Optimizing Capacity Utilization in Queuing Systems: Parallels with the EOQ Model

V. Udayabhanu
San Francisco State University, San Francisco, CA

Sunder Kekre • Kannan Srinivasan
Carnegie-Mellon University, Pittsburgh, PA

Congestion in queuing systems has serious consequences, so that it is never optimal to operate at 100% utilization levels. We develop an expression for the optimal utilization level for an M/D/1 queue, and demonstrate its similarity to the EOQ model of the inventory literature. The model can be used to achieve an optimal mean arrival rate, or to appropriately adjust the available capacity so that the desired utilization level is attained.

I. INTRODUCTION

The congestion of facilities is a key factor to be considered in decision-making, since it has serious negative implications in both the manufacturing and the service sectors. The immediate consequence of congestion is that it leads to increased operational costs; further, the degradation in the quality of service results in customer dissatisfaction and eventual loss of market share. The phenomenon of congestion has been analyzed through the use of queuing models. A brief summary of recent research in the area is presented below.

Zhang et al. (2009) look at the effects of congestion in designing a preventive healthcare facility network. They propose a nonlinear model to minimize the sum of travel, waiting, and service times, and develop heuristic solution methods to determine the appropriate number of facilities and their locations. Chand et al. (2009) also examine congestion in a healthcare environment. They build a simulation model of an outpatient clinic, and evaluate modifications to clinic operations that result in higher physician utilization. Pang and Whitt (2009) focus their attention on large-scale service systems, where server utilization levels close to 100% can be achieved. In such a setting, they show that service interruptions can result in severe congestion; indeed, the larger the scale of the system, the more disastrous the consequences are. Osorio and Bierlaire (2009) employ a queuing network model to identify the causes and effects of congestion in a finite capacity queue, and analyze how it is propagated throughout the network. Vidyarthi et al. (2009) consider the impact of congestion in designing supply chains for make-to-order and assemble-to-order systems. They demonstrate that substantial reductions in customer response time can be achieved with only marginal increases in costs. Zhang (2009) presents a “congestion-based staffing” policy, wherein the number of staffers is adjusted based on the queue length during a planning period, so that the mean queue length is maintained between prescribed limits. Kim and Uzsoy (2008) develop computational procedures to develop optimal capacity expansion plans for manufacturing resources subject to congestion. Kumar and Krishnamurthy (2008) model a customer’s decision-making process when faced with a choice between service providers based on different service times as well as congestion.
levels. They show that the uncertainty regarding congestion levels has a greater impact on a customer’s decision making process than the variability in service times.

The above research conclusively demonstrates that congestion has serious consequences in both manufacturing and service systems. It is never appropriate to achieve one hundred percent capacity utilization, since inordinately long waiting times will result. Some amount of slack capacity is therefore desirable. To the best of our knowledge, determination of the “optimal” slack capacity in the context of congestion of facilities has not been adequately addressed in the past. In what follows, we attempt to deal with this issue by proposing the concept of an “Economic Busy Period” (EBP). In section II, we develop the EBP model and discuss its resemblance to the familiar EOQ model of the inventory literature. Similarities between queuing and inventory models have been noted in the past; for example, Prabhu (1965) presents a unified treatment of the two systems. We develop a numerical example in section III, with concluding remarks in section IV.

II. THE EBP MODEL

The main thrust of this model is to provide a means of assessing whether a facility is operating at an economically appropriate congestion level. This is assessed in terms of the total contribution per unit time. We assume that customers arrive in a single stream in Poisson fashion at a mean rate $\lambda$. Service times are deterministic with mean $\mu$. The average time each customer spends waiting in line is $W$. The profit per customer is represented by $R$, and the cost of waiting per customer per unit time is $h$. The cost of waiting includes loss of customer goodwill and accounts for the long run negative impact on demand. The total contribution per unit time is represented by $\pi$. Then we may write

$$\pi = \lambda[R-(W+\mu)h]$$  \hspace{1cm} (1)

The maximal contribution with respect to the arrival rate $\lambda$ is obtained by setting

$$\frac{d\pi}{d\lambda} = 0.$$  \hspace{1cm} (2)

For an M/D/1 queuing system, the mean waiting time is given by

$$W = \frac{\lambda \mu^2}{2(1-\lambda \mu)}$$  \hspace{1cm} (3)

so that

$$\frac{dW}{d\lambda} = \frac{\mu^2}{2(1-\lambda \mu)^2}$$  \hspace{1cm} (4)

Substituting in (2), we find that

$$\frac{\mu}{(1-\lambda \mu)} = \sqrt{\frac{2[R-(W+\mu)h]}{\lambda h}}$$  \hspace{1cm} (5)

The above represents the EBP formula. It may easily be verified that $\frac{d^2\pi}{d\lambda^2} < 0$ at this point, thus yielding a maximum. Note that $\lambda \mu$ is the mean utilization, so that $(1-\lambda \mu)$ represents the average idle capacity. The left hand side of equation (5) is the average length of a busy period for an M/D/1 queue. Its optimal value (as computed from equation (5)) is called the “economic busy period”, in line with the EOQ terminology. Note that $[R-(W+\mu)h]$ is the contribution rate per customer; $(1/\lambda)$ is the average length of an idle period; and $h$ is the cost of waiting per customer. Since idle time is caused by non-arrival of customers, the contribution per customer may be viewed as the setup cost for each idle period. This can be estimated as follows:

Contribution Loss per idle period

$$= [\text{Arrival Rate}] [\text{Expected Length of Idle Period}] [\text{Expected Loss/Customer}]$$

$$= \frac{\lambda}{\lambda} [R-(W+\mu)h]$$

$$= [R-(W+\mu)h]$$

The average length of the idle period may be interpreted as the mean demand (in time units).
for each idle period, while the customer waiting cost may be treated as the holding cost. Thus the EBP model has an interpretation that is very similar to the EOQ model. The foregoing may be formalized as follows:

Proposition 1: The total contribution is maximized by an appropriate choice of the customer arrival rate; the corresponding optimal busy period (called the EBP) is given by equation (5).

It is true, of course, that there are many situations (for example, the arrival of customers at a bank) in which the arrival rate is an exogenous variable, outside the control of the decision-maker. But equally, there are other situations (for example, a plant operating in a make-to-order environment) where the decision-maker can choose from a portfolio of customers to achieve the optimal arrival rate, and consequently, the EBP. The choice of customers may be made using an appropriate priority rule. For example, customers may be selected based on their relative “importance”. Alternatively, the model may be used to change the available capacity, so that the optimal utilization level is achieved. The model is especially appropriate under such conditions. Note that equation (5) has been presented in terms of the EBP in order to draw a parallel with the EOQ model. The optimal arrival rate \( \lambda^* \) may be obtained by substituting for \( W \) in equation (5) and rearranging terms, yielding the following:

\[
\lambda^2 (2\mu^2 R - \mu^3 h) - \lambda (4\mu R - 2\mu^2 h) + (2R - 2\mu h) = 0
\]

Solving, we get

\[
\lambda^* = \frac{A \pm \sqrt{A^2 \mu h}}{A\mu},
\]

where \( A = 2R - \mu h \)

Since the utilization \( \lambda^* \mu \) must be less than 1, it follows that

\[
\lambda^* = \frac{A - \sqrt{A^2 \mu h}}{A\mu}
\]

Clearly, a real solution exists if \( 2R > \mu h \). This may be stated as follows:

Proposition 2: An optimal customer arrival rate exists if the profit per customer is at least half the cost of keeping a customer waiting during the mean service time.

The situation is represented pictorially below. The two parts of the curve pertain to the positive and negative signs in the formula for \( \lambda^* \). Note that as \( R \) grows without bound, \( \lambda^* \) converges to \( 1/\mu \).

![FIGURE 1: OPTIMAL ARRIVAL RATE](image)
The opportunity cost of idleness may be easily computed as shown below:

**Proposition 3:** Expected opportunity cost of an idle period = $R - (W + \frac{1}{\mu})h$.

This is true since the average length of an idle period is $\frac{1}{\lambda}$. Note that the cost function is concave with respect to $\frac{1}{\lambda}$, as shown below.

The concavity of the function may be explained by noting that the opportunity cost of an idle period is merely the lost contribution per customer. With an increase in the customer arrival rate, the waiting time increases, resulting in a decrease in the contribution.

To further pursue the analogy with the EOQ model, recall that the optimal order quantity results in an exact balance between the setup cost and the inventory holding cost; (i.e.) the two costs are equal at the optimum. The trade off involved in the EBP model is between the opportunity cost of an idle period and the customer waiting cost. While the two costs are not equal at the optimum, the following result is true:

**Proposition 4:** At the optimum, the cost of waiting per customer equals the opportunity cost of an idle period multiplied by the fraction of idle capacity, (i.e.)

$W_h = (R - (W + \mu)h)(1 - \frac{\lambda}{\mu})$  \hspace{1cm} (8)

The above result follows easily from (5).

**III. NUMERICAL EXAMPLE**

We provide a numerical illustration of the EBP model. Consider a facility operating in a make-to-order environment. The profit $R$ per customer is $50. The cost $h$ of keeping a customer waiting for one month is $100. The mean time $\mu$ to service a customer is 0.01 month.

For different values of the monthly customer arrival rate $\lambda$, we can compute the expected waiting times $W$, and consequently the total monthly contributions $\pi$, using equations (3) and (1) respectively. The results are shown in Table 1.

The contribution is seen to attain a maximum when the arrival rate $\lambda$ is 90 per month. This yields a utilization of $\lambda/\mu = 90\%$. The same result could be obtained by using equation (7):

$\lambda^* = \frac{A - \sqrt{A^2 + \mu h}}{A \mu}$,

with $A = 2R - \mu h = 99$. 

California Journal of Operations Management, Volume 8, Number 1, February 2010
The mean length of a busy period is \( \frac{\mu}{1 - \lambda \mu} = 0.1 \), and the same value may be obtained from the right side of equation (5).

The example shows that the maximal total contribution is obtained not when the facility is fully loaded, but at a ninety per cent utilization. This is achieved by controlling the inflow of customer orders so that the monthly arrival rate has a mean of 90. In a further parallel with the EOQ model, it may be noted that the profit function is relatively flat around the optimum, as shown in Figure 3 below. This is testimony to the robustness of the model; it is sufficient to adjust arrival rates/capacity levels so that the utilization level is “fairly close” to the optimum.

**IV. CONCLUSIONS**

The EBP model provides a simple criterion for choosing from a portfolio of customers so that the desired arrival rate is achieved. The model yields the optimal capacity utilization level by achieving a tradeoff between the opportunity cost of idleness and the cost of waiting. At the very least – in cases where the arrival rate is not a controllable variable – the facility is functioning near the optimal utilization.
model provides a means for checking whether the level. If the operating point is far from the optimum, other measures (like increasing available capacity) could be considered. The model is intuitively appealing because of its close resemblance to the familiar EOQ model.

V. REFERENCES


