Exploiting Simulation for Call Centre Optimization

Salman Akhtar and Muhammad Latif

Abstract—**The global trend in developed economies from manufacturing towards services has led to an explosion in the call centre industry. With constant advances in the enabling technologies allied to changing business strategies, call centre management has become a critical business success area. Call centres can be thought of as stochastic systems with multiple queues and multiple customer types, resulting in great challenges associated with managing these systems. It is very difficult to understand the dynamics of call centres using purely analytical techniques due to the operational and mathematical complexities involved. This paper introduces and gives a detailed overview of call centre functions and their operations as the basis to promote the use of Discrete-Event Simulation (DES) for modelling purposes. The effects of calls routing and prioritizing to specific agents with multiple skills in an inbound call centre are discussed and modelled using the Witness simulation software.**

Keywords—**Call centres, Call routing, Discrete-Event Simulation, Tele-Queues.**

I. INTRODUCTION

Call centre managers are increasingly expected to deliver both low operating costs and high quality of service. To meet these potentially conflicting objectives, call centre managers are challenged with deploying the right number of agents with the right skills to the right schedules in order to meet an uncertain, time-varying demand for service [1]. Despite many analytical approaches for modelling call centres, the gap between these models and the call centre's reality is still quite large. These analytical approaches cannot be accurate enough, as they do not mimic randomness. Therefore, DES appears to be the most viable option for accurate performance modelling and subsequent decision support. The demand for decision support models is continuously increasing, due to the increased complexities in call traffic management and the use of multiple channels in call centre operations. Although many researchers have already explored the use of DES in call centre environment, they have not directly addressed the issues of effective routing policies that often incorporate priority rules for the calls and agents. Specifically, the number and types of agents, who handle the calls and the working schedules of these agents under constraints on the quality of service and on admissible schedules is one of the main optimization problems encountered in managing these multi-skill call centres [2]. Without proper DES models, it is very difficult for the managers to deal with such problems and to explore 'what-if' scenarios on a daily basis. In absence of such models, managers cannot visualize the consequences of different process changes before they are implemented.

Salman Akhtar is the researcher in Department of Engineering and Technology, Manchester Metropolitan University United Kingdom, email: salmanakh@hotmail.co.uk and Muhammad Latif is the academic at the same university, email: M.Latif@mmu.ac.uk

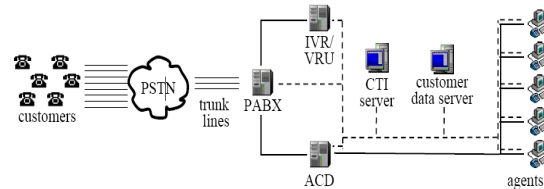## II. OVERVIEW OF INBOUND CALL CENTRE OPERATIONS

Figure 1: Schematic diagram of a call centre [3]

An inbound call centre constitutes of a set of resources known as agents, computers and telecommunication equipment, etc. which enable the customers to get the services required. To improve customer relationships and customer services, inbound call centres can be physically housed across several different locations, time zones, and countries and they can be managed by distributing the load of incoming calls through proper routing and prioritizing strategies. Many organizations use 'Interactive Voice Responses' (IVRs) or 'Voice Response Units' (VRUs) in their inbound call centres operations in addition of providing services through agents. IVRs help customers to resolve their queries and to 'self-serve' without wasting any time in subsequent queuing delays and they can leave the system successfully. Research has proven that up to 80% of customers can serve themselves using IVRs typically in banking industry [3]. Interestingly, the process by which customers who wish to speak to an agent identify themselves using IVRs, can average 30 sec(s) even though subsequent queuing delays often reach no more than a few seconds. Consider schematic diagram of call centre technology in fig 1. When customers call 0845 and 0800 numbers in the United Kingdom, the 'Public Service Telephone Network' (PSTN), sometimes called 'Long-Distance' Network connects the call to a privately owned switch of an organization known as 'Private Automatic Branch Exchange' (PABX) through available 'Trunk-Lines' with the help of two vital pieces of information about each caller: the number from which the caller originates the call, often called the 'Automatic Number Identification' (ANI) number and the number dialled by the caller, often known as 'Dialled Number Identification Service' (DNIS) number. If there is no 'trunk-line' which is free, the caller will receive a busy signal and the call will be blocked. When the customer calls an inbound call centre, various call handling and routing technologies attempt to route the call to an available agent with the help of the combination of ANI and DNIS numbers to the nearest of all call centres based on a customer location [4]. At first, the calls are connected to IVR through PABX, where customers provide information, such as 'Customer Reference Numbers' (CRNs) by using their telephone key pads or voices which helps them receiving the desired service. Infact, the latest generation of 'Speech-Recognition' technology allows IVRs to interpret complex user commands, so customers can enable

'self-service'. In response, the IVR unit uses a synthesized voice to report information, such as anticipated waiting time in queue or the call back option within the stipulated period of time, if available [5]. With continued interaction with IVRs, customers can complete their service needs without needing to speak with agents. In case, if customers do need to speak to agents, IVR unit hands over the calls to a specialized switch known as 'Automatic Call Distributor' (ACD) designed to route the calls to individual agents based on different criteria within the call centre and that operation accomplishes through PABX, which connects IVR to ACD. However, there are often no agents available to immediately answer the call, in which case the customer is typically put on hold by ACD and placed in a queue. Delayed customers can judge that the service they seek is not worth waiting and become impatient. In turn, they may abandon the queue or renege by hanging up, either immediately after being placed on hold or after waiting for some random amount of time without getting served. Customers that do not abandon are eventually connected to an agent. Once connected to an agent, a customer will speak with that agent for some random time after which either the call will be completed or the customer will be handed off to another agent or queue for further assistance. While speaking with the customer, the agents work via a 'Personal Computer' (PC) or 'Terminal' with a shared access to a centralized 'Corporate Information System'. They may spend some time on wrapping-up activities such as an updating of the customers' contact history or the processing of an order that the customer has requested [3]-[5]. To integrate telephony with information system, the middleware which can be used is called 'Computer Telephony Integration' (CTI). It helps identifying the callers ANI and uses it to search the customer database in the organizations' information system. If ANI belongs to an existing customer, CTI routes the customer to the most appropriate agent by retrieving the information from the customer data base and it appears on the screen in the form of a dialog-box, which is often called as 'Screen-Pop', that saves agent time and reduces the calls' duration. In more complicated settings, CTI can be used to integrate a special information system called a 'Customer Relationship Management' (CRM) System into call centres' operations.

CRM systems track customers' records and allow them to be used in operating decisions. With the help of CRM system, a 'screen-pop' can be sent to an agents' computer screen with the history of the customers' previous calls and if relevant, the details of past sales customer has generated. CRM System may suggest an agent 'cross-selling' or 'up-selling' opportunities, or it may be used to route the incoming call to an agent with special 'cross-selling' skills [6]. Finally, the fraction of callers who do not receive service or abandoned the queue may try to call again, and such calls become retrials. The calls that are blocked due to busy signals and the calls that are abandoned due to customer impatience are referred as lost calls by most of the authors. Callers who speak with agents, but are unable to resolve their queries at the first point of contact may also call again, and become returns. Sometimes, satisfactory service can also lead to returns for additional services that generate new revenues, and as such they may be regarded as good, or they may be in response to problems with the original service, in which case, they may be viewed as bad. Effectiveness of service goes parallel to the notion of rework. Among call centres in UK, a call without rework is sometimes referred to as 'One & Done' [7].

III. CONCEPTUAL MODEL OF A MULTI-SKILLED INBOUND CALL CENTRE

Conceptual model of a simple multi-skilled inbound call centre is illustrated in fig 2 that includes lost calls due to busy signals and abandonments. The conceptual model building is an activity in which the analyst tries to capture the essential features of the system that is being modelled. For example, if DES is being used then the aim will be to identify the main entities of the system and to understand the logical ways in which they interact. Call centres are one of the main areas, where DES can be used. The resulting model from this attempt of capturing the essentials of the system is known as a conceptual model [8].

The model depicted in fig 2, represents call centre operations in company X. Fig 2 illustrates four different call types A, B, C and D enter the call centre via IVR.
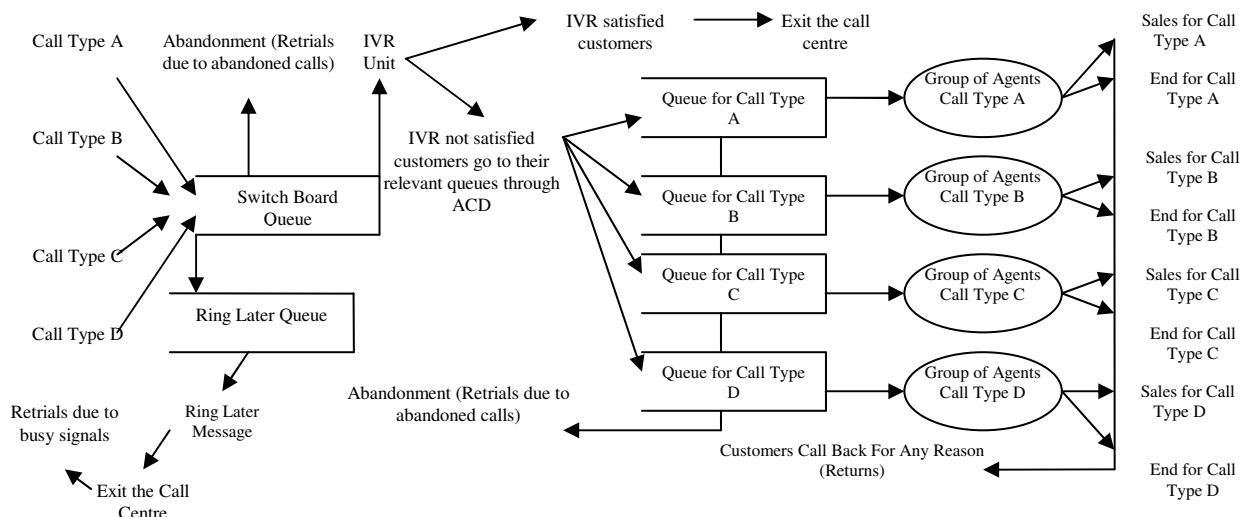

Figure 2: Conceptual model of a multi-skilled call centre (company X)

It is assumed for simplicity that the calls are only abandoned in 'switch-board' queue and queues for calls type A, B, C and D, after some random interval of time i.e., they are not being abandoned as soon as they join the queue. By monitoring the incoming calls over a period of one complete day, it has been determined that the 'inter-arrival' time on average follows a negative exponential distribution with details shown in table I.

Table I: Arrival data for calls in company X

Inter-arrival time (mins)	Type of Call	First Arrival of Call	Maximum Arrivals
Negexp(1)	type A	10 min(s) after opening	499
Negexp(2)	type B	2 min(s) after opening	649
Negexp(1.5)	type C	8 min(s) after opening	374
Negexp(1.2)	type D	5 min(s) after opening	399

After picking up the next call from the 'switch-board' queue, IVR decides which department to direct call to, this activity usually takes 30 sec(s), but could be as short as 10 sec(s) or as long as 1 min(s). The 'switch-board' can queue up to 90 calls in addition to the one being answered. Any calls that are made to the call centre that cannot be handled because the 'switch-board' queue is full, automatically receives a message informing 'to ring later' and is considered to be a lost call. Once the caller has been routed through to the relevant queue, the following occurs:

Group of agents call type A: Six agents are available to take the calls. It has been found that on average calls last between 4 and 6 min(s). There is a capacity for up to 27 calls to be queued (20% of such calls result in a sale).

Group of agents call type B: Eight agents are available to take the calls. It has been found that on average calls last between 3 and 5 min(s). There is a capacity for up to 30 calls to be queued (43% of such calls result in a sale).

Group of agents call type C: Five agents are available to take the calls. It has been found that on average calls last between 7 and 9 min(s). There is a capacity for up to 25 calls to be queued (13 % of such calls result in a sale).

Group of agents call type D: Five agents are available to take the calls. It has been found that on average calls last between 5 and 11 min(s). There is a capacity for up to 19 calls to be queued (15% of such calls result in a sale). Once a call has been processed the agent is free to accept the next awaiting call. The call centre is open from 08:00am to 8:00pm every week day and from 08:00am to 06:00pm on Saturdays.

IV. LOGIC FLOW DIAGRAM FOR A MULTI-SKILLED INBOUND CALL CENTRE

In a multi-skilled call centre, calls can be divided into different types, and can be served by agents partitioned into different groups on a 'First-In, First-Out' (FIFO) basis. Each agent group can be skilled to deal with different or all types of calls. When all agents are skilled in dealing with all call types, the call centre becomes efficient, as waiting times reduce and less abandonment, if it is assumed that the service time distributions for different call types do not depend on the agents' skill set. However, this assumption proofs to be wrong
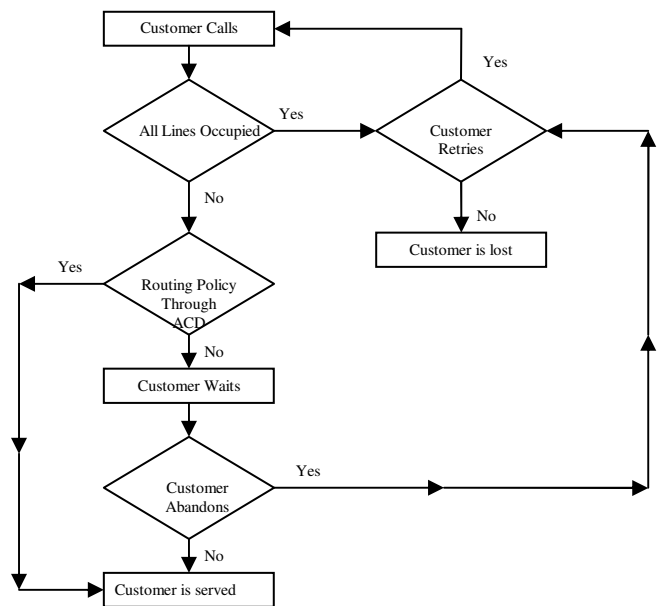

Figure 3: Flow diagram of customers call

in practice as agents work faster when they handle a smaller set of call types, even if their training gives more skills to them. Agents with more skills can be more expensive as their wages depend on their skill sets. Thus, for large volumes of different call types, a good practice is to dedicate number of single skilled agents (specialists) to handle most of the load of incoming calls. A small number of agents with two or more skills can cover the fluctuations in the proportion of calls of each type in the arriving load. The logic flow diagram is provided in fig 3 to explain the service process of a customer proposed by [9].

V. CALL ROUTING AND PRIORITIZING

ACD controls calls to agents or agents to calls assignment in a multi-skilled inbound call centre. This phenomenon is known as routing in a call centre industry. Calls to agents policies, prescribe actions at the arrival epochs of the calls, and agents to calls policies, prescribe actions at service completion. When a particular type of call arrives and there are two or more appropriately skilled free agents, then there has to be a decision, which agent the call should be routed. In a similar way, when an agent becomes free and one or more calls for which the agent has the required skill are waiting to be served, the agent has to choose which call to serve first. These selection problems are often dubbed as call selection and agent selection problems [10]. An arriving call type A can be assigned to the first group in its list through ACD having an available agent. If no agent is available from the groups in its list, the call should queue until the agent is available. Each call type has its own FIFO queue in fig 2, similarly each agent group has an ordered list of calls type to pick from the first non-empty queue. The call centre of fig 2 may need 20 'cross-trained' agents to deal with all call types at a certain time interval and to give better SL (Service-Level) targets, instead of 24 agents dealing with their own call types with separate parallel queues. The objective for call centre managers is twofold: fewer agents and good routing and prioritizing policies under SL constraints. It is vital to avoid unnecessary agents in order to minimize the operating cost of the call centre which is 60 to 70% of the overall budget and to

maximize the business profitability [9],[10]. Different scenarios are presented to explain realistic routing and prioritization problems in this multi-skilled call centre environment and they represent the 'building-blocks' for the DES model building process.
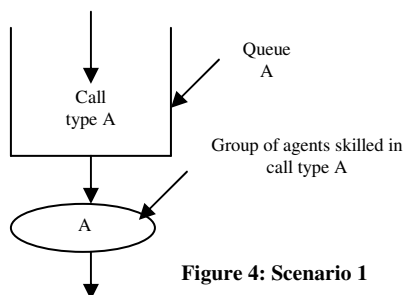
Scenario 1

Figure 4: Scenario 1

In fig 4, a single group of agents handle only one call type A on a FIFO basis, as there is no customer differentiation. Therefore, it is a good representation of single skilled inbound call centre as all the agents are equally skilled. In fig 2, different groups are dealing with different types of calls, so they can be considered as separate call centres. Suppose a type A call arrives randomly at different time epochs, and there are two or more free agents, the call will go to the agent with longest waiting time, if available to take the call. In case of a customer differentiation, serve as a FIFO queue, but VIPs enter the queue with a virtual 15 seconds wait i.e., as if they had joined the queue 15 seconds earlier. VIPs are the high revenue generating customers. An arriving customer that finds no free agent and no space in a queue is put on hold, and generates a cost to the organization. This also gives negative experience to the customers. Waiting customers have patience for waiting, which is modelled as stochastic variable and is called patience threshold. If the waiting time exceeds the patience threshold, the customer abandons the queue by hanging up the call. Callers who are blocked or abandoned the queue may redial, when the call centre is less busy and become retrials. Callers who speak with agents but are unable to resolve their problems may call again, in which case they become returns. Satisfactory service can also lead to returns. It is assumed that due to redials, the time epochs at which customers call are correlated to each other, which is in conjunction with the assumption that calls arrive according to a Poisson process. Every call which is not blocked or abandoned eventually determined by the routing policy and can exit the call centre successfully. It can occur that for some reason the call cannot be helped and is directed to another agent in a different queue. This can be an agent from the back-office team, depending on the difficulty of the callers' query. The dynamics of the inbound call centre is modelled as calls are served exactly by one agent. Thus, redirections of calls do not have to be explicitly considered. However, they can be modelled by adjusting the arrival rates of the incoming calls [11].

Scenario 2

In fig 5, two types of calls are both handled by a single group of agents and the agents are skilled to deal with both types of calls.
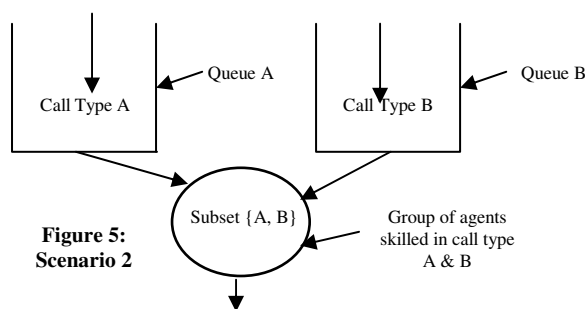

Figure 5: Scenario 2

For the agent and call selection problem, a rule must be specified for the order in which the agents will serve those two call types. Say call type A are VIP callers, therefore they are given higher priority than call type B callers. If calls of both call types are in a queue and an agent becomes free, then that agent selects a type A call to serve next. Such a group of agents is classed as homogeneous agents by [12] as they are equally skilled in both call types. To simplify this control problem, static priority policies are implemented in the model and VIP callers are assigned attributes as 1 so they can be served as soon as any agent becomes free at different time epochs. Each VIP caller has a patience threshold and the caller with highest patience threshold will be served first.
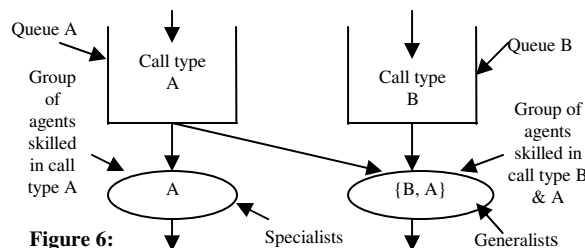
Scenario 3

Figure 6: Scenario 3

In fig 6, two types of calls are being served by two groups of agents. Group of agents skilled in call type A only deals with call type A as their primary skill and group of agents skilled in call type B and A serves both call types, B as primary skill and A as secondary. Such groups are classed as specialists and generalists by [13]. This design is preferable, when type A calls are VIPs, but there are not enough agents in specialists group to serve them, so group of generalists helps group of specialists by giving priority to type A calls over type B calls to maintain an adequate SL requirement, but generalists take more time to serve VIPs. Similarly, the same design can be used when type B calls are VIPs and generalists group's capacity is in excess. Here, acceptable efficiency for both groups of agents can be achieved by routing call type A to free agents in generalists group. Calls can be served according to priority policies and they can be of constant nature over different time intervals. Adding little flexibility in routing strategies can provide minimal requirement for agents in conjunction with simple static priorities [13]. If both call types find agents of their respective groups free, then they will be served by their own groups.

Scenario 4

In fig 7, both call types A and B are served by either of groups 1 and 2 agents to represent full flexibility. It also reflects the fact that skill groups may be defined on a relative, rather than
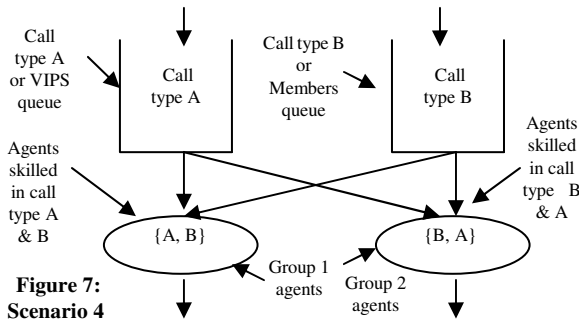
Figure 7: Scenario 4

absolute basis [11]-[13]. Group 1 is assigned call type A as primary skill and call type B as secondary and group 2 is given call type B as primary skill and call type A as secondary. If there are waiting calls for type A and B, then group 1 gives priority to type A calls over type B and group 2 gives priority to type B calls over type A. Groups 1 and 2 only serve their secondary skill calls when deem necessary, say when they have free agents and their relative queues are fully occupied. There is another setting which is possible for fig 7, if call type A queue is represented as VIPs and B as members, serve VIPs queue, if it is not empty plus all members waiting for more than 30 seconds as single FIFO queue, but if VIP queue is empty, serve the first in members queue. Similarly, if members queue is not empty, serve the members first plus all VIPs waiting for more than 5 seconds, as a single FIFO queue, but if members queue is empty serve the first in the VIP queue.
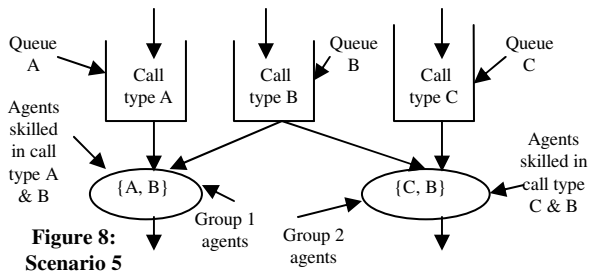
Scenario 5

Figure 8: Scenario 5

In fig 8, two groups of agents deal with 3 types of calls. Group 1 serves call type A as primary and call type B as secondary. Similarly group 2 serves call type C as primary and call type B as secondary. Say, call type B are VIPs, therefore they are prioritized over call types A and C. That means call type B will jump in front of the queue as soon as they arrive. Agents of group 1 and 2 will take the calls from the first non-empty queue in their list. The VIP callers are important corporate asset, and they can be served by in-house agents that the organization trains and manages itself. Agents specialized in handling two different call types give same performance as agents trained in all call types depending if the routing strategy is effective [14].
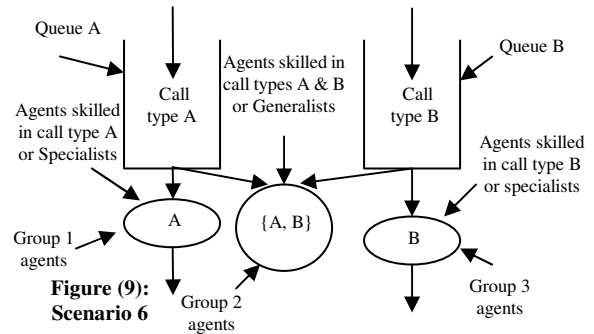
Scenario 6

Figure (9): Scenario 6

Fig 9 is the most popular scenario in the call centre environment specially when used in conjunction with scenario 5. Two types of calls A and B are arriving randomly at different time intervals and 2 groups of specialists and 1 group of generalists that have both skills are dealing with them. It is optimal to give higher priority to specialists then generalists as they work faster and their call proficiency can be equal to 1.0 as their AHT (Average Handling Time) is equal to the forecasted AHT for each time period of the day [15]. Similarly, the calls requiring a particular skill should first be routed to an agent of that skill as a primary skill. Only if all agents with the skill as a primary skill are busy, should agents with that skill as a secondary skill be considered. Typically, a high productivity can be obtained by scheduling exactly the number of agents required to handle all arriving calls. However, the SL will be low and waiting times will be longer. Hence, to meet a SL requirement, it is necessary to schedule additional agents, called 'safety-agents' by the call centre managers, considering the cost constraints at the same time [14],[15].
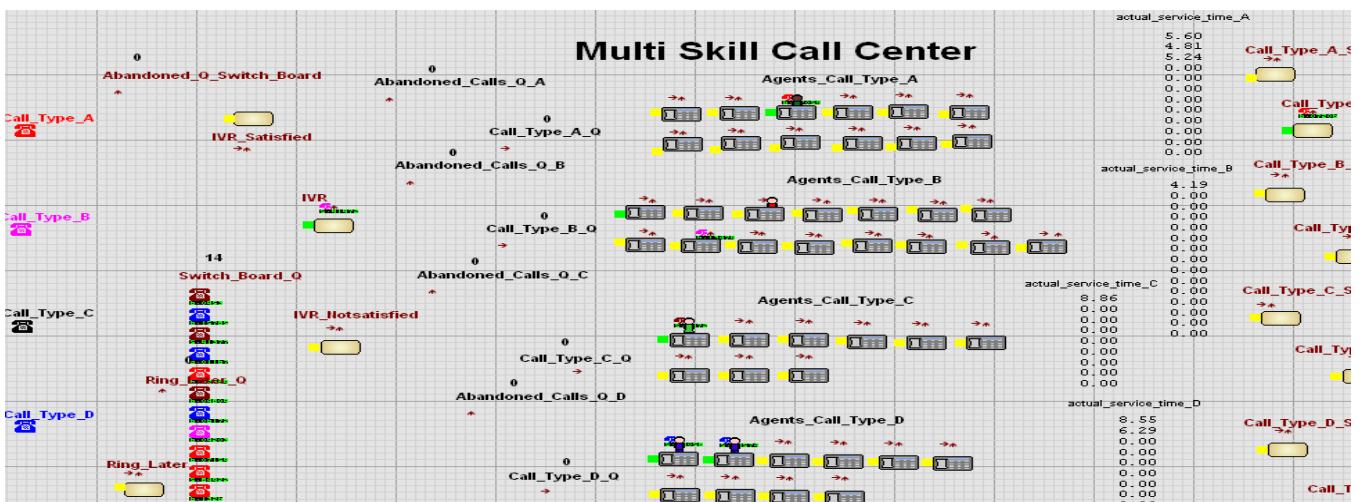
VI. SIMULATION MODEL

Figure 10: Partial view of the constructed model

Figure 11: Display of dynamic variables for strategy 3

At this early stage, a simulation model using Witness has been constructed which is partially displayed in fig 10. The model in fig 10 essentially incorporates scenarios 1 & 2. The agent prioritizing strategy was modelled on the basis of selecting an agent from the available pool for a particular skill set that has the:

(1)   lowest utilization.
(2)   lowest number of jobs completed.
(3)   or a combination of (1) and (2).

Verification and validation of the model is in process hence discussions on results obtained are inappropriate at this stage. Fig 11 illustrates a simulation run using strategy 3.

VII. CONCLUSION

In this paper, call centre functions and operations have been introduced, and described with different routing and prioritizing strategies to underline the use of DES for modelling purposes. The first two strategies have been implemented in a Witness simulation. At this stage of development, validation and verification is being performed on the model constructed. During the modelling process considerable effort has gone into exploring and modifying how resources (agents) are actually allocated. Understanding how Witness assigns resources, it was realized that this logical method of allocation needed alterations to suit a specific strategy that blended with the scenarios devised and discussed. Two important strategies were developed and implemented based on agent utilization and jobs completed. It is realized that additional call routing and prioritizing strategies can be built to model a much bigger and complex setup by following the principles discussed and illustrated, but the liability of testing or verifying lies with the person responsible. The scenarios put up for this paper and their assumptions are the main 'building-blocks' for the call centre managers and can be very helpful in decision making.

REFERENCES

[1]   O. Z. Aksin, M. Armony, and V. Mehrotra (2007), "The Modern Call Centre: A Multi-Disciplinary Perspective on Operations Management Research", Production and Operations Management Journal, 16(6), pp 665-688.

[2]   V. Hlupic, and G. J. D. Vreede (2005), "Business Process Modelling Using Discrete-Event Simulation: Current Opportunities and Future Challenges". In: International Journal of Simulation and Process Modelling, 1(1-2), pp 72 – 81.

[3]   N. Gans, G. Koole, and A. Mandelbaum (2003), "Telephone Call Centres: Tutorial, Review, and Research Prospects". In: Manufacturing and Service Operations Management Journal, Vol. 5, pp 79-141.

[4]   A. N. Avramidis, A. Deslauriers, and P. L. Ecuyer (2004), "Modelling Daily Arrivals to a Telephone Call Centre". In: Management Science, 50:7, pp-896-908.

[5]   M. Armony and C. Maglaras (2004), "Contact Centres with a Call-Back Option and Real-Time Delay Information". In: Operations Research, 52(4), pp 527-545.

[6]   E. L. Ormeci and O. Z. Aksin (2006), "Revenue Management through Dynamic Cross-Selling in Call Centres". In: Department of Industrial Engineering Koc University 34450, Sariyer-Istanbul, Turkey.

[7]   V. Mehrotra and J. Fama (2003), "Call Centre Simulation Modelling: Methods, Challenges, and Opportunities". In: Proceedings of the 2003 Winter Simulation Conference.

[8]   K. Kotiadis and S. Robinson (2008), "Conceptual Modelling: Knowledge Acquisition and Model Abstraction". In: Proceedings of the 2008 Winter Simulation Conference.

[9]   R. B. Wallace and W. Whitt (2005), "A Staffing Algorithm for Call Centres with Skill-Based Routing". In: Manufacturing & Service Operations Management Journal, 7(4), pp 276-294.

[10] P. Chevalier, R. A. Shumsky, and N. Tabordon (2004), "Routing and Staffing in Large Call Centres with Specialized and Fully Flexible Servers". In: International Journal of Electronics and Communications (AEU), 60(2), pp 95-102.

[11] G. Koole and A. Pot (2006), "An Overview of Routing and Staffing Algorithms in Multi-Skill Customer Contact Centres". In: To be published paper, Department of Mathematics, Vrije Universiteit Amsterdam, The Netherlands.

[12] M. E. Sisselman and W. Whitt (2006), "Value-Based Routing and Preference-Based Routing in Customer Contact Centres". In: Production and Operations Management Journal, to be published.

[13] P. Chevalier and V. J-C. Schrieck (2006). "Optimizing the Staffing and Routing of Small Size Hierarchical Call Centres". In: Productive and operations Management Journal, to be published.

[14] N. Gans and Y. P. Zhou (2007), "Call-Routing Schemes for Call Centre Outsourcing". In: Manufacturing and Service Operations Management, 9(1), pp. 33-50.

[15] H. Shen and J. Z. Huang (2008), "Interday Forecasting and Intraday Updating of Call Centre Arrivals". In: Manufacturing and Service Operations Management Journal, 10(3), pp 391-410.