Queue Dynamics, and Call-centre Answer Time Targets

Michael Tanner, Mitan Ltd., CMath, FIMA

1/November/2018

Abstract

Planners and managers of call centres and retail services need an understanding of the dynamics of queues, and how various factors affect waiting time. This article explains the effect of various factors on waiting time, and a number of other characteristics of queues. The emphasis, and much of the terminology, is for call centres and help desks, but managers of queues in other environments, such as retail and banking, will also find useful guidance in this article. The T-Calc package can be used to display many of the charts in this document, and to edperiment with changing parameters.

Contents

List of Figures

List of Tables

1 Introduction

This article is intended to give call-centre planners and managers an understanding of the dynamic behaviour of queues, and how various factors affect waiting time. Managers of queues in other environments, such as retail and banking, will also find useful guidance in this article.

Figure 1 shows the basic framework for call-centre planning (or any type of queue management planning). This document is about how mathematical queueing theory contributes to this planning process, which it does in two distinct ways. First, in setting performance targets, allows us to explore the balance between answering calls quickly and staff costs. Doing this assumes that the way queueing theory leads us to specify a performance target, typically saying something like 80% of calls must be answered within 15 seconds, is a reasonable description of how callers react to having to wait. Secondly, having establised performance targets, and obtained a forecast of the workload expected, queueing theory is used to calculate how many agents are needed in each time interval.

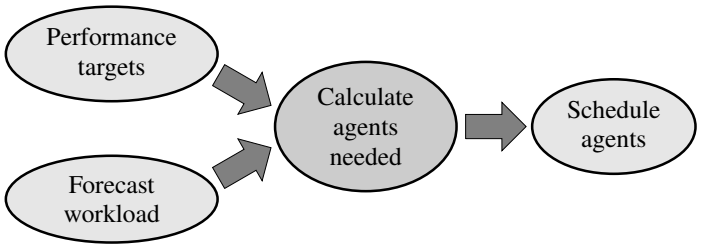


Figure 1: Simple overview of call-centre planning

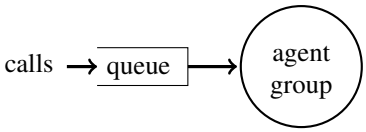


Figure 2: The simplest call centre

This document deals only with the simplest call centre consisting of a single queue of incoming calls being answered by a single group of agents, as depicted in figure 2. In this introduction we look at how queueing situations, such as call-centres verus supermarkets, differ, and the volatility of queue behaviour, which is important to understand when interpreting queue measurements. The ”Lessons” section (2) explores how each factor (for example call-rate, or variability of call duration) affect performance. Then ”Call scheduling methods” () looks at how queue limits or the order in which calls are answered affects performance. Important points, as well as being explined in detail in the main text, are highlighted as ★ *Important point*

Queueing theory is a branch of mathematics that provides insight into queue behaviour. Much like weather forecasting can tell us how likely it is to rain, but not guarantee exactly what will happen, queuing theory cannot predict exactly how long a customer or caller will have to wait but can tell how likely their wait will be, say, more than 30 seconds.

The article has to be a little mathematical and statistical, but results are presented as graphs, showing, for example, how average wait varies with the call rate, or number of agents. With each graph there is an explanation of what is being shown. Queueing theory has limitations, and where some effects cannot be analysed with queueing theory, simulation is used.

1.1 Parameters and metrics

Table 1 lists the basic parameters that define a queue. A number of additional parameters will be introduced later. The most widely used mathematical model of this situation is called Erlang-C. A mathematical model is a set of formulae or equations which describe the behaviour of the system. Mathematical models use simplifying assumptions, but still capture the essential behaviour of real-world system. Table 1 also shows the mathematical symbols used for the parameters in queueing-theory formule.

Parameter	Symbol	Comments
Agents	m	The number of servers sharing the commom queue. A server could be a person answering phone calls or a till operator, or a self-servcie till.
Customer arrival rate	λ	The average number of calls or customers arriving in a time interval of, say. 30 minutes. Arrivals are assumed to be random, so they will sometimes bunch together and sometimes have significant gaps between arrivals.
Average service time	T_S	The average time taken to serve a customer or handle a call. This includes "wrap" which is server activity related to a particular customer but taking place after the customer has left the till or the call has ended.

Table 1: Basic queue parameters

Metrics that can be used to compare queueing situations are listed in Table 2. The rest of this section uses average waiting time and service level as the basis for comparison of queues. The service-level target time is assumed to be 15 seconds, typical of call-centres. The Erlang-C model assumes that customers/calls that join the queue only leave once they have been served. Other models of queues deal with customers abandoning the queue, or being ejected from the queue after some waiting-time threshold, or not being allowed to join the queue if the queue has reached some threshold size. In these cases an important additional measure is the percentage of calls that are lost without being answered.

Metric	Formula	Comments
Traffic (workload)	$u = \lambda T_S$	"Traffic" is a term from the origins of queueing theory in telephoine systems. In this article "workload" is also used. Workload can be interpreted in several ways: for systems with no lost customers it is the average number of servers busy. Note that λ and T_S must be expressed using the same time units.
Agent utilisation	$\rho = u/m$	In calculations ρ ranges from 0 for completely idle to 1 for fully occupied. Often utilisation is scaled to a percentage from 0% for completely idle to 100% for fully occupied. Important because it is uneconomic to provide servers/agents unnecessarily.
Queue size		The most visible aspect of queues in retail or banking, and closely related to waiting time.
Waiting time		Obviously very important to customers. Average waiting time has weaknesses as a way of representing the experience of customers.
Service level		The percentage of calls/customers that are answered/served within a specified target time. The preferred basis for planning staffing levels in call centres, but still an imperfect representation of customer experience.

Table 2: Queue metrics

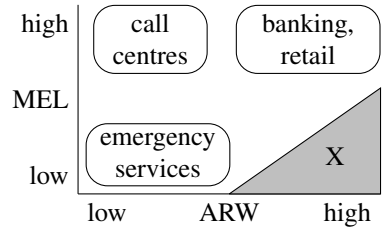
1.2 Queueing situations

A useful way of categorising queueing situations is the "acceptable relative wait" (ARW). This is just the ratio of what is considered an acceptable waiting time to the average service time, expressed as a percentage. ARW is defined in (1) where T_s is the mean service time and T_A is some estimate or opinion of what is an acceptable average waiting time: only a very approximate value is needed.

$$\text{ARW} = \text{acceptable relative wait} = \frac{\text{acceptable wait}}{\text{average service time}} = \frac{T_A}{T_s} \times 100\% \tag{1}$$

Another metric is "minimum economic loading", MEL, which is the minimum percentage utilisation of agents or servers. Example values for ARW and MEL are given in Table 3. These are based on approximate figures from various sources, and are not rigorous but simply intended to show how queueing situations differ. MEL values for emergency services are given as zero, since the services would have to be provided even if they were hardly ever used. Note also that for emergency services we are concerned with the arrival of the ambulance or fire appliance at the incident, not simply the answering of the call requesting help. In contrast, commercial retail/banking situations have a high MEL since they would not be profitable at low utilisation, while customers appear to be tolerant of a moderate waiting time.

Figure 3 shows the different queueing situations. A low ARW can be achieved if MEL is low, since it is economic to provide more capacity. Conversely if MEL is high then capacity will be limited, but a high ARW means that customers will tolerate the longer queues. There may be examples of high ARW with low MEL, marked X in the diagram, but if MEL is low then high capacity could be provided and customers would not tolerate ong waiting times. What makes call centres different, and difficult to manage, is the low ARW in combination with a high MEL, although a mitigating factor is that activity can be centralised in larger units, which are more efficient.



★ *The combination of low ARW and high MEL makes call centres an especially difficult queueing situation to manage.*

Figure 3: Queueing situations ARW and MEL

Service	ARW	MEL	Comments
Ambulance	7%	0%	Responses are prioritised according to patient condition, and there is wide variation, but for life-threatening conditions a response time of 8 minutes seems representative (see [2]). Service time is hard to determine, but appears to be of the order of 2 hours (see [3]).
Call centre	10%	85%	Wait 10-20 secs for an 180 sec. call.
Supermarket belted tills	100-200%	90%	Customers typically wait for one or two preceding customers to be served.
Supermarket express tills Self-checkout tills Retail department store Bank teller	$> 300\%$	90%	Customers typically wait for several preceding customers to be served.
Special event	infinite?	100%	Customers will queue for a long time for a unique sports or musical event.

Table 3: Acceptable relative wait, minimum economic load

1.3 Volatility of queues

★ *Queue size is very volatile.*

★ *A non-overloaded queue is empty most of the time*

★ *Measurements, particularly averages, may not be "representative" of customers experience.*

The most visible aspect of a queueing situation is the size of the queue. Figure (**??**) is an extract from a simulation showing the moment-by-moment queue size. Transient build-ups of queue size happen when arriving customers bunch together, or several adjacent customers have longer service times. The dashed line shows the mean queue size. Waiting time is closely related to queue size.

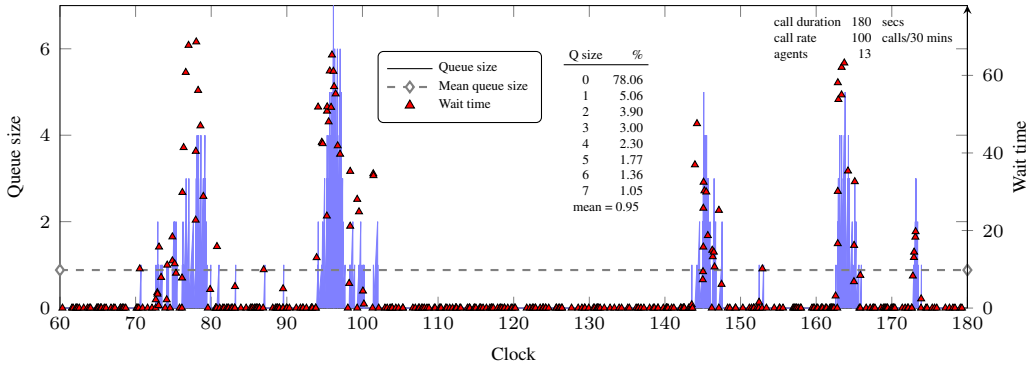
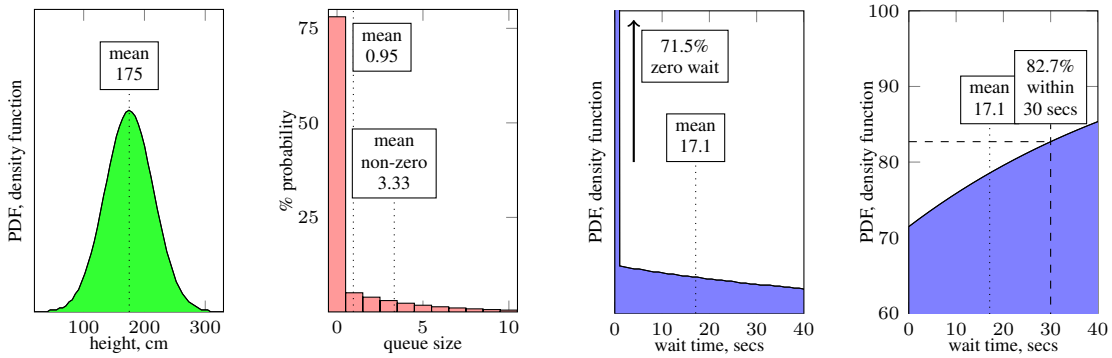


Figure 4: Queue volatility

When interpreting call-centre measurements it is important to understand the nature of the statistical distributions of waiting-time and queue-size. A natural assumption is that the average queue-size, say, is representative of what callers encounter, because the classic statistical distribution is the bell-curve shown in in Figure 5a, which is the distribution of the heights of adult males in the UK. The mean is 175cm, and this is "representative" in the sense that many people's height is near 175cm, and the mean is the most common value. On the other hand Figure 5b is the distribution of queue size for the situation used for the above simulation. The mean queue size is 0.95, but by far the most common value is zero. For customers arriving to find a non-zero queue the average queue size is 3.33. (A zero queue may still mean the customer has to wait, if all agents are busy). So the average queue size is not at all represetative of customers experience. While it may be tempting to conclude that the service is good because on average there is only one customer waiting, this may be a false conclusion. Similarly, waiting times are polarised, between a large majority of customers who are answered immediately, and a minority of customers who may have to wait a very unreasonable time to be answered.



(a) Heights) (b) Queue size (c) PDF (d) CDF

Figure 5: Wait time distribution?

1.4 Erlang

The word "Erlang" crops up a lot when discussing queues. Agner Karup Erlang was a Danish mathematician and engineer working on early automatic telephone systems, and in 1917 developed his famous formula known as Erlang-C (there is also Erlang-A and Erlang-B, which do not concern us here.) The term "erlang", without a capital letter, is a unit of "traffic" (see later). The maths for Erlang-C, and other queueing models, can be found in any text-book on queueing theory, such as [1], [4], [7]).

1.5 Terminology

A queue might be a queue of calls at a call-centre waiting to be answered, or a waiting line at a supermarket checkout, or at a bank till. In a call-centre the people answering the calls are referred to as "agents". In a face-to-face queueing environment, a supermarket or bank, the people serving will be referred to as "servers". The terms "calls" and "customers" are largely synonymous. "Service time" means the time it takes to deal with a customer or call. "Duration" means the same as service time, more commonly used fir call-centres. "Workload" means the arriving calls/customers, meaning both the arrival rate and the average service time.

1.6 T-Calc and T-Lib

T-Calc is a package for queueing related calculations, designed for educational as well as planning purposes. A number of the charts in this document can be displayed in T-Calc and the user can experiment by varying parameters such as call rate, duration, targets, and so on. The T-Calc package can be used without charge, and can be downloaded from www.mitan.co.uk. T-Lib is a library of queueing calculation functions, packaged as a Windows DLL, for us with Excel or interfaced to other programs. (T-Calc does not currently include a simulation facility, so analyses described in this document which need simulation were done with a different package not publicly available.) Contact the author for further information about T-Calc or T-Lib.

1.7 Simulation

Some characteristics of a queue cannot be analysed by mathematical formulae, or not without great difficulty. Simulation provides a powerful alternative or complementary method for evaluating queue performance, but has it's own drawbacks and limitations, see [10]. A number of situations described in this article that diverge from the Erlang-C assumptions were analysed by simulation.

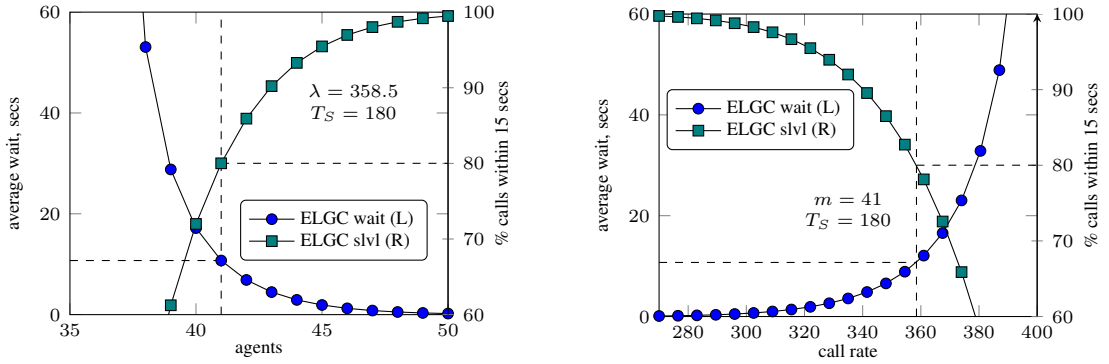
2 Queue dynamics

This section considers the behaviour of a single queue with multiple servers, and the effect of various factors. A single queue of customers or calls waiting to be served by a single group of tills or agents is very common, such as in call centres, banking and retail. As each factor is considered, essential points are highlighted as "★ *Important point*".

2.1 Non-linearity

★ *As the number of agents is reduced, or call rate increases, then waiting time increases and service level decreases. The deterioration in service is more rapid the closer the workload is to the capacity of the servers or agents.*

If the load is doubled, say from 40% of capacity to 80% of capacity, then average waiting time and queue size will more than double. Waiting time increases ever more sharply as the workload increases towards full load. This is the most important characteristic of queues where the arrival of calls, or customers, is random and there is no feedback mechanism to restrict or divert arrivals when the agents, or servers, are heavily loaded. Servers or machines can operate at full, or near-full, load only when long waiting times do not matter.



(a) Performance vs agents

(b) Performance vs call rate

Figure 6: Non-linearity

Figure (6a) shows how waiting time and service-level deteriorate as the number of agents is reduced. The dashed lines highlight values for 41 agents, showing a service level of 80% for a target answer time of 15 seconds. Figure (6b) shows waiting time and service level as call rate is increased for constant agents constant, along with the agent utilisation. Table 4 lists the factors used for this illustration.

Factor	(6a)	(6b)
Number of agents	variable	41
Calls per 30 mins	358.5	variable
Mean call duration	180 secs	
Service level target	80% in 15 secs	
Evaluation method	ElgC	

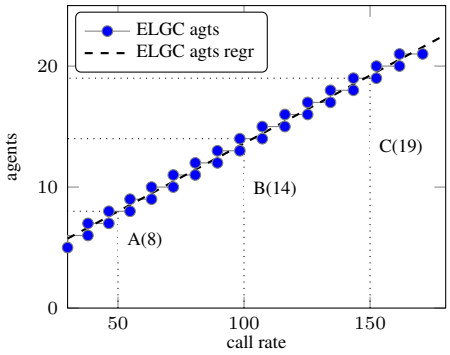
Table 4

.

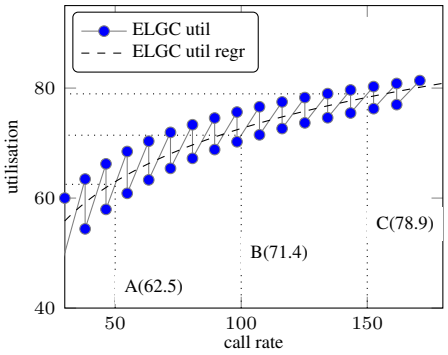
2.2 Agents Required

★ *For same service-level at high workload, relative to the number of agents, can be loaded more for lower workload.*

The number of agents required to provide a specified service level or average wait is not simply proportional to the workload. If the workload is doubled then the number of agents needed will be less than double. Large groups of agents are more efficient at providing service than smaller groups of agents.



(a) Agents



(b) Utilisation

Figure 7: Agents required vs call rate
(This chart is available in T-Calc

Figure (7a) shows the agents needed to meet meet the target service level as call rate is increased. The points A, B, and C, with associated results in Table 5, show that at 50 calls/30mins 8 agents are needed, but doubling the call rate, at B, or tripling the call rate, at C, does not requires less than double of triple the agents. Note how the number of agents is a stepped line, since agents cannot be fractional. Figure (7a) shows the corresponding agent utilisation.

	A	B	C
Calls per 30 mins	50	100	150
Agents required	8	14	19
Utilisation %	62.5	71.4	79.0

Table 5

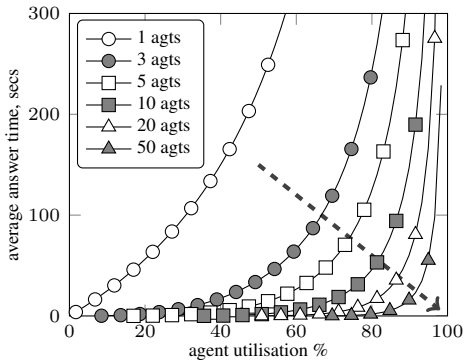
Factor	Value
Mean call duration	180 secs
Service level target	80% in 15 secs
Evaluation method	ElgC

Table 6

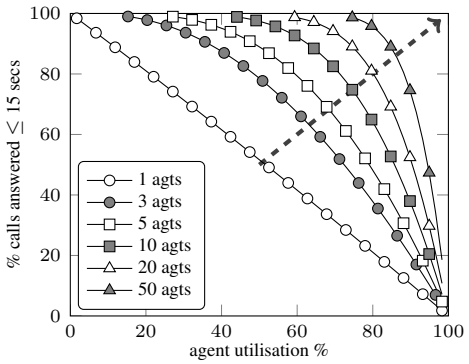
2.3 Effect of group size

★ *Larger groups can be loaded more heavily than smaller groups for the same service level.*
★ *A larger group shows a sharper drop in service level as workload approaches capacity.*

A larger group has a different service-level characteristic than a smaller group, given that both groups are loaded to the same utilisation. The larger group is less affected by bunching of arrivals or longer service times than the smaller group, so that for a given target service-level a larger group can be loaded more heavily than a smaller group. However big the group, waiting time ultimately rises dramatically as 100% load is approached, and for the larger group deterioration of service level starts at a higher load but is then more rapid.



(a) Average answer time



(b) Service level

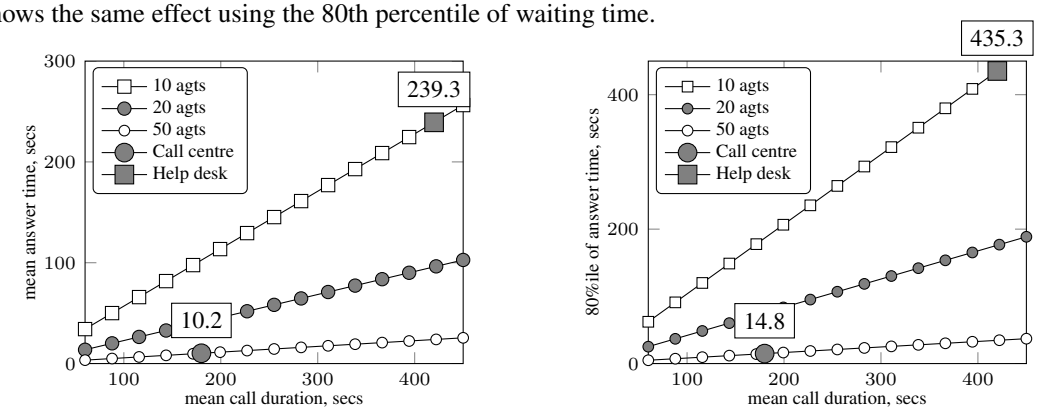
Figure 8: Effect of group size and variability of duration
(This chart is available in T-Calc)

Figure (8) shows results for different numbers of agents in the group, plotted against agent utilisation. As the number of agents increases, the shape of the waiting-time curve changes. For a bigger group, the waiting time remains lower for longer as load increases. But however big the agent group is, waiting time ultimately rises dramatically as 100% load is approached, so that the larger the group, the sharper the deterioration in service as load increases. The background arrow in Figure (**??**) emphasises this effect, with the waiting-time curve pushed further into the bottom-right corner

2.4 Call duration

★ *A workload made up of fewer and longer calls will result in a lower service-level than the same amount of work made up of more and shorter calls.*

Figure (9) shows the effect of call duration on waiting time, given a constant workload. Figure (9a) shows the mean wait time for groups of size 10, 20, and 50 agents, with agent utilisation constant at $\rho = 88.85\%$, chosen so that 50 agents with $T_S = 180$ exactly achieves a target of 80% answered within 15 seconds. Mean wait is directly proportional to the mean call duration. A simple example helps explain why this is so. Consider the next call at the head of the queue, with all agents busy on calls. The waiting customer must wait for one of the calls in progress to finish, and with all agents busy the rate at which calls finish is T_S/m. So the wait is proportional to T_s. Figure (**??**) shows the same effect using the 80th percentile of waiting time.



(a) Mean wait (b) 80 %ile wait time

Figure 9: Effect of call duration

2.5 Call-centres and help-desks

★ *Help-desks are different from call-centres, usually having a relatively small number of agents and long call durations.*

Help desks are, in practice, a different situation to call centres. Typically a help desk has fewer agents and longer call durations. The mathematical theory for analysis of help desks is the same as for call centres, but the real situations are different. Figure (**??**) shows why. Suppose we have a call-centre with 50 agents and a mean call duration of 180 secs that is achieving a target of 80% calls answered within 15 secs. Compare this with a help desk with the same agent utilisation, but a mean call duration of 420 secs, as shown in Table **??**. Also, since a help desk may have a small number of very experienced or highly trained agents, it may be important to make the best use of their time, with little time spent waiting for calls, and a higher utilisation. These factors mean that a help-desk will need to use some kind of call-back or voice-mail system for users to request help, so that the long waiting times will be acceptable.

	Call-centre	Help-desk
Agents	50	10
Mean call, secs	180	420
Agent utilisation, %	88.9	

Table 7: Call-centre and help-desk parameters

	Call-centre	Help-desk
Mean wait, secs	10.2	239.3
80th %ile wait, secs	14.8	435.3
Prob(zero wait) %	68.4	36.5

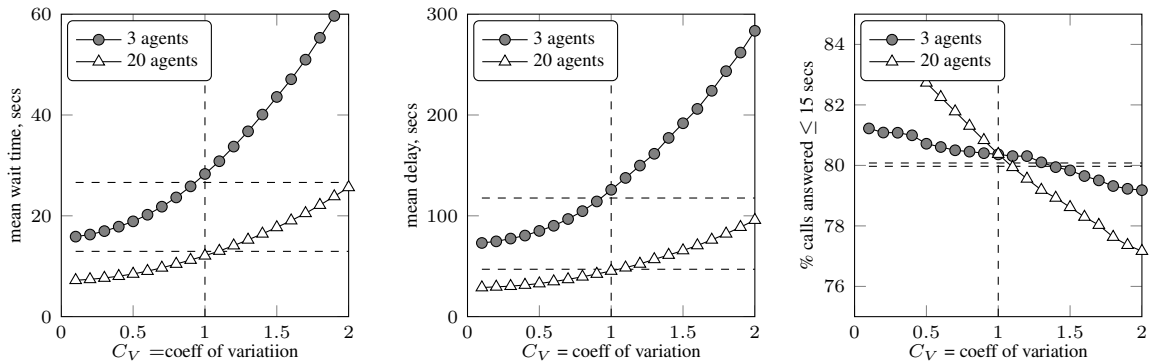
Table 8: Call-centre and help-desk performance

2.6 Variability of call duration

★ *More variability of call duration increases waiting time and decreases service level.*
★ *Variability of call duration has more effect on smaller agent than larger groups.*

As well as average call duration, variability of call duration also matters. If call duration varies a lot, then there can be a bigger impact of waiting times when some especially long calls arrive closely together. As a general rule-of-thumb, constant service times halve the mean wait compared to exponential service times. It isn't usually possible to make service times constant, but reducing the variability will improve waiting time. Variability is measured by the "coefficient of variation", which is simply the ratio of standard deviation to the mean, as in (2). The larger C_V, the more variable the service time. $C_V = 0$ corresponds to constant service times and $C_V = 1$ to "exponential" service times as assumed for the Erlang-C formula. The effect of variation is illustrated in Figure (10)

$$\text{coeff. of variation} = C_V = \frac{\text{standard deviation}}{\text{mean}} = \frac{\sigma_S}{T_s} \tag{2}$$



(a) Mean wait time (b) Mean delay time (c) Service level

Figure 10: Effect of variability of call duration

Figures (10a) and (10b) show how answer time and mean increase with C_V, and Figure (10c) shows service level decreasing. The Erlang-C formula assumes $C_V = 1$, so the results shown here were obtained by simulation..

2.7 Randomness of arrivals

★ *Linkage or correlation between arrivals means increased queue size and lower service level.*

Random arrivals means that no linkage between the arrival times of different customers, such as there would be if customers arrived in groups. "Random" has a precise technical statistical meaning. Random arrivals may, by chance, bunch together at times, and at other times be widely spaced. This is not the same as the underlying rate of arrivals varying with the time of day. We measure randomness by the coefficient of variation (see above) of the inter-arrival times. Completely regular arrivals would have a constant time between arrivals, with $C_V = 0$. Arrivals in large groups would mean that inter-arrival times are a mixture of zero (between members of the same group) and long intervals between groups, and this would make C_V very large. The assumption in Erlang-C, random arrivals, is that $C_V = 1$, and is justified by measurements of real systems.

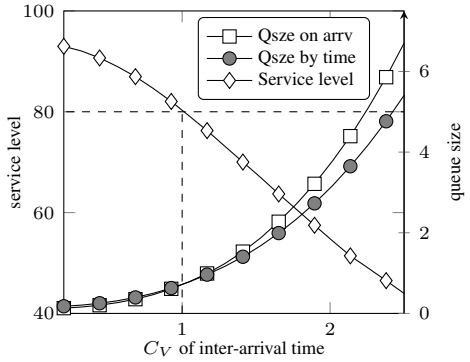


Figure 11: Randomness of arrivals

Figure 11 shows how the average queue size and service level are affected by arrival pattern, characterised by C_V. Several interesting effects occur. With $C_V = 1$, completely regular arrivals, the average queue size is minimised. Queues still happen because of variation in service times. There are two separate curves for average queue size, and these will be explained below. Service level is affected significantly by non-randomness in the arrival pattern. If calls bunch together more, then some calls will encounter a bigger queue and longer wait than otherwise.

2.8 Customers view and servers view

★ *Linkage between customer arrivals, such as arrining together in small groups, then queue size will be greater and service level worse than for random arrivals.*

★ *The perception of how busy the queue is can be different between servers and customers, and is only the same for purely random arrivals.*

Figure XX shows two different curves for average queue size, because there are different ways of measureing average queue size. The curve marked "QARV -queue on arrival" is measured by recording the queue size at the time each customer arrives. The curve maked "QTME - queue by time" is obtained by recording each change in queue size (so not only arrivals, but also departures from the queue), calculating the total time for which the queue was $0, 1, 2, \cdots$, and calculating the time-weighted average.

QARV is the queue size as seen by arriving customers. QTME is the queue size as seen by the servers. It is interesting that QARV is not the same as QTME, though both are valid measurements. QARV=QTME only when

$C_V = 1$, that is for pure random arrivals. This can partly explain why staff in, say, a bank not understand the custo,ers view. The staff see QTME, because they are aware of the intervals when there is no queue, while the customers see QARV. The same difference of view will arise from busy and quiet times, though the cause of that is varying underlying arrival rate rather than buching of arrivals.

3 Call scheduling methods

3.1 Introduction

A call scheduling method is how calls are selected from the queue to be answered. These methods can be divided into several categories, as shown in Table 9, and are explored in the rest of this section.

Category	Abbreviation	Meaning
Lost-call methods	LQSZ	Limited-Queue-SiZe
	LWTM	Limited-Wait-TiMe
	LWTM	Limited-Wait-TiMe
Independent of duration	FCFS	First-Come-First-Served
	LCFS	Last-Come-First-Served
	RSS	Random selection for service
	WPTY	Wait priority
Duration -dependent	SPF	Shortest-Process-First
	LPF	Longest-Process-First

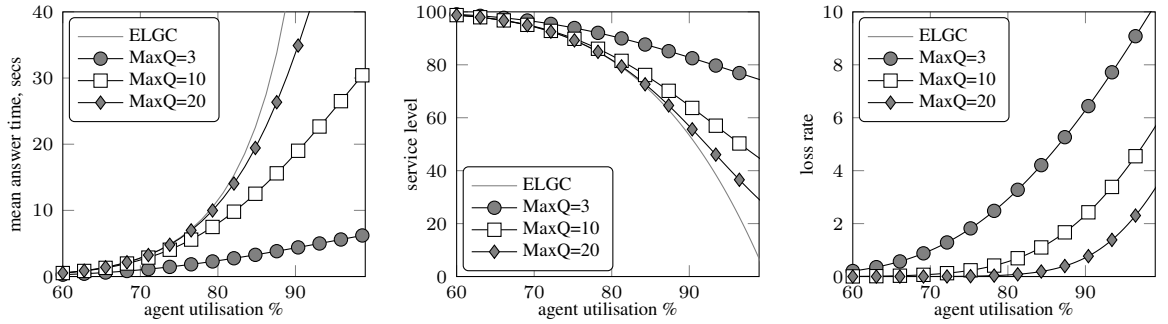
Table 9: Scheduling methods

3.2 Lost call methods

Calls may be refused or disconnected at times of transient overload. It may be better to deliberately refuse a call than subject the caller to a long wait, and by removing calls when a transient peak load occurs, the time that the peak persists can be reduced. The LQSZ and LWTM models assume lost calls are never retried, but can be extended to include retries. (There is an important difference between callers retrying after a significant delay, which is easily modelled by increasing the offered call rate an appropriate amount, and near-immediate retries which do not reduce the transient peak load and are more complicated to model).

3.2.1 LQSZ - Limited queue size

With LQSZ a maximum queue size is set, and when the maximum is reached new calls are rejected, perhaps a recorded message saying try later. The practical message from LQSZ is that penalising a small proportion of callers can improve both answer time and service level.

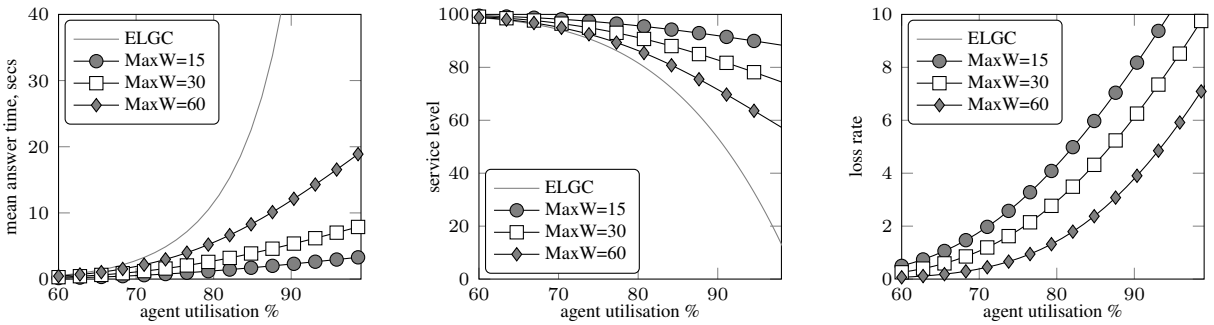


(a) Mean waiting time (b) Service level (c) Loss rate

Figure 12: Limited queue size

3.2.2 LWTM - Limited waiting time

LWTM is similar to LQSZ, but with LWTM a call is rejected after the caller has already waited some time, compared to LQSZ where the caller is at least spared a long wait. LWTM may be appropriate with call-back systems, where instead of being rejected, the caller is given the choice of being called back later. (Callbacks can be offered either with a LQSZ scheme, or with a LWTM, or both.)



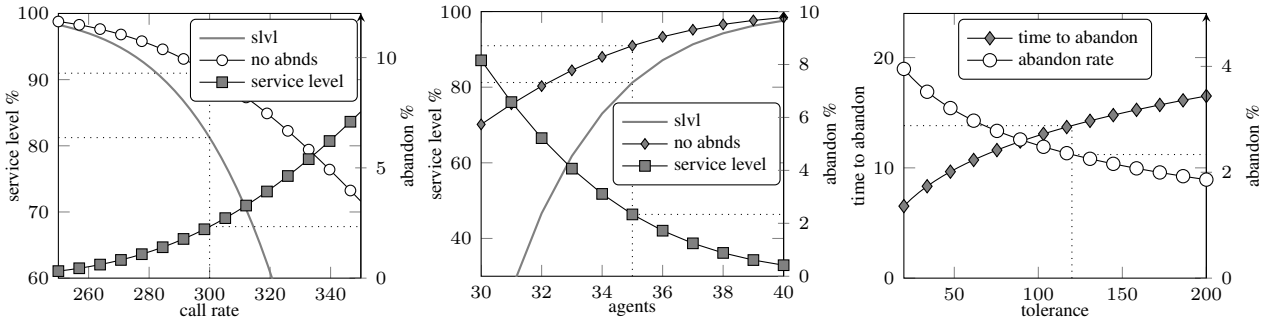
(a) Mean waiting time (b) Service level, 15 secs (c) Loss rate

Figure 13: Limited waiting time

3.3 Abandoned calls

Abandoned calls are not a sceduling method but are included in this section because their effect is similar to queue size or waiting time limits. A small number of calls are lost, and since those calls are lost during a transient peak, service level is improved. The difference with abandoned calls is that the tendency to abandon is a characteristic of the callers, and cannot be controlled in the same way as setting a maximum queue size limit or waiting time. The ”tolerance” factor is the average time a caller will wait before abandoning the call. Tolerance is a random value for each call, so is different from a fixed waitin-time limit. Figures reffig:AbndVsCrte and 14b show examples of how service level and abandon rate relate to call rate and number of agents, assuming a value for tolerance.

Figure 14c shows the relationship between the tolerance parameter and ”time to abandon”. Tolerance is an average over all callers, those that abdon and those that do not, so includes callers with high tolerance that do not. thereforem actually abandon. Time to abandon is the average tolerance over those callers which do abandon, and so only included relatively low values.



(a) Call rate (b) Agents (c) Tolerance and abandon time

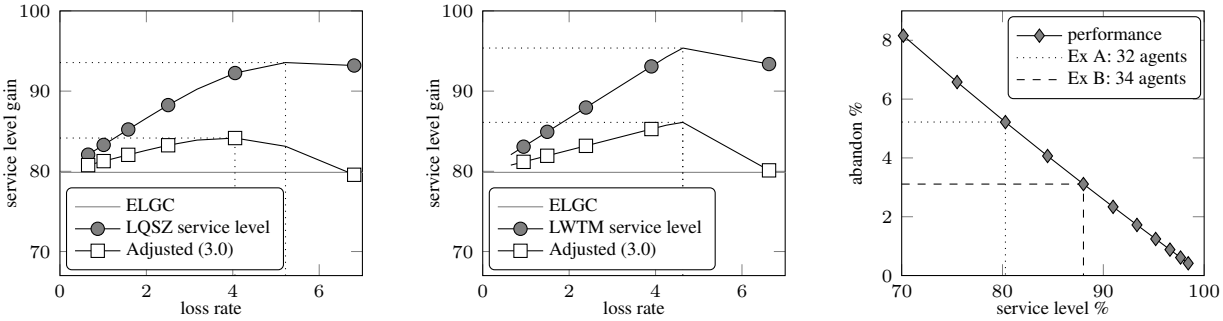
Figure 14: Abandon calls vs call rate and agents

3.4 Lost calls and service level

Queue limits and abandoned calls act as a safety valve, rejecting calls during transient peaks in the workload. This will reduce average waiting time and increase service level for most callers, at the cost of a relatively small number of customers getting no service at all. The queueing-theory models of these mechanisms assume that calls once lost are not later retried, although in practice a large proprtion of rejected callers can be expected to retry. The models can be extended to take account of retries, but for simplicity the basic models are used here.

Management of a call-centre must judge what level of lost calls is worth accepting in order to preserve good service for most callers. Figures 15a and 15a show examples of the relationships between service level and loss rate. The service level is significantly improved compared to the ”ELGC” curve which is for no lost calls. The loss rate increases as the queue limits are made more stringent, but beyond a certain loss rate service level does not improve, so there would be no point in using limits which result in higher loss rates. While a long wait is poor service to a caller, being rejected would seem much worse service, so we can gain some insight by giving more weight to a lost call than to a call that waited longer than the target time. In Figures 15a and 15a the ”Adjusted” curves are the service level adjusted by countin a lost call as 3 times (a purely arbitrary vale) the value of a late-answered call. The result emphasises that it is only worth accepting a quite small percentage loss rate.

The situation for abandoned calls is shown in Figure 15c. For queue limits management has two controllable input factors, the queue limit and number of agents, so that a trade-off between servic elevel and loss rate can, within limits, be made. With abandoned calls the only controllable factor is the number of agents, which determines both the service level and the abandon rate.



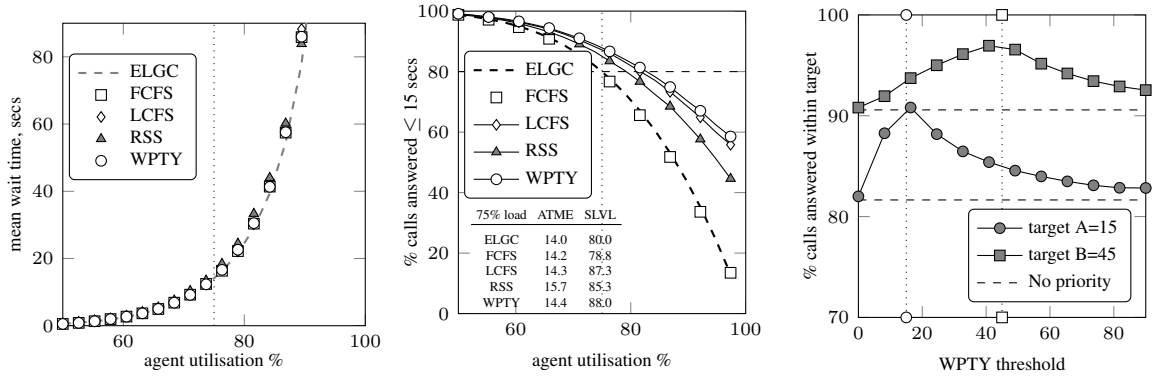
(a) Limited queue size (b) Limited waiting time (c) Abandon rate and service level

Figure 15: Service level and loss rate for LQSZ and LWTM

3.5 Independent of call duration

★ *Call-scheduling methods which do not depend on individual service times produce the same average waiting time, but can give different service levels.*

An important result from queueing theory is the ”conservation law” (see [5]) which means that if calls are selected for answering in a way that is independent of their duration, or any measure of their duration then the average waiting time will be unaffected by the order of service. This doesn't imply that the distribution of waiting time will be the same, just the average. Figure 16a shows how mean answer time is the same for very different scheduling methods.



(a) Mean waiting time (b) Service level (c) WPTY threshold

Figure 16: LCFS, RSS and WPTY, simulation results

3.5.1 FCFS - First come first served

It is often taken for granted that calls should be answered in the order in which they arrive, since that seems ”fair” and fairness is known to be an important factor in how people react to queueing. Exceptions to FCFS are accepted where there is an obvious justification, such as hospital Accident and Emergency, (a UK term, or ER to US readers) where life-threatening conditions get seen first. In supermarkets, basket-only checkouts are accepted or expected. These situations are perceived as fair over the long term. Surprisingly, FCFS is relatively ineffective for meeting typical service-level targets. Figure 16b shows that even RSS outperforms FCFS.

3.5.2 LCFS - Last come first served

LCFS seems a strange scheme, and far from fair, but it can have surprising benefits in improving service level and reducing abandon calls. Apart from call centres LCFS is effective in managing perishable products, including storage of blood for transfusions.

3.5.3 RSS - Random selection for service

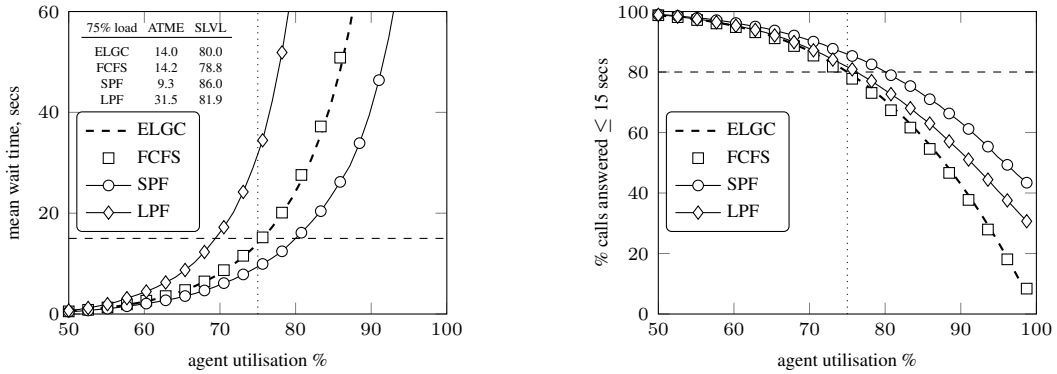
With RSS the next call to be answered is selected at random from among the calls waiting. RSS applies where calls that cannot be answered immediately are rejected with a busy tone or recorded message, rather than held in a queue, and the callers expected to call back. Typically this happens only in very small CC's, or small groups of people taking calls that do not think of themselves as a call-centre. One example is calling a health centre to make an appointment with a doctor. There may be only two of three people taking calls, without any call queueing system, and if a caller gets a busy tone then they must simply retry.

3.5.4 WPTY -Waiting time priority

With WPTY an arriving call is queued at normal priority, but if it waits for more than a specified threshold, the call is relegated to a lower priority. Service level is maximised when the threshold is set to the target answer time. Once a call can no longer contribute to the achieved service level it is relegated to be answered only when no ”non-expired” calls are waiting. This seems a ruthless approach to maximising service-level, but Figure 16c shows that WPTY improves the overall service level compared to FCFS. In a busy system a relegated call may wait a long time to be answered, but this can be mitigated by setting a second, longer, wait time threshold at which a relegated call is reinstated to normal priority.

3.6 Duration dependent methods

When a priority scheme is based on, or indirectly related to, the call durations, then the conservation law no longer applies so the average waiting time may change.



(a) Mean waiting time (b) Service level

Figure 17: SPF and LPF, simulation results

3.6.1 SPF - Shortest process first

The supermarket basket-only checkouts have already been mentioned. Similarly there may be ”quick-tills” in banks. These are not strictly SPF, since a customer with just one item in a basket will not be served ahead of someone with 10 items. But the basket-only till is a practical approximation to SPF. SPF can be shown to minimise the average waiting time in a queue. Since in a call-centre the duration of a particular call cannot be known until it is answered (and not even then) SPF cannot be applied. However, SPF can give us an indication of the performance of priority schemes which implicitly give priority to shorter calls. SPF also provides a lower bound on the average waiting time that can be achieved with any priority scheme.

3.6.2 LPF - Longest process first

LPF does not seem a likely scheme in practice, as intuition and experience tells us that service will be worse with LPF than with most other schemes. The relevance of LPF is to indicate the possible performance of priority schemes based on criteria that implicitly give preference to longer calls.

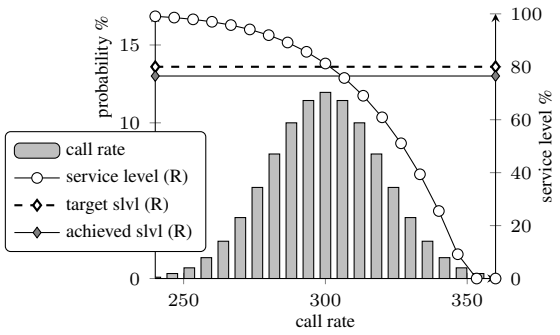
4 External factors

This section looks at forecast error and agent abscence. Both sources of uncertainty increase the number of agents that have to be scheduled in order to meet, on average, whatever answer-time target is set. In practice a call-centre would have agents scheduled to work on background tasks, who could then be switched to call handling to cover absent agents or to handle a greater than forecast workload. So while it is mathematically possible to include forecast and agent uncertainty in the queueing models used for call-centre planning, it is not usually done in practice.

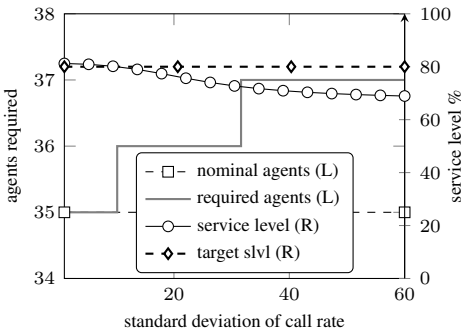
4.1 Forecast error

The call rate used for planning will be a forecast value, with margin of error, and can be treated as a random variable with a Normal distribution[1]. In the example shown in Figure **??** the mean is 270 calls per interval, with the standard deviation of 5% of the mean. For simplicity we represent this as a number of discrete values.

Represented in this way, a positive error is just as likely as a negative error of the same magnitude. Looking back at XXX means that an overestimate of call-rate may result in more agent capacity being assigned than in reality is needed, and a higher achieved service-level. Conversely an under-estimate of call rate may result in fewer agents, and a lower achieved service level. The non-linearity of service-level vs call rate means that the over-estimate will improve achieved service-level less than the under-estimate will decrease service-level. So a calculation, such as that below, which takes account of call-rate as a random variable will, in general, assign more agent capacity than just using the central, most likely, call rate.



(a) Service level

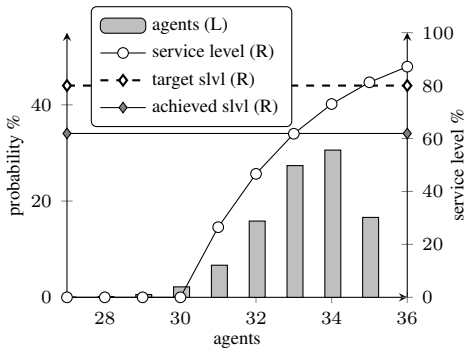


(b) Agents needed

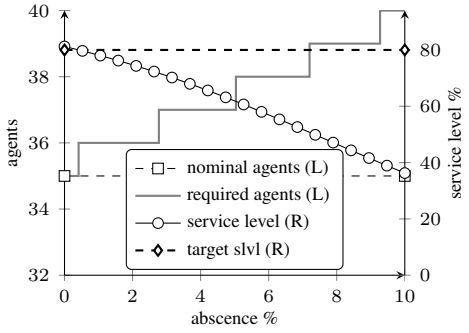
Figure 18: Forecast error

4.2 Agent absence

A particular number of agents will be scheduled, but some may be absent. Figure shows the example of 41 agents scheduled with each agent attending with 95% probability. It is not accurate to assume the expected number of agents is $0.95 \times 41 \approx 39$, since there is a significant chance that fewer than 39 will attend, and this will impact the expected service level.



(a) Attendance



(b) Required agents

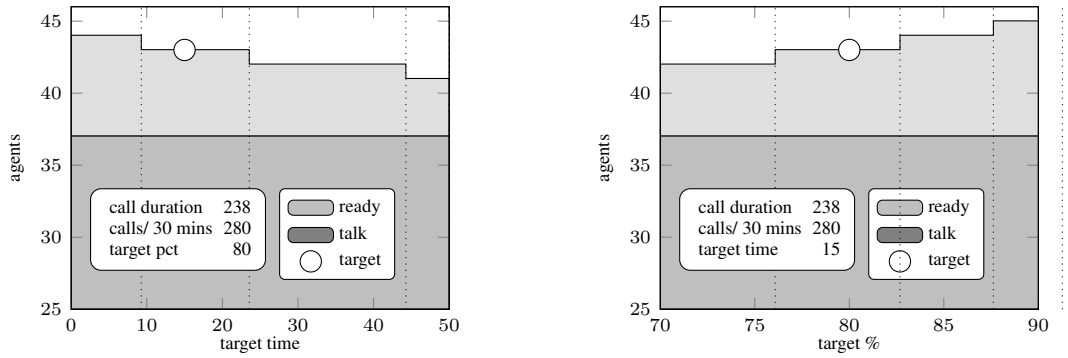
Figure 19: Agent attendance

[1] Strictly this should be a truncated Normal distribution, but since, with the parameters chosen, low values have an almost zero probability, we can ignore this complication.

5 Comparing targets

5.1 Sensitivity to target parameters

The purpose of a target is to decide how many agents are needed. How much does the target have to change in order to change the number of agents? Figure 20 shows, for a particular workload, the effect of target time and target percentage separately. The target is 80% within 15 seconds, requiring 43 agents. Figure 20a shows the number of agents remains the same as long as the target answer time is from 9.3 to 23.6 seconds. Similarly Figure 20b shows the target percentage can range from 76.1 to 82.7% without altering the number of agents.



(a) Target time (b) Target %

Figure 20: Effect of targets

5.2 Basic Erlang-C targets

The chart in Figure 21 shows how the number of agents responds to the target parameters. The horizontal axis is the target percentage of calls, and the vertical axis is the target answer time. The diagram applies to a single workload of 280 calls per 30 mins with an average call duration of 238 seconds, that same as in Figure 20. The target shown is 80% calls within 15 seconds, marked A on the chart, which requires 43 agents. An alternative target of 90% answered within 45 seconds is marked B, and results in the same number of agents. The targets A and B are simply alternative ways of expressing an identical target, even though they may sound rather different. Anywhere on the dashed line through A and B is also an identical target to A and B. Figure 20 shows that the target parameters can vary significantly without changing the number of agents needed. The arrows from A to the band edges in Figure **??**fig:trgtband correspond to Figure 20, and the shaded band containing target A is the range of targets that will still result in 43 agents being needed. Moving out of this band will give a different number of agents. Each band on the chart corresponds to a particular number of agents. The example has been chosen so that the target is located near the centre of one of the bands, but in other cases the target could be close to a band edge.

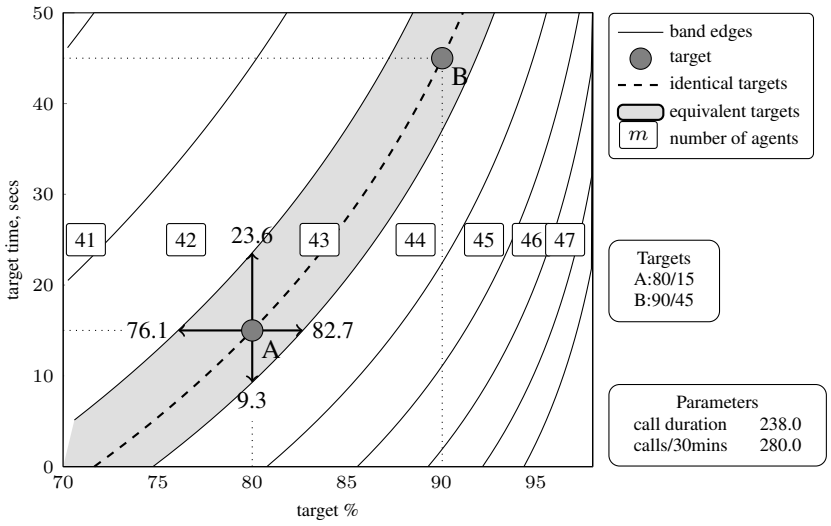
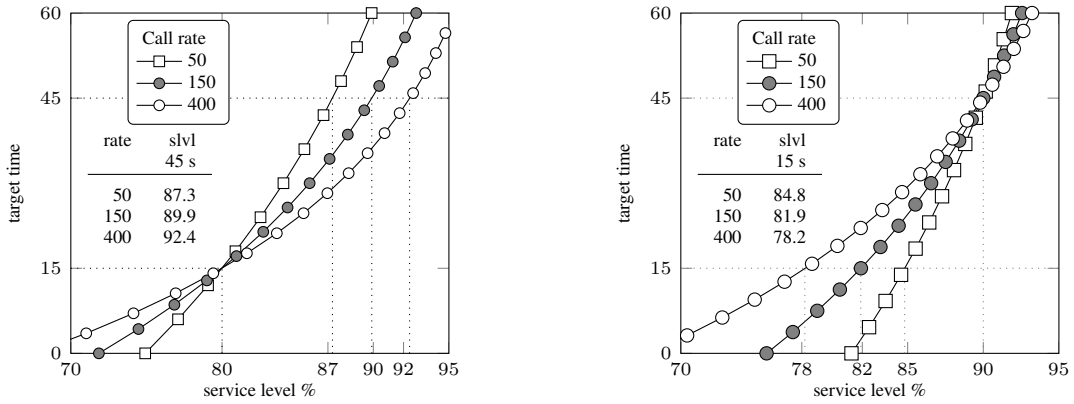


Figure 21: Targets bands

5.3 Comparing targets



(a) 80/15

(b) 90/45

Figure 22: Comparing targets

Targets are identical or equivalent only for a particular workload. Figure **??** shows, for a range of call rates (with the same mean call duration) the targets that are identical to 80/15. For this analysis we assume that fractional agents can be used: the same effects would be seen using integer agents, but the graphs would be less clear. If, for a particular call rate, we assign agents to achieve 80/15, then we will get a 95th percentile of waiting time. For 100 calls/interval the 95th percentile is 102 seconds. For 30 calls/interval, the 95th percentile is different, at 155 seconds, while for 400 calls/interval the 95th Percentils is 66 seconds.

If 15 seconds is believed to be an important threshold for callers reaction to waiting, then using an X/15 target makes sense, and the different 95th percentiles at different call rates are not important. However, if seems just as likely that callers reactions are related to a higher threshold, a waiting time considered definitely unreasonable and only to be suffered exceptionally, say by 1 in 20 callers, or a 95th percentile. So although an 80/15 target implies some 95/Y target, the target time chosen matters. There seems no objective basis for defining, say 15 seconds, as a ”reasonable” waiting time. There is also little objective basis for saying what is an exceptional and unacceptable waiting time.

Another way of looking at targets is to observe that call centres have to be efficient, and the utilisation of agents cannot be too low. Recognising that long waiting times mean poor caller satisfaction, and bearing in mind the relationship between utilisation and waiting times demonstrated by Erlang-C and other queuing models, there is a limit to how high utilisation can be. So the service-level targets used may be simply implicit utilisation targets. An explicit agent utilisation target, decide against a background of queueing theory service-level predictions, might be just as effective and simpler to manage.

References

[1] Tanner M. *Practical Queueing Analysis*. McGraw-Hill 1995.

[2] The Health Foundation. *Ambulance Response Times*. 2018. www.qualitywatch.org.uk.

[3] National Audit Office. *NHS Ambulance Services*. 2017.

[4] Kleinrock L. *Queueing Systems Volume 1: Theory*. John Wiley & Sons 1975.

[5] Kleinrock L. *Queueing Systems Volume 2: Computer Applications*. John Wiley & Sons 1975.

[6] Gnedenko B V and Kovalenko I N. *Introduction to Queueing Theory, 2ed*. Birkhauser Boston 1989.

[7] Gross D and Harris C. *Fundamentals of Queueing Theory*. John Wiley & Sons 1998.

[8] Tijms H. *New and old results for the M/D/c queue*. Int. J.Electron Commun. Vol 60 Pages 125-130. Elsevier 1998.

[9] US Patent 6044355. *Skills-based scheduling for telephone call centers*. US Patent Office, 2000.

[10] Tanner M.. *Introduction to Simulation*. Mitan Ltd. 2018.

9