**8.5 Buffer Capacity, Blocking, and Abandonment**

Thus far we have assumed that all arrivals get in and are eventually processed, as in the security checkpoint at the Vancouver International airport, so the throughput rate of the process is limited only by its inflow rate and the maximum processing rate. In this section, we consider situations in which some of the arrivals may not be able to enter the process at all, while some who do enter may choose to leave because of long delays before being served. We will then evaluate the process performance in terms of the throughput rate, waiting time, and queue length.

In many applications, there may be a *limit on the number of customers that can wait before being served*, which is called the**buffer(or** **waiting room) capacity**, to be denoted as *K*. When the buffer is full*, any new arrivals are turned away, which is called* **blocking**. For example, waiting space in a restaurant, barber shop, or the drive-in facility at a bank, storage bins for purchased parts, or telephone lines at a call center all have limited buffer capacity to accommodate customers waiting for service. In the L. L. Beancall center if there are two CSRs and sixtelephone lines, then at most four callers can be put on hold, so the buffer capacity is *K* = 4. Once the buffer is full, any new caller will get a busy signal and cannot enter the system. These blocked arrivals represent loss of business if they do not call back.

Moreover, even if they are able to join the queue, some of the *customers who have to wait long for service may get impatient and leave the process before being served,* which is called **abandonment**. Again, if they do not return, it means lost business.

To analyze these situations, we need to introduce some additional notation. *The average fraction of arrivals blocked from entering the process because the input buffer is full* is referred to as the **proportion blocked** (or **probability of blocking**)and is denoted by *Pb.* Thus, even though the potential customer arrival rate is *Ri*, only fraction (1 – *Pb*) gets in, so the net rate at which customers join the queue is *Ri*(1 – *Pb*)*.* Moreover, out of those customers who do get in*, a certain fraction Pa may abandon the queue, which is referred to as the* **proportion abandoning**, denoted as *Pa.* Thus, the net rate at which customers actually enter, get served, and exit the process is *Ri*(1 *– Pb*)(1 − *Pa*), and the resulting throughput rate can then be calculated as

*R* =min[*Ri*(1 *– Pb*)(1 − *Pa*), *Rp*] **(Equation 8.12)**

Thus, fractions of customers blocked and abandoned are important measures of process performance because they affect the throughput rate which in turn impacts financial measures of process performance*.* With blocking and abandonment, *Ti* now refers to the waiting time of only those customers who get into the system and are served. Note that with limited buffer capacity, regardless of the magnitudes of inflow and processing rates, the queue will never exceed *K*, thus assuring that the process will be stable.

**8.5.1 Effect of Buffer Capacity on Process Performance**

With finite buffer capacity, but without abandonment (*Pa* = 0), the Finite Queue worksheet of **Performance.xls** can be used to calculate various performance measures for given values of the number of servers *c*, buffer capacity *K,* arrival rate *Ri* and the processing rate of each server 1/*Tp*. Specifically, the spreadsheet calculates the probability of blocking *Pb,* the average number of customers in queue *Ii* and in the system *I,* the average waiting time of a customer in queue *Ti* and in the system *T,* the capacity utilization *u*, and so forth. We illustrate these computations for the call center application.

**Example 8.8**

Suppose that the L. L. Bean’scall center is currently staffed by one CSR who takes an average of 2.5 minutes to handle a call and suppose that calls come in at an average rate of 20 per hour. Furthermore, suppose there are five telephone lines, so that, at most, four customers can be put on hold. L. L. Bean would like to estimate the proportion of callers who will get a busy signal and are thus lost to the competition. They would also like to know the average time that a customer has to wait for a CSRto become available. Finally, they would like to know the effect of adding more telephone lines on various performance measures.

In this case, we have a service process with finite buffer capacity, and we are given the following information:

Number of servers *c* = 1

Buffer capacity *K* = 4

Arrival rate *Ri* = 20 per hour

Processing time *Tp* = 2.5 minutes or the processing rate of each server 1/ *Tp*  = 1/2.5 = 0.4 per minute or 24 per hour. With this data input into the Finite Queue worksheet of **Performance.xls** spreadsheet, we get the following measures of performance:

Probability of blocking *Pb* = 10.07%

Average number of calls on hold *Ii* = 1.23

Average waiting time of a caller on hold *Ti* = 0.06835 hours = 4.1 minutes

Average total time that a caller spends in the system *T* = *Ti* + *Tp*  = 4.1 + 2.5 =6.6 minutes

Average total number of customers in the system *I* = 1.98

Thus, on average, about 10% of all callers will get busy signal and go elsewhere, and about 90% get through. If no one abandons the queue (*Pa* = 0), the throughput rate will be

*R =* min[*Ri*(1 – *Pb*), *Rp* ] = min[20(1 – 0.1007), 24] = 17.99 calls/hour

and the average server utilization will be

*u*= **= 17.99/24 = 0.7495

Thus, the CSR is busy only about 75% of the time and idle for about 25% of the time. Because of variability, however, there will be an average of 1.23 callers on hold and 10% of all callers (or two per hour) will get a busy signal, resulting in lost sales.

To study the effect of adding more telephone lines, we simply change the value of *K* and see how it affects key performance measures. Table 8.1 summarizes the results (rounded up to two decimal places).

**Table 8.1 Effect of Buffer Capacityon Process Performance**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Number of telephone lines *n* | 5 | 6 | 7 | 8 | 9 | 10 |
| Number of servers *c* | 1 | 1 | 1 | 1 | 1 | 1 |
| Buffer capacity *K* = *n –* c | 4 | 5 | 6 | 7 | 8 | 9 |
| Blocking probability *Pb* (%) | 10.07 | 7.74 | 6.06 | 4.81 | 3.85 | 3.11 |
| Throughput *R* (units/hour) | 17.99 | 18.46 | 18.79 | 19.04 | 19.23 | 19.38 |
| Average number of calls in queue *Ii* | 1.23 | 1.52 | 1.79 | 2.04 | 2.27 | 2.48 |
| Average wait in queue *Ti* (minutes) | 4.10 | 4.95 | 5.73 | 6.44 | 7.09 | 7.68 |
| Capacity utilization *u* | 0.75 | 0.77 | 0.78 | 0.79 | 0.80 | 0.81 |

Note that, as the buffer capacity (number of telephone lines) is increased, the blocking probability declines, and more callers are able to get into the system. Interestingly, however, the average waiting time of the callers who do get in increases*.* Thus, increasing the buffer capacity has two opposing effects: increasing the process throughput but also increasing the average waiting time of customers served. The optimal buffer size should take into account the financial impact of both, as we study in the next subsection.

**8.5.2 The Buffer Capacity Decision**

Customers blocked from entering the call center cost the retailer potential revenue if they do not call back. If they do enter they may have to wait on hold, during which the call center may have to pay telephone charges. Long waits also mean customer dissatisfaction and some customers abandoning the queue, again resulting in lost potential revenue. Thus, each of the operational performance measures,, that is, blocking, abandonment, queues, and delays has a direct bearing on economic measures, which are, revenues and costs. Capacity investment decisions should balance them all. We illustrate with L. L. Bean’s problem of choosing the number of telephone lines to lease.

**Example 8.9**

Continuing Example 8.8, suppose that any caller who receives a busy signal hangs up and orders from a competitor. L. L. Beanestimates the average cost of lost sales to be $100 per customer.

Furthermore, suppose that after a customer call gets in, each minute spent waiting on hold costs the retailer $2 in terms of lost goodwill (which may affect future sales). If leasing each telephone line costs $5 per hour, how many lines should the call center lease?

Note that the call center incurs four types of costs:

1. Cost of the CSR’s wages, say, $20 per hour
2. Cost of leasing a telephone line, assumed to be $5 per line per hour
3. Cost of lost contribution margin for callers getting busy signals, assumed to be $100 per blocked call
4. Cost of waiting by callers on hold, assumed to be $2 per minute per customer

Now, with one CSRand five telephone lines, *c*= 1 and *K* = 4, the cost of the server is $20*c* = $20 per hour, and the cost of leasing telephone lines is $5(*K* + *c*) = (5)(4 + 1) = $25 per hour.

We determined in Example 8.8 that the average number of customers blocked because of busy signals is

*Ri* *Pb* = (20)(0.1007) = 2.014/hour

The contribution margin lost because of blocking is therefore

$100 *Ri* *Pb* = (100)(2.014) = $201.40/hour

**Performance.xls** gave the average number of customers on hold as *Ii* = 1.23. If each waiting customer costs $2 per minute, or $120 per hour, the hourly waiting cost will be

$120 *Ii* = (120)(1.23) = $147.6/hour

The total operating cost, therefore, is

$(20 + 25 + 201.4 + 147.6) = $394/hour

Increasing the number of telephone lines increases the buffer capacity *K,* and, as above, we can compute the total cost per hour, as summarized in Table 8.2.

**Table 8.2 Effect of Buffer Capacity on Total Operating Cost**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Number of telephone lines *n* | 5 | 6 | 7 | 8 | 9 | 10 |
| Number of CSRs *c* | 1 | 1 | 1 | 1 | 1 | 1 |
| Buffer capacity *K* = *n* – *c* | 4 | 5 | 6 | 7 | 8 | 9 |
| Cost of CSR’s wages ($/hour) = 20 *c* | 20 | 20 | 20 | 20 | 20 | 20 |
| Cost of telephone lines ($/hour) = 5 *n* | 25 | 30 | 35 | 40 | 45 | 50 |
| Probability of blocking *Pb* (%) | 10.07 | 7.71 | 6.03 | 4.78 | 3.83 | 3.09 |
| Margin lost due to blocking ($/hour) = 100 *Ri Pb* | 200.8 | 154.2 | 120.6 | 95.6 | 76.6 | 61.8 |
| Average number of calls waiting on hold *Ii* | 1.23 | 1.52 | 1.79 | 2.04 | 2.27 | 2.47 |
| Average cost of waiting ($/hour) = 120 *Ii* | 147.6 | 182.4 | 214.8 | 244.8 | 272.4 | 296.4 |
| Total cost of service, blocking, and waiting ($/hour) | 393.4 | 386.6 | 390.4 | 400.4 | 414 | 428.2 |

Thus the total cost is minimized when the number of telephone lines is *n* = 6 or the optimal buffer capacity is *K* = 5. Leasing one more line not only costs more but also increases the cost of waiting time experienced by callers who do get in. In this instance, the waiting time of a caller is so expensive that the firm is better off not serving some customers at all than first admitting them and then subjecting them to long waits. Conversely, leasing one fewer line is also nonoptimal because it leads to more blocking and a greater loss of contribution margin on customers that are turned away than the saving in the cost of leasing the telephone line or customer waiting. The optimal buffer size thus correctly balances these costs.

It is interesting to note that, although limiting the buffer capacity denies access to new arrivals when the buffer is full, these customers would have had to wait long if they were allowed to get in, so they may be better off not getting in at all! An approach to imposing buffer limitation would be to inform callers that their wait may be long and hence they should call back later or, better yet, the service provider will call them back later. That would improve service to customers who do get in, without affecting those who are blocked.

**8.5.3 Joint Processing Capacity and Buffer Capacity Decisions**

In section 8.4 we determined optimal processing capacity *c* assuming unlimited buffer. In section 8.5.2 we determined optimal buffer capacity *K* for a given processing capacity *c*. In this section we determine both the processing capacity *c* and the buffer capacity *K* to minimize the total cost, which consists of the cost of servers and the buffer capacity as well as the loss due to customer blocking and waiting. We illustrate by determining the optimal number of CSRs and telephone lines to install in the call center example.

**Example 8.10**

As before, suppose the call center has an average of 20 incoming calls per hour. Each caller who gets a busy signal is blocked for an opportunity loss of $100, and each minute spent by a customer on hold costs $2 in terms of lost goodwill. Recall that each CSR takes 2.5 minutes to process one call and is paid $20 per hour. Suppose leasing each telephone line costs $5 an hour. The problem is to determine the optimal number of CSRs andtelephone lines.

The total hourly cost, then, consists of the following:

* Cost of CSR’s wages: $20*c*
* Cost of line charges:$5(*K + c)*
* Cost of lost sales due to blocking: $100 *RiPb*
* Cost of waiting: $120*Ii*

The problem is to determine *c* and *K* that minimizes

Total hourly cost = $20*c* + $5(*K + c)* + $100 *RiPb* + $120*Ii*

With *Ri* = 20/hour, 1/*Tp* = 24/hour and different values of *c* and *K*, the spreadsheet **Performance.xls** provides values of *Pb* and *Ii*, as summarized in Tables 8.3 and 8.4. Substituting them in the total hourly cost formula above yields Table 8.5.

**Table 8.3 Effect of Buffer and Processing Capacity on the Blocking Probability**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Pb* | *K =* 1 | *K =* 2 | *K =* 3 | *K =* 4 | *K =* 5 | *K =* 6 |
| *c =* 1 | 27.47% | 18.63% | 13.44% | 10.07% | 7.74% | 6.06% |
| *c =* 2 | 6.22% | 2.53% | 1.04% | 0.43% | 0.18% | 0.07% |
| *c =* 3 | 1.16% | 0.32% | 0.09% | 0.02% | 0.01% | 0.00% |
| *c =* 4 | 0.18% | 0.04% | 0.01% | 0.00% | 0.00% | 0.00% |
| *c =* 5 | 0.02% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |

**Table 8.4 Effect of Buffer and Processing Capacity on the Waiting Line**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Ii* | *K =* 1 | *K =* 2 | *K =* 3 | *K =* 4 | *K =* 5 | *K =* 6 |
| *c =* 1 | 0.27 | 0.60 | 0.92 | 1.23 | 1.52 | 1.79 |
| *c =* 2 | 0.06 | 0.11 | 0.14 | 0.16 | 0.17 | 0.17 |
| *c =* 3 | 0.01 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 |
| *c =* 4 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *c =* 5 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

**Table 8.5 Effect of Buffer and Processing Capacity on the Total Hourly Cost**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Total Cost | *K* = 1 | *K* = 2 | *K* = 3 | *K* = 4 | *K* = 5 | *K* = 6 |
| *c* = 1 | $612.42 | $479.11 | $419.06 | $393.94 | $387.41 | $391.35 |
| *c* = 2 | $186.91 | $123.88 | $102.79 | $97.60 | $98.60 | $102.02 |
| *c* = 3 | $104.65 | $93.60 | $94.27 | $98.10 | $102.78 | $107.70 |
| *c* = 4 | $108.86 | $111.07 | $115.50 | $120.38 | $125.35 | $130.35 |
| *c* = 5 | $130.51 | $135.12 | $140.05 | $145.04 | $150.04 | $155.04 |

It follows the lowest total hourly cost of $93.60 is attained at *c* = 3 and *K* = 2, so L. L. Beanshould hire three CSRs and lease five telephone lines.

In this economic analysis, we have assumed specific values for the cost of lost sales due to blocking and the cost of customer’s waiting time in queue. In practice, these costs are usually difficult to estimate. For example, how can one place a dollar value on the physical and mental pain suffered by a patient waiting in a hospital? Even at business level, it is difficult to estimate future sales lost due to a customer dissatisfied with long waits who in turn may share his experience with friends. In such situations, instead of estimating and minimizing costs, the process manager could choose to set limits on the proportion of arrivals blocked and the average waiting time of a customer as policy variables and look for a combination of the buffer and processing capacities that would provide acceptable values of these performance measures.