Cross Training - Is it a Panacea for all Call Center Ills?

Benny Mathew and Manoj K. Nambiar

Abstract—**Arriving at an optimal schedule for the staff and determining their required skills in a call center is imperative to balance the conflicting requirements of delightful customer experience, high employee satisfaction and low cost. Due to the complex nature of contact centers, researchers have been moving from analytical modeling to simulation modeling. We have modeled a call center using our in-house discrete event simulation tool called DESiDE.**

In this paper we take a multi-skilled call center and carry out sensitivity analysis of the centers performance to cross training, handle time distribution and call-to-agent allocation criteria.

Index Terms—**Call Center, Contact Centre, Discrete Event Simulation, DES, Workforce Scheduling, Cross Training, Manpower Planning, Schedule Optimization, multi-skill**

I. Introduction

COMPANIES have realized the importance of service in order to attract and to retain customers. Over the years, call centers have become the preferred channel in providing service to customers. Arriving at optimal level of staffing, their schedule and skills of a contact centre is essential for achieving high level of customer satisfaction and also in keeping costs low. Call centers use analytical models developed by Erlang and Palm to arrive at staffing requirements. These are models are ideal during initial operations of a call centre where there is little measured data available. However, once more data is 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 behavior of call centers. In this paper we examine use effect of some of these factors like cross-training, handle time distribution and call-to-agent allocation criteria. For simulation we are using in-house developed tool called DESiDE.

II. 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*.

A. Call Centre Components

Fig. 1 shows the components of a typical call centre. Inbound calls are those initiated by customers calling in to the center [1]. If all trunk lines are busy, the call may 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 not, 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 center, we distinguish various call types and we distinguish agents by their skill group. The skill group is defined as the subset of call types that an agent can handle. An agent handling only single type of call is called *specialist*, an agent handling more than one type of call is called *cross-trained*. *Skill-Based Routing (SBR)*, or simply routing, refers to rules (programmed in the *ACD*) that control in real time the agent-to-call and call-to-agent assignments.

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 section IV where modeling of *Agent Teams* is described.

B. Call Center Metrics

Though there are many Contact Centre metrics, only those influencing staffing size are listed below:

Blockage: Indicates what percentage of customers will not be able to access the center at a given time due to insufficient network facilities in place. Most centers measure blockage by time of day or by occurrences of “all trunks busy” situations [2].

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

Manuscript received December 08, 2011; revised January 30, 2012.

Benny Mathew is with Tata Consultancy Services, Quadra II, Hadapsar, Pune 411028 INDIA. He is also a member of IAENG (phone: +91 20 6622 4422; fax: +91 20 6622 4499; e-mail: benny1.m@tcs.com).

Manoj K. Nambiar is with Tata Consultancy Services, Gateway Park, MIDC, Andheri(E) Mumbai 400093 INDIA (e-mail: m.nambiar@tcs.com).

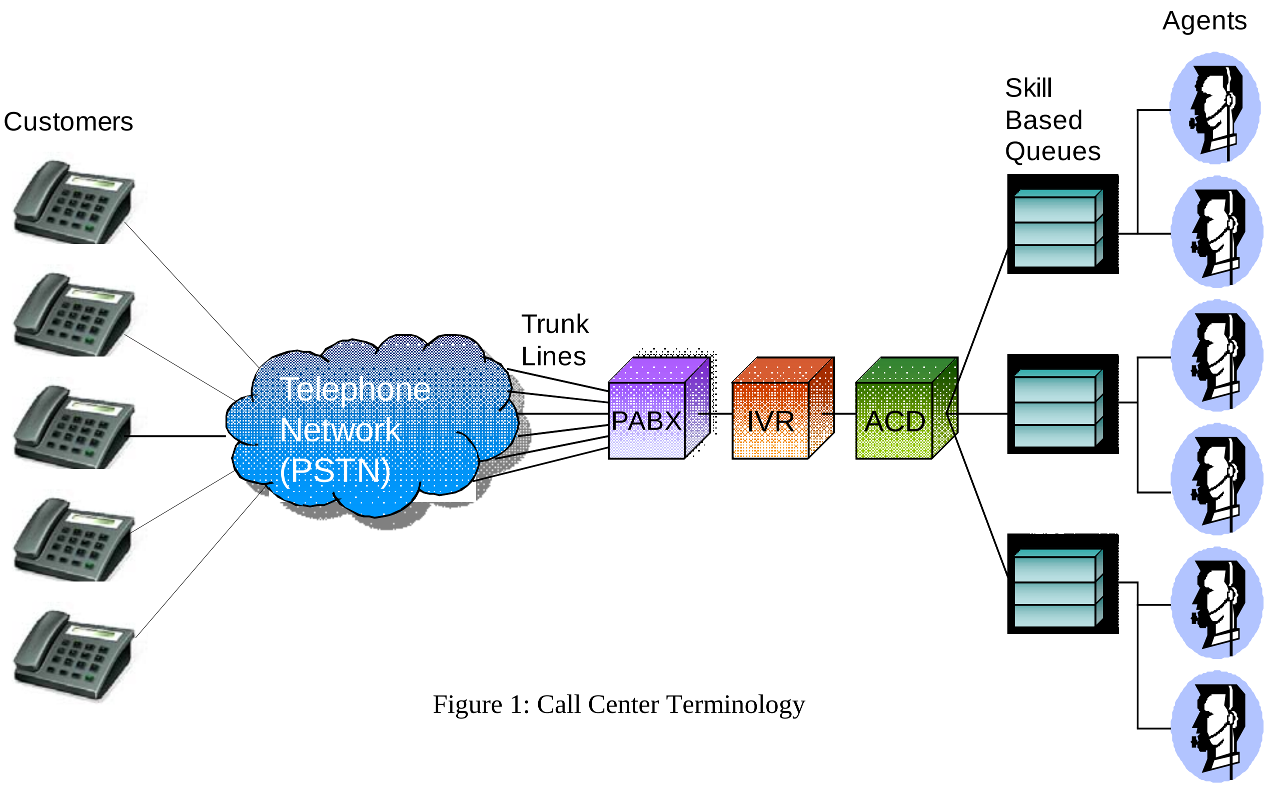

Figure 1: Call Center Terminology

customer service level is concerned. The percentile value of SOA is also sometimes referred as *Time Service Factor (TSF).* 80/20 *TSF* means that 80 percent of the customers have less than 20 second *SOA*.

Service Level: Percentage of calls answered within a definite timeframe.

Agent Occupancy/Utilization: Agent occupancy is the measure of actual time an agent is busy on customer contacts compared with available or idle time, 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 centers may also include other costs like cost of telecom infrastructure, power and other rents.

III. Limitations Of Erlang/Analytical Models

Erlang[3] and Palm[4] models have served the telecommunications industry well since publication of Erlang’s paper in 1917. These are models are ideal during the initial operations of a call centre where there is little measured data available. When more information is available, one needs to move to simulation in order to improve upon the accuracy of the predicted results. Using simulation one can remove the assumptions made in these models as well as account for more complex factors affecting performance of a call center.

The assumptions in Erlang/Palm models are: the arrivals follow Poisson distribution and service time and abandonment follow exponential distribution, there is not priority and service discipline is first-in-first-out (FIFO), the call handling time is independent of customer and also of the agent receiving the call, all traffic received is accepted and there are no retrials.

Researchers [5],[6] have analyzed actual data from call centers and found that: calls do not always arrive in Poisson distribution and the analyzed data points to Uniform, Lognormal and other distributions. Call handling times service-times too have been found to be Log-normally distributed and in some cases may even be bi-modal. The error due to Poisson arrivals assumption is to some extent addressed by specifying the average number of arrivals in 15 minute or 30 minute periods. However, since the state at the end of one period is not preserved as initial condition for subsequent period, this approach also leads to inaccurate results.

Some of these assumptions have been addressed through advancement [7],[8],[9] of analytical models. However the advanced analytical solutions only solve a subset of the limitations listed out in this section. Also, simulation gives more detailed statistics (not just averages) and gives ability to study more complex factors like call allocation strategy. As a result, there has been a gradual shift towards simulation as a means of resolving Contact Center staffing problem [10],[11],[12].

IV. Simulation Modeling of Call Centers

We at TCS are also using simulation to model and predict various Call Centre metrics using an in-house tool call DESiDE. Though DESiDE is a generic discrete event simulation tool, a customized version has been made in order to carry out Call/Contact Centre simulation.

A. *Call Center Resource Models*

Each component of a call centre is modeled as discrete event resources in DESiDE. During simulation the attributes of these resources can be changed. The attributes (also called parameters in certain simulation tools) of each resource influences its time-wise behavior. 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 center. The various resources of modeled in DESiDE and their attributes are listed below:

Customer: The Customer resource generates the entities that represent incoming calls to a Call Center. A Call Center 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 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.

PABX: This resource 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.

IVR: This resource emulates time spent by the customer at the IVR. The IVR has a single attribute

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.

ACD: The attribute of ACD is the routing matrix based on which the calls are routed to the correct SBQ's. 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.

SBQ: The ACD forwarded calls wait at the 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.

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

The attributes for the entire team are *agent handle time#* and *agent selection method*. The handle time will be different for each class of call. If more than one agent is available then the agent selection method comes into picture. Though not supported by all brands of call center hardware, the DESiDE model supports the following options for *agent selection method*.

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 center is. For example at low loads (say less than 60% of agents are busy), UCD can be used and when the call center load increases we could use we could more advanced strategy like EAD.

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: Defined as the subset of call types an 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

# These attributes are used to generate time required for various activities. These can be selected from a rich set of distributions in DESiDE like Normal, Exponential, Lognormal. Weibull, Gamma, Uniform, Erlang. In case the time data is available in a text file, then the file can also be used as input for the simulation.

The distributions in DESiDE also supports transformations like bound, translate and scale.

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).

Cost: One can assign different per hour cost for each agent representing their wages earned.

B. *Call Centre Simulation*

After the attributes of all the resources are set and number of repetitions is also set, the simulation can begin. At the end of simulation, the various metrics of call centre like time to answer, agent utilization, total cost, and number of blocked and abandoned calls are reported. DESiDE takes the schedule of one week and carries out multiple repetitions of the schedule. In the report the confidence level is mentioned. In case the required confidence level is not met the number of repetitions is increased and simulation is rerun.

Simulation can be used to carry out complex what-if analysis. The various attributes of the resources can be changed in order to do generate many scenarios. For example one can examine the impact on cost and customer service level if 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.

V. Sensitivity Analysis

Since the simulation clock progresses in virtual time, we can run easily change attributes and check the impact in a short period of time. By increasing the number of iterations, the tests run longer but it improves the confidence level in the results. Let's study the effects of changing some of the important attributes of Call Center. The metric chosen to represent performance of call center is 80 percentile speed of answer (SOA). This metric has been chosen since many call centers have the SLA that the 80% of the time, SOA should be less than 20 seconds. This is also called 80/20 TSF. Before the simulation test results are explained, the scenario is described. Here all the attributes that are kept invariant for the simulation runs are described.

A. *The Scenario*

Consider a call-centre with incoming calls that requires three different types of skills on the part of call center agent. There are 24 call centre agents working who are available to answer the incoming calls. For the sake of simplicity we have made assumptions that are listed below. Note that simulation modeling follows software programming life cycle and requires intensive testing and debugging. One of the methods to test accuracy of simulation models is to make same assumptions as analytical models and to compare the results. Hence, initially the same assumptions are made for simulation models as in case of analytical models. Listing down the scenario details:

- All 24 agents follow the same 8-hour schedule and there incoming calls only during these 8 hours.
- The incoming calls follow Poisson process (inter-arrival times follow Exponential distribution) and have equal priority.
- The exchange has infinite capacity and hence there is no blockage and no retrials.
- No time is spent at the IVR and calls are straightway directed to the skill based queues.
- A cross-trained agent can handle call requiring any of the three skills. This means that in case there are 25% cross-trained agents, 6 agents have all three skills. The three skills are equally divided among the remaining agents who are single-skilled.
- The workload in the graphs shows the average number of incoming every half hour per skill. So in case the workload shows 60, it means in half hour duration there will be total of 180 calls (60 per skill).

B. *Simulation Results*

1. *Cross-Training*

First let's examine the performance of call centre as we increase the arriving call rates and check the impact of cross-training.

Keeping all assumptions mentioned at the beginning of this section, we keep a handle time with average of 180 seconds following exponential distribution. We see that (Fig. 2) when the workload is 60, the 80/20 TSF is met with 4 (12.5%) cross-trained agents, while at higher workload of 66, the 80/20 TSF is met only when 18 (75%) agents are cross-trained.

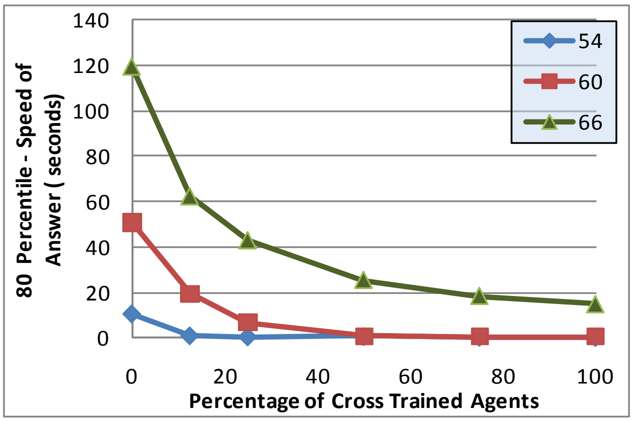

Figure 2: Impact of Cross Training

2. *Cross-Training with changing call-mix*

In earlier case, the average number of calls of the three different call-types was the same and the variability in arrivals was only due to Poisson arrivals. Now we'll examine the impact of varying the average number of arrivals every half hour in such a way that the total arrivals is constant but the percentage of mix of the three skills keeps varying by a specified percentage. Let's examine whether the additional variability is also absorbed by cross-training.

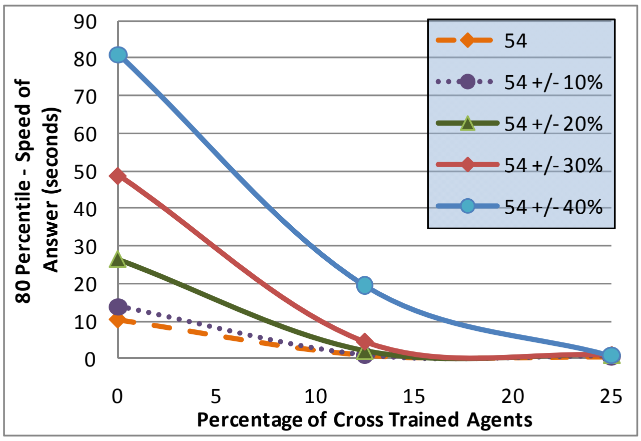

Figure 3: Impact of Cross Training
(Changing call mix)

We can see that when there is no cross-training, the impact of this variation is quite severe. The 80 percentile SOA increases from 10 seconds for no variation to more than 80 when mean arrivals vary by 40%. However we see that (Fig. 3) with cross-training, the problem is very easily tackled. Only 4 cross-trained agents bring down the 80 percentile SOA to meet 80/20 TSF.

3. *Cross-Training with Lognormal Service Times*

For these tests also, the scenario is the same as described at beginning of this section, with the following additions: There is variation of 20% in efficiency of each agent, there is 30% variation in the call mix every 30 minute and incoming calls follow Poisson with average of 60 calls every half hour for each skill.

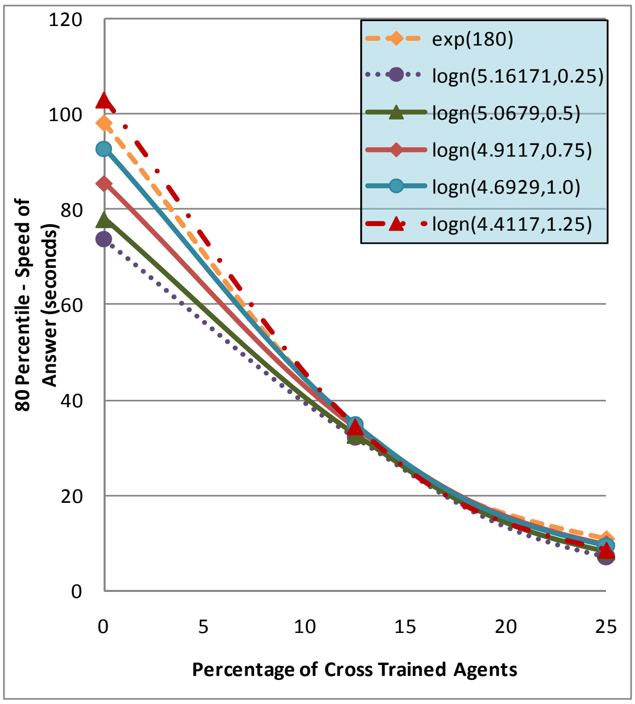

Figure 4: Impact of Cross Training
(Lognormal Service Times)

As mentioned in Section III, researchers have found that service times distribution in call centers have often been known to follow Lognormal distribution.. Hence we carry out simulation tests comparing Exponential service times with Lognormal service times. The Lognormal parameters are varied in such a way that the average handle time is 180 seconds. From the results we see (Fig. 4) that there is significant difference in SOA depending upon the service time distribution and it increases when the distribution has higher variance. However with cross-training, there is little impact of service time distribution on the performance.

4. *Call Allocation Strategy*

In section IV various strategies for work allocation has been described in the description of *Agent Team*. These are the strategies which can be programmed into an ACD so direct a call to an agent when more than one agent with requisite skill is available. In this section, we compare results of using couple of advanced strategies over the default UCD strategy.

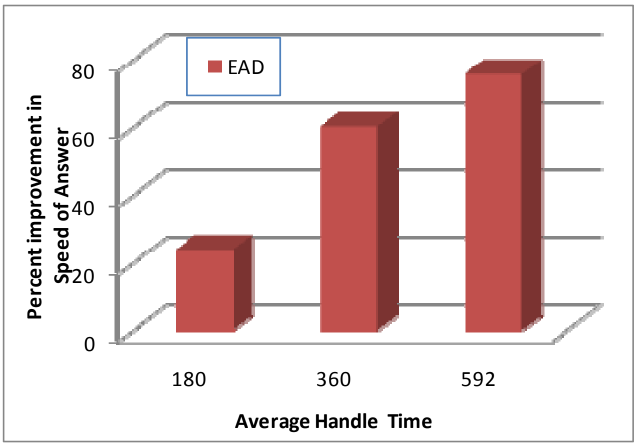

Figure 5: Call Allocation Strategy
(No Cross Training)

For these tests, the scenario is the same as that is described earlier in this section, with the following additions: There is variation of 30% in efficiency between agents, there is 30% variation in the call mix every 30 minutes, incoming calls follow Poisson, handle times are Exponential. Now we take 3 cases with average handle time of 180, 360 and 592 seconds. For each of these cases, we take baseline call allocation strategy and adjust the call arrival rate so that 80/20 TSF is just met.

First, let's take the case when there are no cross-trained agents and check the advantage of using EAD over UCD (Fig. 5). Here the Y-axis shows percentage reduction (improvement) in 80 percentile SOA. We can see that there is about 20% improvement in performance when we use the EAD strategy when service time is 180 seconds. When the agent service times are higher, the improvement is also higher.

Next, we take the case when there are 25% cross-trained

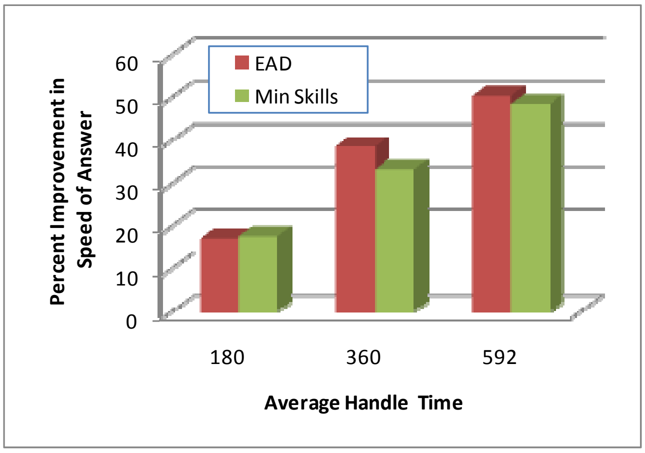

Figure 6: Call Allocation Strategy
(With Cross Training)

agents (Fig. 6). Here, apart from EAD strategy we can also check whether there is an advantage in using Minimum Skills strategy over UCD. Here we see that both strategies give good improvement as compared to UCD.

VI. Conclusion

This paper explained why and how simulation is becoming popular means of predicting call centre performance. Using our own simulation tool, we carried out various simulation experiments to check sensitivity of call center performance to various conditions like changing service time distribution and call mix ratios. We also see the beneficial effect of cross-training agents in different scenarios. We could also prove the benefit of using advanced call allocation strategies in improving call center performance.

While we have been able to get valuable insights by using synthetic data, the real power of simulation can be realized when real measurements are used along with the what-if capability to arrive at number of agents, their skills and schedule.

References

[1] N. Gans, G. Koole and A. Mandelbaum, “Telephone Call Centers: A Tutorial and Literature Review”, Manufacturing and Service Operations Management (2003), vol. 5, pp. 79–141

[2] P. Reynolds, Call Center Staffing: The Complete, Practical Guide to Workforce Management (2003), ISBN-13: 978-0974417905

[3] I. Angus, “An Introduction to Erlang B and Erlang C”, Telemanagement #187 (2001), pp. 6-8

[4] A. Mandelbaum and S. Zeltyn, “Service Engineering in Action: The Palm/Erlang-A Queue, with Applications to Call Centers”, Advances in Services Innovations, Springer-Verlag, pp. 17-48

[5] L. Brown, N. Gans, A. Mandelbaum, A Sakov, H Shen, S. Zeltyn and L. Zhao, “Statistical Analysis of a Telephone Call Center”, Journal of the American Statistical Association (2005), pp. 36-50

[6] L. Franzese, M. Fioroni, R. Botter and P Filho, “Comparison of Call Center Models” , Proceedings of 2009 Winter Simulation Conference (2009), pp. 2963 – 2970

[7] O. Garnett and A. Mandelbaum, “An introduction to skills-based routing and its operational complexities”, Technion (2000), Israel.

[8] O. Garnett, A. Mandelbaum and M. I. Reiman, “Designing a call center with impatient customers” Bell Laboratories(2000), Murray Hill, N. J.

[9] O. Jouini, A. Pot, G. Koole and Y. Dallery, “Real-Time Scheduling Policies for Multiclass Call Centers with Impatient Customers”, 2006 International Conference on Service Systems and Service Management (2006), pp. 971 - 976

[10] A. Brigandi, D. Dargon, M. Sheehan and T. Spencer, “AT&T's call processing simulator (CAPS): Operational design for inbound call centers”, Interfaces 24 (1994), pp. 6-28

[11] S. Akhtar and M. Latif, “Exploiting Simulation for Call Centre Optimization”, Proceedings of the World Congress of Engineering 2010, Voll III (2010), pp 2963 – 2970

[12] V. Mehrotra, J. Fama, “Call Centre Simulation Modeling: Modeling, Challenges, and Opportunities”, Proceedings of 2003 winter Simulation Conference (2003), pp. 135-143