



Why Can't We Manage Our Workforces By Occupancy?

Dr. Leonard J. Forys, Dr. Ashok Errmailli, Dr. Jonathan Wang

It's so easy

It's so easy to determine workforce requirements if we simply decide to maintain a desired occupancy for our agent workforce. Occupancy in this case is the proportion of time that agents are busy. All we have to do is forecast the work volume for our period of interest, divide by the time interval and the desired occupancy and we are done!

For example, suppose that our desired occupancy is .9 (agents busy 90% of the time) and that we forecast that for a given 15 minute period we will have 100 calls with average hold time of 180 seconds. This gives us 18,000 seconds of work volume (call-seconds). Divide this by 900 seconds (the number of seconds in 15 minutes) and .9 and we get $18,000 / (900) (.9) = 22.22$ agents. If we round this number up we get 23 required agents. This type of calculation is so simple that we can make up an Excel spreadsheet to do it and we are done. Also, it is easy to see the impact of changes. If we double the workload, we double the number of required agents. It's just so easy!

The table below is a sample Excel spreadsheet that does the required calculations for some sample values.

Number of calls	Forcing Interval (seconds)	Average Hold Time (seconds)	Target Occupancy	Required Agents	Rounded Required Agents
25	900	180	0.9	5.6	6
50	900	180	0.9	11.1	12
100	900	180	0.9	22.2	23
200	900	180	0.9	44.4	45
300	900	180	0.9	66.7	67
400	900	180	0.9	88.9	89
500	900	180	0.9	111.1	112
600	900	180	0.9	133.3	134
700	900	180	0.9	155.6	156
800	900	180	0.9	177.8	178
900	900	180	0.9	200.0	200
1000	900	180	0.9	222.2	223

Table 1: Spreadsheet Example



So what's wrong?

Although the approach is appealing, it has some problems. The main problem is that it does not provide for the same quality of service throughout the day, and across days. Suppose we take the number of "required" agents from Figure 1 and calculate what the resulting average speed of answer (ASA) would be using the algorithms of Irene. The results are graphed in Figure 1.

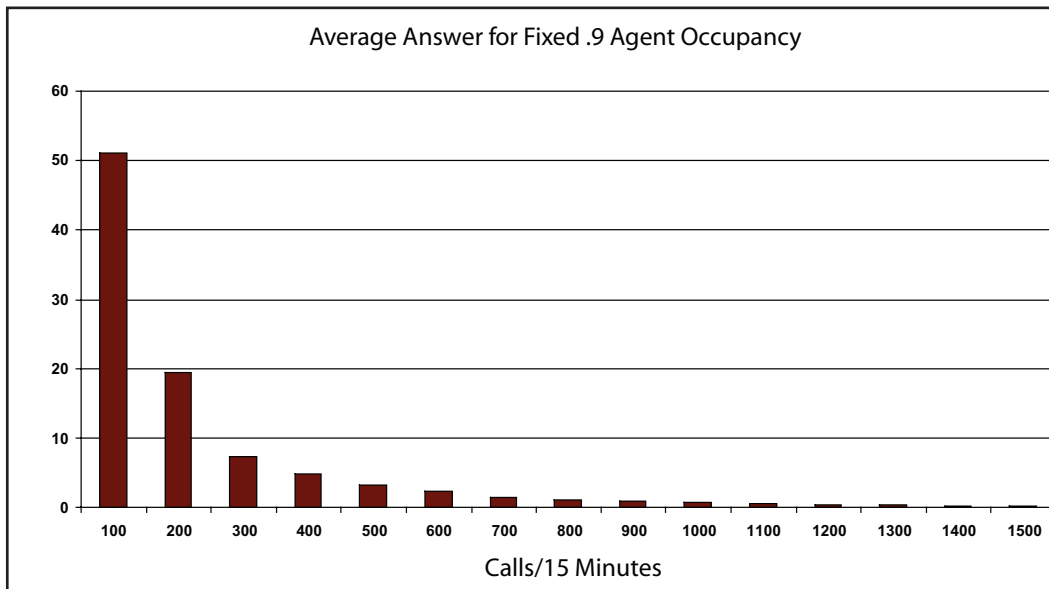


Figure 1: Service Consequences of Forcing by Occupancy

As we can see, for situations where the number of calls is relatively small, the average speed of answer is very high. In this example, the average answer was over 50 seconds when there were 100 calls in 15 minutes. On the other hand, when the number of calls is larger (e.g., 1500 calls in 15 minutes) the average answer is only about half of a second. As a result of this our customers would see widely differing service depending on when they make their calls. Since individual customers often call at the same times, those customers calling during low traffic periods will see the same poor service each time they call. Also, the service during high traffic periods may be too good, and thus cost too much.

What's the reason?

The basic reason for the results is that when the number of calls is large, we get the advantage of cancellation of peaks and valleys of call demand. The call arrival pattern within the 15-minute period is likely to be more regular than when the number of call arrivals is small. Also, when you have a large number of call arrivals, the differences in their hold times tends to average out. When you have only a few calls, it is possible that you can get bad luck and have all of your calls be long, resulting in large delays.

The opposite side of the coin

The opposite side of the coin is to set our service objectives so that we get constant average speeds of answers, independent of call volumes. That way the customers would see a uniform level of service, independent of when they called. Customers will not experience periods of poor service during low traffic periods and we will not be inefficient during high traffic periods. Sounds good. But is there a downside to this? Figure 2 illustrates the occupancy that will result if we fix the average speed of answer to be 20 seconds. The average hold time is 180 seconds in this case, the same as the previous example.



For small number of calls, the agent occupancies are modest, but reasonable. However, for high call volumes, the occupancies are close to 1.0. While this may be great for cost effectiveness, it has potential dangers. Small errors in forecasting could cause large changes in the speed of answer. Likewise, small deviations in work assignments could have similar consequences.

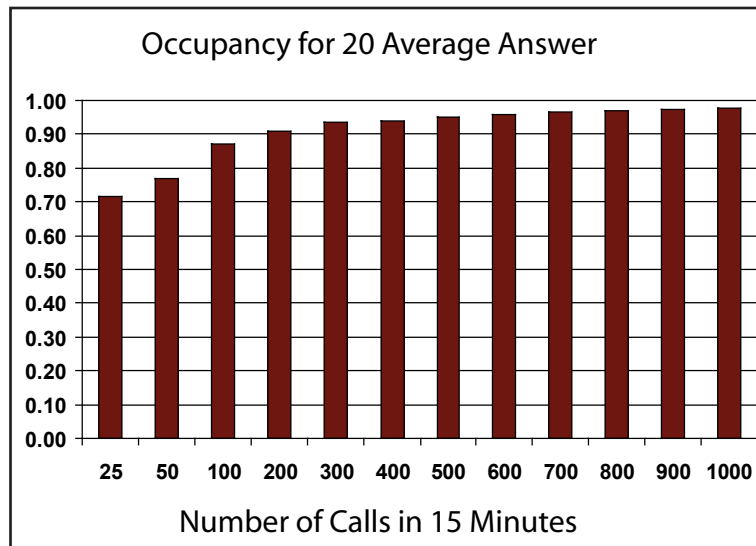


Figure 2: Occupancy for fixed average answer

A compromise – Occupancy Limits

There is a lot of folklore in the field that suggests that if agents are too busy, then the average hold time will increase. Agents will get their “breathing room” by working slower. The studies that we have seen that justify this however, are flawed. Our own experience suggests that for short periods of time if occupancies are high, agents will in fact often rise to the challenge and work more efficiently. However, if occupancies are chronically too high, then agents’ behavior is uncertain. Some of deleterious effects of high occupancies can be overcome by providing breathing space between calls by providing automated initial announcements made in the agent’s own voice. A reasonable compromise is to set a limit on the occupancy that agents should be subjected to.

There are a number of factors that influence selection of maximum occupancy. A typical value used in the industry has been around .95. However, this needs to be modulated by a number of factors. Some of these include:

- ✓ If pre-announcers are used, maximum occupancy can be set higher.
- ✓ If traffic volume is unpredictable, use a modest setting for maximum occupancy.
- ✓ If work force is inexperienced, maximum occupancies should be modest.
- ✓ If absenteeism is unpredictable, maximum occupancies should be modest.
- ✓ If managers are experienced and capable, maximum occupancies can be set higher.

Generally, for new sites and inexperienced agents, a modest value of maximum occupancy should be used. As experience is gained, this can be set higher. We have seen mature sites where the maximum occupancies of .97 and .98 have been used.



So how does Maximum Occupancy come into play?

In a workforce management application, such as Irene, two sets of calculations are made for each forcing period. One calculation is made to determine the required force based solely on service objectives such as average speed of answer or service levels. Another calculation is made based solely on occupancy. The two force requirements are compared and the larger of the two is used. What this means in practice is that for low traffic intervals, the force requirement will be that governed by the service objectives, while for high traffic intervals, the force requirements will be that governed by maximum occupancy. What this means is that for high traffic periods, the service will be better than that targeted by the user. Suppose we use the same traffic parameters as before, 180 average hold times, and target our average speed of answer to be 15 second and impose a maximum occupancy limit of .95. Then we find that for 1,000 calls/15 minutes, the average speed of answer will be five seconds.

Conclusion – Occupancy does play a role, but it’s a little bit more complicated

A workforce management tool such as Irene from ISC provides users with the capability to trade off risks versus agent efficiencies through the use of maximum occupancy. Users can effectively disable the maximum occupancy limit by using 1.0 (100%) as their inputs to the program and it will compute agent requirements solely based on service objectives. On the other hand, users can put in a fictitiously high value of target average answer objective (say 1000 seconds) and the software will compute agent requirements based solely on occupancy objectives as set by the maximum occupancy limit. Or, more advisedly, a compromise can be achieved by using a prudent value of occupancy until we better understand how our situation is behaving.

About ISC

ISC provides Irene, the most advanced workforce management system available to contact centers today. Irene forecasts customer service demand and delivers schedules that support performance targets, agent preferences, and business goals. Irene reduces payroll costs, improve service levels, and increase employee satisfaction. Whether you are managing thousands of agents globally or several dozen agents from one site, Irene meets your needs.

ISC was founded in 1973 to provide training development and consulting services to the call and contact center industry. From the beginning, ISC has been dedicated to providing measurable, sustainable improvements in the performance of people, processes, and technologies that shape the customer experience. In 2000, ISC introduced Irene. This award-winning software uses innovative technology that provides unparalleled scalability and dramatic advancements in forecasting and scheduling capabilities for contact center managers.

