

Analysis of Call Routing Rules for Improving Call Centre Operations in Nigeria

Babalola Gbemisola O.
Department of Computer Science
Afe Babalola University, Nigeria

Oguntimilehin Abiodun
Department of Computer Science
Afe Babalola University, Nigeria

Odejayi Adeniyi
Department of Computer Science
University of Ibadan, Nigeria

ABSTRACT

Routing rules plays a very important role in the services offered by call centres in a competitive environment. For example, a call centre whose priority is to reduce overall mean time to service completion, one might think it best to route calls to agents who can handle it the fastest sometimes even holding a call in queue to wait for that agent to free up rather than routing it to a slower agent. However, this rule does not account for the increase in congestion resulting from repeated phone calls associated with unresolved issues. On the other hand, for a call centre that is primarily focused on call resolution, it seems optimal to route each call type to the agent who can handle it the best, thus holding that call in queue even if other agents are idle and/or become available earlier. However, in an environment where there is significant variability across different agents' resolution probabilities. Routing rules that are based solely on these rates are likely to lead to long queues.

This work attempts to determine whether average handling time and call resolution rate are true determinants of operational success of a call centre to reduce waiting queue. It also aim at examining whether emphasis should be on reducing handling time or effective call resolution including the trade-offs between these two criteria. The result emphasizes the trade-offs between Average Speed of Answer (ASA) and Call Resolution (CR) rates and also shows that neither waiting-time nor resolution oriented rules are superior to each other; it is subjectively dependent on the value the call centre places on either of the rule..

Keywords

Call Centre, Routing rule, Call Centre, Simulation Analysis.

1. INTRODUCTION

The call centre service has grown a great deal with its application in all sectors of the economy. It serves as a primary contact between businesses and clients. But in recent times, customers waiting for so long in order to lodge a complaint or make an enquiry have become a worrisome phenomenon in the call centres.

A customer's experience during a service encounter consist of two parts namely: the time spent waiting for the service and the service itself. Call centres give priority to the two criteria with emphasis on one more than the other. Those that place more emphasis on time spent waiting for the service are more concerned with reducing the average time involved in handling a call while those that are concerned with the service itself aims at effective resolution of customer issues.

[1] says for a call centre to reduce waiting lines with emphasis on the reduction of time spent, its best to route calls to agents who can handle customer issues the fastest, sometimes even holding a call in queue to wait for that agent than routing the call to a slower agent. But this might lead to further increase

in congestion, repeat calls from unreceptive issues and undue burden on some agents. [2], states that for a call centre to reduce waiting lines, emphasis should be on the service itself that is; call resolution. Its best to route calls to agents who resolve customer issues, sometimes holding a call in queue to wait for such agent. This might also lead to increase in congestion and undue burden on some agents.

After a customer has received service from a call centre agent on a particular issue, a subsequent call from that customer about the same issue is a clear sign that the issue had not been resolved during the previous service encounter, and this lack of resolution is a strong sign of customer dissatisfaction. Thus, Call Resolution rates are very important customer-oriented operational metrics in most telecommunication companies in Nigeria. As data collection and analysis technologies for accurately measuring Resolution Probability values begin to emerge, call centre managers are increasingly focused on managing the Call Resolution metrics. Higher Call Resolution rates result in reduced system congestion (due to decreased call-backs and hence lower total call rates) and subsequently lower staffing costs. As such, these metrics have been attracting more attention from call centre leaders.

In this work, different strategies for routing multiple types of calls to a large group of agents were explored, where these assignments are made dynamically based on the specific attributes of the agents and/or the current state of the system. We believe that this study will make several important contributions to the call centre operations management literature.

2. RELATED WORKS

Various attempts have been made by several authors and organizations to find a comprehensive and universally accepted definition for the term call centre. Each group defining it as it appears to her. A call centre is a system that offers complete management of all communication channels between a business and its customers, optimizing process, eliminating duplicated work and making better use of time. [3]. Call centre was defined as a centralized office used for the purpose of receiving and transmitting a large volume of request by telephone [4]. A call centre was also defined as a set of resources (communication equipment, employees, computers etc.) which enable the delivery of services via the telephone [5].

Call routing is the sequence of path taken to convey a customer's call to a service agent. Call routing also known as call distribution relates to a set of rules which are applied to isolate the most appropriate resource for a specific called. Call routing is experience by the customer as being guided through a decision tree [6]. By progressing through that tree the system provides information to and collects user inputs from the caller. The corresponding realization is often referred to as routing path. However having reached the leaf of the decision

tree, the collected information is considered as being sufficiently complete and call distribution takes over to determine the most appropriate agent based on agent properties, user input and system load to route the call.

All routing techniques or algorithms used in call distribution follows a baseline routing rule which serves as a benchmark for routing calls [7]. The benchmark routing rule usually followed is the first-come, first serve or longest wait rule. Here the rule states that the first customer to arrive on a queue or the customer that has waited the longest on the queue and it follows the sequence until all calls are attended to.

In [8], the author summarizes an analysis of a unique record of call center operations. The data used comprised of a complete operational history of a small banking call center, call by call, over a full year. Taking the perspective of queuing theory, the author decomposed the service process into three fundamental components: arrivals, customer patience, and service durations. Each component involved different basic mathematical structures and required a different style of statistical analysis. Some of the key empirical results are sketched, along with descriptions of the varied techniques required. Several statistical techniques were developed for analysis of the basic components. One of these techniques is a test that a point process is a Poisson process. Another involves estimation of the mean function in a nonparametric regression with lognormal errors. A new graphical technique is introduced for nonparametric hazard rate estimation with censored data. The models were developed and implemented for forecasting of only Poisson call arrival rates.

Call centers are important channels of communication within the consumer relationship and a point of integration between suppliers and their customers. Correctly sizing the capacity of a given Call Center can bring benefits not only in terms of improved customer service (efficacy), but also in terms of reduced operating costs (efficiency). However, specifying the capacity of a Call Center is not a trivial task, but one that demands a significant knowledge of mathematics, in particular of analytical models. This author presents the Erlang B, Erlang C and Simulation models followed by a comparison based on a case study, in order to identify the advantages of using simulation. This work is limited to the comparison of call center model while ignoring analytical methods and soft computing methodologies [9].

The works of [10] was to establish analytical methods (such as Queue theory) to experimental methods (such as simulation) and discussing their adequateness to complex operations – set up in the matter of dimensioning the handling capacity of a large brazilian call centers company. The experimental approach is suggested to be implemented as an alternative methodology to deal with the issue, instead of the analytical method in use. The results obtained are used to justify the adequacy of the experimental approach to the modern call centers operation, as long as it is possible to have the model closer to reality. The main implication points to a better understanding of the operation achieved with the new approach. This work did not explore other call transference process during a client attending operation before being handled by the correct agent; (ii) conferences amongst the client and more than one operator at the same time; (iii) conditional call detours towards specialized services; and (iv) other queue disciplines than the traditional FIFO.

The author considered the problem of minimizing staffing costs in an inbound call center, while maintaining an

acceptable level of service in multiple time periods. The problem is complicated by the fact that staffing level in one time period can affect the service levels in subsequent periods. The author presented a simulation based analytic center cutting plane method to solve a sample average approximation of the problem. The authors establish convergence of the method when the service level functions are discrete pseudo concave. An extensive numerical study of a moderately large call center shows that the method is robust and, in most of the test cases, outperforms traditional staffing heuristics that are based on analytical queueing methods. The problems solved in this work were fairly simple instances of a call center staffing problem, but since no assumptions are made on the arrival and service processes and simulation is used to evaluate performance, it seems that the method would also apply in more complicated settings. Call abandonments, skill-based routing and prioritizing multiple customer classes are problems that call center managers commonly face and it would be interesting to incorporate those in the algorithm [11].

The author of [12] show via concrete illustrations how the variance can be reduced in the simulation of a telephone call center to estimate the fraction of calls answered within a given time limit. The author examined the combination of a control variate and stratification with respect to a continuous input variable, and find that combining them requires care, because the optimal control variate coefficient is a function of the variable on which they stratify. In a setting where they compared two similar configurations of the center, they examined the combination of stratification with common random numbers. The authors show that proper use of common random numbers reduces the convergence rate of the variance of the difference of performance measures across the two systems. This work is limited to the comparison of two similar configurations of a call center, Cases of dissimilar configurations, and stratification with random numbers will be difficult.

3. METHODOLOGY

This work begin with a review of related works as it forms the basis for this research and provides sources to scientific papers that give insight into routing rules and call centre operations. Having understood the call centre operations, a request was made for call centre data from a call centre in Nigeria. The data was obtained from the automated data logging system comprising of agent identity, calls attended to, call handling time, call status, etc. This data was used to carry out a simulation analysis for the call centre while testing the operations of different routing rules.

Each routing rule was used independently with the collected data to simulate the call centre operation. In analyzing the data gathered, simulation was carried out by using the data gather from the above call centre to estimate parameters needed to characterize the model and when this was conceptualized, simulation was carried out using a collection of JAVA programs on each of the routing rules. At the end of each simulation analysis, the results obtained from using each routing rule was compared so as to answer the research questions and to make recommendations to call centres. The analysis provided the basis for the answer to the research questions and the conclusion.

3.1 System Design

This overall system approach is presented in figure 1 below:

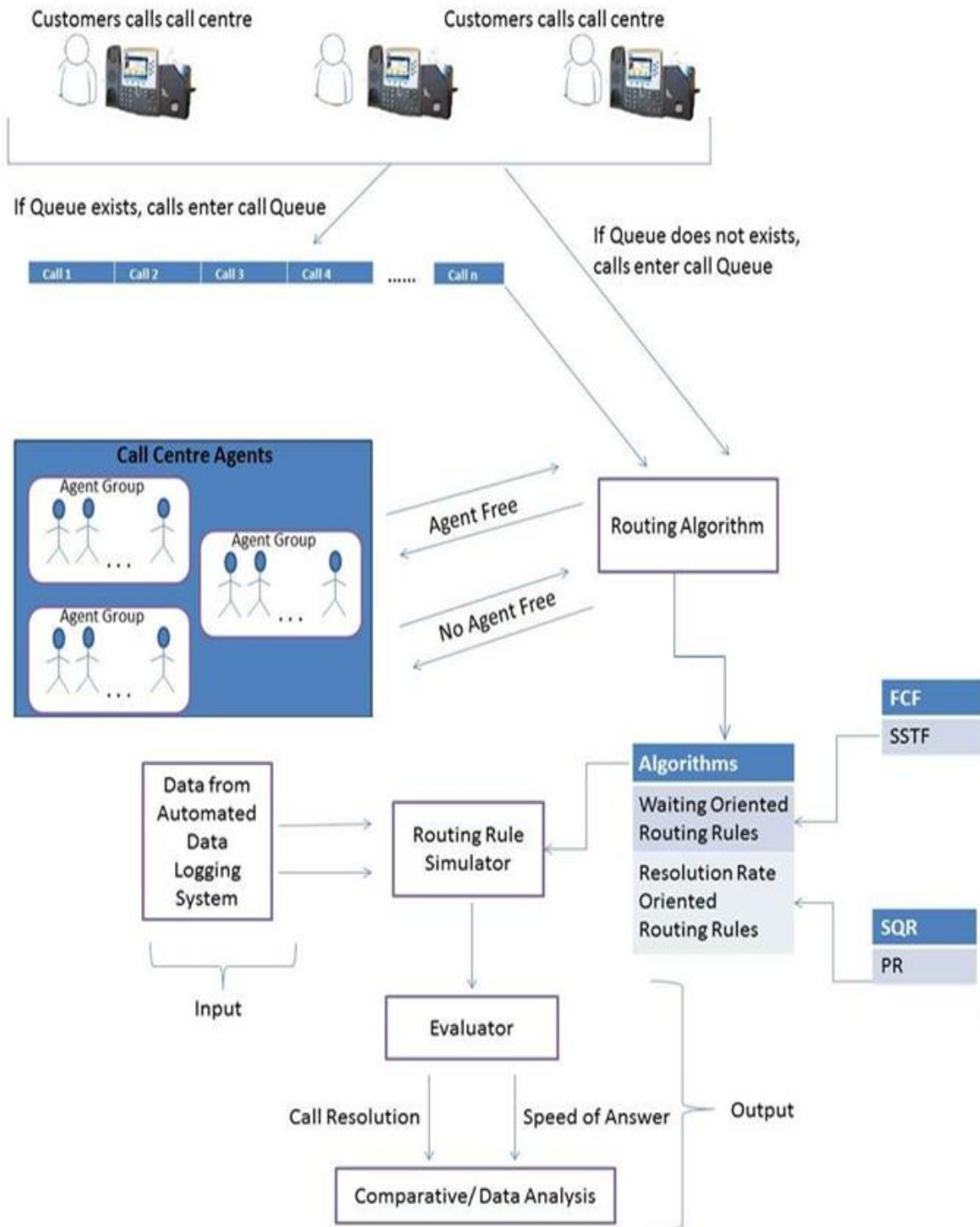
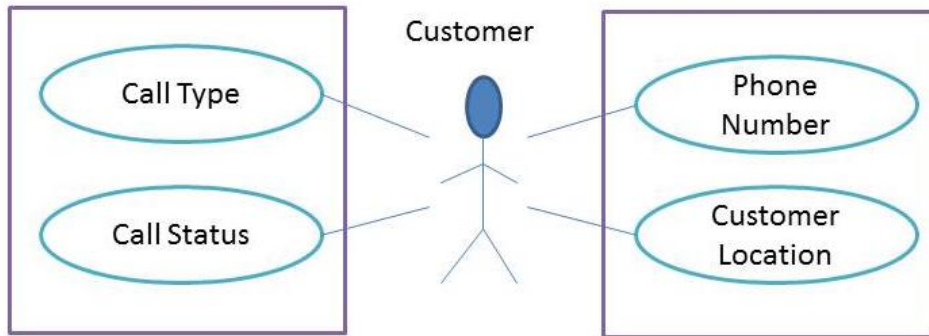


Figure 1: Overall System Approach

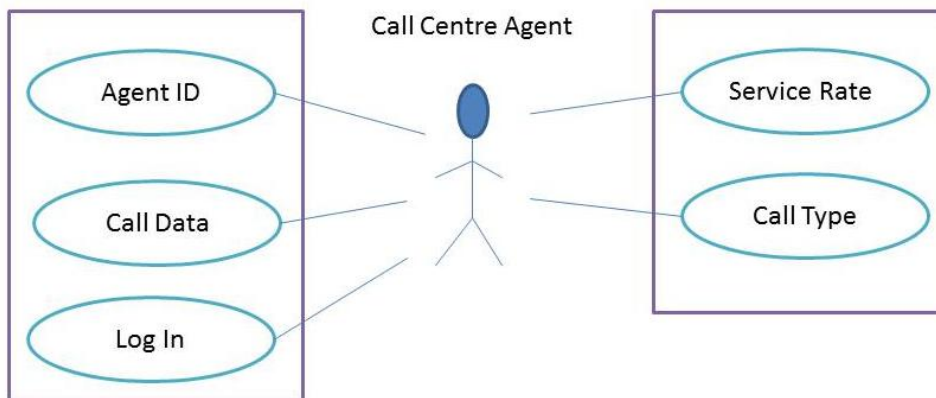
3.2 Use-Case Diagrams

The figures 2 and 3 below shows the main actors in our call centre study which comprises of the customer, call centre agents and the system itself. Customers essentially make specific call types to call centres to make complain or

enquiries. The call is considered to be a new (fresh call) or a call back. On receiving the call, call centre agent requires the customer location and Phone number for record and authentication purposes.



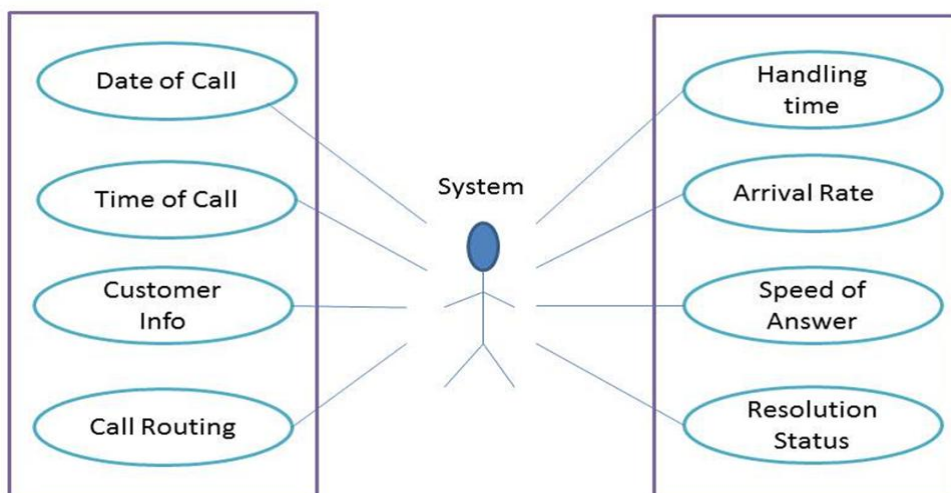
Figures 2: Customer actor and Attributes



Figures 3: Call Centre Agent actor and Attributes

The use case below in figure 4 show the system functions as it is responsible for routing customer calls to available agents using predefined routing rules. These rules will be highlighted later in subsequent section. The system also records the

handling time for each call, each call arrival rate, resolution status, speed of answer, etc. which are required for computational analysis in order to test the viability of the routing rule.



Figures 4: Call Centre System Actor and Attributes

Call centre agents are saddled with the responsibility of attending to customer issues. Due to the volume of customer calls, most call centres employ multiple agents to attend to customer issues. Every call centre agent has a unique identification number which helps managers to monitor the progress of each agent and for regular appraisal. Call centres have agent groups who comprises of agent with special

trained skill set for handling specific problems ranging from device platform issues to service related issues. The service rates of agents are also recorded. Call centre agents are expected to observe that they are logged into the system and that the system is recording call data such as call date, call time, etc. The conceptual model shows more call data attributes.

3.3 Conceptual Model

A conceptual model illustrates abstract and meaningful concepts in the problem domain. The aim of this step is to

decompose the problem in terms of individual concepts or object. Figure 5 below depicts the system conceptual model:

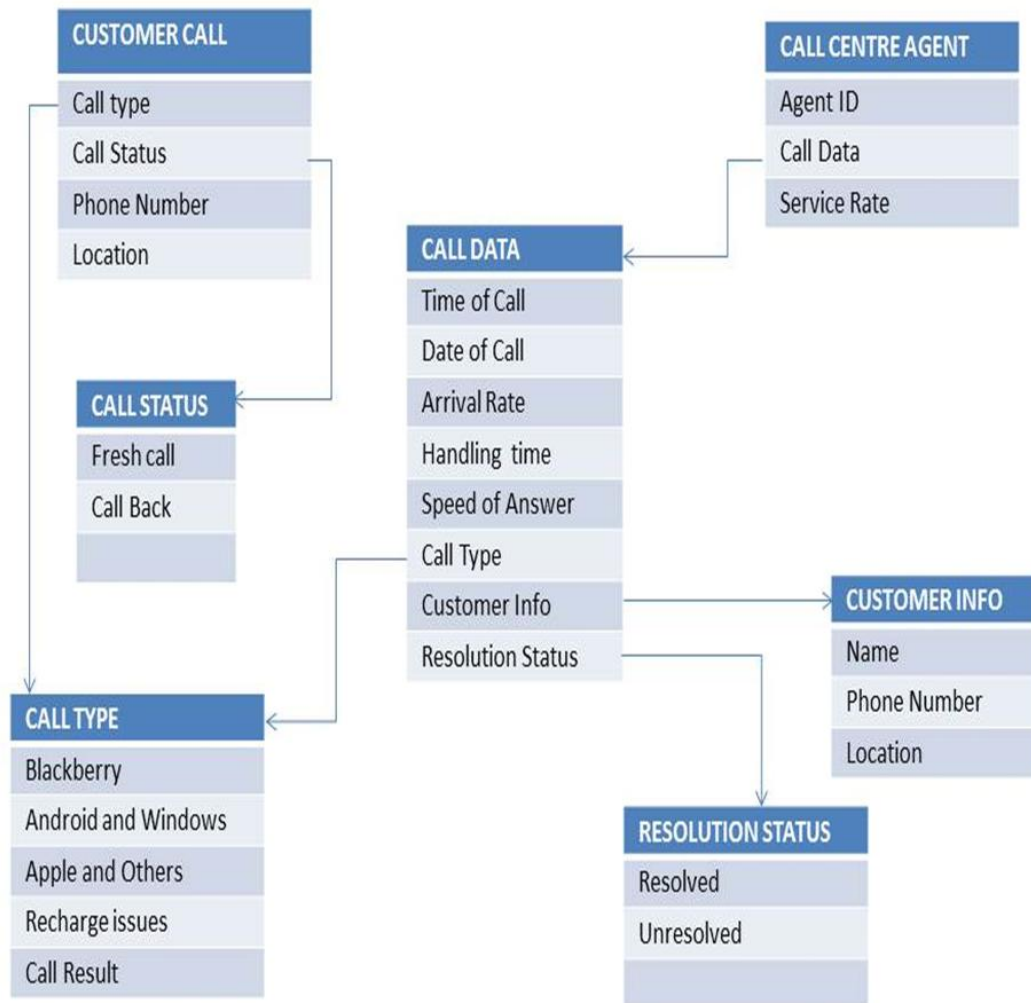


Figure 5: Conceptual model showing attributes

3.4 Definition of Routing Rule

As adapted from [7], our benchmark routing rule will be the First-Come-First-Served/Longest-Wait (FCFS/LW) rule, which we specify as follows.

(1) First Come First Serve/ Longest Waiting (FCFS/LW): When a call arrives and finds no calls of that type in queue and agents of one or more matching group available assigns that call to the agent who has been free the longest, regardless of his/her group.

Below, we introduce several other routing rules whose performance we will compare to that of FCFS/LW.

3.4.1 Waiting-Time Routing Rules

(2) Fastest Call First Rule (FCF): A call of a particular type that arrives when agents of multiple matching groups are free will be routed to a matching agent group that has the highest service rate for that call type.

(3) Shortest Service Time First (SSTF): A call of a particular type that arrives when agents of multiple matching groups are free will be routed to a matching agent group that has the relative Shortest Service Time for that call type.

3.4.2 Resolution Probabilistic Routing Rules

(4) Shortest Queue Routing (SQR): A call of a particular type that arrives when multiple agents are free will be routed to an agent from the group that has the shortest queue for that call type.

(5) Probabilistic Routing (PR): A call of a particular type that arrives when multiple agents are free will be routed to an agent from the group that has the highest resolution probability for that call type.

4. RESULTS AND DISCUSSION

Having proposed diverse set of routing rules in section 3, the performance of these routing rules is defined in terms of the two key performance metrics of overall average speed of answer (ASA) and aggregate call resolution (CR) rate. This work attempts to identify which of these rules delivers the best operational performance knowing full well that different call centre managers are likely to put different weights on each of the two key performance measures.

4.1 Simulation Process

The simulation analysis was carried out by implementing the routing rules using a collection of Java programs was used for simulation. The simulation was executed for 2000 calls by multiple agent group and multiple call types. Each routing rule was implemented separately and the handling time for each call is noted. The required output from each routing rule is the speed of answer for various calls and the call resolution. The results are then aggregated by weighted averaging method to obtain the Average Speed of Answer (ASA) and the Call Resolution rate (CR). The call resolution rate is defined by the ratio of the total number of calls resolved by an agent to the runtime of the simulation Program.

4.2 Data Collection and Database Preparation

The data collected for this study is automated and machine generated from the Call Centre data logging system. Data drawn from the organization’s electronic database was for a period of one (1) month, September 2015. This data contained information about the volume of calls received, who handled the calls and how they were handled

4.3 Program Structure

The application is a standalone application. On executing the program, the screenshots below in figure 6 reveal the results from various routing rules.

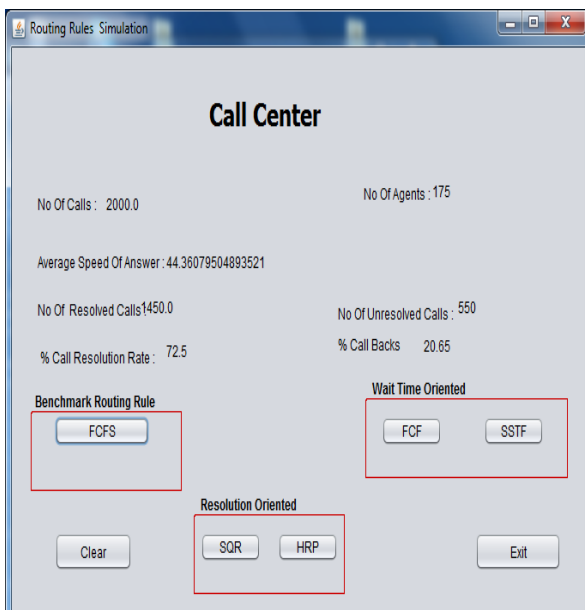


Figure 6: Screen shot of simulation analysis

4.4 Simulation Results and Analysis

Because of the number of rules examined in our simulation study, the results were organized into several sets of numerical comparisons. These comparisons are based on mean ASA and CR rates that are computed as weighted averages across the call types.

4.4.1 Comparisons within Rule Groups

Waiting-Time Routing Rules: Fig 7 and 8 presents the ASA values and CR rates for each of the various waiting-time rules described in chapter 3 as well as the benchmark **FCFS/LW** rule. While each of the waiting-time rules results in significantly lower ASA values than the **FCFS/LW** value, there are significant differences in CR rates across these rules.

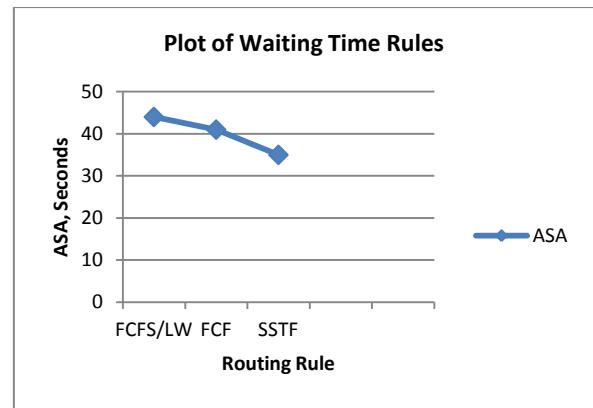


Figure 7: ASA Based Waiting-Time Routing Rules

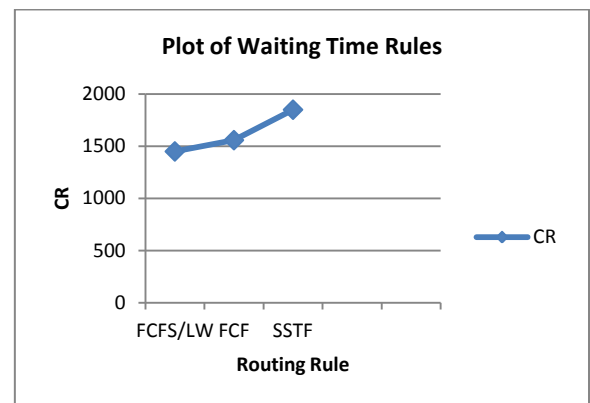


Figure 8: CR Based Waiting-Time Routing Rules

The focus of the **FCF** rule is clearly on getting calls out of the system as quickly as possible. However, this rule is myopic in the sense that it completely neglects the resolution probabilities. As a result, the CR rate associated with **FCF** rule is lower than the **SSTF** rules, a non-trivial difference which translates to a significant gap in customer satisfaction and loyalty. We note that this lower CR rate results in an increase in system congestion that drives up the mean waiting time under the **FCF** rule.

Resolution probability Routing Rules: Figure 9 and 10 presents the ASA values and CR rates for each of the various resolution probability rules described in chapter 3 as well as the benchmark **FCFS/LW** rule.

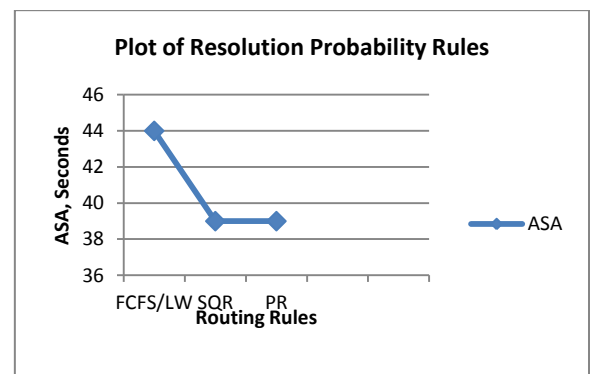


Figure 9: ASA Based of Resolution probability Rules

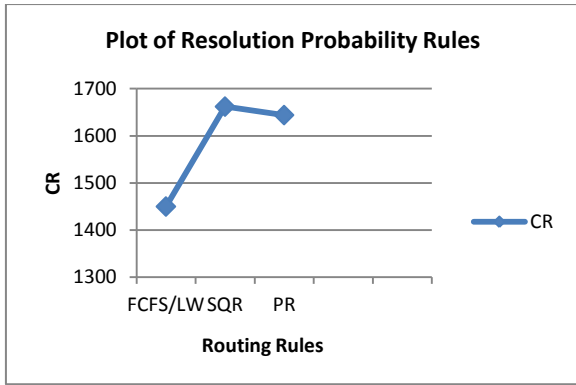


Figure 10: CR Based Resolution probability Rules

The results in this graph lead to several insights. First of all, it is interesting to note that the **SQR** and **PR** rule produces nearly identical results to one another. In addition, **SQR** and **PR** rules produces ASA values far lower than **FCFS** and they also dominate the benchmark **FCFS/LW** rule.

On the surface, the **SQR** and **PR** rules are greedy in the sense that they route according to the maximum resolution

probability with no forward-looking consideration. As such it was hypothesized that from a resolution probability perspective, there may exist situations where it may be better to hold a call and wait for a better matching (in terms of call resolution) agent to become idle.

In addition, because these two rules make their routing decisions based solely on the resolution probability (RP) without consideration of the service rates, they run the risk of routing calls to agents with long call handling times, which would have the effect of increasing waiting time. A mitigating factor is that by aiming to maximize call resolution, these rules also reduce the number of customer callbacks, which has the effect of reducing the overall system load and thus dampening the average waiting time.

4.5 Objective Actualization

4.5.1 Objective 1

How can routing rules achieve a balance between the two goals of low handling time and high call resolution rate?

Table 1: Weighted Average Results obtained from simulation Analysis

RULE	CR	ASA	Non CR	RESOLVED CALLS	CALL BACKS	% resolved calls	% Call backs
FCFS/LW	1450	44	550	0.402777778	0.152777778	67.12962963	25.46296296
FCF	1558	41	442	0.432777778	0.122777778	72.12962963	20.46296296
SSTF	1850	35	150	0.513888889	0.041666667	85.64814815	6.944444444
SQR	1662	39	338	0.461666667	0.093888889	76.94444444	15.64814815
PR	1644	39	356	0.456666667	0.098888889	76.11111111	16.48148148

Table 1 above presents the result for various waiting-time and resolution oriented routing rules as well as the benchmark **FCFS/LW** rule. While each of the waiting-time rules results in significantly lower ASA values than the **FCFS/LW** value, there are significant differences in CR rates across these rules.

The focus of the **FCFS** rule is clearly on getting calls out of the system as quickly as possible. However, this rule completely neglects the resolution probabilities. As a result, the CR rate associated with **FCFS** rule is lower than the **SSTF** rules, a difference which translates to a significant gap in customer satisfaction and loyalty. It was observed that this lower CR rate results in an increase in system congestion that drives up the mean waiting time under the **FCFS** rule.

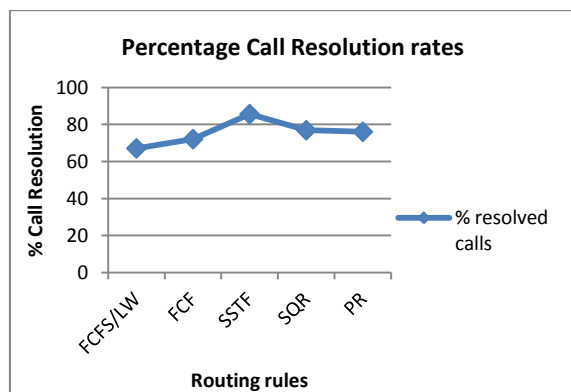


Figure 11: System Call Resolution rates of all the rules

First of all, the **SSTF** rule produces the highest CR rates. This result is not surprising as it seeks to explicitly maximize the overall CR rate. However, comparing **SSTF** with **PR** which has higher CR rate, it was observed that **PR** results is a far higher ASA value than **SSTF** and any of the others. For call centers that place a much higher value on successful call resolution than on customer waiting time, **PR** may be an attractive rule; for all other call centers, however, the incremental gain in the CR rate comes at a significant cost in terms of ASA.

Next, it is interesting to note that the **SQR** and **PR** rules produce nearly identical results to one another. In addition, these rules not only produce ASA values far lower than **RRPR**; they also dominate both the benchmark **FCFS/LW** rule. On the surface, the **SQR** and **PR** rules are greedy in the sense that they route according to the maximum resolution probability. As such from a resolution probability perspective, there could be situations where it may be better to hold a call and wait for a better matching (in terms of call resolution) agent to become idle. This is what rule **PR** does, producing a CR rate that is higher than either **SQR**.

In addition, because these two rules make their routing decisions based solely on the resolution probability (RP) without consideration of the service rates, they run the risk of routing calls to agents with long call handling times, which would have the effect of increasing waiting time. A mitigating factor is that by aiming to maximize call resolution, these rules also reduce the number of customer callbacks, which has

the effect of reducing the overall system load and thus dampening the average waiting time.

Finally, it is worthy of note that while the **SSTF** rule features the lowest ASA, it also results in a higher CR rate than **SQR**, which suffers only slightly higher ASA values. Taken together, these results clearly demonstrate that intelligent routing decisions can have a significant positive impact on operational performance regardless of a call center's relative weighting on call resolution rates and customer waiting times. Hence, by focusing on the wait time and resolution parameters, the **SSTF** and **SQR** rules actually perform well on both ASA and CR metrics. In fact, we infer that **SSTF** is more optimal in both ASA and CR than **SQR**. These results imply that for our dataset, **SSTF** as a better balance between minimal handling time and call resolution and hence provides much more important parameters for making routing decisions.

4.5.2 Objective 2

Are wait-time oriented routing rules superior to resolution rate oriented rules or vice versa?

With the foundation of research question 1, focusing on the call resolution parameter as the primary basis for routing decisions, **SSTF** and **SQR** has two possible consequences, which are that the overall system CR is high and that call-agent matches based on call resolution alone may result in long average service times, leading to long waits and a higher ASA value.

It is clear that the first objective is achieved by **SSTF** and **SQR** with a CR of 1935 and 1795 respectively. Moreover, a side benefit of high system CR is that callback volume is reduced. This lowers the effective load on the system and tends to offset the second consequence of long service times; therefore the overall ASA does not increase much.

It is also obvious that the overall callback rate to the system, as a percentage of the original arrival rate (0.6/Sec), is lower under **SSTF** than under **SQR**. Table 2 below shows the overall callback rate to the system, as a percentage of the original arrival rate for all the rules.

Table 2: System CR and callback rates of considered routing rules

RULE	CR (%)	CALL BACKS RATES (%)
SSTF	85.64814815	6.944444444
SQR	76.94444444	15.64814815

Considering **SSTF** and **SQR**, it suffice to say that wait time oriented rules are superior to resolution oriented routing rule. It should be noted that considering other rules, **SQR** is a better routing rule compared to **FCF** thereby suggesting that resolution oriented rules are superior to wait-time oriented routing rule. It is worthy of note that **SQR** results in lesser call backs compared to **FCF**. Therefore, the results show that neither waiting-time nor resolution probability rules are superior to each other; it is subjectively dependent on the value the call centre places on either of the rule. Every call center manager must decide where her priorities lie in terms of customer waiting times and call resolution.

5. CONCLUSION

The experimental results deliver several important insights and at the end of collating and analysing the data, it was discovered that several of our routing rules dominate the benchmark **FCFS/LW** rule, revealing that there is considerable value to making use of detailed agent performance information to drive routing decisions. It can also be inferred that neither waiting-time nor resolution oriented rules are superior to each other; it is subjectively dependent on the value the call centre places on either of the rule. Either of **SSTF** OR **SQR** routing rules on its own would not produce the desired result but a combination of both would help produce the best operating result. Finally, by comparing routing rules, demonstrations were presented to help managers understand the trade-offs between ASA and CR rates and to identify the routing rules that will most effectively produce the desired results.

It is recommended that every call center manager must decide where their priorities lie in terms of customer waiting times and call resolution. The result shows that several of our routing rules dominate the benchmark **FCFS/LW** rule. It should be noted that either of the routing rules- Waiting time or Resolution probability on its own would not produce the desired result without some measure of bias. It is recommended that a combination of both would help produce the best operating result; hence a hybrid rule is proposed. Managers should understand the trade-offs between handling time and call resolution rates so as to identify the routing rules that will most effectively produce the desired results.

In closing, several extensions to the work carried out in this research, is proposed as this is a very promising research direction. For environments with multiple call types, there are also clearly issues about which agents to train to handle which types of calls when both customer waiting times and call resolution rates are considered. While there has been a significant amount of research on skill-based routing and agent pooling, future research can be done to consider the impact of such rules on CR rates when different agent groups have different Average Handling Time (AHT) and Resolution Probability (RP) values for different call types.

6. REFERENCES

- [1] Armony .M 2005, Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Systems*, 51(3-4):287–329.
- [2] Francis de V´ericourt and Yong-Pin Zhou. 2005. A routing problem for call centers with customer callbacks after service failure. *Operations Research*, 53(6):968–981.
- [3] Brizola, N. , S. W. Costa, T. A. Pazeto, and P. J. F. Freitas. 2001. Planejamento de Capacidade de Call Center. In : ICIE, Flo-riano´polis.
- [4] Buist E. and L'Ecuyer P. 2005. A java library for simulating contact centres. In *Proceedings of the 2005 Winter Simulation Conference*, pages 556–565, Orlando, Florida, USA.
- [5] Gans, N., Koole, G., & Mandelbaum, A. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* , 5(2):79-141.
- [6] Ryder .G. 2009. Routing to Develop Expertise in Customer Contact Centres. PhD thesis, University of California, Santa Cruz.

- [7] Mehrotra .V, Ross .K, Ryder .G and Zhou .Y 2009. Routing to Manage Resolution and Waiting Time in Call Centers with Heterogeneous Servers.
- [8] Lawrence Brown, Noah Gans, Avishai Mandelbaum, Anat Sakov, Haipeng Shen, Sergey Zeltyn, and Linda ZHAO. 2005. Statistical Analysis of a Telephone Call Center: Journal of the American Statistical Association. Vol. 100, No. 469, Applications and Case Studies DOI 10.1198/016214504000001808
- [9] Luiz Augusto G. Franzese, Marcelo Moretti Fioroni Rui Carlos Botter Paulo José de Freitas Filho. 2009. Comparison of call center models. Proceedings of the 2009 Winter Simulation Conference
- [10] Marco Aurélio Carino Bouzada. 2009. Journal of Operations and Supply Chain Management 2 (2), pp 34 - 46, C International Conference of the Production and Operations Management Society
- [11] J'ul'ius Atlason, Marina A. Epelman, Shane G. Henderson. 2005. Optimizing Call Center Staffing Using Simulation and Analytic Center Cutting-Plane Methods.
- [12] Pierre L'Ecuyer. Modeling and optimization problems in contact centers. 2006. Proceedings of the Third International Conference on the Quantitative Evaluation of Systems - (QEST'06), pages 145–154.