

Better Service Levels At No Extra Cost?

How changes to call scheduling rules can help, managing the abandoned call rate, and why you may be using more agents than you need.

Michael Tanner, FIMA CMath

Copyright © Mitan Ltd. 1997-2004

Introduction

Service levels can be improved by simple changes to the way calls are scheduled, without increasing the number of agents. In order to predict performance we shall need to take account of abandoned calls, so we'll also see how to predict and manage the abandoned call rate. Then we can explain why some call-centre planning tools may recommend more agents than you may really need.

An Example

The ideas in this article apply to all call centres, big or small, simple or complex, but to illustrate the discussion we shall use the example in Fig. 1. Here a single group of agents needs to handle 360 calls per hour. Each call lasts on average 3 minutes, and the target is to answer 80% of calls within 15 seconds. The usual Erlang-C planning formula tells us that we need 22 agents.

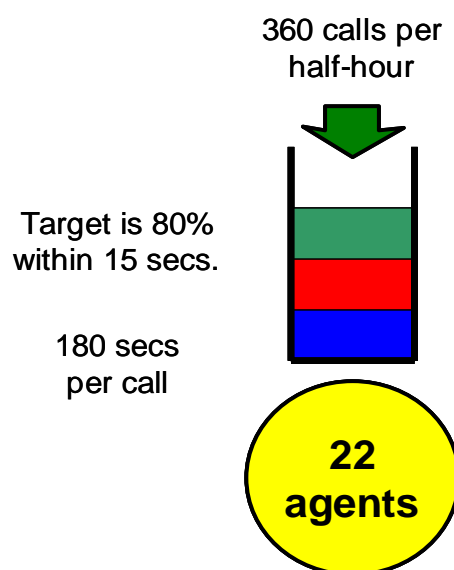


Figure 1. Example

The column labelled "Normal" in Fig. 2 shows the resulting performance. "Normal" here means that calls are answered in the order they arrive, and no calls are deliberately refused or disconnected. The charts show the percentage of incoming calls that are answered immediately, answered within 15 seconds, answered after 15 seconds, lost (deliberately removed), or abandoned. What Fig. 2 doesn't show is that for Normal scheduling nearly 6% of calls will be waiting 45 seconds or more. Why does this happen when we have the recommended number of agents and a "reasonable" service level target is being met?

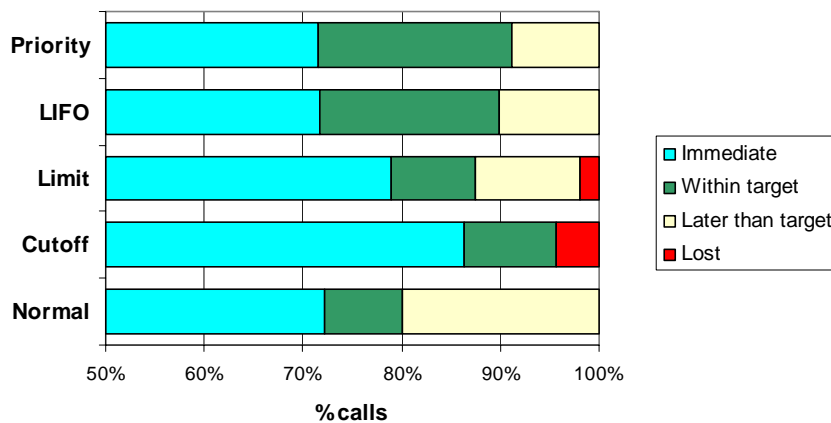


Figure 2. Performance for 360 calls/hour, no abandons.

What Causes Long Waiting Times?

The glib answer is too many calls and not enough agents. A more useful response is that some callers will get long waits because of the random way in which calls arrive. Sometimes calls bunch together, causing a short-term queue build-up. Once a queue has built up, it will affect the waiting times of calls for quite a while, just as a minor hold-up on the motorway causes tailbacks long after the original hold-up has disappeared. Can anything be done about these queue build-ups apart from deploying more agents?

More Agents Is Not The Only Way Of Getting A Better Service Level

We can improve the service-level either by preventing queues building up in the first place, or by protecting subsequent calls from the effect of the build-up. Queue build-up can be prevented by limiting the number of calls waiting, or by limiting the length of time a call may wait. This means deciding that a few calls will either be rejected with a busy tone or disconnected, in order to give the vast majority of calls a better service.

Protecting subsequent calls is done by handling calls in a different order, not the order in which they arrive. This favours some callers at the expense of others, which is in one sense unfair. But again we penalise a few calls so that a higher proportion of calls get good service.

Abandoned Calls

Some callers suffering long waits will be impatient enough to abandon their calls. Clearly this will happen mostly during temporary congestion. For now we'll assume that callers have unlimited patience and never abandon. But abandoned calls are an important factor in service levels and call scheduling. Later on we shall look at how abandoned calls, service level, and call scheduling interact.

Preventing Queue Build-Up

Queue build-up can be prevented either by disconnecting calls that have already waited too long, or by giving busy tone to further calls as soon as a small build-up occurs. Most ACDs let you specify such limits, but often they are set very high to correct unusual situations, rather than to manage the service level.

The "Cutoff" column in Fig. 2 shows what happens when calls are disconnected if they have waited 15 seconds without being answered. The service level is over 95%, but with over 4% of calls lost. "Limit" is when new calls are refused if there are 5 calls already waiting. The service level in this case is 87% with 2% of calls lost.

In both cases service level is improved, but at the expense of turning away some calls. The particular values of 15 seconds and 5 calls have been chosen for illustration. Other values could be used to give a different trade-off between lost calls and service level.

Protecting Calls From Queue Build-Up

Disconnecting or refusing calls when minor congestion occurs may seem a little drastic. What if we don't turn away any calls, but handle them in a different order? The "LIFO" column in Fig. 2 is for answering calls "last in first out". In other words the most recently arrived call is answered first. The tailback of calls then does not affect the waiting time of later calls. Service level improves to almost 90%, with no lost calls.

The LIFO method is clearly unfair, and could be difficult to justify. In any case, most ACDs do not provide for LIFO call handling! What the LIFO case does show is the potential for improving service level by changing the way calls are scheduled.

More acceptable, and just as effective, is the "Priority" method, where calls that have waited more than 15 seconds are given a lower priority. Most of the time this means calls are answered in the order in which they arrive, but during short-term congestion the method is a bit like LIFO. Fig. 2 shows that Priority gives a service level of 91%, a little better than LIFO, with the advantage of being more comprehensible.

Call Scheduling Can Improve Service Level

So it is clear that changing the way calls are scheduled can improve the service level. It is true that some calls get a worse service than they would with Normal scheduling, but the few calls that are penalised are far outweighed by the many calls that get a better service. Later, when we look at abandoned calls, we'll see that changing the call scheduling has practically no effect on the abandon rate. This shows that very few calls do in fact suffer a worse service.

When The Forecast Is Wrong.

So far we've looked at what happens when the call rate forecast is correct, and we have the right number of agents in place. But what happens if the call rate is higher than expected? Is the best call scheduling method under normal load still the best when things get more difficult? Figure 3 shows the performance at 440 calls per hour. If all calls were answered, this would represent 100% agent occupancy.

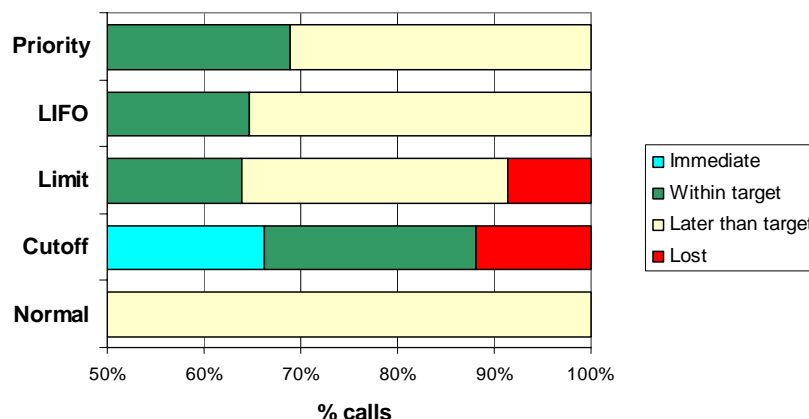


Figure 3. Performance with 440 calls/hour, no abandons.

Normal scheduling gives zero service level, with all calls having to wait a long time and the agents only barely able to handle the workload. The Cutoff scheme is the only one to meet the service level target, but with 12% of calls lost. Limit loses fewer calls than Cutoff, while giving nearly 65% service level. Priority out-performs Limit, delivering nearly 70% service level, with no lost calls. The choice seems to be between Cutoff and Priority, depending on the relative importance of service level and lost calls in your particular business.

Abandoned Calls And Call Scheduling.

How readily callers abandon affects which call-scheduling scheme is best. Later we take a good look at abandoned calls. For now look at Figure 4, which shows the performance when callers may abandon. The differences in service levels between call scheduling schemes is less, but still significant. You can choose, to some extent, between calls abandoning and calls being refused or disconnected.

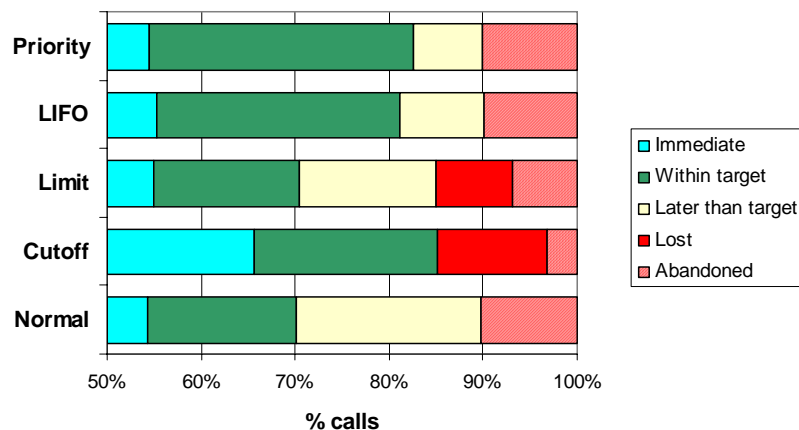


Figure 4. Performance with 440 calls/hour, calls may abandon

The Priority scheme looks attractive since it meets the service-level target with no rejected calls. Apart from Cutoff, all the schemes give the same level of unanswered (lost plus abandoned) calls. We might prefer Limit, reasoning that it's better for a caller to get busy tone than become impatient and abandon.

The Overloaded Call Centre

Sometimes it just isn't practical or economic to deploy enough agents. Something then has to give. Even the most patient callers will eventually abandon, and the size of the queue will ultimately be limited by the circuits installed. If you just let things sort themselves out, then you will deliver a terrible service level and your agents will get demoralised.

The alternative is to give good service to as many callers as you can, and to reduce the aggravation caused to callers who cannot be answered. A busy tone is probably less annoying than being left to abandon. A message after a short delay maybe better if the caller isn't paying for the call. If you take control in this way you will satisfy more callers and give your agents a better sense of achievement. In practice the Priority scheme might be combined with Limit or Cutoff in order to cope with overload.

Your Best Call-Scheduling Scheme

What is the best call-scheduling scheme for you? It depends on several factors. The most important is how you think your callers are affected by busy-tone, disconnection, long waits, and being provoked to abandon. Each scheme gives different proportions of these outcomes. The accuracy of your forecasting is relevant. The Cutoff and Limit schemes may seem to cause unnecessary lost calls, but could be the best choice if you often suffer from overloads. How quickly do your callers abandon? With very impatient callers the differences between the schemes is less than if callers are more patient.

Our service level is good, so why do we get so many abandoned calls?

Obviously service level and abandon rate are related, but the relationship is not as simple as it seems at first sight. Abandoned calls have more impact on service level than the simple fact that abandoned calls reduce the agent workload.

Calls often arrive in bunches, and it is then that the queue builds up and answer times lengthen. Impatient callers will abandon at these times of temporary congestion. Abandoned calls reduce the workload just at the right time, improving service level markedly. It may be the abandoned calls that enable you to meet your service level target. Without abandoned calls your service level would be much worse. The effect of abandoned calls can be seen in Fig 3. Remember we used standard Erlang-C to find out we needed 22 agents to get an 80% service level. Now we can see that the actual service level is 90%, with an abandon rate of 3.3%. Of course, we have made an assumption about how patient callers are, but a 3% abandon rate is not untypical, so our assumptions are reasonable.

If we are concerned only with the service level target, then we can meet that with 20 agents rather than 22. The abandon rate will go up to 6.4% with 20 agents, which seems rather high.

Managing The Abandon Rate

Suppose we wanted to get the abandon rate below 2%. This would take 24 agents instead of 22. To get below 1% abandoned calls we would need 25 agents. In general it takes more agents to get an acceptable abandon rate than it does to achieve what seems a reasonable answer time target. One reason for this is a lack of awareness that typical service level targets mean that about 1 in 20 calls will wait a minute or more.

Managing the abandon rate needs a planning tool based on a proper analysis of the complex interaction between service level and abandoned calls - a "super Erlang-C" formula. Mitan's research has produced the necessary theory to explain and predict abandon rates.

Why does our planning software recommend too many agents?

Most, probably all, call-centre planning tools use Erlang-C to calculate how many agents are needed. Erlang-C is a very useful formula, but assumes that callers never abandon, and calls are never refused or disconnected.

Abandoned calls, queue size limits, and waiting time limits act as safety valves when the pressure of work builds up. These safety valves improve the service level significantly. So a planning tool based on Erlang-C will often recommend more agents than are really needed to meet the target answer-time.

Remember that Erlang-C can tell you nothing about abandon rates. Despite this, one approach is to use Erlang-C then assume that, say, 5% of calls answered later than the target time will abandon. This is based on the fallacy that service level determines abandon rate.

Planning Methods

There is really no substitute for planning methods that take proper account of abandoned calls or queueing limits. Simulation can be used, and is indispensable for many purposes, but simulation is unwieldy and prone to misuse. Simulation was used for some of the results in this article, but many were obtained using formulae developed by Mitan during on-going research into call-centre planning. (These are built into the "PhoneCalc" package.)

Conclusion

Service levels can often be improved without more agents. The abandoned call rate can be managed more effectively. To do so you need a clear understanding of your business objectives, and planning tools that are based on effective research into how callers and queues behave.