Service Quality and Customer Abandonment: a System Dynamics approach to Call Center Management.

Massimiliano Caramia
Istituto per le Applicazioni del Calcolo "M. Picone"
caramia@iac.rm.cnr.it

Stefano Armenia, Riccardo Onori
DISP - Faculty of Engineering - "Tor Vergata" University, Rome
armenia@disp.uniroma2.it - onori@disp.uniroma2.it

Valerio Giannunzio
Faculty of Engineering - "Tor Vergata" University, Rome
v.giannunzio@virgilio.it

Abstract

The ability to profitably manage the level of resources in a service system can be considered a strategic skill in all those organizations, including no-profit ones and Public Administrations, that aim at providing an added value service to customers as well as balancing the level of service (in terms of quality) with costs. In this paper we will focus on a typical service system inside of which, in every moment, management struggles in order to reach that balance, because of the extremely dynamic behavior of the entire system: a Call Center. Our aim is to show that an efficient management of the customers abandonment and the quality of service offered to customers, can positively impact on a correct resource leveling in the system, which may otherwise be found by means of typical Operations Research or Queuing Theory methods. In particular, we want to show that this can be more easily inferred and understood by resorting to simulation. After introducing some preliminary aspects by means of Queuing Theory (Erlang's formula), we'll first study the problem of customer abandonment (balking and reneging) by simulating a Call Center simple model by means of a process-oriented discrete-time simulation tool (Arena), and then explore a more complex model taking into account customer satisfaction an the quality of the service offered approaching the modelling process of the system by means of System Dynamics. Results show that the level of resources can be further reduced, and that the customer (often thought as an entity external to the system) plays instead an important role on the performance of the whole system, both operationally and economically.

1. Introduction

Since the beginning of the so called civilized society, man has always had to deal with queues, e.g. in order to receive food, for entering into buildings, or generally in processing customers orders in a shop. As customers, we are usually quite disappointed with long waiting times, but also managers of the establishments at which we wait do not like us to wait for too long, since it may cost them a lot in terms of business, money, customer fidelity, and so on.
The evolution in the ways of doing business is today greatly based on the fast growing awareness that managing the relationships with customers has become one of the key factors which can lead a modern organization to success. That is why organizations need to understand that correctly managing their Service Systems may in the long run reveal itself as one of their best competitive advantages.

In this work we will describe how to take into account customer satisfaction in order to describe queue abandonment and how a qualitative service can reduce a system resources usage, thus improving those results which could have been drawn with a typical mathematical approach, Queuing Theory (QT) or Operations Research (OR). In particular, we will show some of the issues which the management of a Call Center must traditionally confront with, and we will try to explain how simulation tools and System Dynamics (SD) may be used to better analyze the performance of a generic Service System and its management policies.

Specifically, in the first part we will try to put in evidence how a strategic issue like resource management in a Service System like a Call Center may suffer from critical aspects, e.g. customer abandonment or quality of the service offered, and how simulation tools may help to overcome some of the problems encountered in the analysis. We will start with an introductory scenario, with hard data, and with an analysis drawn according to Erlang's formulas. Then we will see how simulation with an Event Driven Simulation (EDS) software like Arena fits better in some situations where mathematical analysis may become too difficult because of the complexity of the system, and we will give an example based on the same scenario. Simulation results will also be provided. We will see how System Dynamics may help to show more evidently the strict relationship between Service Level and Quality of Service, Customer Satisfaction and Agent Burnout, as well as how these are connected to staffing issues or to workload forecasting and handling. In particular, we will show how a SD model can improve the performance of a Call Center by comparing with the results obtained by Erlang's formulas and with Arena.

2. Where simulation fits in the analysis of a Service System

Whilst queuing theory can be used to analyze simple service systems, more complex service systems are typically analyzed using simulation. In fact it turns out that it is not possible to develop analytical models for service systems because of the characteristics of the input or the service policy, the complexity of the system itself, the nature of the queue discipline or combination of the above. For example, a multi-station multi-server system with some recycling, where service times are normally distributed and a complex priority system is in effect, it becomes almost impossible to model analytically (Hillier, Lieberman, 1990). Furthermore, if one were interested in the transient behavior of a service system or if the probability distributions were to vary with time according to various relationships between dynamic parts of the system, it might not be possible to develop analytical solutions or efficient numerical schemes.

More in detail, there are three major reasons that justify the use of simulation tools in a Service Center analysis:

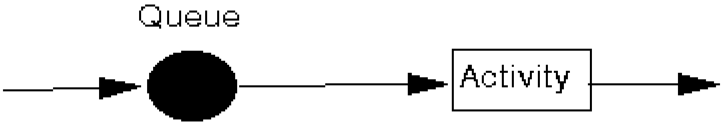
1. **Uncertainty:** arrival times are highly unpredictable and arrivals tend to come up in bunches, making it very difficult to properly assess system resources without the right tool. Simulation is designed, as one of its core values, to handle uncertainty;
2. **Complexity** of the Service System structure: other tools like spreadsheets do not allow to put in due evidence the many interactions and interdependencies occurring in the system;
3. **Dynamic** environment: most service centers are today going through dramatic changes in technology and work flow processes.

Some very general applications of simulation in a service system may be to perform evaluations and/or comparisons between policies or strategies as well as also between simulation and other analytic tools, to predict outcomes, optimize solutions, analyze functional relationships and perform sensitivity analysis. No matter what type of analysis is carried out, the latter generally begins with a series of ***"what-if"*** questions like, for instance, what would be the impact on a Call Center's efficiency and effectiveness if more staff is added or a Voice Response Unit (VRU) is installed, calls are rerouted to another center after a certain hour in the day, demand increases/decreases, calls are rerouted on a skill-based basis, several Call Centers are consolidated into a single one (Anton-Bapat, 1999).

We can generally group simulation applications to Call Center Management into short-term (Service Level setting, forecast analysis, allocating staff, budget analysis, customer affectivity analysis, agent scheduling, impact of new technologies, call-flow pattern analysis, introduction of new products/services, productivity analysis, turnover analysis) and long-term planning horizons (adding a new CC, consolidation, CC assessment and benchmarking, becoming a virtual Call Center, adding electronic access, outsourcing, integration, multifunctional Call Center).
Some of these issues could have never been analyzed using traditional methods, while other get better and more consistent results when analyzed through simulation. In the past, many poor decisions about procedural change were probably taken in such a light mood that the disruption they provoked went unnoticed until after the companies had already paid a high price in term of lost customers and tarnished reputation. No customer-conscious company can afford to take such risks anymore in this competitive age.

2.1 Setting the context: the call center

Recall from the queuing theory that in essence all service systems can be broken down into individual sub-systems consisting of entities queuing for some activity (as shown below).



To analyze these sub-systems we need information relating to six basic characteristics: the arrival process (how customers arrive, how arrivals are distributed in time and whether the population is finite or infinite), the service mechanism (base resources, service time distribution, number of available servers, whether they are in parallel or in series, whether pre-emption is allowed, etc…), queue discipline (FIFO, LIFO, etc…) and characteristics (is there any balking or reneging?), system capacity, number of service channels and number of service stages (Gross-Harris, 1998).

A quick look inside a typical Call Center (denoted throughout the paper as CC) reveals complex interactions between several "resources" and "entities", where the former are the operators answering the phone and the latter take the form of calls or, rather, customers calling the CC in order to receive a certain service. These calls, usually classified by call types, then navigate through the various CC structures; this means that while traversing through the CC, calls occupy trunk lines, wait in one or several queues, abandon queues and are redirected through interactive voice response (IVR) systems until they reach their destination, an agent or some predetermined self-service feature. Once the call is handled, or the customer has received service, it then leaves the call center. During all of these transactions, another critical resource is consumed, time.

By referring to the queuing theory, a call center can be sketched as a parallel-server birth-death (Kleinrock, 1975) model, also known as a M/M(G)/c/K service system, based on:

1. an exponential or Poisson call arrival process (M)
2. an exponential or general service time distribution (M or G)
3. a finite number of parallel servers (c)
4. a finite system capacity (K, in our case the capacity of the queue)

Notwithstanding these general concepts, we could model the arrival process by imposing that calls do not arrive simultaneously, but one by one and with a given expected inter-arrival time ("birth process"). We know that calls tend to bunch up especially in certain moments of the day, but we are not too wrong in our limitation, especially because calls enter the queue according to a certain discipline and thus any two (or more) of them cannot occupy a queue slot at the same time. It is also necessary to know the reaction of the customer upon entering the system. A customer may decide to stay in line no matter how long he has to wait or may decide not to enter the system. If he/she decides not to enter the queue upon arrival, he is said to have balked. In a CC, a customer calling may not have any idea of how long the queue is until he really is in the line (and in the latter case, only if he receives information about expected waiting times), thus we cannot talk about balking but we can model the fact that the customer may find a busy line because of trunks already full to their limit. On the other hand, a customer waiting may decide to leave the queue (that is reneging, also known as abandonment) either because he has been waiting for too long or because he received information on how long the waiting time is in that particular moment. If we model this aspect only by taking into account that customers may all the time retain the prerogative to renege if their estimate of the total wait is intolerable, we may run the risk to underestimate (or overestimate, depending on the situation) abandons (refer to the queuing theory for the aspect of queues with impatience; Gross-Harris, 1998). This is one of the main reasons (even if not the most important, as we will see later on) that would account for the use of simulation in such a context.

Moreover, we assume that calls get service according to the FIFO discipline and as soon as there is a free agent. This brings us directly to the issue of the service process. We can divide it in two phases: the first, during which customer and agent talk and the customer gets service (thus talk time equaling service time by the customer point of view), the second, during which the agent is required to make a sort of after call work, in order to store all the information he did not have the time to write down during the conversation or to perform those actions requested by the customer. The expected mean service time, by the service process side, is then given by the sum between talk time and after call work (in such a context, this is the so called "death rate"). This means that every time an agent gets free again, he picks up a call, thus getting busy, and becomes free again only after a certain time.

Another important issue is about the finite system capacity, which in our case may be simplified by assuming a finite queue. From the CC point of view, this translates in finite number of trunks, each with a finite number of call slots available, even if we will further simplify the situation by assuming that the system has only a single service channel and a single stage of service. The chance that customers might want to call back if they were not satisfied with the service they received should carefully be accounted for, as well as callback probability in the chance that a customer finds a busy line or after having abandoned.

2.2 The advantages and disadvantages of simulating a Call Center

In such a dynamic and volatile environment as a service system, the intrinsic value of simulation resides in that it may allow a top senior manager, or just an analyst or advisor, to take better decisions than those eventually derived from traditional analytical approaches, as well as virtually cutting any risk associated to an improper or useless business strategy. The objective of the CC manager or analyst is twofold: first, to achieve a high service level (SL), i.e., to get the caller to an agent in the shortest amount of time, and second, to provide the caller with the appropriate information in the most efficient manner (measured in terms of call talk time and handle time). The net objective is to minimize the time spent by the caller in the CC while providing the best possible service. Ultimately, these issues generally come down to a trade off of better customer service versus the expense of providing more service capability, that is determining the increase in investment of service for a corresponding decrease in customer delay (and sometimes also in costs due to a tool-free service; Cleveland-Mayben, 1997-1999). These primary measures and objectives usually reflect the performance of a CC, and balancing these objectives can be a challenging task for call center analysts. Furthermore, there exists a great deal of sensitivity in the cause and effect of the performance parameters involved (Anton-Bapat, 1999). For example, a small adjustment in call routing may have a significant debilitating change on customer, or a minor reduction in trunk-line capacity may cause too many “busies” and raise the potential for lost customers (which, in turn, according to the well known Customer Based View, may in the long run lower the revenues and profits of the company). Moreover, also very important, an incorrect staffing may cause long waiting times, frustrated customers and exasperated agents. Such circular relationships must carefully be defined and analyzed in order to achieve peak performance for the call center.

A disadvantage in simulating a Service System is that it is sometimes difficult to find optimal solutions, unlike linear programming where, for example, we may have different algorithms that will automatically find an optimal solution. One way to attempt to optimize using simulation is to make changes to the model (by setting its parameters) and run the simulation computer program to see if an improvement has been achieved or not, and repeat. This process can consume large amounts of computer time. Nowadays there are fortunately some simulation packages which also support some optimization tools. In our paper, we didn’t resource to any optimization since it goes beyond the basic scope of this article. In the last paragraph we will also address this issue among those to be further explored in future works.

Instead, the advantages of using simulation, as opposed to analytical methods, are that it can more easily deal with time-dependent behavior; that the mathematics of queuing theory is hard and only valid for certain statistical distributions - whereas the mathematics of simulation is easy and can cope with any statistical distribution; that in some situations it is virtually impossible to build the equations that queuing theory demands (e.g. for features like queue switching, queue dependent work rates) and, finally, that simulation is much easier for managers to grasp and understand than queuing theory or difficult analytical methods.

2.3 Modeling a Service System like the Call Center with System Dynamics

By means of a systemic approach (i.e., System Dynamics) to the analysis of those phenomena present in any service system, it is possible to consider in the model also those soft variables which in other modeling methods would only be accounted for as external factors, and not as strictly correlated with the behavior of the so called internal system's variables. By simulating a SD model of the system, in which such "soft" factors have been included (as for example the so called Agent's Burnout) it is possible to see how the overall behavior may dramatically change as well as also the requested resource level. And in the particular context of our study, we will show how this is particularly true when accounting in the model for service quality, operator burnout and customers motivations, as well as for other causal relationships (Sterman J., 2000).

In order to develop our model, we only partially referred to existing System Dynamics literature on similar models. Other authors (Oliva, 1996), in fact, report SD modeling of waiting line based systems which only in part have helped us in understanding the absolutely peculiar dynamics of a telephone service system, mostly because they deal very specific issues in fields like the Health System (Gonzales Busto-Garcia,1999; Van Aeckere-Smith,1999; Emergency Room Dynamics Model, HPS Inc. 2002), the restaurant business (Fung 1999 and Fung 2001) and others (see also the well known Hanover Insurance model). Even if we haven't found clear evidence of previous specific literature on CC modeling with SD, we have however found many interesting issues on this topic in the queuing theory (Gross-Harris 1998, Kleinrock 1975, Naor 1969) as well as in the OR (Hillier-Lieberman 1990) and of course specific CC Management literature (Cleveland-Mayben 1999, Anton-Bapat 2000). We have thus sketched our own maps of the most evident causal relationships in such an environment and then drawn our SD model by also keeping well into account the typical processes in a system based on waiting lines (call flows and agent flows).

3. Call Center structure and dynamics

In this paragraph we will illustrate some of the issues connected to the analysis of particular problems in a typical CC. More in detail, we will analyze the impact of Abandonment and Quality on Staffing and Service Level. Towards an improved understanding of the issues we are going to deal with, we will develop some simple cause/effect diagrams in order to elicit and show the various systemic structures that drive the behaviors of a CC system.

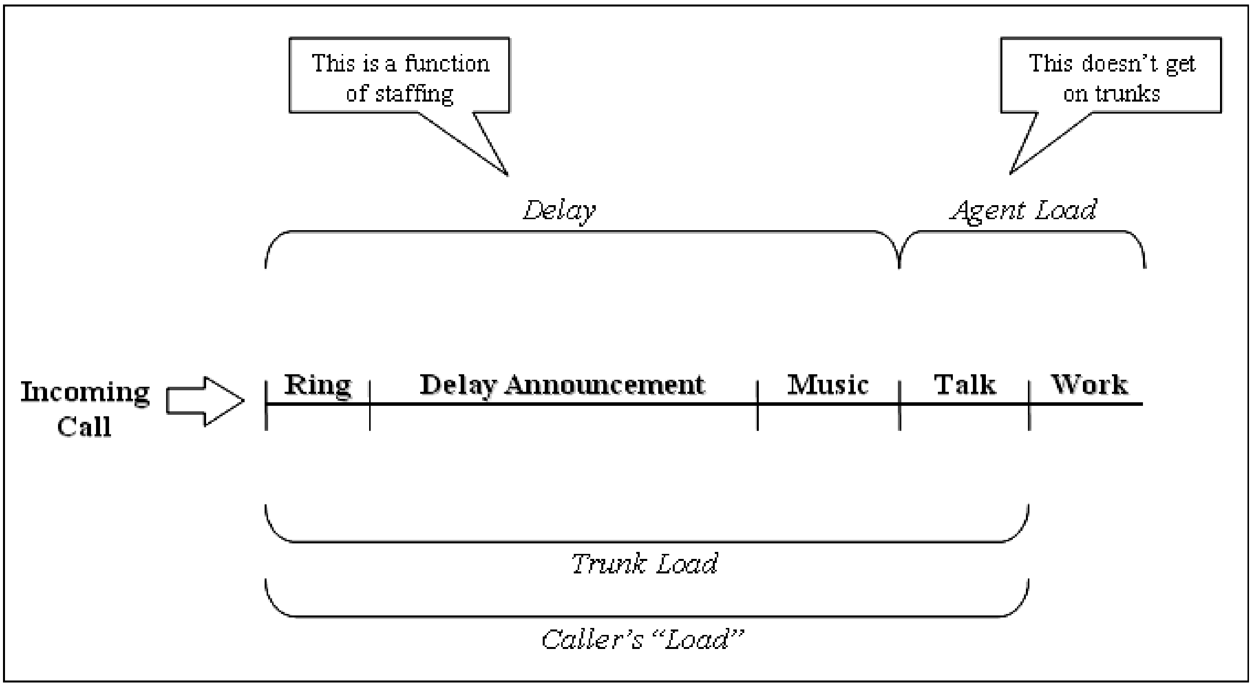


Figure 1: A view of a call flow inside the call center.

3.1 Some definitions

Let us introduce some of the acronyms and definitions used during the following pages (Cleveland-Mayben 1999, AA.VV. 2000):

Service Level (SL): X % of calls answered in Y seconds.
Average Speed of Answer (ASA): (from Little's formulas of the queuing theory: W_q, the expected waiting time in queue), also called Average Delay or Waiting Time (when seen from a customer point of view). It represents the average delay of all calls in a queue and can be calculated by dividing the total delay by the number of calls in a queue. It is the average waiting time of a customer in a queue.
Average Handling Time (AHT): it is the sum of the *Average Talk Time* (ATT) and the average After Call Work time (ACW). Also called "Agent Load" (from Little's formulas of the queuing theory: μ, the expected mean service time).
Agent Occupancy (OCC): it is the percentage of time that agents (also referred as TSR) spend in AHT with respect to the total time of their shift in schedule (from Little's formulas of the queuing theory: $\rho=\lambda/(\mu*c)$, the expected measure of traffic congestion for c-server queues). The inverse of OCC, represents the time agents are available and waiting to handle calls.
Telephone Service Representative (TSR): the generic human server who answers the phone. Also called operator, rep, agent.
Trunk: also called a line, a telephone circuit linking two switching systems.
Trunk Load (TKLD): it includes all aspects of the transaction other than the ACW. Can be viewed as ATT +ASA. Trunk load carries the delay (ASA). It is measured in erlangs (hours of trunk traffic), that is: $(ASA+ATT)_{in\_secs}$ * $N°\_of\_Calls_{(in\_1\_hour)}$

In general, a call flow inside a CC follows the scheme represented in Figure 1 (Cleveland Mayben 1999)). A customer dials the CC number and if he/she gets a "busy", depending on some factors affecting his degree of tolerance (his actual mood, motivation, the availability of substitutes, and so on) he decides whether to retry or not. Once he gets connected, in most of the situations, he then enters into a waiting line, characterized from ASA. Once again, with reference to the so called "*seven factors affecting callers tolerance*" (which include: service level of the CC, the time being in a queue and the possibility of "perceiving" its length, customers degree of motivation, availability of substitutes, competition's service level, level of expectations, time available to wait, who's paying for the call, human behavior; Cleveland-Mayben 1999) the customer may then decide to hold on and stay in the queue or to hang up ("abandon" the queue). Once at last a customer reaches an agent, the call gets handled. Handling a call requires time (AHT), in terms both of talk time between agent and customer and of wrap-up work time. From an agent point of view, the process does not stop after having handled a call. Instead it goes on in circle, because as soon as an agent becomes free again, he suddenly picks up the next waiting call.

3.2 Basic Call Center dynamics: a classical analytical method

In this paragraph we will make some general considerations on usual CC management practices as well as describe a classical analytical method, both in order to put the basis for understanding which ones are the high leverage points to act upon with the aim of improving the system's performance. We will put in evidence some limitations of the classical methods, thus setting the context for an approach with system dynamics.

The key to achieving Service Level objectives, ultimately comes down to having the right people in the right places at the right times and doing the right job (which means quality!). Base staffing and trunking calculations cannot be separated from a reasonably accurate forecast of call loads, which in the following, we will assume as being the best possible one. Note that, in the example shown in this paragraph, trunking should be calculated in conjunction with staffing because staffing impacts delay (or ASA), which, in turn, affects the load that trunks must then carry (see Figure 1). In fact, as also can be seen from the definitions in the previous paragraph, as a rule of thumb, the more staff handling a given call load, the less delay the callers will experience, and therefore it directly impacts how many trunks are required. There is no way to know base trunking needs without knowing how many agents will be handling the forecasted call load. Moreover, in general, there is no staff-to-trunk ratio or formulas that can be universally applied. In fact, if SL is low, the trunks will have to carry more delay and, consequently, more trunks would be needed. Staff and trunks are a classic example of the need to look at "the big picture", especially when speaking about the issue of integrating their related budgets.

Still, many CC managers calculate base staffing by using some ratios or formulas; even though these methods may sound logical, they are dead wrong, since they do not relate the outcome of staffing to the desired SL. That is mostly because the desired "targets" are moving. For example, staff productivity (calls that a group of agents can handle) is not a constant factor, rather it is continuously fluctuating because it is heavily influenced by vacillating call loads and SL objective. The biggest problem of these operative approaches is that they are quite simplistic, because they ignore one of the most important driving laws in incoming call centers: *calls bunch up.*

As seen in Paragraph 2.1, the whole CC service process may be simply summarized, without taking into account complicate interrelationships between the different parts of the system, only by a few key factors and equations. First, the arrival process can be modeled with a Poisson process based on a given inter arrival constant, λ; second the service process can be modeled either with an exponential or a general distribution with a mean value of 1/μ=AHT, third, the system capacity is given by the product of the number of trunks and the number of calls that each trunk can hold; fourth, the system has *c* operators, that manage customer calls according to a FIFO discipline. The queue has a single channel and the service has a single stage. This kind of models were studied in particular by Erlang, in 1917, and the original physical situation which motivated him to devise analytical formulas was of course the telephone network. Later on, Erlang's original formulas were adapted to staffing calculations as Erlang C and to trunking purposes as Erlang B. In particular, Erlang C was so formulated:

$$P(>0) = \frac{\frac{A^N}{N!}\frac{N}{N-A}}{\sum_{x=0}^{N-1}\frac{A^x}{x!} + \frac{A^N}{N!}\frac{N}{N-A}}$$

where:

A = total traffic (in erlangs)

N = number of active servers

P(>0) = probability of delay greater than 0

This formula calculates predicted waiting times (ASA) based on the number of reps, the number of customers waiting to be served and the average amount of time it takes to serve each customer; it can also predict the resources required to keep waiting times within targeted limits, and that is why it is useful for staffing.

But, as with any mathematical formula, Erlang C has built-in assumptions that do not perfectly reflect real circumstances. One problem is that by assuming that all incoming calls are anyway staying in queue, it means that customers will wait until they get an answer, i.e. they will never abandon. Moreover, as said, it assumes that there is an infinite trunking and system capacity, thus that nobody will ever get a busy signal. The result is that Erlang C may *overestimate* the staff really needed. Erlang C is not of exclusive use in the telecommunications world; it can be used to determine resources in any situation where people might wait in a queue for service, and thus there are tables to make it easier to use and more accessible than the formula itself. Of course there are also computer programs.

In the following example, we show Erlang C results based on the use of an Erlang C program provided by ICMI (AA.VV. 2000). It basically requires four variables as input: Average Talk Time (ATT), Average After Call Work (ACW), number of calls (the projected volume for the time unit - say, typically, half an hour - we are analyzing), Service Level objective in seconds (i.e., 90 calls in 20 seconds, the input 20). Once the numbers are fed as an input to the program, the output provides a wealth of information and insight into the dynamics of a CC.

TSRs	P(0) (%)	ASA	DLYDLY	Q1	Q2	SL (%)	OCC (%)	TKLD
30	83	209	252	29	35	24	97	54.0
31	65	75	115	10	16	45	94	35.4
32	51	38	74	5	10	61	91	30.2
33	39	21	55	3	8	73	88	28.0
34	29	13	43	2	6	82	86	26.8
35	22	8	36	1	5	88	83	26.1
36	16	5	31	1	4	92	81	25.7
37	11	3	27	0	4	95	79	25.4
38	8	2	24	0	3	97	77	25.3
39	6	1	21	0	3	98	75	25.2
40	4	1	19	0	3	99	73	25.1
41	3	1	18	0	2	99	71	25.1
42	2	0	16	0	2	100	69	25.0

Table 1: Erlang C for Incoming Call Centers.

In Table 1 we reported values for the following factors and variables:

TSRs: number of reps required on the phone
P(0): probability of delay greater than 0 secs
ASA: Avg. delay of "all" calls
SL: X% of calls answered in Y seconds
DLYDLY: Avg. delay of delayed calls
Q1: Avg. number of calls in queue at any time, even when there is no queue
Q2: Avg. number of calls in queue when all reps are busy or when there is a queue
OCC: Percentage of agent occupancy: the pct. of time agents will be spending while handling calls
TKLD: hours (erlangs) of trunk traffic. The actual traffic carried by trunks in a half-hour will be, in each row, half of what is given

Note that in our application of Erlang C, we assumed the following parameters, which are typical values in a CC environment:

Average Talk Time in seconds:	180
Calls per half hour:	250
Average after call work in seconds:	30
Service Level in seconds:	20

The first interesting column to focus our attention on in Table 1, SL (Service Level), represents the percentage of calls to be answered in a given number of seconds. If, for example, your objective is 80/20 (which means 80 percent of calls answered in 20 seconds), keeping going down the column, we pass from 73% to 82%. Since the program is calculating staff required, some rounding is involved (people are obviously integer numbers), and since 82% meets our 80/20 standard, then that is the row we will concentrate on. Each column provides insight and information into the chosen service level.

With a concluding remark on analytic methods, we can say that Erlang C is fairly accurate for good service levels, while for bad ones it cannot truly show how bad they really are. As pointed out in the previous paragraphs, it has however some disadvantages: for instance, it assumes no abandoned calls or busy signals as well as "steady state" arrival (traffic does not increase or decrease beyond random fluctuation within the considered time period). It also assumes a fixed number of staff handling calls throughout the time period and that all agents within a group can handle the calls presented to the group, thus neglecting peaked traffic or the need for skill-based routing, different groups of agents or complex network interflow. There is where computer simulation enters the game.

3.3 The impact of customer abandonment: a simulation with Arena

Arena Call Center can be used in conjunction with standard Arena constructs (a widely-used, general purpose simulation tool) to generate models of specific CC architectures. Building a model entails developing flowchart-style scripts that depict the current and proposed call routing process. For that purpose we use the process illustrated at a higher level in previous sections. The model generates streams of arriving calls that are held by the Call Center. As soon as they enter the CC, calls are assigned to a trunk line and routed through the center to agents who will eventually serve them.

There are other complicating factors that are built into the model. Calls can be blocked if trunks are busy, or if a certain limit, based on the ratio of calls in progress to the number of agents available, is exceeded. Once a call is blocked, it may contact back, depending on a specified distribution. Moreover, abandons may occur when the caller terminates the contact before reaching an agent. For each call, abandonment is modeled by a distribution for the amount of time a customer will wait (ASA) prior to abandoning the CC. For each call, a value is generated from this distribution to determine how long a customer will wait before abandoning, if not yet connected with an agent. Once a call abandons the CC, it may as well contact back.

Data were collected from a subset of 30 runs per each value of the main leverage parameter, with a "warm up" (transitory) period of 900 seconds. We show the results in the following table:

TSRs	ASA (sec)	Handled	Abandoned	SL (%)	OCC (%)
30	29	210	45	42	97
31	26	223	26	56	94
32	19	233	18	68	91
33	12	241	6	81	89
34	8	243	4	88	86
35	6	243	2	90	82
36	4	247	1	92	81

Table 2: Simulation Data obtained with Arena 7.0 ®

In Table 2 we reported values for the following factors and variables:

TSRs:	number of reps required on the phone
ASA:	Avg. delay of “all” calls
Handled:	Total calls handled
Abandoned:	Total calls abandoned
SL:	X% of calls answered in Y seconds
OCC:	Percentage of agent occupancy: the pct. of time agents will be spending while handling calls

Note that in our simulations, we assumed the following parameters:

Average Talk Time in seconds:	180
Calls per half hour:	250
Average after call work in seconds:	30
Service Level in seconds:	20
Simulation Time:	2700
Warm Up period:	900

Note how taking into account queue abandonment has allowed us to reach our target Service Level with only 33 agents instead of 34. Also ASA improves, even if occupancy gets a little bit higher, but this depends on one of the so called five “immutable laws” in incoming CC (see next chapter), that is: for a given service level, larger agent groups are more efficient than smaller groups. With 34 agents, we see how the situation dramatically improves. Whether choosing to have 33 or 34 agents could then seem obvious if agent’s burnout (depending on high occupancy) is not taken into account.

Burnout, errors and rework, and the effort for a qualitative service are some of the issues we will focus on in the next paragraph in order to extend our understanding not only of the processes acting in a CC but also of other existing relationships between physical processes and “soft” variables like agent stress or burnout and customer satisfaction.

3.4 A system dynamics approach: the impact of Quality and Service Level

Let us start from the five fundamental principles that govern a call center and characterize its dynamics (Cleveland-Mayben 1999):

1. For a given call load, when service level goes up, agent occupancy goes down (mostly because the call load is evenly distributed among them)
2. By keeping improving the Service Level, a point of diminishing returns will be reached (limit to the growth)
3. Given a SL, larger agent groups are more efficient than smaller ones (agents show a higher occupancy percentage)
4. All other things being equal, pooled groups are more efficient than specialized ones (handle more calls with same number of agents, same call load with fewer agents, same call load with same number of agents at a higher SL)
5. Given a call load, if staff is increased, then ASA will decrease and trunk load will go down

Moreover, we have found evidence, also in the literature (Busacca-Valdani 1999, Sterman et al. 1997), that the quality of the service offered depends both on factors connected with the effort and commitment of the management towards a qualitative service, as well as also with customer expectations and satisfaction (e.g. calls handled qualitatively by agents), and on variables which are typical of a call center structure, i.e. Service Level (SL), Average Handling Time (AHT), Average Speed of Answer (ASA), free line, and so on.

In particular, we can say that an effort of the management towards making agents handle calls in a qualitative way may be viewed as a tendency to skill agents so that they can meet customer expectations, thus improving the customer satisfaction index. In fact, reducing errors and rework has been seen to have a positive impact on service level, morale, customer satisfaction and, last but not least, costs. Typically, customers expect a CC to be accessible, to be treated courteously, to promptly do what they ask, not to deal with poorly trained agents, to be responsive to what they need and want, to do it right the first time, to be socially responsive and ethical, and so on. As said before, a call is defined as “qualitative” when: the customer is satisfied with the received service and of his/her experience in the CC, the TSR captures all needed/useful information, all data entry in the ACW phase is correctly done, the TSR provided correct response and the caller received clear and correct information in a time not perceived as too long, the caller does not get transferred around the CC or does not get rushed, the customer has confidence that his call was effective, the caller does not feel necessary to verify, check-up, repeat or even call back, the TSR is satisfied of how he just handled the call, the caller did not get any busies or was not placed on hold for too long, and, in the end, the CC mission is accomplished. In the end, the equation that puts in strict relationship quality effort and highly skilled or experienced agents seems not to be too far from reality.

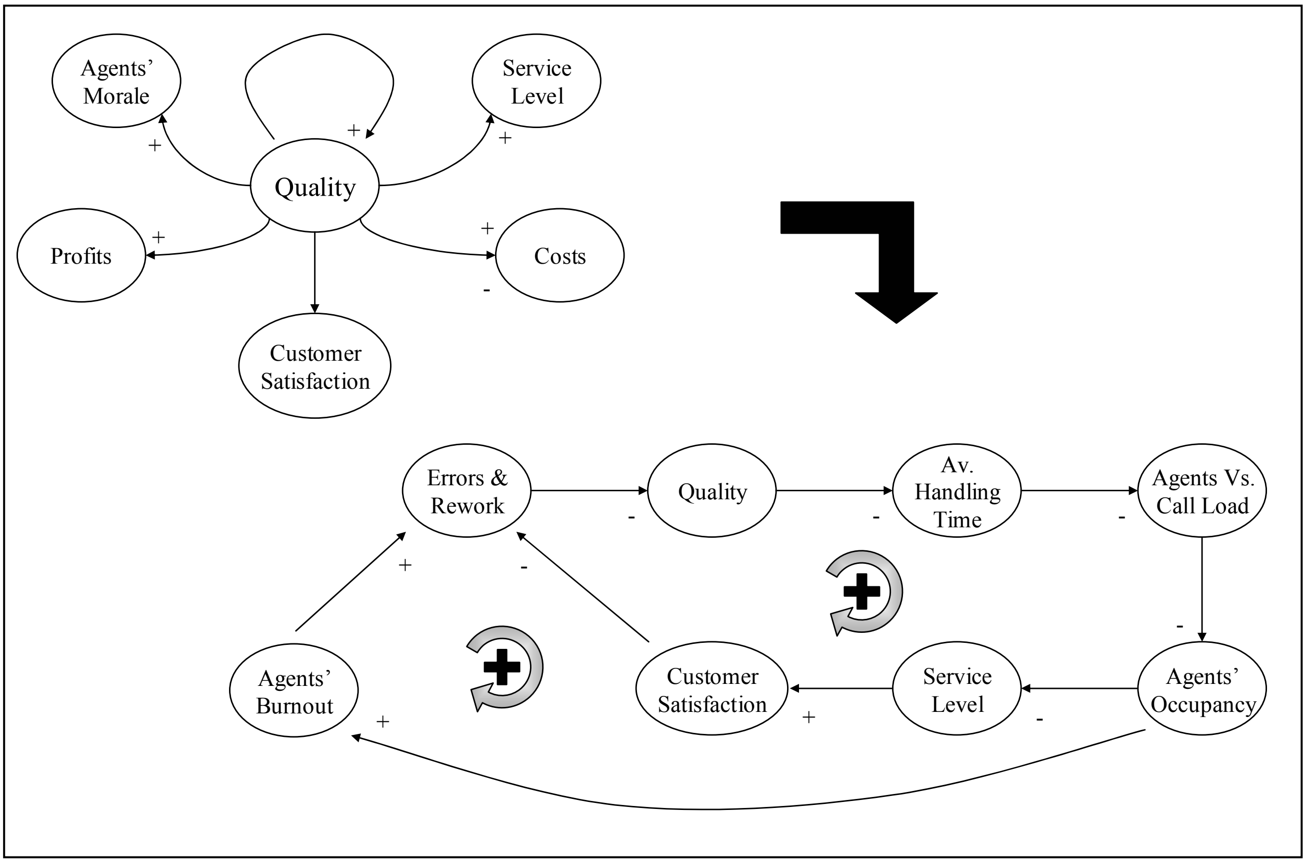


Figure 2: Positive (+) Feedback loop between Quality and Service Level.

We can thus draw a first qualitative causal map which consists of a main positive feedback loop between Quality and Service Level. This means that if SL is improved (because of a better quality in service), OCC generally decreases and then also staff burnout goes down. Thus, agents do not need to frequently take breather, have a higher schedule adherence and can then manage call handling more qualitatively (fewer errors). On the other hand, when SL deteriorates, also Quality (perceived and effective) gets worse, thus starting a dangerous degenerative snowball effect, which may put the performance of the entire structure at stake.

In the first part of Figure 2, we want to show how a quality service positively influences both customer satisfaction and agent satisfaction (pride in their job), thus acting also positively on service level. This of course also brings to the result of increasing profits for the company, since increasing customer satisfaction means also increasing customer fidelity and, in the long run, revenues, thus having a positive return on higher investments in quality.

In particular, when quality goes down, customer complaints drive up average talk time, thus disrupting AHT and increasing network costs. Moreover, as AHT goes up, the number of reps becomes insufficient to handle the call load at a desired Service Level. This, in turn, increases agent OCC, deteriorates SL and lowers customer satisfaction, thus increasing rework, which negatively act back on quality, lowering it further and so on. However, it must be noted that as AHT grows, also average delay (ASA) goes up. This causes abandons to grow larger and the queue to shorten. This effect acts in a balancing way (negative loop) on the entire structure of the system (thus showing the presence of a "Limits to the Growth" archetype).

In light of what has just been explained, and developing further the causal loop devised in Figure 2, we can finally draw the causal loop diagram of Figure 3, where:

- **Call Load**: this one has been considered as an exogenous variable since we're modeling our system on a very short time window. Thus, it would be meaningless to model the call load as an endogenous variable (a sort of beginning population to serve). Its value is based on real forecasts.
- **Staff Size** and **Effort in Quality**: decision leverages. We have already said about the need to consider, at this stage, the latter as a policy leverage. As long as instead the Staff Size is concerned, once again, we have considered it as an exogenous variable since the simulation runs over such a short amount of time which would make it almost impossible to appreciate any change in the level of staff (if we simulated the performance of a call center over an entire day, it would then make sense to analyze, for example, schedule adherence. In that case, the staff size should be modeled as endogenous).
- **B1**: Balancing Loop.
- **R1**, **R2**, **R3**: Reinforcing Loops.

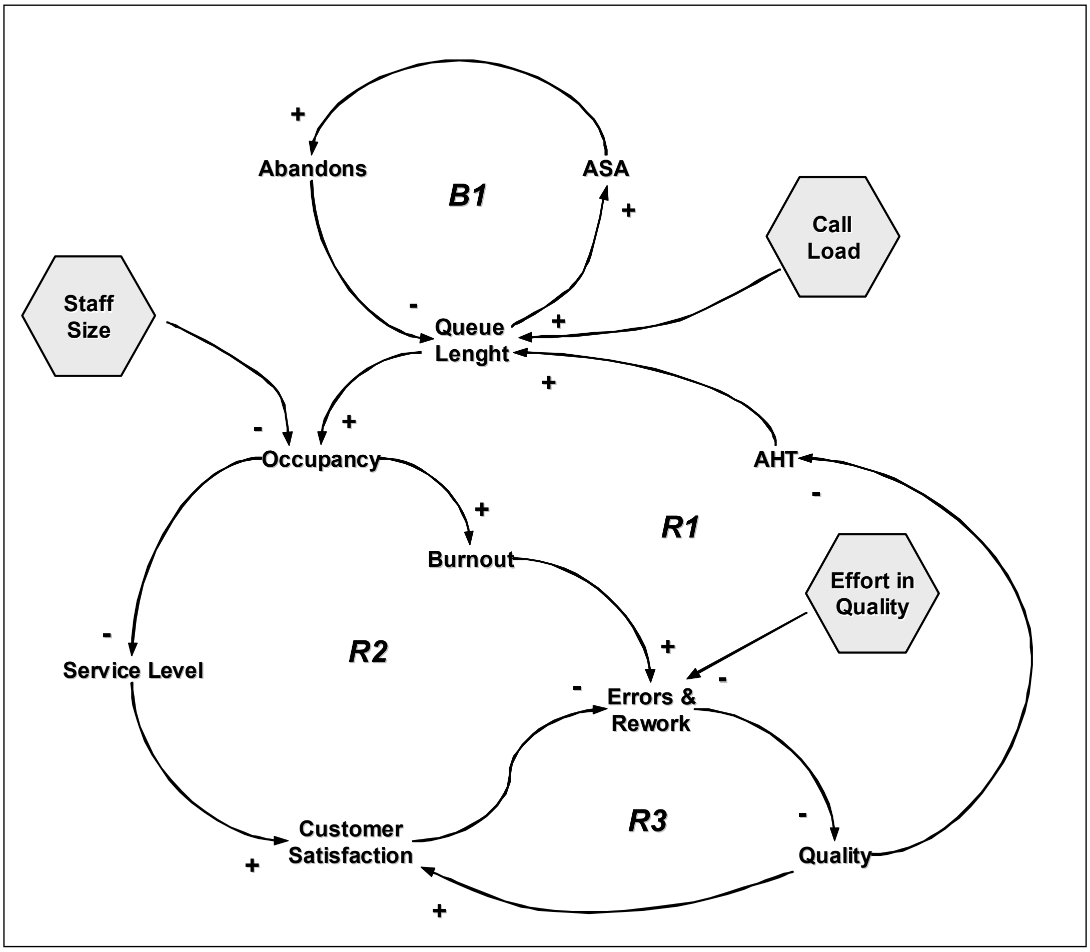


Figure 3: Causal Loop Diagram of a Call Center system model.

Looking at the balancing loop (B1), we see that if *queue_length* increases, then *ASA* increases: in fact, a new customer arriving at the back of the queue would have to wait, on average, more than if the queue would have been shorter. This in turn drives up *Abandons* (in particular, a customer may abandon also in dependence of the customers tolerance factors, cited in Paragraph 3.1, but these are not depicted in the CLD of Figure 3) which of course tend to reduce the length of the queue.

The first positive loop (R1) shows instead how *queue_length* drives up the agents *occupancy*, then causing them stress. This in turn causes more *errors_and_rework*, thus negatively influencing the *quality* of the service offered. If quality decreases and there is not any effort in order to prevent quality deterioration, then the necessary time to handle a call (*AHT*) becomes longer: this is due because of operators stress. Of course, if *AHT* goes up, then on average an operator sets free in more time. This means that, on average, fewer customers are drawn from the queue into service: so *queue_length* increases (even with the assumption that the number of working agents remains constant over the whole simulation runtime).

The second reinforcing loop (R2) mostly consists of the same elements of R1, but focuses on the influence of *Service_Level* (SL) over Quality (meant as the quality of service). In fact, if *queue_length* decreases, this in general drives up SL. With a certain delay, a good SL is known to have a positive effect on the opinion that customers have of the organization, thus driving up their satisfaction. A someway satisfied customer, tends to believe more favorably that the operator he was on the phone with has understood his/her needs, and then does not make him waste time on useless conversations. This can be seen as to drive down the *errors_and_rework* factor, thus allowing the CC to offer a better service.

The third snowball effect loop (R3) takes into account the immediate effect that a qualitative call may have on the conversation and thus on customer satisfaction. In our model, the quality effort tends to drive down errors and rework, thus driving up the overall quality performance and in the long run, customer satisfaction.

Note that the quality of service may be defined according to several indicators that identify how the call has been managed all through the process. As said before, there are several aspect that may help define a call as “qualitative”. For each of these aspects of quality, a custom key indicator may be developed or identified. In this paper we want to show how acting on a high leverage point as pushing on the quality of service, may help to improve a CC performance. We will then just use such a parameter as an aggregate of key indicators which may for sure be developed in many ways.

3.5 The modeling process

The model has been implemented with Ithink 7.0® (Demo Version) and has been run on an Intel® PC Processor Pentium IV, 2.0 GHz. The modeling process has taken into account the real process that basically consists of two main flows: a customer flow and an agents flow.
The customer flow has been modeled by taking into account four main states for the incoming calls: a call can, in fact, be: in the queue state (Queue Level), or being served (in the Service Center Level) and then either abandon (Abandoned Level) or being handled (Handled Level). This is shown in detail in Figure 4: the incoming traffic can enter the queue only if the capacity of the system allows for it, and lost calls due to *busies* are calculated by taking into account the difference between the whole amount of arrived calls and the sum of calls in queue, in service and handled and abandoned calls. Note that the balancing loop B1 has been modeled here with the following variables: Queue (a level), Queue length, ASA (the delay experienced by customers), Leakage Fraction (which basically accounts for the people who abandon) and the outflow driven by Abandons. Customers, who are not satisfied with the service received, have been modeled as an inflow back from the service center level into the queue.

The agents co-flow has been modeled by considering the two possible states for any operator, which can be either “Free” or “Working”. As soon as there is free agent, he picks up a call as long as there is one waiting in the queue, thus flowing into the *Working_Agents* level, modeled here as a conveyor. In fact, it takes a working agents a time which is equal to the stochastic variable *Handling_Time* (sum of two exponential distributions: *Talk_Time* and *After_Call_Work*) in order to set free again and flow back into the *Free_Agents* level (Figure 5).

In Figure 6 we can see how the constant *Quality_Effort* represents the policy of a qualitative service, then directly influencing the amount of “errors” during a conversation. This in turn has an effect on the level of reworked customers, thus driving the overall quality of the service (*Quality_Factor*) and hence customer satisfaction.
Service Level (Figure 7) has been defined as the fraction (percentage) of handled calls that have been answered into the service level objective (that is, in our case, 20 seconds). We have the following equation:

Service Level = Calls answered in SL obj. / (Calls Answered + Calls Abandoned)

Note that with Customer Satisfaction (CS) we mean here the ability to delight customers; it is in some way also connected to how customers perceive the quality of the offered service. We have modeled it as dependant on two main key factors: Service Level and Quality Factor (representing the percentage of correctly handled calls with respect to the overall call load experienced by the CC in the considered time window). Since both variables may assume values in the range from 0 to 1 (percentage factors), also Customer Satisfaction, being modeled as their product, will assume values in the same range.

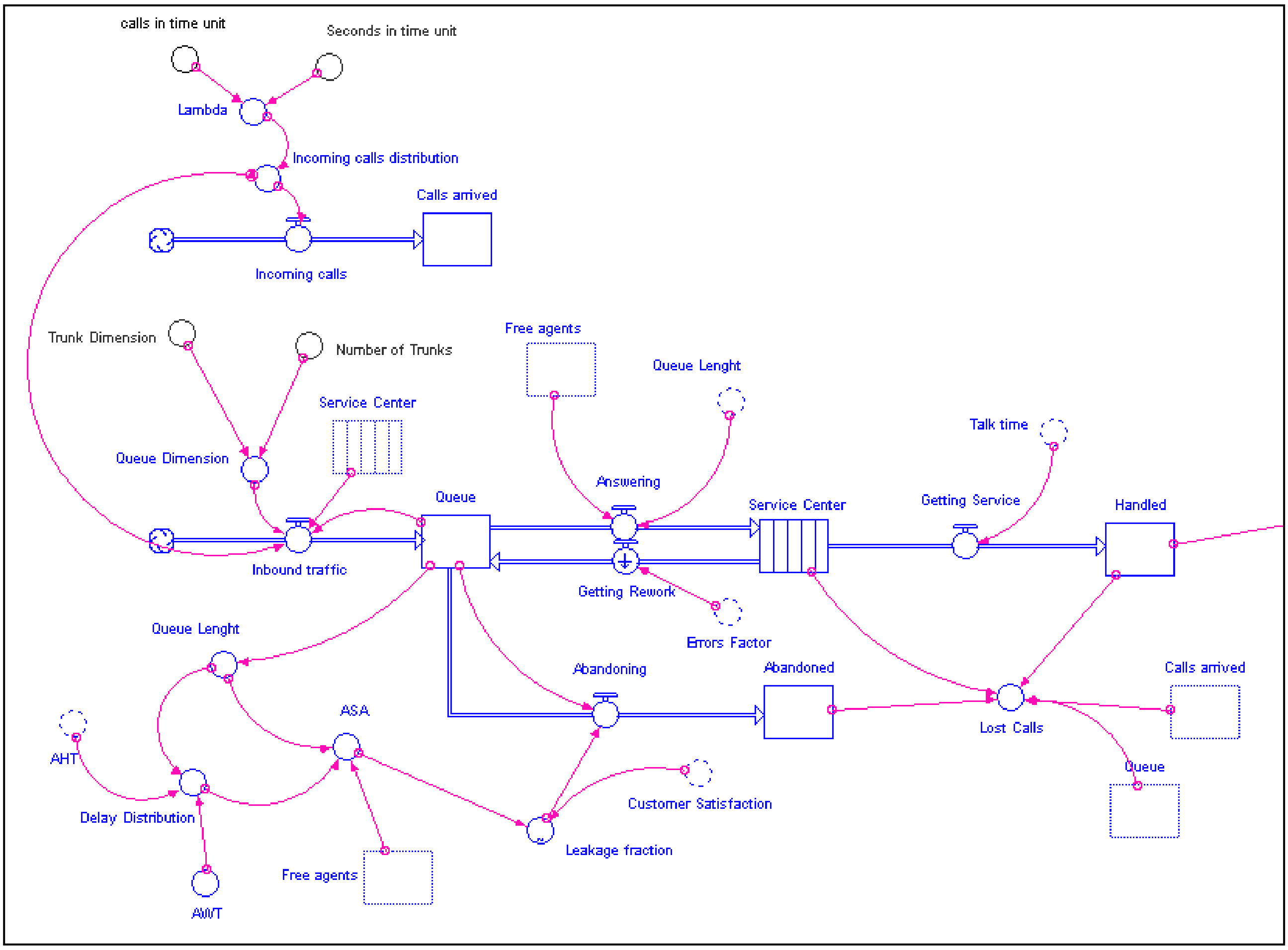


Figure 4: The Customers co-flow.

It is necessary to take into account also abandoned calls at the denominator because it is then possible to have a better idea of the accessibility of the call center and because an increase in abandonments gives a key measure also of the deterioration of the overall service level.

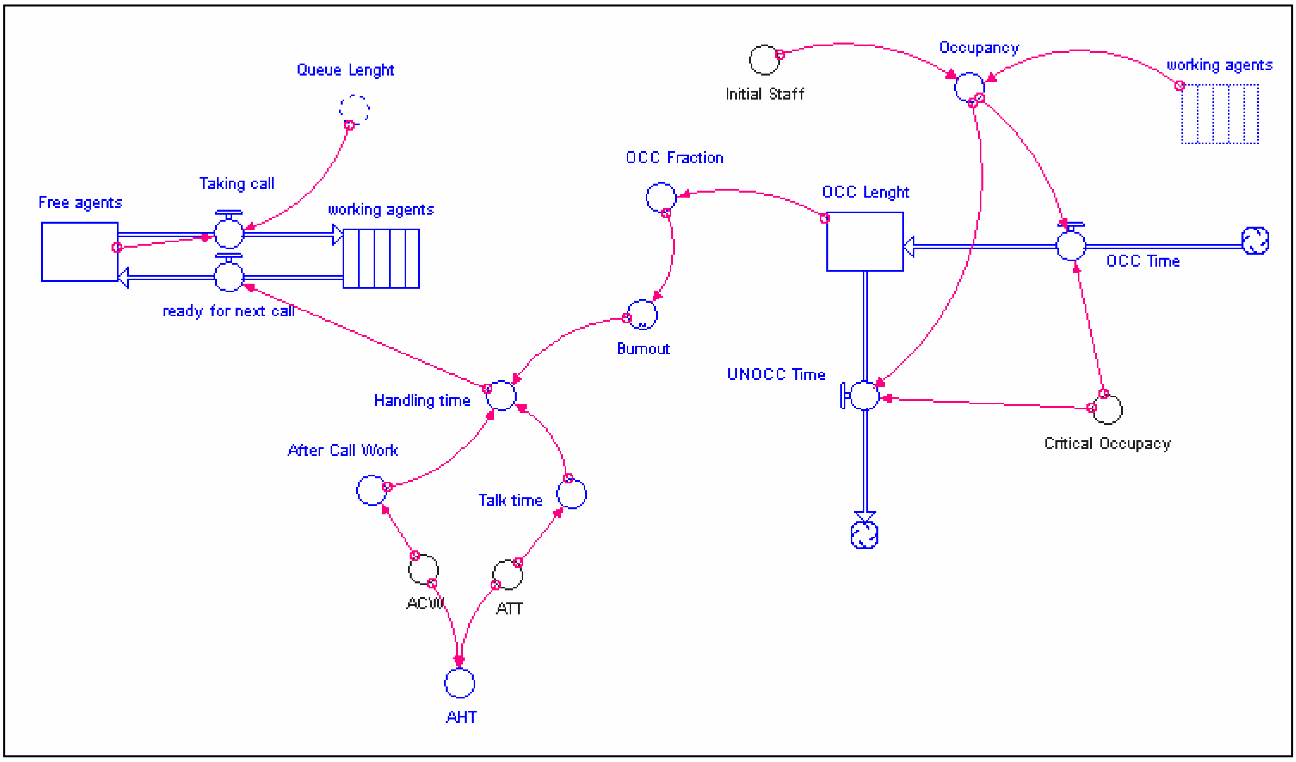
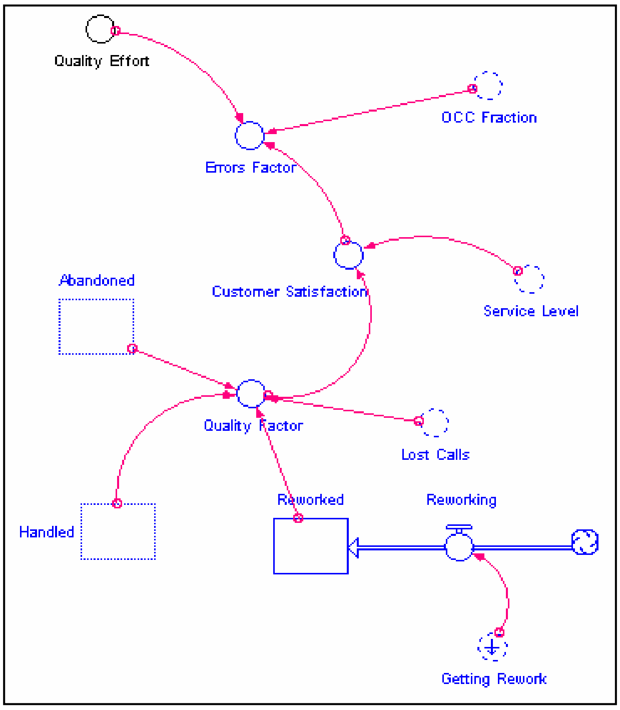


Figure 5: The Agents co-flow.



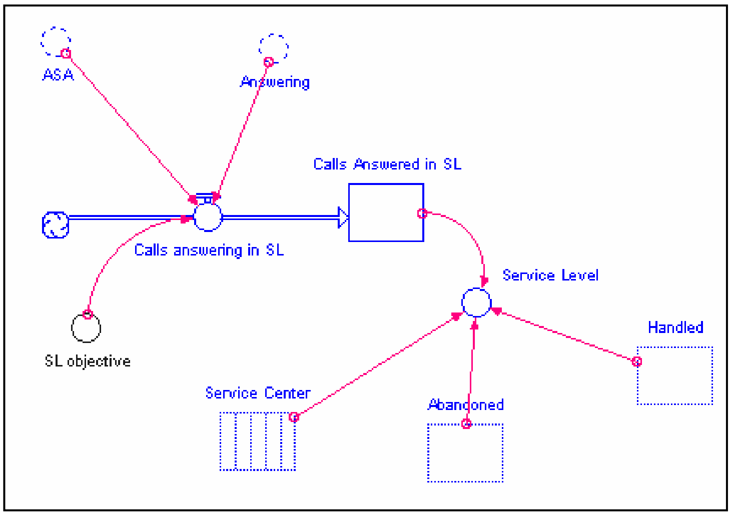


Figure 7: Service Level

Figure 6: Effect of Quality

Getting into details, the variable *Leakage_Fraction* is a graphical auxiliary that returns the probability that a customer will abandon the queue. It has been modeled as a function of the fraction

ASA/Customer_Satisfaction and the graphical function was obtained by real data analysis and by evaluating how the percentage of abandons was varying according to ASA and CS. As ASA remains constant, it is possible to appreciate the influence of a worse CS on Abandonments. As said, the graphical function has been sketched by referring to some real CC data: by analyzing it, we noticed a correlation between ASA and abandoned calls, relatively to the total of incoming calls. Interpolation of point data has the allowed us to sketch the above function. In order to model ASA, we first noticed how such a factor is highly variable, even for a constant number of customers in the waiting line: that is why it has been modeled with a normal distribution. In order to evaluate the σ, we observed a correlation between expected values of waiting times and queue length. This allowed us to calculate the expected value of the waiting time for a call in the queue which has undergone an *Average_Handling_Time* (AHT) of just one second: that is a percentage factor estimating the average waiting time in the queue (AWT). In this way, the average value of ASA comes from the product *AWT*AHT*Queue_Lenght*. Moreover, the Gaussian distribution has a variance set so that 99.7% of calls have an expected ASA belonging to the following interval:

$$[AWT*AHT*(Queue\_Lenght-1); AWT*AHT*(Queue\_Lenght+1)]$$

We also want to put here in evidence some other basic assumptions and simplifications adopted in the modeling process. First, we have assumed that customers not satisfied with the service received, directly go back into the queue, instead of being considered as a factor to be added to the incoming traffic. This is a worst case situation in which all unsatisfied customers are able to get into the queue. On the other hand we have completely neglected the situation for which those customer getting a busy signal or those abandoning the queue may decide to call back. For good service levels, this did not seem to alter considerably the obtained results (the percentage of reworked calls was quite low), but the risk could be to overestimate the number of agents. Second, we have not directly taken into account the seven customer tolerance factors in a separated way, but only considered the most important ones for simplification purposes. Third, we have modeled the effect of service quality on customer satisfaction as instantaneous, and not with a delay that could have cut it out of the simulation time span. Though this approach may not seem very systemic, our intent was to model the effect that a good offered service has on the customer calling and the immediate feedback that the latter may put in the dialoguing process (i.e., as soon as the customer feels that the service is not that good, he immediately start asking for more information or clarifications that make the agent loose time and stress more). Fourth, we have neglected the effect of trunk load on system capacity. An improvement of the model could also take well into account an optimization of such technological resources (number of trunks, trunk load, and so on). Fifth, we haven't optimized the model. At this introductory level of this subject, we just wanted to show some basic behaviors of a CC environment.

3.6 Simulations of the model and results

We have run two different sets of simulation according to the following policies:

1) Fixed Effort in Quality and variation of the leverage connected to the Initial Staff;
2) Fixing the Initial Staff value found in the first set of simulations, according to which the objective Service Level has been met, we show how ASA improves by varying the effort in Quality;

In both cases, basic data have been chosen in order to compare results with the previous method. Data were collected from a subset of 30 runs per each value of the main leverage parameter, with a "warm up" (transitory) period of 900 seconds. We show the results in the following tables:

TSRs	ASA (sec)	Handled	Abandoned	Reworked	SL (%)	OCC (%)	Avg. Queue
30	27	210	48	7	53	95	4
31	23	223	29	5	62	93	3
32	18	233	17	4	70	91	2
33	11	241	5	2	81	89	2
34	8	243	4	2	87	85	1
35	6	243	3	1	90	82	1
36	4	247	1	1	93	81	1

Table 3: Simulation Data obtained with IThink 7.0 ® .- SET (1): Fixed Quality (0,85) – Variable Agents (TSRs)

In Table 3 and Table 4 we reported values for the following factors and variables:

TSRs:	number of reps required on the phone
ASA:	Avg. delay of "all" calls
Handled:	Total calls handled
Abandoned:	Total calls abandoned
Reworked:	Total of calls reworked
Avg. Queue:	N° of calls which on average form the Queue
SL:	X% of calls answered in Y seconds
Qty Effort:	effort in order to provide a qualitative service
OCC:	Percentage of agent occupancy: the pct. of time agents will be spending while handling calls

Note that in our simulations, we assumed the following parameters:

Average Talk Time in seconds:	180
Calls per half hour:	250
Average after call work in seconds:	30
Service Level in seconds:	20
Simulation Time:	2700
Warm Up period:	900

Comparing results shown in Table 3 with those in Table 1, we can see how the situation sensibly improves when the initial staff is composed of 34 agents. SL is now 87% (compared to 82% of the Erlang's method), OCC goes down to 85% (it was 86%) and most of all, ASA falls to 8 seconds against a previous value of 13. Even more interesting, if we just wanted to meet our service level objective of 80/20, we will only need 33 reps instead of 34, still with a better ASA and with an occupancy slightly higher (89% versus 86%).

Comparing results shown in Table 3 with those in Table 2, we can see how, even if not in an extraordinary way, the situation also improves when the initial staff is instead composed of 33 agents. We can see that, if both ASA and Abandons improve marginally, we also have here some reworked calls which brings the total "served" calls up to 243 instead 241. In general, however, the situation improves.

For the second set of simulations (Table 4), we fixed the value of 34 for the group of agents and simulated the model with different values of the Quality Effort parameter. It is interesting to observe how a deterioration of the service quality may bring to a deterioration of ASA due to a higher agents Occupancy: this, in turn, does not allow us to meet our SL objective. Notice also how the level of abandons and reworks increases, especially as we fall under the 0.5 critical quality threshold.

Furthermore, the results returned by each simulation show how the higher values of Abandons and Reworked calls, as well as the ones of Service Level and Occupancy, are a consequence of the reinforcing feedback loops connected with a low quality of the service that bring the system to uncontrollable behaviors; more in detail, during our simulations the number of runs that showed such an atypical behavior is higher for low levels of Quality effort (9 out of 30 runs for a Qty Effort level of 0.35) and lower for high quality (1 out of 30 runs for a Qty Effort of 0.85, and no one at all in our 30 runs for a quality of 0.95).

Qty Effort	ASA (sec)	Handled	Abandoned	Reworked	SL (%)	OCC (%)	Avg. Queue
0,35	13	235	21	14	78	88	2
0,5	12	235	15	7	82	87	2
0,65	9	240	9	5	84	86	1
0,75	9	241	6	3	85	85	1
0,85	8	243	4	2	86,9	85	1
0,95	7	247	3	1	86,9	85	1

Table 4: Simulation Data obtained with IThink 7.0 ® .- SET (2): Fixed N°of Agents (34) – Variable Quality

3.7 Conclusions and future work

In this paper we have shown how simulation may considerably help in the analysis and evaluation of resources in a service system. In particular, we have seen the case of one of the most classic service systems: the call center. We have applied different techniques to a very simple and basic CC model, representing the main processes and aspects that drive its basic behaviors and dynamics.
We have also seen the typical limitations of a classical queuing theory approach and shown how simulation may help overcome such limitations (though maybe introducing other and new ones). The System Dynamics approach has allowed us to further improve our understanding of the system's behaviors as well as to appreciate the impact that a qualitative service may have on Service Level and ASA (and on the management of resources).
Some of the simplifications described in Paragraph 3.4 are to be removed in future works, and other aspects will be further developed. In particular, we plan to distinguish between a strategic level and a tactical level of the model. According to the tactical point of view, we are looking forward to extend our model by taking into account aspects like skill-based routing, call transfers, agent groups, impacts of a VRU installation and the extension of the simulation period to one day. At a strategic level, we want to explore aspects connected to the flow of human resources in the call center, analyzing how their learning curve may influence the Average Handling Time and how a qualitative service, described in terms of those quality effort indicators which we haven't specified at this stage, may in the long run influence both customer satisfaction and the economic result of the organization which the call center refers to (cash flow, profits, return on investments, etc…). We will also try to investigate issues related to the trade off between optimization of resources and costs (i.e., trade off between agent costs or trunk load costs) as well as the simulation of Graph Partitioning algorithms concerning the problem of skill-based agent routing.

4. References

Cleveland Brad, Mayben Julia, *Call Center Management on Fast Forward* (1997-1999), Call Center Press - A division of ICMI, Annapolis, Maryland

AA.VV., *Call Center Forecasting and Scheduling* (2000), Call Center Press - A division of ICMI, Annapolis, Maryland

Anton Jon, Bapat Vivek, Hall Bill, *Call Center Performance Enhancement – Using Simulation and Modeling* (1999), Ichor Business Books – Purdue University Press, West Lafayette, Indiana

Bianchi C., *Modelli contabili e modelli dinamici per il controllo di gestione in un'ottica strategica* (1996), Giuffré Editore SpA, Milano

________, Bivona E., Landriscina F., *Promoting entrepreneurship through open-distance-learning management flight simulators: ecoroll educational package.* (2000) International System Dynamics Conference Proceedings, System Dynamics Society, Bergen

Busacca B., Valdani E., *Customer Based View* (1999), Finanza, Marketing e Produzione, Anno XVII

Forrester, J.W., Industrial Dynamics. (1961), Productivity Press.

___________ , Principles of Systems. (1968), Productivity Press.

___________ , Urban Dynamics. (1969), Productivity Press.

Gonzales-Busto B., Garcia R., *Waiting lists in Spanish public hospitals: a system dynamics approach.* (1999) System Dynamics Review 15(3), 201-224

Donald Gross, Carl M. Harris, *Fundamentals of queueing theory*. (1998), New York, Chichester, John Wiley & Sons.

Fung K.K., *It is not how long it is, but how you make it long – Waiting lines in a multi-step service process*, (2001) System Dynamics Review 17(3), 333-340

________ , *Follow the laggard? – not all bottlenecks are created equal*, (1999) System Dynamics Review 15(4), 403-410

Hillier, Lieberman, *Introduction to Operations Research*. (1990), McGraw Hill

Homer J., *Macro and micro-modeling of field service dynamics.* (1999) System Dynamics Review 15(2), 139-162

High Performance Systems Inc., *Ithink Software, Version 7.0 Demo* (2000), Emergency Room Dynamics model

Jennings O.B., Mandelbaum A., Massey W.A., Whitt W., *Server staffing to meet time-varying demand*. (1995), AT&T Bell Laboratories

Kleinrock, *Queuing Systems*. (1975), John Wiley & Sons

Milling, P. M., Quality Management in a dynamic environment. *The Cybernetics of Complex Systems – Self Organizations, Evolution and Social Change*. Geyer F (ed.). InterSystems Publications: Salinas, CA, 125-136

Morecroft J., Sterman J., Modeling for Learning Organizations. (1994) Productivity Press, System Dynamics Series.

Naor P., *The regulation of Queue Size by Levying Tolls.* (1969), Econometrica 37(1), 15-24

Oliva R., *A dynamic theory of service Delivery: implications for managing Service Quality.* PhD Thesis, Sloan School of Management, MIT

Read B. Brendan, *Designing the Best Call Center for your Business* (2000), CMP Books, NY

Roberts, E. B. (ed.), Managerial Applications of System Dynamics. (1978) Productivity Press.

Roberts N., Andersen D.F., Deal R.M., Garet M.S., Shaffer W.A., *Introduction to Computer Simulation: The System Dynamics Approach*. (1983) Addison-Wesley.

Ruth, M. Y Hannon, B. Modeling Dynamic Economic Systems. (1997) Springer Verlag.

Senge, Peter M. *The Fifth Discipline: The Art and Practice of the Learning Organization*. (1990) Doubleday/Currency.

Sterman J.D., Business Dynamics. Systems Thinking and Modeling for a Complex World. (2000) Mc Graw Hill.

_________, Repenning N., Kofman F., *Unanticipated side effects of successful quality programs: exploring the paradox of organization improving.* (1997) Management Science 43(4), 503-521

Van Ackere A., Smith P.C., *Towards a macro model of National Health Service waiting lists.* (1999) System Dynamics Review 15(3), 225-252

Vennix J. A. M., Group Model Building. Facilitating team Learning Using System Dynamics. (1996) John Willey & Sons.

Ward Whitt, *Dynamic Staffing in a telephone call center aiming to immediately answer all calls*. (1999) Operation Research Letters 24, 205-212