 = 14.425, x = 10\%$.

Case 2 (90% utilization):
$k = 4, \lambda_0 = 0.4, \mu_0 = 1/9, c_a^2 = 0.25, c_s^2 = 0.2, \rho = 0.9$ & $T_0 = 30.975, x = 10\%$.

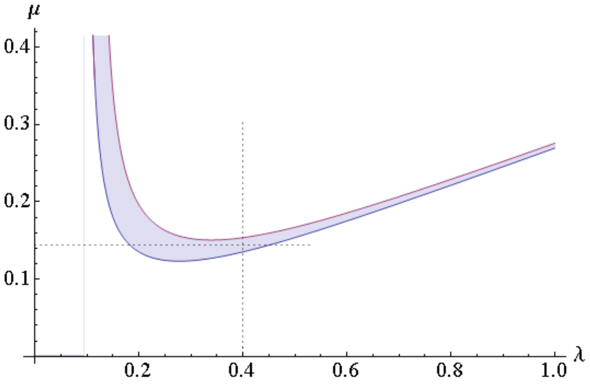

Fig. 7. 10% tolerance region for (λ, μ) at 70% utilization

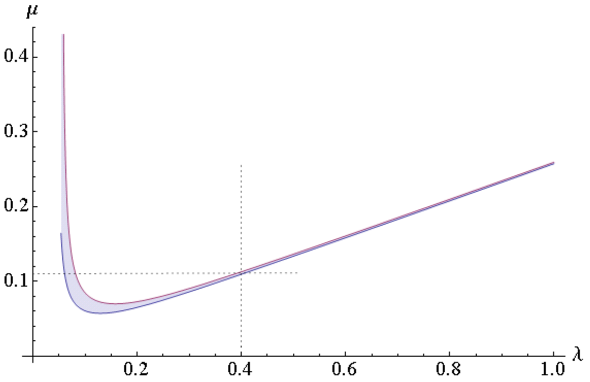

Fig. 8. 10% tolerance region for (λ, μ) at 90% utilization

When λ is small, the batching delay component dominates the average time in system, so μ has to be large to keep the waiting time and pick time small. 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 Figs. 7 and 8, 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, the tolerance range 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.
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.

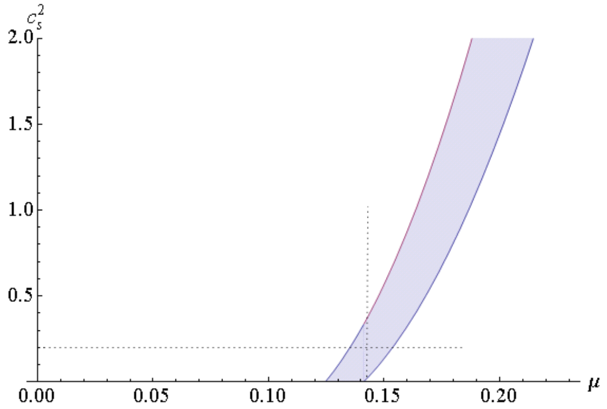

Fig. 9. 10% tolerance region for (μ, c_s^2) at 70% utilization

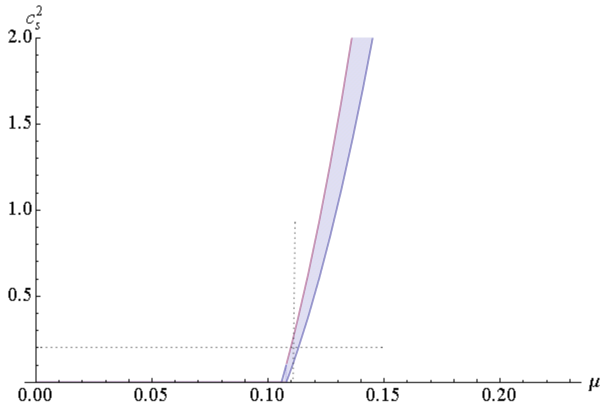

Fig. 10. 10% tolerance region for (μ, c_s^2) at 90% utilization

4.2. Feasible Region in (μ, c_s^2)

To get the feasible region of (μ, c_s^2), we similarly solve Eq. (11). This allows us to develop some insight into the role played by the variability in the picking operation. Figs. 9 and 10 show the feasible region of (μ, c_s^2), for the two example configurations defined above. From Figs. 9 and 10, we see that the feasible region becomes much tighter as the utilization increases. As the picking rate increases, the tolerance region for c_s^2 becomes wider as indicated by the length of the vertical line within the feasible region at a particular μ. In both plots, the batching delay component remains fixed as μ and k are held constant. 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 μ reduces both the waiting time and picking time, the system can tolerate higher levels of variability and still stay within the SLA.

5. Concluding Remarks

We introduced a form of sensitivity of the performance of a production or service operation, as modeled by a queue, by finding the feasible region of selected model parameters that would result in an acceptable range of a given performance measure. Such an analysis provides complementary information to traditional sensitivity analysis, which usually takes the form of gradient estimation. We call the type of analysis performed tolerance analysis. As we have seen in three examples of progressively higher complexity, the shape and size of the feasible regions are not always intuitive and the analysis adds value to the decision making process in system design.

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 shape of the feasible region computed in a tolerance analysis will give valuable insights on the relative risks caused by uncertainties in different parameters. Tolerance analysis can also be used as a way to measure the volume flexibility of an operation. 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. When comparing alternative system designs, the size of the feasible region can be used to rank the designs in terms of performance risk or volume flexibility. A larger feasible region typically implies higher volume flexibility and lower performance risk.

While we believe that tolerance analysis will give important information for operational risk management, many challenges remain to be studied. Many analytical models are approximate and hence the feasible region derived by the proposed approach is also approximate. However, we believe that the shape and size of the feasible region derived from an approximate model will give valuable insights on the relative risks caused by uncertainties in different parameters, or relative risks in comparing different system designs. For models that are not analytically solvable, finding a feasible region will 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 (Shanthikumar and Yao 1989), we can use an efficient search technique such as a binary search in these cases. Known 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 be a fruitful area for future research.

6. 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,” Proc. of the 1980 International Symposium on Computer Performance Modelling, Measurement and Evaluation, 3-9.

[6] Gupta, D. (2013), “Queueing models for healthcare operations,” Chapter 2 in Handbook of Healthcare Operations Management: Methods and Applications, B.T. Denton (Ed.), Springer Science+Business Media, New York, 19-44.

[7] Kleijnen, J.P.C. (1997), “Sensitivity analysis & related analyses: A review of some statistical techniques,” J. Stat. Comp. 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.

[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), 374-388.

[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] Mahdavi Pajouh, F. and Kamath M. (2010), “Applications of queueing models in hospitals,”

Proceedings of the 2010 Midwest Association for Information Systems (MWAIS) Conference, Paper 23.

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

[13] Robbins, T., Medeiros, D.J., and Dum, P. (2006). “Evaluating arrival rate uncertainty in call centers,” Proceedings of 2006 Winter Simulation Conference, Perrone, L.F., et al. (Eds.)

[14] 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.

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

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

[17] 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.

[18] 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.

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

[20] Whitt, W. (1993), “The Queueing Network Analyzer,” Bell System Technical Journal, Vol. 62, No. 9, 2779-2815.

[21] 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.

[22] Wolfram Research (2010), Inc., Mathematica, Ver. 8.0, Champaign, IL.

[23] Wustenhoff, E. (2002), “Service level agreement in the data center,” Sun Microsystems BluePrints Online.