# Improve Customer Satisfaction Through Dedicated Service Channels

**Sheldon D. Goldstein**
**Indiana Institute of Technology**

*This study presents results of an analysis that optimizes the use of multi-channel service providers to reduce the wait time that customers must endure before a service can be performed. Despite the fact that airlines and supermarkets, among other businesses, have made use of quick check-out lines to speed the delivery of service, many companies do not implement multiple service channels and therefore do not enjoy the benefits of dedicated, multi-channel service to improve customer satisfaction.*

*In fact, the results of this study show those customers with long-term project needs and those with short-term service needs both benefit substantially by dedicating some portion of the service providers to only short-term projects. This is counterintuitive, since it is natural to believe if we take someone away from the long projects to "fight fires", it will delay the long projects. This paper shows that dedicated services can shorten the wait time for both short and long-term projects.*

## BACKGROUND

As an academic and consultant, it is common for me to see customer satisfaction surveys from my clients with comments like, "service takes too long". Using a simple reassignment of service staff, a company can substantially decrease the wait time that all customers experience when they request service. Reassigning only a few people to work on short-term problems can decrease the wait time dramatically for both long and short-term projects. The method works best at companies that have a variety of service problems and project opportunities; service that is varied as to the time it takes to complete the task. For instance, the IT department that needs to develop a new report as well as deal with giving new users passwords.

It is often the case that lots of thought goes into the tactical delivery of a service, but little effort is expended on the process of service delivery. As an example, the IT department of a company has to provide many different types of support to their internal customers. They fix printers, unlock computers, recover files, develop new reports, and introduce product upgrades or roll out new programs, to name a few. These services differ in many ways, but a distinguishing feature of these services is the amount of time it takes to complete each of the tasks. If it takes a technician 15 minutes to unlock a computer or 2 hours to fix a printer problem, those customers do not expect to wait 2 days until they are served. The urgency of the job often matches the time it will take to complete the job.

Dealing with these conflicts usually means "fighting fires" because planning for unspecified problems that arise randomly is not easy. And, as with challenges of this type, if we can find a way to do it efficiently, we can increase the satisfaction level of our customers. It is natural to believe that taking resources away from the long-term projects will compromise our commitments to those activities. This analysis shows that by segregating service channels into dedicated short-term service and dedicated long-term projects, we can improve the performance of both.

## THE MODEL

The solution to our dilemma requires more than simply prioritizing incoming jobs. Prioritization leads to favoritism. No one likes to wait for a service. It is imperative that a solution which improves the waiting experience for short-term service jobs must not disadvantage the wait for long-term projects. At the same time, we want to minimize cost, which means adding no more labor content to the company.

Texts concentrate on improving the waiting line experience for customers by either reducing the service time (get more work out of the existing workforce) or increasing the number of service channels (throw money at the problem and add more service technicians). However, the solution must also account for the cost of the operations, and minimize it for the method to be accepted by management.

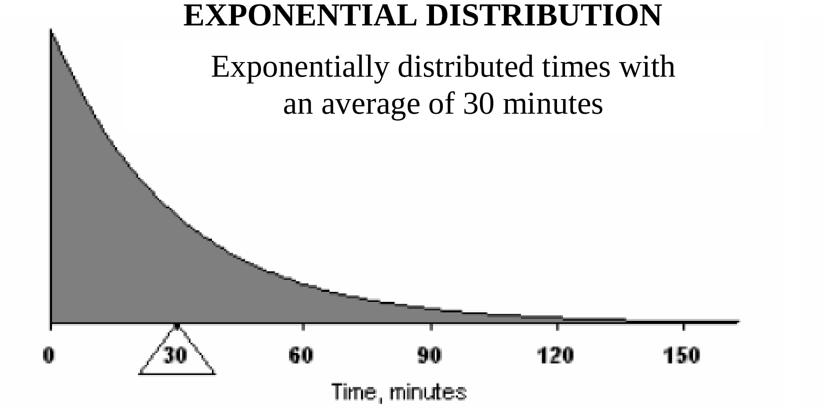
### Waiting Line Models

One way we could address this problem is to see it as a waiting line (queue). These are often modeled with Poisson arrival patterns. The Poisson arrival distribution is particularly appropriate for service jobs since they show up randomly and while we can predict how many we will receive each year, we can never predict exactly when the next one will arrive.

The Poisson process shows that the time between events is distributed in an exponential fashion. Service times are also modeled using an exponential distribution. This weighs shorter service times more heavily than longer service times. Other models exist, and if your situation would dictate a different arrival or service distribution to fit your needs, it would be best to modify your analysis to incorporate that information.

The exponential distribution looks like this:

**FIGURE 1**
**EXPONENTIAL DISTRIBUTION**



While arrivals and service (from a practical perspective) do not have their highest probability at zero time, an instruction to "reboot your computer" comes as close as possible. Whatever the assumptions are, our motivation is to improve service to our customers while at the same time minimizing cost to the organization. The model presented here proposes to segregate service to dedicated providers based on the length of time needed for each service.

Figure 2 shows the traditional queuing system with one waiting line and multiple servers. Consider that all customers are directed to wait in this line regardless of their specific need for service.

**FIGURE 2**
**TYPICAL MULTI-SERVER QUEUE**

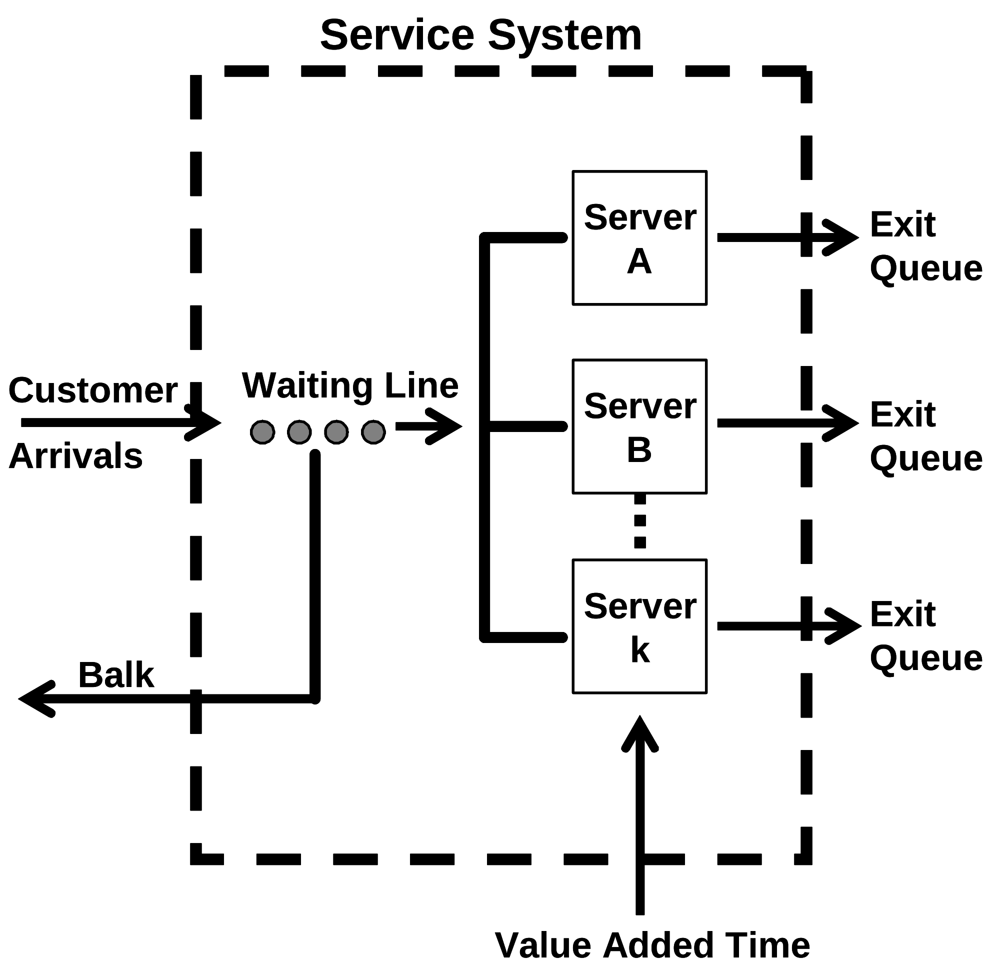
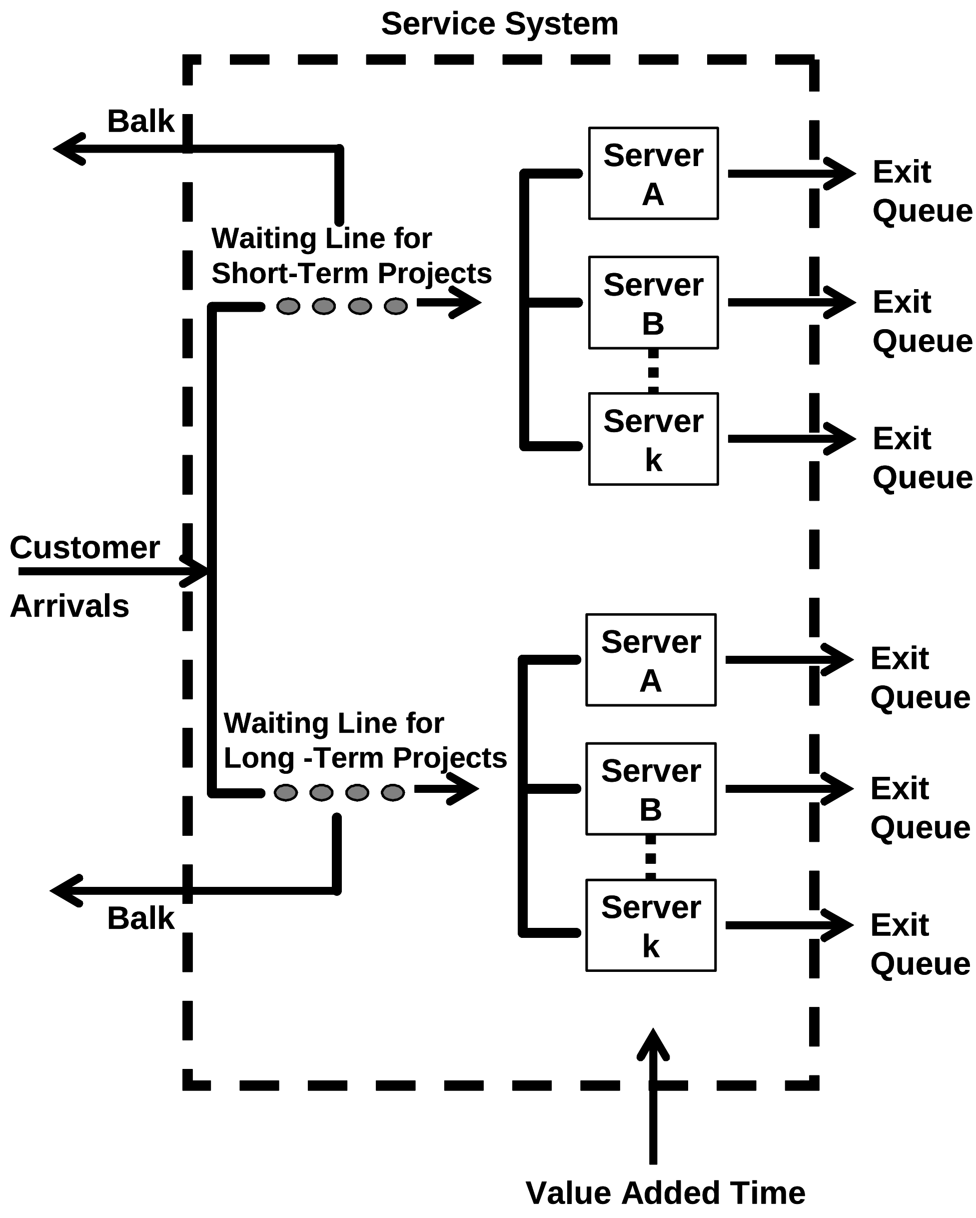



Figure 3 shows the proposed model of having two waiting lines, one for jobs of short duration and one for jobs of long duration. Customers would be directed to either the short-job line or the long-job line immediately upon entering the queue.

**FIGURE 3**
**PROPOSED MULTI-SERVER QUEUE**




## CASE STUDY

Let's say we have a service department that handles 200 jobs a month. Of these jobs, 150 are short jobs, each taking an average of 2 hours to complete. The remaining 50 jobs take 41.6 hours to complete on average. These numbers are an average from one of my clients. Then, in a month, we would need 2380 hours to complete these jobs, and that translates into about 14 employees, all of whom work at near 100% utilization. While this is an unrealistic assumption, any upper limit on employee utilization will work in this analysis. You simply must add the appropriate number of employees to cover the workload, given the discrepancy between the hours paid to employees and the hours of work received.

This is the only input data we need to completely analyze a waiting line model for this service department. However, these service times and arrival times are averages and not a constant rate for every job. There is a highly non-linear pattern to the arrivals and service distribution. Think of sitting in your favorite frozen yogurt shop and watching the arriving customer pattern. Customers arrive in pairs, sometimes a few pairs at a time, and sometimes there may be a group of 8 that comes in at once, or sometimes only one person will arrive. The average number of arrivals does not fully describe the dynamics of the waiting line in the shop. Similar arguments hold for the service pattern. Waiting lines build and decline in size as the day progresses, and for that reason we look at the probabilistic nature of the queue.

The first step is to analyze this problem as if all jobs went into a single queue as shown in Figure 2, regardless of the time it takes to complete a job. It is a single queue with 14 service channels. If we assume that the average time to complete a job comprised of 75% short and 25% long jobs is 11.9 hours, then the single queue, multi-channel service model gives us the following results (for an M/M/14 model in the Kendall notation):

- Probability that a customer has to wait = 95.0%
- Average wait time until service is provided = 69.4 hours

Imagine that you are served by this department and you have a locked computer. You would have to wait (on average) 69.4 hours until someone could address your need. That might not be an unexpected wait time for someone who has a job that takes 40 hours to complete, but a simple service ticket really should be attacked in a time that responds to the urgency of the situation.

## THE LINK TO CUSTOMER SATISFACTION

When conducting satisfaction interviews, two of the most common complaints about service are, “it takes too long to get anything done”, and “we need better communications”. These are often related. The likely refrain is, “It’s fine once they get to it” and “I would be more understanding of the delays if they would keep me informed as to their progress.” Customers want it when they want it, how they want it, and where they want it. Since this is not always realistic, and customers know that, the closer we can come to meeting customers’ needs, the higher their level of satisfaction. Everyone knows that a 2 hour job takes 2 hours to complete. That is the only “value-added” time. And, it is very annoying to wait for 2 hours to get it started. This is wasted time, or time that does not add value. Cutting out the wasted time means getting “lean,” and that is the benefit of optimizing waiting lines.

## THE IMPROVED MODEL

We never want our waiting line to grow large enough that it entices customers to exit the queue, an action called "balking." When a customer balks, they arbitrage their satisfaction by finding another supplier to serve their needs, and in this way, you have encouraged your customer to become acquainted with your competitor. If a customer can't balk because they are captive in terms of being required to use internal service providers, then they become disgruntled. Either way, making customers wait is never desirable, and we should do whatever is financially feasible to minimize the wait time for our customers. This results in improved customer satisfaction, which is a very desirable outcome for a service department.

The question then is, would it be better to provide two distinct service lines (potentially, each with multiple service channels) to address the long wait times resultant from our mixed-service, single waiting line model?

The intent is to meet the following objectives:

- Minimize the probability that a customer has to wait
- Minimize the wait time before service can be provided
- No increase in the number of service providers
- No pressure to improve the service time
- Choose a staffing mix that keeps employees busy

Under these assumptions, the model of Figure 3 applies. This is analyzed as a parametric study to assess the best mix of the number of providers for the different service channels. In addition, the second parameter is the percentage of short jobs that are given to the short job employees. For instance, instead of 14 providers in one waiting line (arrivals), the model is analyzed as a series of cases that consider providers in different combinations between service lines as follows:

- 1 server in the short-job line, 13 servers in the long-job line, (1, 13)
- 2 servers in the short-job line, 12 servers in the long-job line, (2, 12)
- 3 servers in the short-job line, 11 servers in the long-job line, (3, 11)
- 4 servers in the short-job line, 10 servers in the long-job line, (4, 10)

The reason for considering these cases is to evaluate the impact that they have on the utilization of each set of employees (we don't want one group sitting around with lots of time on their hands, while the other group is working every minute of every day). Also, we want to see how much work we should transfer to each of the queues to minimize the wait for all customers. Notice that all these options use only 14 total employees.

In our case we have 150 short jobs and 50 long jobs. When we move (say) 20% of the short-term jobs to its dedicated queue, we are shifting 30 of the 150 short-term jobs to that dedicated queue leaving 170 jobs (120 short and 50 long jobs) in the queue that handles mixed jobs.

Why does this help us solve the problem? It is because the long jobs block up our system. With 50 long jobs entering the system per month, and each long job taking 41.6 hours to complete on average, if four long jobs come in on any particular day, that will dedicate four employees to those jobs, and with around 12 long jobs entering the system every week, all the service employees might be working on those long jobs at any one time, leaving no service employees to work on the short jobs. Therefore, dedicating employees to the short jobs only, gives them the ability to get those completed, and still leave most of the employees free to work on the long jobs.

## Queuing Theory Analysis

The equations that describe a multi-channel system are:

**Operating Characteristics**

The following formulas can be used to compute the steady-state Operating characteristics for multiple-channel waiting lines, where

$\lambda$ = the mean arrival rate for the system
$\mu$ = the mean service rate for *each* channel
***k*** = the number of channels

1. **The probability that no units are in the system:**

$$P_0 = \frac{1}{\sum_{n=0}^{k-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^k}{k!}\left(\frac{k\mu}{k\mu - \lambda}\right)}$$

2. **The average number of units in the waiting line:**

$$L_q = \frac{(\lambda/\mu)^k \lambda\mu}{(k-1)!(k\mu - \lambda)^2} P_0$$

3. **The average number of units in in the system:**

$$L = L_q + \frac{\lambda}{\mu}$$

4. **The average time a unit spends in the waiting line:**

$$W_q = \frac{L_q}{\lambda}$$

5. **The average time a unit spends in the system:**

$$W = W_q + \frac{1}{\mu}$$

6. **The probability that an arriving unit has to wait for service:**

$$P_w = \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k \left(\frac{k\mu}{k\mu - \lambda}\right) P_0$$

## Results of the Analysis

There are five dimensions to our problem.

- How long is too long for a customer to wait for service?
- How many people should we dedicate to short-term jobs?
- What percentage of the short-term jobs should we shift to those dedicated workers?
- What is the probability of a wait and how long would that wait be? We might have a high probability of waiting, but if we only wait 15 minutes on average, then that might not be bad.
- What is the probability of an empty system? In other words, is there a high probability that workers will be idle in one queue and very busy in the other? The corollary to that is the question, “have we over-utilized one group of employees giving them more work than they can handle”?

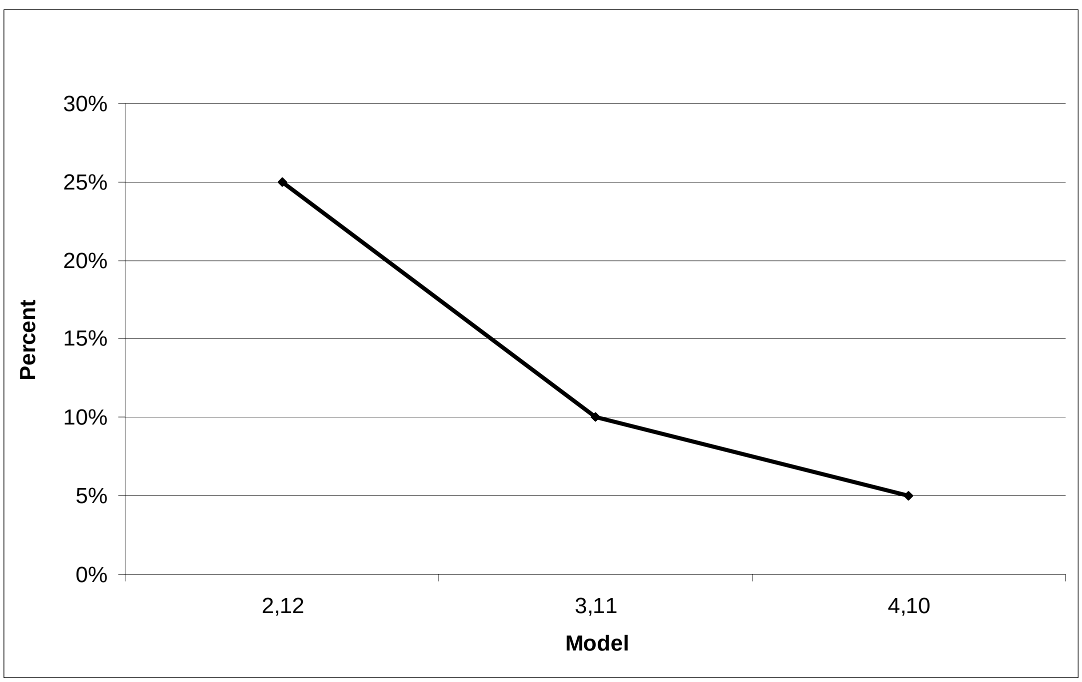
Figures 4 through 7 display the results of the analysis in graphical form. We can quickly dispense with the 1, 13 model since it is very easy to over-work or under-work employees given the metrics in this example. This condition is noticed when a scenario results in an arrival rate that exceeds the service rate for that case. In general, whenever the probability of a wait is more than 95% or less than 5%, we can expect workers to be over- or under-worked. That means we should be deciding on the trade-offs represented by the 2, 12 model, or the 3, 11 model, or the 4, 10 model.

In our efforts to minimize the wait time for customers, we might consider matching the probability of a wait for both short and long jobs. They will take different times to complete, but getting jobs into the service channel after experiencing the same wait time probability shows no favoritism to either service request. While matching the probability (likelihood) of waiting for the short and long projects is arbitrary, this gives us a good place to start our decision process.

The parametric study has several dimensions, as shown in the detailed tables 2 through 5 in the appendix. The four figures shown here result from our choice of equal probability of a wait for short- or long-jobs.

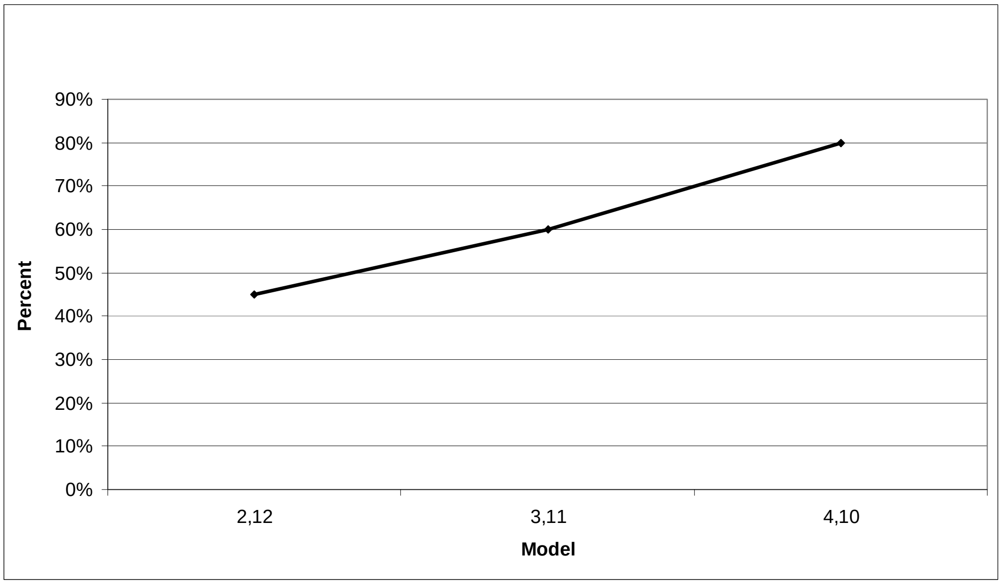
In Figure 4, we show the graph of the probability of a wait (assumed to be the same for long and short jobs) and in Figure 5, the percent of short jobs sent to the dedicated short-job providers that match with this probability of a wait.

**FIGURE 4**
**PROBABILITY OF A WAIT**




It is apparent that as we dedicate more employees to short-jobs, the likelihood of a wait for either short-or long-term jobs goes down. However, as we increase the number of employees dedicated to short-jobs, we must be sure that higher percentages of those short-jobs are diverted to those providers.

**FIGURE 5**
**% SHORT JOBS TRANSFERRED TO DEDICATED PROVIDERS**




Figures 6 and 7 show the predicted wait time corresponding to each of these models and the probability of zero jobs in the queue. It is apparent that as we move short jobs to their dedicated providers, both the probability of a wait and the actual wait time decreases for both long and short job customers.

**FIGURE 6**
**WAIT TIME**

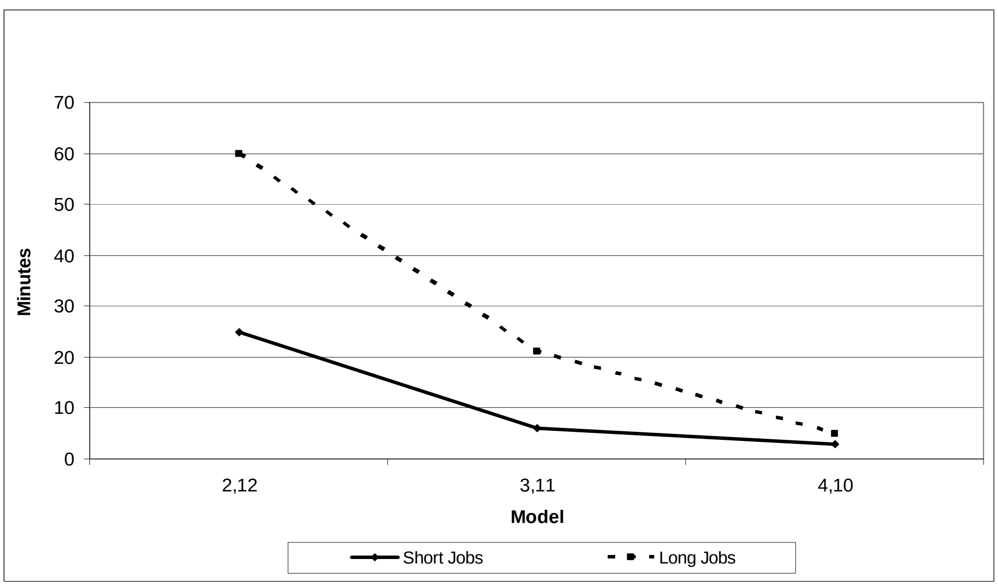
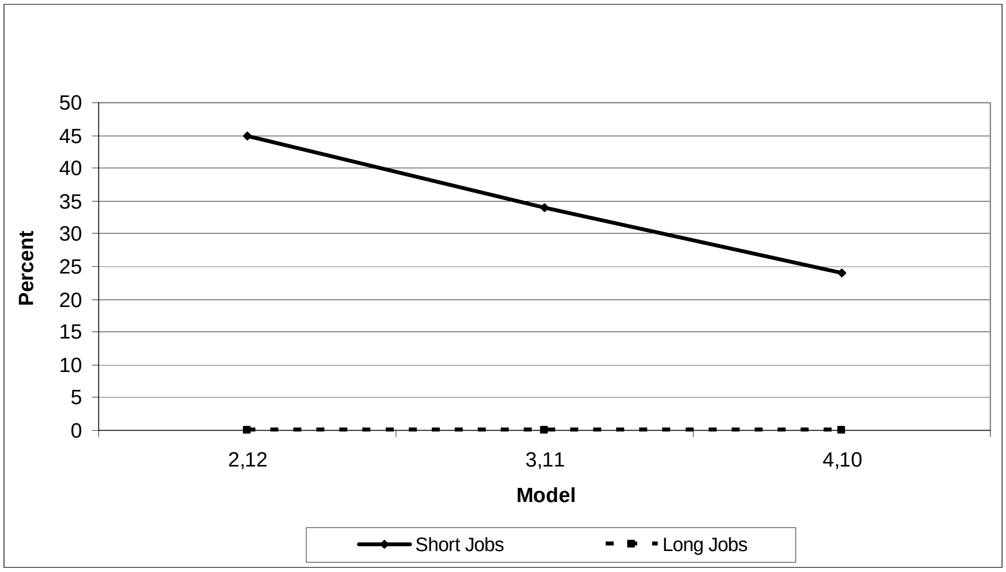


**FIGURE 7**
**PROBABILITY OF ZERO JOBS IN THE QUEUE**




From a practical standpoint, in the 3, 11 model for instance, shifting the optimum number of jobs to the short job employees (60%), there is a 6 minute wait for a short job to be started and a 21 minute wait for a long job to be started. In the worst case for the 3, 11 model, if all 100% of the short jobs were shifted to the short job providers, there would be a 33% chance that a short job would have to wait for 32 minutes, and a 1.2% chance that a long job would have to wait for 1 minute. That is a very large range over which the outcomes are all acceptable.

The probability of a wait, the average wait time at that condition, and the probability of idle workers may be summarized as follows:

**TABLE 1**
**SUMMARY OF RESULTS**

| | **2, 12 Model** | **3, 11 Model** | **4, 10 Model** |
|---|---|---|---|
| Probability of a wait for either short or long jobs | 25% | 10% | 5% |
| % of short jobs sent to dedicated providers | 45% | 60% | 80% |
| Wait time<br>Short jobs<br>Long jobs | <br>25 minutes<br>60 minutes | <br>6 minutes<br>21 minutes | <br>3 minutes<br>5 minutes |
| Probability of zero jobs in the queue:<br>Short jobs<br>Long jobs | <br>45%<br>0% | <br>34%<br>0% | <br>24%<br>0% |

At this point, management decision-making comes in. There are several trade-offs to consider before making the final choice of system design. First, is there some work we can give to the short-job employees when they are idle? Second, when we choose a model, does it give sufficient latitude to accommodate normal variation in arrival rate without under- or over-working any employees?

## CLOSURE

In all these models, the employees who are working on the long projects never get a break. They are mostly utilized at full capacity principally because the short-term jobs represent only 300 average hours per month out of the 2380 hours of incoming work per month. The natural question is, "why do we need 3 or 4 employees (21% to 29% of the workforce) dedicated to those short jobs when they represent only 12.6% of the work?" The reason is that we are trying to serve customers with as little wait time as is rational, and with a Poisson arrival rate that represents a non-uniform distribution, if it so happens that 4 jobs enter the system at one time, the wait time is not only determined by the ability of employees to service the account, it is affected also by the probability of a high arrival volume that is the nature of the exponential distribution. For this reason, we need a larger number of servers dedicated to the short jobs, because while they represent only 12% of the work hours, they represent 75% of the volume of jobs entering the system.

As we choose the best alternative, any of the cases studied resulted in substantially improved wait times for both short and long jobs. Customers will take note that you get to their work more quickly, in a matter of minutes, not hours or days. That leads to improved customer satisfaction.

Every case is different, and each company should perform this analysis using their own metrics that describe the service they provide. A simple waiting line model is easy to simulate and find the right combination of labor force and number of jobs to shift to dedicated service providers. It is the dedication of service providers to the short jobs that brings the benefits. It is only an application of the "express service" model we all know.

## REFERENCES

Waiting line models can be found in any text on quantitative methods; i.e.,

Anderson, Sweeney, Williams, (2004), Quantitative Methods for Business ,Thomson South-Western.

Modeling was done with free software downloaded from: www.bus.ualberta.ca/aingolfsson/QTP/.