A TUTORIAL ON MODELLING CALL CENTRES USING DISCRETE EVENT SIMULATION

Benny Mathew
Innovation Lab – Performance Engineering
Tata Consultancy Services
Quadra-II, Hadapsar, Pune, India 411028
Email: benny1.m@tcs.com

Manoj K. Nambiar
Innovation Lab – Performance Engineering
Tata Consultancy Services
Gateway-Park, Andheri, Mumbai, India 400093
Email: m.nambiar@tcs.com

KEYWORDS

Discrete Event Simulation, Workforce Planning, Call Centre Modelling, Manpower Planning, Skill-Based Routing, Multi-Skill Call Centre

ABSTRACT

Arriving at an optimal schedule for the staff and determining their required skills in a call centre is imperative to balance the conflicting requirements of delightful customer experience, high employee satisfaction and low cost. Due to the complex nature of modern call centres, simulation modelling is increasingly being used to predict their performance. We have modelled a call centre using our in-house discrete event simulation tool called DESiDE.

This paper describes how every component of call centres were modelled as simulation resources. This paper also describes the changes that had to be made to DESiDE in order to handle the special requirements of call centre modelling and also the metrics used by call centres.

1. INTRODUCTION

Companies have realized the importance of service in order to attract and to retain customers. Over the years, call centres have become the preferred channel in providing service to customers. At a call centre, arriving at optimal level of staffing, their schedule and skills is essential for achieving high level of customer satisfaction and at the same time it is necessary to keep costs low. In current practice, most call centres use analytical models developed by Erlang and Palm to arrive at staffing requirements. These models are ideal during initial operations of a call centre when there is little information available and it is necessary to make assumptions. However, as more and more information of the call centre gets available, use of simulation will yield more accurate results. Using simulation one can remove assumptions made in analytical models and one can also factor in more complex behaviour of call centres. In this paper we describe how this complexity was handled while modelling each resource of a call centre. For simulation we are using an in-house discrete event simulation tool called DESiDE. This paper also describes some of the changes/additional components that were required to make DESiDE handle modelling and simulation of call centre and also report the call centre metrics.

2. CALL CENTRE TERMINOLOGY

A *Call Centre* is a centralized office used for the purpose of receiving and transmitting a large volume of requests by telephone. A call centre is operated by a company to administer incoming product support or information inquiries from consumers. Outgoing calls for telemarketing, clientele, product services, and debt collection are also made. In addition to a call centre, collective handling of letters, faxes, live chat, and e-mails at one location is known as a *Contact Centre*.

2.1 Call Centre Components

Figure 1 shows the components of a typical call centre. Inbound calls are those initiated by customers calling in to the centre (Gans et al. 2003). If all trunk lines are busy, the call will be *blocked*, else the call is first answered by an Interactive Voice Response (IVR) unit. IVR is a technology that allows a computer to interact with humans through the use of voice and keypad inputs. Customers may be able to complete the service interaction at the IVR. If this case, the calls are passed from the IVR to an Automatic Call Distributor (ACD). An ACD is a specialized switch designed to route each call to an individual agent; if no qualified agent is available, then the call is placed in a queue. A queued customer may *abandon* without receiving service.

In a multi-skill call Centre, we distinguish various call types, and we distinguish agents by their skill-set. Skill-set is the set of call types which an agent can handle. Skill-Based Routing (SBR), or simply routing, refers to rules (programmed in the ACD) that control the agent-to-call and call-to-agent assignments in real time.

If more than one agent with requisite skill is available, agent selection criteria comes into picture. The selection criteria can be programmed in the ACD. These methods are described in detail in section 4.1.6.

2.2 Call Centre Metrics

Though there are many Call Centre metrics, only those influencing number of staff and skills required are listed below:

Proceedings 27th European Conference on Modelling and Simulation ©ECMS Webjørn Rekdalsbakken, Robin T. Bye, Houxiang Zhang (Editors)
ISBN: 978-0-9564944-6-7 / ISBN: 978-0-9564944-7-4 (CD)

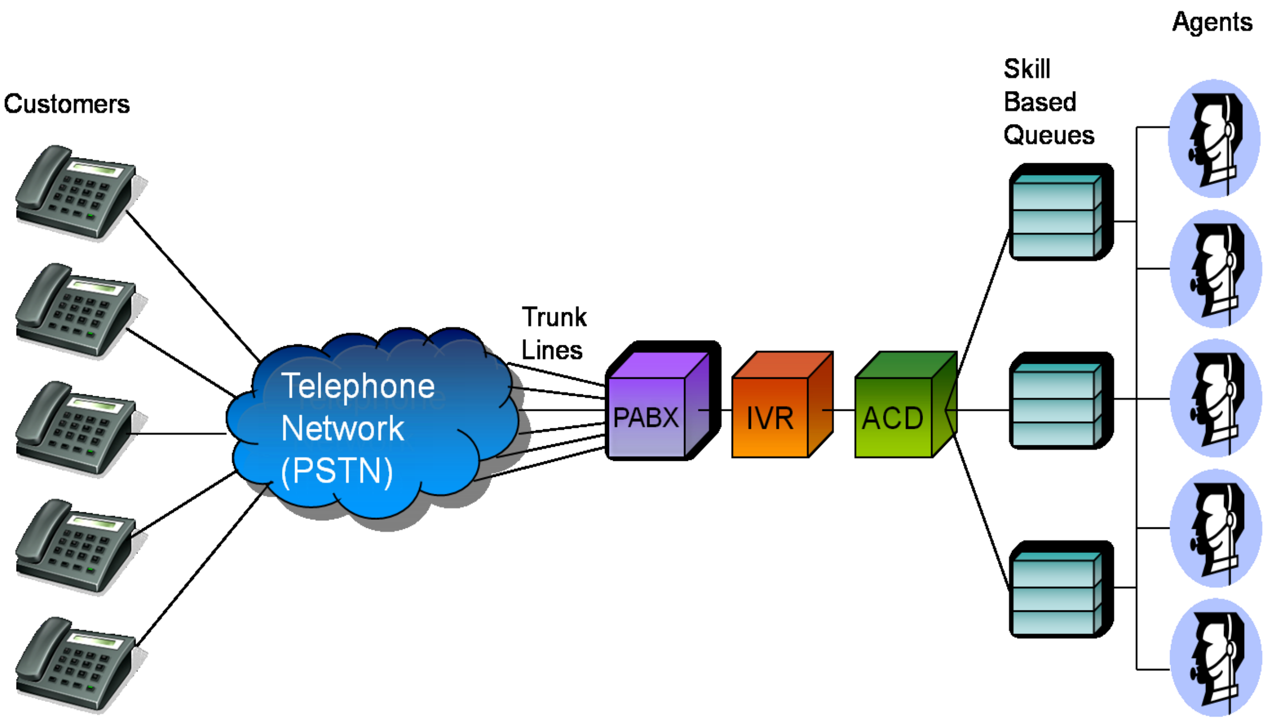

Figure 1: Call Centre Terminology

Blockage: Indicates what percentage of customers will not be able to access the centre at a given time due to insufficient network facilities in place. Most centres measure blockage by time of day or by occurrences of "all trunks busy" situations (Reynolds 2003).

Abandon Rate: Percentage of calls abandoned while waiting to be answered. Abandon rate is not typically a measure associated with e-mail communications, as e-mail does not abandon the "queue" once it has been sent, but it does apply to web-chat interactions.

Average Speed of Answer (SOA): Average time (usually in seconds) it takes for a call to be answered by the service desk. This is one of the most important metrics as far as customer satisfaction is concerned. The percentile value of SOA is called as *Time Service Factor (TSF)*. 80/20 TSF means that 80 percent of the customers have to wait for less than 20 seconds to speak to an agent.

Service Level: Percentage of calls answered within a defined timeframe. TSF is a hence a service level metric.

Agent Occupancy/Utilization: Agent occupancy is the measure of actual time an agent is busy on customer related work as compared with available time. This is calculated by dividing workload hours by staff hours.

Staff Shrinkage: The amount of time staff is unavailable for handling calls due to training, time off, breaks, etc.

Average Call Handle Time: Average time taken by agent to complete a call.

Cost Per Call: This is usually the cost of staff cost per call. However, some call centres may also include other costs like cost of telecom infrastructure, power and other rents.

3. ANALYTICAL AND SIMULATION MODELLING

Analytical models of call centres developed by Erlang (Angus 2001) and Palm (Mandelbaum and Zeltyn 2005) have served the telecommunications industry since the publication of Erlang's paper in 1917. Several enhancements of these models have also been made in order to address the complexities of a modern call centre (Garnett and Mandelbaum 2000, Garnett et al. 2000 and Jouini et al. 2006). These models are used for planning number of agents required, their skills and the schedules they should be following to meet acceptable service levels for projected call volumes.

Simulation is then used to validate the analytical model. Simulation is also used to conduct what-if analysis so as to manage and improve the call centre operations and to plan ahead, in the face of potential scenarios (Anton et al. 2002). The reasons for using simulation is effectively summarised in Bouzada's (2009) paper. In this paper he concludes with the following reasons for using simulation: (i) it is possible to include more details of the operation, to use statistical distributions more compatible with the input data and to have the model closer to reality, assuring the collection of more accurate results; (ii) the service level computed by Erlang formulas is usually underestimated, mainly because these formulas ignore the calls abandonment; (iii) other performance indicators (not available while using analytical approaches, as the abandonment rate) can be evaluated, presented and analysed; (iv) minimum and maximum

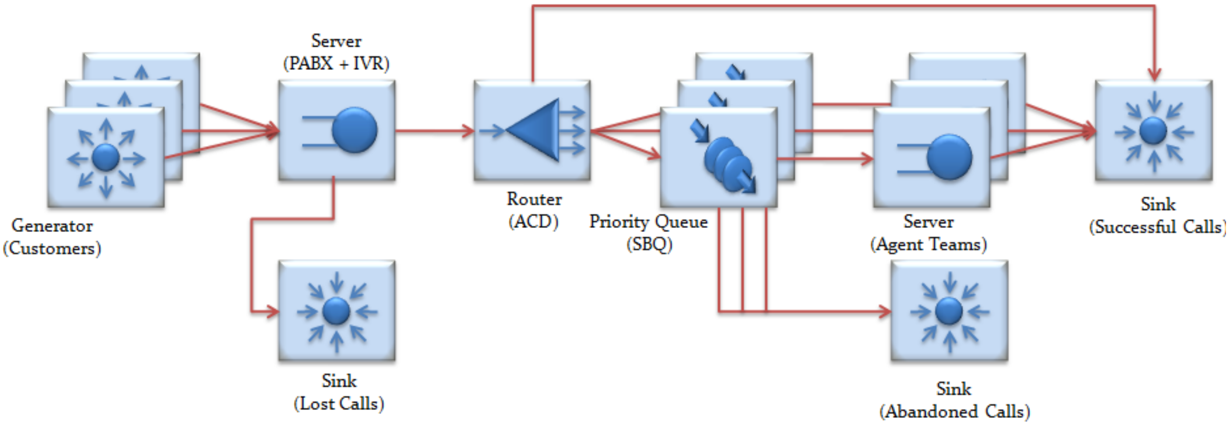

Figure 2: Queuing Network Diagram of Call Centre

values of each important indicator can be obtained, the analyst not being restricted to the average values as when using the Queue Theory; (v) a better understanding of the operation is achieved with the adoption of the experimental approach, which provides the possibility to dynamically follow up the system behaviour and its performance indicators behaviour and, therefore, understand why the queues are being formed and the reason why the waiting time is high, for example, while throughout the Erlang methodology it is possible to see only the generated outputs (numeric indicators) in relation to the provided inputs, making more difficult the complete comprehension of the operation; (vi) the communication can become easier via the use of graphic animations.

4. SIMULATION MODELLING OF CALL CENTRES

While there is sufficient literature on how simulation modelling has helped in improving accuracy of call centre performance predictions (Brigandi et al. 1994, Akhtar and Latif 2010, Mehrotra and Fama 2003), there is very little or no material on how the complexities of a call centre have been tackled in the modelling environment. Also, not all factors that affect performance of a call centre have been covered. For example, there is little or no literature that deals with modelling of misdirected calls.

We at TCS are using DESiDE to model and predict various call centre metrics. Though DESiDE is a generic discrete event simulation tool, a customized version has been built to carry out call centre simulation.

4.1 Call Centre Resource Models

A queuing network representation of a call centre is shown in Figure 2. Each component shown in the figure is modelled as discrete event resource in DESiDE. During simulation the attributes of these resources can be changed. The attributes (also called parameters in certain simulation tools) of each resource influence the resource's behaviour. This enables creating different scenarios and carrying out what-if analysis. On completion of the simulation, a report is generated that comprises the various metrics of a call centre. The various resources modelled in DESiDE and their attributes are listed below:-

4.1.1 Customer

The Customer resource generates the entities that represent incoming calls to a Call Centre. A Call Centre simulation will have several resources of Customer type. Each one generates entities that demand one particular skill on the part of the agent. This is represented by class of entity. The Customer resource also has the following attributes:

Inter-arrival Time[#]*:* Represents time between arrival of calls and is expressed as random distribution. In case number of arrivals is specified for 15/30 minute durations, then the inter-arrival distribution is limited to options of Deterministic or Exponential. Note that in simulation, the end state of one period automatically becomes the starting state of the subsequent period.

Percent High Priority: Represents percentage of calls made by high priority customers.

Retrial Percent: Percentage of customers that will retry calls in case previous attempt did not get

[#] These attributes are used to generate time required to generate events or time required for completion of service. They can be selected from a rich set of distributions in DESiDE like Normal, Exponential, Lognormal, Weibull, Gamma, Uniform, and Erlang. The distributions in DESiDE also support transformations like bound, translate and scale.

In case data has not been fit to any distribution, the measured data can be provided to DESiDE in textual format. DESiDE will generate numbers that match distribution of data in the file.

through because all lines were busy

Retrial Time$^{\#}$*:* The time between retrials expressed as a random distribution.

Patience Time$^{\#}$*:* The time that a customer will spend on hold before abandoning the call expressed as random distribution.

4.1.2 PABX

This resource standing for Private Automatic Branch Exchange (PABX) keeps track of total number of live calls in the exchange. The attribute of PABX is the *number of trunk lines*. In case all trunk lines are busy, subsequent calls will get blocked till at least one of the lines is freed.

4.1.3 IVR

This resource emulates time spent by the customer at the IVR. The IVR has a single attribute, namely *IVR Time*$^{\#}$. This is distribution of service time or time spent by customer before being forwarded to agent. In certain cases calls may be completed at the IVR itself. Based on the user interactions with IVR the skill required on the part of the agent is decided and this interaction time will be different for different skill requirements.

4.1.4 ACD

One attribute of ACD is the *routing matrix* based on which the calls are routed to the correct Skill Based Queues (SBQ). This matrix maps class (skill) and priority of call to a particular team. The ACD does not directly route the call to team but to its associated queue. In case there is no mapping for incoming task, it is assumed to have been completed at the IVR itself.

Another attribute of ACD is *percentage of misdirected calls*. A certain percentage of incoming calls may get directed to an agent who does not have the requisite skill. This can happen due to wrong option selected by the caller or due to wrong interpretation by the speech recognition software in the IVR.

4.1.5 SBQ

The ACD forwarded calls wait at the Skill Based Queues (SBQ) till an agent with requisite skill is available. The check for free agent is triggered whenever a new call is queued at SBQ or when an agent arrives either at time his/her shift starts or after a break and also when an agent is freed after completing a call. An SBQ does not have any user settable attributes.

4.1.6 Team

The attributes for team are of two types, one applicable for all the entire team and another set that has to be provided individually for each agent.

The attributes for the entire team are *agent handle time*$^{\#}$*, redirection time*$^{\#}$ and *agent selection criteria*. The handle time will be different for each class of call. If more than one agent is available then the agent selection criteria comes into picture. DESiDE supports the following options for agent selection criteria.

Uniform Call Distribution (UCD): An incoming call is routed to the agent has been idle for the longest time.

Expert Agent Distribution (EAD): An incoming call is routed to the agent who is best qualified to handle the call.

Least Occupied Agent (LOA): An incoming call is routed to the agent whose utilization is the least.

Least Skills: An incoming call is routed to the agent who has the least number of skills. Calls are allocated preferably to single skilled agent so as to preserve availability of agents who can handle more than one skill

Least Cost: Calls are allocated to available agent costing the least in terms of wages drawn.

Mix of agent selection methods is also possible based on how busy the call centre is. For example at low loads (say less than 60% of agents are busy), UCD can be used and when the call centre load increases, we could use a more advanced strategy like EAD. Note that only few of the above options are usually available in standard call centre hardware.

The attributes that need to be provided for individual agents are:

Schedule: Each agent can follow different schedule in terms of arriving and leaving from work as well as having breaks.

Skillset: An agent can be specialised or cross-trained. A specialised agent will take only one type of call (corresponding to one SBQ). A cross-trained agent can handle more than one type of call and hence answer calls coming to more than one SBQ. Skillset for an agent is defined as the subset of call types that the agent can handle.

Expertise: To account for variation in the level of expertise of agents, one can set an expertise factor for each agent. For arriving at service time distribution, one can take data of average performing agents. These agents are then given expertise factor of 1. Better performing agents are given expertise factor greater than 1 and lower performing or trainee agents are given expertise factor lower than 1. This is particularly useful when a batch of inexperience agents join (expertise factor less than 1) or a batch of agents undergo special training (expertise factor greater than 1).

Hourly Wages: The hourly salary that the agent receives.

4.2 Call Centre Simulation

Before the simulation can begin, values need to be assigned to the attributes of all the call centre resources. Apart from all the attributes, the simulator also requires the number of repetitions.

Since DES depends upon the generation of pseudo-random numbers, the longer the simulation runs more confident we have on the simulation results. The solution takes input data of one week and carries out the simulation. This simulation is repeated multiple times in order to improve the confidence in the simulation results. At the end of the simulation, the various metrics of call centre like time to answer, agent utilization, total cost, and number of blocked and abandoned calls are reported. In the report, the confidence level of each metric is also mentioned. The confidence level is calculated using method of batch means (Law 1977). In case the desired confidence level is not met, the simulation is rerun after increasing the number of repetitions.

Simulation can be used to carry out complex what-if analysis. The attributes of the resources can be changed in order to generate many scenarios. For example one can examine the impact on cost and customer service level if the schedule is modified (through change in shift timing or staggering of breaks), if a few agents are cross-training, if agents efficiency changes (through training or as result of new agents joining), if agent selection criteria is changed and also if there is a dedicated team to handle high-priority customers.

Using the simulator, a study (Mathew and Nambiar 2012) was conducted to find how call centres performance metrics are affected by agent efficiency, handle time distribution and call-to-agent allocation criteria on the call centres performance and some interesting findings were reported.

5. DESiDE CUSTOMIZATION

Given that DESiDE is a generic DES tool, and call centres have very different metrics requirements, there were many changes required to make modelling and reporting easier. Mainly the changes required were related to metric and calculations. Other changes were made to enable data entry and reuse and libraries were added to support the complex call to agent allocation logic. Some of these customizations are listed below.

5.1 Excel Based Input and Reporting

Section 4 describes the various resource models and their attributes. Given that the attribute set is so large, it is difficult to enter such a large volume of data every time through a graphical interface (GUI). Hence, instead of developing a GUI, an excel spreadsheet serves to provide input data for simulation. To begin with, workforce managers/analysts are already familiar in using Excel spreadsheets with built-in Erlang calculators and hence it is easy for them to adapt. Using Visual Basic for Applications (VBA) scripts makes it possible to validate user entered data. When reporting time, work duration, breaks, meetings and skills available of each agent is provided, VBA scripts automatically display how many agents and with what skills are available at 30 minute intervals. This aids the user in arriving at the initial roster for agents.

Once the simulation is completed, the copy of excel input file is created and new tabs are inserted into this excel file. This new file serves as the report file with the results available in the newly inserted tabs. As a result, both inputs of simulation as well as results of the simulation are available in the same file making it easier to refer later.

5.2 Quantile Reporting

One important metric that represents performance of a call centre is the percentile value of speed of answer (SOA). Unlike mean and standard deviation that requires only single pass of data, calculations of quantiles like percentiles require several passes of the data. Hence individual observations are required to be stored till the simulation ends. Since we require to report quantiles every 30 minute for each skill, this not only/ would have resulted in large memory overhead, but also lot of CPU time for calculation at the end of the simulation.

This problem has been addressed using the P^2 algorithm (Jain and Chlamtac 1985). This algorithm has been used when 30 minute quantiles are required. For weekly quantile calculations, DESiDE provides choice of using either fixed-bin or variable-bin histogram approaches.

5.3 Call to Agent Routing

Since incoming calls from a customer and agents both have certain set of attributes to match, the complexity of code that allocates calls to the right agent in DES is on the higher side. Calls have priority and also require certain skill on the part of the agent. The agent side is even more complex since agents can have one or more skills, they work in shifts, and they can be on leave or take breaks

To handle calls waiting in SBQ, priority queues are used [Figure 3]. Separate queues are used for each skill. The calls are ordered first on priority and then by time of entry to the SBQ. These queues are in a wrapper class which provides means to insert and retrieve waiting calls. The wrapper call provides APIs to insert calls into corresponding skill queue and retrieve the first/head call from all the queues. The retrieved calls are in sorted order of priority and SBQ arrival time

Priority queues are used to handle free agents too. Free agents are agents who are currently available to take incoming calls. Here too, like with incoming calls, separate queues are used to keep information of free agents available for each skill. Since an agent can have multiple skills, an agent can be present in multiple skill queues. Within each queue, the order in which the agents are maintained depends upon the call allocation strategy. If the strategy is UCD, the agent who has been idle for the longest will have highest priority, while if the strategy is EAD, the priority will be based on the agent's efficiency. The different ordering of queues based on call

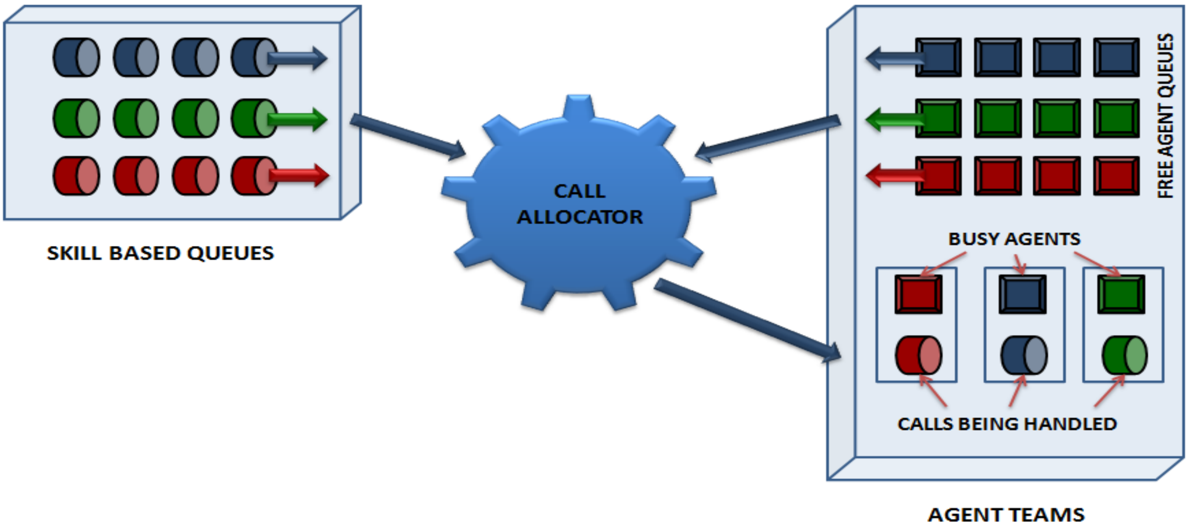

Figure 3: Call Routing

allocation strategy is achieved by using the *compareTo* (Naftalin and Wadler 2006) function available in Java. The priority queues with information of the free agents are maintained by a wrapper class. This wrapper class provides APIs to set allocation strategy, insert/remove the agent information from the priority queue. The wrapper class ensures that agent's information is available in all queues corresponding to the agent's skill. When an agent is no longer available to take new incoming calls, either because the agent is already attending a call or because it is time for break or meeting or end of shift, the wrapper class ensures that the agent is removed all skill queues

The call allocator [Figure 3] is responsible to allocate calls to free agents. The call allocator retrieves the calls in head of queue of each skill and checks corresponding free agent queues. In case there is a match, the call gets allocated. The call allocator carries out its function whenever there is a trigger. These triggers are new call insertions into wait queue and agent insertions into free agent queue. An agent will enter a free agent queue at the start of the days shift, after each break and on the completion of a call.

Note that Figure 3 is for representation purpose only. If there are agents in free list with a particular skill, there will not be any waiting calls requiring that particular skill.

6. CONCLUSION AND FUTURE WORK

This paper explains the reason why simulation modelling is increasingly being used to predict call centre performance. While there are many papers on theory behind modelling of call centres, the focus of this paper is more on implementation of the same in a DES environment. The complexities of modelling call centre resources in DES and the handling of call centre metrics are explained in detail in this paper.

While we have developed the solution using our own DES tool, the same can be implemented in any tool that allows flexibility in adding custom resource models and metrics.

Our call centre model supports a large set of attributes and some of these were covered in a sensitivity analysis study (Mathew and Nambiar 2012). We would like to extend this study to cover the remaining attributes. Such an extensive study will help in arriving at a guide which will list the most influential attributes in resolving a given call centre problem scenario.

REFERENCES

Akhtar, S. and M. Latif. 2010. "Exploiting Simulation for Call Centre Optimization". *Proceedings of the World Congress of Engineering, Vol III*, 2963 – 2970

Angus, I. 2001. "An Introduction to Erlang B and Erlang C". *Telemanagement #187*, 6-8

Anton J.; V. Bapat and B. Hall. 2002 "Call Centre Performance Enhancement Using Simulation and Modelling". ISBN-13: 978-1557531827, 32-35

Bouzada, M. A. C. 2009. "Dimensioning a Call Centre: Simulation or Queue Theory?". *The Flagship Research Journal of International Conference of the Production and Operations Management Society, Vol. 2 No. 2*, 34-46

Brigandi, A.; D. Dargon; M. Sheehan and T. Spencer. 1994. "AT&T's call processing simulator (CAPS): Operational design for inbound call centres". *Interfaces 24,* 6-28

Gans, N.; G. Koole and A. Mandelbaum. 2003. "A. Telephone Call Centres: A Tutorial and Literature Review". *Manufacturing and Service Operations Management,* vol. 5, 79–141

Garnett, O. and A. Mandelbaum. 2000. "An Introduction to Skills-Based Routing and its Operational Complexities". *Technion*, Israel.

Garnett, O.; A. Mandelbaum and M. Reiman. 2000. "Designing a call centre with impatient customers". *Bell Laboratories*, Murray Hill, N. J, 208-227

Jain, R. and I. Chlamtac. 1985. "The P^2 algorithm for dynamic calculation of quantiles and histograms without storing observations". *Communications of the ACM, Vol. 28 No. 10*, 1076-1085

Jouini, O.; A. Pot; G. Koole and Y. Dallery. 2006. "Real-Time Scheduling Policies for Multiclass Call Centres with Impatient Customers". *International Conference on Service Systems and Service Management*, 971-976

Law, A. M. 1977. "Confidence intervals in discrete event simulation: A comparison of replication and batch means". *Naval Research Logistics Quarterly, Volume 24, Issue 4*, 667-678

Mandelbaum, A. and S. Zeltyn. 2005. "Service Engineering in Action: The Palm/Erlang-A Queue with Applications to Call Centres". *Advances in Services Innovations,* Springer-Verlag, 17-48

Mathew, B. and M. K. Nambiar. 2012. "Cross Training - Is it a Panacea for all Call Centre Ills?". *Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol II, IMECS*, 1543-1548

Mehrotra, V. and J. Fama. 2003. "Call Centre Simulation Modelling: Modelling, Challenges, and Opportunities". *Proceedings of 2003 winter Simulation Conference*, 135-143

Naftalin, N. and P. Wadler. 2006. "Java Generics and Collections". *O'Reilly Media*, ISBN-13: 978-0596527754, 31-36

Reynolds, P. 2003. "Call Centre Staffing: The Complete, Practical Guide to Workforce Management". ISBN-13: 978-0974417905, 180-196

AUTHOR BIOGRAPHIES

BENNY MATHEW is a Senior Scientist with the TCS Innovation Lab – Performance Engineering. Benny Mathew holds a Masters Degree in Reliability Engineering from Indian Institute of Technology Bombay, Mumbai. He has worked for 5 years in Symantec Software and 11 years in TCS in the area of Software Performance Engineering. His expertise lies in discrete event simulation of computer systems, call centres and business processes. He has also worked in the areas of performance measurement, testing and file systems performance, databases and middleware. His email address is `benny1.m@tcs.com`

MANOJ K. NAMBIAR is Chief Scientist with TCS Innovation Lab – Performance Engineering. Manoj holds a Bachelors Degree in Computer Engineering from the University of Mumbai. At TCS, he worked from 1994 for seven years in the design, implementation and support of ISDN BRI products of NORTEL Networks. Later he started work in the area of Performance Engineering of enterprise systems as a consultant before moving in 2006 to the Corporate Technology Office - R&D Unit where he is now serving as a council member in the Systems Research area and leading a research group in the field of Performance Engineering - monitoring, analysis & modelling. His e-mail address is: `m.nambiar@tcs.com` and his Web-page can be found at `http://www.tcs.com/about/research/researchers/Pages/Nambiar-Manoj.aspx`