

# OPTIMAL DETERMINATION OF SERVICE CAPACITY FOR SYSTEMS WITH SENSITIVE ARRIVAL RATE

Parag Dhumal, Angelo State University, [Parag.Dhumal@angelo.edu](mailto:Parag.Dhumal@angelo.edu), (325) 942 2383,  
Management and Marketing Department, Angelo State University, San Angelo, TX 76909

Vikas Agrawal, Fayetteville State University, [VAgrawal@uncfsu.edu](mailto:VAgrawal@uncfsu.edu), (910) 672 1338,  
Department of Management, Fayetteville State University, Fayetteville, NC 28301

## ABSTRACT

We develop the model for optimal determination of service capacity. Assuming poisson arrivals and exponential service, model minimizes the total cost which is sum of inventory holding and service capacity cost. From sensitivity analysis results we found that underestimation of arrival rate is costlier than overestimation. Still in both cases cost penalty is convex. Thus our model is simple, easy to use, robust with respect to estimation errors, and is of practical significance.

Key words: Service capacity, modeling, optimization, system performance.

## INTRODUCTION

In today's competitive global environment, effective management of company's capacity had gained a center stage that will help organizations to reduce cost and gain competitive advantage. Any kind of wastage of resources is considered as lost opportunities and represents strategic disadvantage to organizations. Companies striving to reduce cost often overlook capacity component which is a large component of the costs of an organization. Managers must improve how the company utilizes capacity and then take extra steps to reduce the capacity or to pay less for the capacity to realize cost benefits [5].

In this paper we develop a model to optimally design the service capacity of the system. Having too much capacity in the system increases cost associated with leasing the equipment or investments made in buying or upgrading equipment. Having too little capacity would increase the waiting time and inventory in the system resulting in increased holding cost. This would also be true in the service scenario where labors/operators costs are associated with addition of service capacity/rate which required to be balanced against customer wait time or cost of losing customer. Considering the tradeoff between these two costs the objective is to minimize the total cost of the system by optimally designing the service rate. We also demonstrated the effect of possible under-estimation and over-estimation of the arrival rate on the holding cost, service cost and the total cost of the system and the parameter that affects the most.

To operationalize model, arrival rate is assumed to be following Poisson distribution. Service rate is assumed to be exponentially distributed. Cost associated with the service capacity is assumed to be linear resulting in M/M/1 queuing system. We use various derivations/properties of M/M/1 system to optimally determine service rate of the system. Our paper is organized as

follows. In the next section we review relevant literature and discuss the importance and need of the problem considered. In the following section we formulate model and derive solution for optimal service rate. Later, we perform cost sensitivity analysis to evaluate robustness of model. In the next section illustrate the derived results using numerical problem. In the last section we present concluding remarks and directions for future research.

## LITERATURE REVIEW

The problem of effective service capacity design has been an active area of study by researchers and practitioners alike for some time. Yu-Lee [6] has pointed out that in order to reduce cost managers need to improve the capacity utilization either by reducing the capacity or paying less for the existing capacity. According to Hertenstein, Polutnik and McNair [4] in a global environment where companies strive to retain their competitiveness companies must understand their existing capacity and use it wisely.

Amiri [1] studied a reliable service system design problem that involves locating service facilities, determining their number and capacities and assigning user nodes to those facilities. The objective was to minimize the cost that consist of costs of accessing facilities by users and waiting for service at these facilities as well as cost of setting up and operating the facilities. Boronico and Siegel [3] had developed mathematical model that will result in significant cost savings through efficient allocation of capacity for toll roadways utilizing manpower planning analysis. The objective function, which is to minimize the expected cost subject to a reliability constraint on service quality, included both operating cost and cost allocated due to customer waiting.

Effective capacity planning has also been area of research in the area of inventory management. Berman and Sapna [2] developed an algorithm that will optimally control the service rate for an inventory system as a function of the number of customers waiting for service. Zanono and Zavanella [7] proposed a mathematical model that will find the optimal production schedule of steel billets for a steel plant, based on the relevant parameters of the productive system (set-up and processing time, demand profile).

In the literature various models have been developed for general to unique situations. While unique situation models are specialized providing better results, they tend to be complex. In this paper we develop a simple easy to use model that can be applied to vast situations. Also we present sensitivity analysis to show that our model is robust to underestimation and overestimation of arrival rate. Our ultimate goal for future study is to test the performance of this model to various unique situations as it is or with simple modifications to see if it provides dual benefits, simple yet providing better results for various situations.

## MODEL FORMULATION

Consider a system with one server having the service capacity of  $\mu$  units per unit time. We assume the service rate is exponentially distributed and the system is subjected to Poisson arrivals  $\lambda$  units per unit time. This system is modeled as M/M/1 and will to reach at steady state under the assumption that service rate is greater than arrival rate ( $\mu > \lambda$ ). The number of units in

waiting in the queue depends on the service rate  $\mu$ . Increasing the service rate reduces the number of units waiting or average queue size and the associated inventory holding cost of system.

At steady state, the average number of units in the system is given by  $\frac{\lambda}{(\mu - \lambda)}$

Let  $h$  be the inventory holding cost expressed in \$/unit/ unit time.

Therefore the inventory holding cost of system per unit time =  $\frac{\lambda}{(\mu - \lambda)} * h$

Other type of cost associated with the system is the cost associated with service capacity. This cost involves cost of buying, upgrading, or leasing the capacity. We assume that this cost is linearly associated with capacity. Let  $a$  be the cost of additional unit of service capacity of the system per unit time.

Therefore the service capacity cost per unit time =  $\mu * a$

The total cost of system per unit time,  $TC = \frac{\lambda}{(\mu - \lambda)} * h + \mu * a$  ..... (1)

Increasing the service rate  $\mu$ , decreases the inventory and associated holding cost at the expense of service capacity cost. In this situation, there is a trade-off between the inventory holding and service capacity cost. To find service rate that will minimize the total cost of system, the second derivative of the total cost equation (1) with respect to service rate,  $\mu$  is obtained.

$$\frac{d^2TC}{d\mu^2} = 2 * \frac{\lambda}{(\mu - \lambda)^3} * h$$
 ..... (2)

In equation (2), the arrival rate  $\lambda$ , service capacity  $\mu$ , and inventory holding cost  $h$  are all positive variables. Under the assumption of  $\mu > \lambda$ , the second derivative of the total cost equation with respect of service rate is positive. This shows total cost of system is convex function and has only one global minimum. To obtain the optimal service rate, we equate the first derivative of the equation to zero.

$$\frac{d TC}{d\mu} = -\frac{\lambda}{(\mu - \lambda)^2} * h + a = 0$$

Solving the equation for  $\mu$  yields,

$$\mu - \lambda = \pm \sqrt{\frac{\lambda * h}{a}}$$

Disregarding the negative value as  $\mu$  is always greater than  $\lambda$ , we have

$$\mu = \lambda + \sqrt{\frac{\lambda * h}{a}}$$
 ..... (3)

The above equation yields the optimal value of service capacity that will minimize the total cost of the system. The total cost per unit time of the system can be obtained by substituting the value of  $\mu$  obtained from equation (3) into equation (1).

$$TC = \frac{\lambda}{\left(\lambda + \sqrt{\frac{\lambda * h}{a}} - \lambda\right)} * h + \left(\lambda + \sqrt{\frac{\lambda * h}{a}}\right) * a$$

Solving the above equation yields,

$$TC = \lambda a + 2\sqrt{\lambda h a} \dots\dots\dots (4)$$

**SENSITIVITY ANALYSIS**

In above analysis, to compute the optimal service rate,  $\lambda$ ,  $h$ , and  $a$ , should be known with accuracy. Inaccurate estimation results in non optimal solution incurring higher costs. Of the above variables, leasing cost  $a$  and inventory holding cost  $h$ , can be easily obtained with accuracy. But estimating arrival rate  $\lambda$  may be difficult. So it is interesting to analyze how sensitive total cost equation is with respect to the estimated value of arrival rate  $\lambda'$ .

Let's say the estimated value of arrival rate =  $\lambda' = p * \lambda$

Where,  $p$  is positive real number.  $p < 1$  represent the cases of underestimation and  $p > 1$  represent the cases of overestimation.

The value of service rate using the equation (3) based on estimated value of arrival rate is:

$$\mu' = \lambda' + \sqrt{\frac{\lambda' * h}{a}} \dots\dots\dots (5)$$

Substituting the value of service rate from equation (5) into equation (1) yields the following total cost of the system per unit time:

$$TC' = \frac{\lambda}{\left(\lambda' + \sqrt{\frac{\lambda' * h}{a}} - \lambda\right)} * h + \left(\lambda' + \sqrt{\frac{\lambda' * h}{a}}\right) * a \dots\dots\dots (6)$$

Taking the ratio of optimal total cost from equation (4) to actual total cost from equation (6) yields:

$$\frac{TC'}{TC} = \frac{\frac{\lambda}{\left(\lambda' + \sqrt{\frac{\lambda' * h}{a}} - \lambda\right)} * h + \left(\lambda' + \sqrt{\frac{\lambda' * h}{a}}\right) * a}{\lambda a + 2\sqrt{\lambda h a}} \dots\dots\dots (7)$$

Dividing the numerator and denominator of the above equation by  $a$ , yields the equation in which the holding and service capacity cost are appeared as ratio. Hence the results obtained do not depend on specific values of holding and service capacity cost, but depends on the ratio of two costs. Thus substituting, the ratio of holding cost to service capacity cost as  $k$  in equation (7) yields:

$$\frac{TC'}{TC} = \frac{\frac{\lambda k}{(\lambda' + \sqrt{\lambda' k} - \lambda)} + \lambda' + \sqrt{\lambda' k}}{\lambda + 2\sqrt{\lambda k}} \dots\dots\dots (8)$$

We can further reduce one more variable, actual arrival rate  $\lambda$  without losing the applicable domain of the analysis by expressing all the variables in the time units such that the value of  $\lambda$  will be unity.

Hence we have,  $\lambda' = p * 1 = p$

Also the equation (8) can be rewritten as:

$$\frac{TC'}{TC} = \frac{\frac{k}{(p - 1 + \sqrt{pk})} + p + \sqrt{pk}}{1 + 2\sqrt{k}} \dots\dots\dots (9)$$

To study the behavior of the cost ratio function obtain from equation (9), we plot the % increase in the cost verses % error in the estimation of arrival rate. Figure 1 shows plot for family of k values from 1 to 5. Figure 2 shows this function for family of k values from 0.01, 0.1, 1, and 10.

For any given value of k, cost ratio curve is of right shape. In case of underestimation the curve sharply reaches the infinite value. That is, even the small of error in underestimating arrival rate, incurs large penalty. Visual examination of the curve for underestimation indicates that it is likely to exhibit convex property. In case of overestimation of arrival rate, cost increases with increase in the error throughout the time. But the beginning of the curve shows convex increase and afterwards tends to be linear. This is likely because after certain point, increase in service rate does not necessarily decrease queue size and associated inventory holding cost. But the cost of service capacity increases linearly with the increase in error. Also the curve is not symmetric that is if we compare the cost incurred for the same percentage of error in overestimation and underestimation, underestimation cost is much higher.

The one more important observation is that the cost ratio function is insensitive to higher k values, i.e. higher the value of holding cost, less the increase in the total cost of system for given value of errors. In case of overestimation this is very likely as increase in error, will add more service capacity and will decrease the % increase in total cost for the cases where holding cost is higher. But in case of underestimation the results are contrary to our intuition. We would expect that with increase in the k or holding cost, cost penalty should be higher for given percentage of error. The reason for this contradiction is that, estimated value of server capacity is positively associated with the k value or holding cost. Therefore impact of underestimating arrivals will be less severe on the computed service capacity for higher values of k. For e.g., if unit arrival rate is underestimated to 0.81(19% error), the percentage of error in the computed service rate is 18.19%, 14.5%, and 10.82% for respective k values of 0.01, 1 and 100.

To verify the convexity claims made above, the second derivate of cost ratio function with respect to estimated arrival rate  $\lambda'$  or p is obtained as follows:

$$\frac{d^2 \frac{TC'}{TC}}{dp^2} = \frac{2k * (1 + \frac{k}{2\sqrt{pk}})^2}{(p-1 + \sqrt{pk})^3} + \frac{k^3}{4(p-1 + \sqrt{pk})^2 * (pk)^{(3/2)}} - \frac{k^2}{4(pk)^{(3/2)}} \dots\dots (10)$$

For all the  $p < 1$  , i.e. for all the cases of underestimation of arrival rate the above equation yields a positive number showing that ratio of costs is convex function. Please refer to appendix for detail proof of convexity. For higher errors in underestimation, cost penalty is much higher. This also implies that with the small percentage of underestimation, there will not be much deviation from optimal cost.

To have better understanding for the cases of overestimation, we plotted the second derivate of cost ratio curve for various values of k. Figure 3 shows this curve for  $k = 1$  . In case of overestimation, overall curve exhibit non-convex property. But interestingly, up to certain percentage of overestimation, the original curve in equation (9) is convex and after that it is non-convex showing linear increase in cost. The point of overestimation at which curve changes its convex property to non-convex, is positively associated with k value. The following table shows for given k value, what is the maximum % of overestimation for which curve remains convex. Beyond this percentage of overestimation, the curve is no more convex. Table clearly shows that, curve is convex for the possible percentage of error. This implies for higher errors in overestimation, cost penalty is much higher. Also the total cost of the system will not be much different from optimal solution if percentage of error in overestimation is smaller.

### CONCLUSION AND FUTURE RESEARCH

In this paper we developed the model for optimal determination of service capacity or service rate. Assuming arrivals as Poisson and service as Exponential, model minimize the total cost of system which is comprised of inventory holding cost and costs associated with capacity. From the results of sensitivity analysis with respect to arrival rate we found that cost of underestimation is much costlier than overestimation. Also the cost penalty associated estimation errors of arrival rate is insensitive to higher ratio of holding to service capacity cost. We also have shown that cost penalty associated with underestimation and most of the practical values of overestimation is follows convex function indicating lower penalty associated with smaller errors in estimation. Thus our model is robust and could be very well used even if arrival rate is not known precisely.

We have presented very simple formulation for optimal determination of service capacity in M/M/1 systems. Our model can be applied to manufacturing as well as service industries. For future study we can check if this model provides better results to models developed for unique situations in the literature. Another interesting extension would be to extend results from one work station to serial system. Both this extensions would provide significant practical implications.

\*NOTE: All figures, tables, references and appendices available upon request from Vikas Agrawal ([VAgrawal@UNCFSU.Edu](mailto:VAgrawal@UNCFSU.Edu), 910-672-1338)