Service Engineering of Call Centers: Research, Teaching, Practice

Avishai Mandelbaum
Professor*
avim@tx.technion.ac.il
(972) 4-8294504

Sergey Zeltyn
Postdoctoral Fellow*
zeltyn@ie.technion.ac.il
(972) 4-8292333

ABSTRACT

A course on Service Engineering has been taught at the Technion for over ten years [19]. Some unique aspects of the course are the incorporation of state-of-the-art research and real-world data in lectures, recitations and homework. Currently, the application focus of the course is telephone call centers, which constitute an explosively-growing branch of the service industry. Indeed, due to their practical importance and the diversity of their operational problems, call centers provide numerous challenges to Service Sciences, Engineering and Management.

In this contribution, we discuss significant research directions in the field of Service Engineering of Call Centers. The role of measurements and data collection at the individual-call level is emphasized. We describe software tools and databases that have been developed at the Technion in order to analyze operational performance of call centers and facilitate their statistical analysis. This prepares the ground for a survey of our "Service Engineering" course, with which we conclude.

INTRODUCTION

Service Engineering is a newly emerging scientific discipline ([11],[17],[19],[20]). As we perceive it, it caters to operational service-challenges that arise in our postindustrial society. To this end, researchers in the area develop scientifically-based engineering principles and tools, often culminating in software, which support the design and management of service operations. Moreover, a multi-disciplinary approach is called for in order to balance service quality, efficiency and profitability from the likely conflicting perspectives of customers, service-providers, managers and society. In our research, and clearly biased by our scientific roots, we focus on methodologies from *Operations Research* and *Statistics*.

In this note, we are concerned with ***Call Centers***. These are service organizations for customers who seek service via the phone. Due to advances in Information and Communication Technology, the number, size and scope of call centers, as well as the number of people who are employed there or use them as customers, grows explosively. Indeed, some estimate that, in the U.S. alone, the call center industry employs several million agents which, in fact, outnumbers agriculture.

The call center environment gives rise to numerous managerial challenges that differ in their nature and time-scale. For example, training and hiring problems should be solved on a yearly and/or monthly scale, staffing and scheduling is typically treated on a weekly/daily basis and Skill-Based Routing (SBR) decisions – matching of customers with telephone agents - should be exercised in real time. In the first part of this note, we discuss some central research directions, active and desirable, that can help to address these challenges.

In the second part, we describe a course on Service Engineering that has been taught at the Faculty of Industrial Engineering and Management, Technion, Israel. This is a compulsory course that is attended by over 120 students per year. Its goal is to provide students with knowledge and skills necessary for their future professional activities, accounting for the fact that they are likely to be employed in service enterprises that constitute the major part of the modern economy. Our experience has been that call centers provide an excellent motivational and training field for service-engineering methods; and that call-center real data, blended into lectures and homework assignments, perfectly complements active learning of these methods. There exists a publicly available Internet site of our "Service Engineering" course [19], and its materials have been used for research and teaching worldwide, both in academia and industry.

* Faculty of Industrial Engineering and Management, Technion, Haifa, 32000, Israel

OPERATIONAL MODELS IN CALL CENTERS: RESEARCH SURVEY AND PRACTICAL CHALLENGES

We focus on operational applications of Service Engineering methods in call centers. We do not discuss strategic problems, such as the development of new services or long-term workforce management. The reader is referred to [7] for a comprehensive survey of the state of research on telephone call centers.

Data Collection - a Prerequisite for Scientific Research

We strongly believe that systematic measurements and data collection are prerequisites for the analysis and management of any service system. In addition, detailed transaction-based measurements provide information that is inaccessible via aggregated (e.g. interval-based) summaries. Call centers are no exception.

Specifically, large call centers generate vast amounts of data. A detailed history of each call that enters the system can, in theory, be reconstructed via the Automatic Call Distributor (ACD) and Interactive Voice Response Units (IVR). However, call centers have not typically stored or analyzed this data, using instead the ACD reports that summarize performance over certain time intervals (say, 30 minutes). We advocate the change of this approach and emphasize the practical and research advantages of call-by-call data analysis. In [4] we applied our approach in a comprehensive analysis of a small Israeli call center and continuing research on larger call centers is currently underway.

DATA-MOCCA - database of call-by-call measurements

Call center data is processed by vendor-specific programs, in formats that are not amenable to operational analysis. DATA-MOCCA (DATA MOdel for Call Center Analysis) [21] has been developed to address these shortcoming. It is a universal model for call center data that, together with a graphical user interface, enables real-time statistical analysis at second-to-month resolutions. Currently, DATA-MOCCA covers call-by-call data of two large call centers, a U.S. bank and an Israeli cellular-phone company, over periods of 2-3 years each. (For example, the U.S. bank data has close to 120 million calls, out of which about 40 million were served by agents and the rest by a VRU – Voice Response Unit.) The raw data for DATA-MOCCA is dumped by commercial routing and call recording systems. Transforming it into our universal format takes a significant data-cleaning effort. This effort has been partially funded by the IBM Academic Fellows program, with the ultimate goal being the creation of a data-repository that is publicly accessible via the Internet, and which draws data from industries such as Financial, Telecommunication, Healthcare, Hospitality, etc. Till then, researchers and practitioners can ask the first author (AM) for the data and its accompanying software.

Forecasting Arrival Rate

The standard model of call arrivals to a call center has been the time-inhomogeneous Poisson process, which accommodates both predictable and stochastic demand variability. Statistical analysis [4,7] shows that this model provides a very good approximation to reality. However, prediction of future arrival rates, being a crucial first step for staffing decisions, turns out to be a complicated statistical task.

Two research directions are important in this regard. First, *time series* prediction techniques should be enhanced. Different methods could be appropriate for predictions that are performed weeks-ahead, days-ahead or hours-ahead. In addition, a specific call center often has unique features for its call arrivals (e.g. monthly bills sent by a cellular-phone company imply surges of incoming calls following billing cycles). Taking these features into account would significantly improve prediction accuracy.

Second, in certain circumstances one should accept the fact that there exist significant uncertainty and temporal correlation in the arrival-rates themselves. Appropriately, models with *random correlated arrival rates* must be employed, in contrast to the classical queueing models where the arrival rate is assumed known (deterministic).

Service Time: Definition and Modeling

The service time in call centers is typically defined as the time that an agent spends handling a call. It must include the talk time between an agent and a customer, as well as times on hold, after-call work, etc. If λ denotes the arrival rate per time-unit and E[S] is the mean service time, their product $R=\lambda \cdot$ E[S] is called the *offered load*. It is the basic quantity needed for staffing decisions, as discussed below. (λ is assumed here constant for simplicity; later we address time-varying rates.) The most widely used parametric model of service times is the *exponential* distribution. However, the *lognormal* distribution seems to provide an excellent fit for the call centers that we have analyzed recently [1]. Since models with exponential service times are much more tractable analytically than their alternatives, and since even seconds of service durations could have significant economic impact, the effect of the service distribution on performance of queueing systems should be carefully studied (see [23]).

Impatience and Abandonment

Until recently, most call centers used the classical M/M/n queueing model, also called *Erlang-C*, in their staffing. Erlang-C assumes Poisson arrivals at a constant rate λ, exponentially distributed service times with a rate μ, and n independent statistically-identical agents. However, Erlang-C does not acknowledge customers' abandonment and consequently can depict a distorted picture of a call center's operation [8,15]. For example, even a minor abandonment rate in a heavily-loaded system can improve waiting times of those who do not abandon by orders of magnitude. This improved operational performance must be traded off

against customers' frustration and lost business due to abandonment. Nowadays, an increasing number of call centers incorporate customers' abandonment in their staffing/scheduling software and performance goals.

The Erlang-A (Palm) Model

The theoretically simplest and practically most feasible way to account for customers' impatience is the following: in addition to the Erlang-C assumptions described above, suppose that each arriving caller is equipped with an exponentially distributed patience time. Customers abandon when their required waiting exceeds their patience. This model, first introduced by Palm [18], will be denoted by M/M/*n*+M and referred to as Erlang-A (A for Abandonment). See [15] for a recent summary and [6] for software that enables calculations and staffing according to Erlang-A.

Operational Regimes

A central challenge in the design and management of a service operation in general, and of a call center in particular, is to achieve a desired balance between *operational efficiency* and *service quality*. Here we consider the staffing aspects of this problem, namely having the right number of agents in place. "The right number" means, first of all, not too many, thus avoiding overstaffing. This is a crucial consideration since personnel costs typically constitute about 70% of the costs of running a call center. "The right number", however, also means not too few, thus avoiding understaffing and consequent costs associated with poor service quality. We now present two approaches to the staffing problem, both within the framework of Erlang-A.

***Q**uality and **E**fficiency **D**riven (**QED**) Regime*

This operational regime is governed by the so-called *Square Root Staffing Rule*:

$$n \approx R + \beta\sqrt{R}\,, \qquad -\infty < \beta < \infty;$$

where $R=\lambda \cdot E[S]$ is the offered load defined above and β is a Quality-of-Service (QoS) parameter. This rule was first used by Erlang (at the Copenhagen Telephone Company) close to 100 years ago. However, a formal QED analysis for various queueing systems appeared much later. The pioneering work is [9] that analyzed Erlang-C (β then must be positive); Erlang-A was considered in [8].

It turns out that if the number of servers *n* is not small, QED staffing enables high levels of *both* Efficiency (utilization of agents, say, around 90-95%) and service Quality (say, 50% of the customers are served immediately upon calling, average wait is 5-10 seconds, and abandonment rates are 1-3%). The QED regime arises also as economically optimal when minimizing the sum of staffing costs and waiting costs [3].

Efficiency-Driven (ED) Regime

Another common operational regime is characterized via the staffing rule $n \approx R - \gamma R, \ (0<\gamma<1)$. In this case, virtually all customers are delayed prior to being served and, approximately, the fraction abandoning is γ. The ED regime is to be used if efficiency concerns dominate those of service quality; for example, this is common practice in not-for-profit environments.

Stationary vs. Time-Dependency

A standard approach to staffing decisions in call centers is to break the day of work into short time-intervals (usually 15 or 30 minutes), assume that the Poisson arrival rate is constant over these intervals and apply stationary queueing models (e.g. Erlang-A) in order to determine how many servers are needed during each interval. Although this approach seems adequate for many call centers, it cannot capture the performance of highly time-varying systems. In the latter case, one should resort to models with time-dependent arrival rates. See [5] for an adaptation of the square-root staffing rule to time-varying arrival rates.

Staff Scheduling and Agents Assignment

As mentioned, staffing problems are typically solved by using steady-state models over short time intervals, separately. In practice, however, individual service agents are typically assigned to shifts (say, 8 hours including breaks) where the duration and location of breaks is constrained by trade-union agreements. This setting gives rise to two separate problems. First, one should determine the timing of shifts and the number of agents working during each shift, while satisfying also the staffing requirements considered above. This problem is typically solved by Integer Programming. Second, individual agents must be assigned to shifts. Here the complexity of the problem renders it analytically intractable and, hence, one resorts to heuristic techniques. (One could also attempt "shift bidding", where the employees themselves state their preferences and are then assigned to shifts according to their ranking, taking into account priorities – for example seniority – and systems constraints.)

Skills-Based Routing (SBR)

SBR technology enables the differentiation of many types of customers/calls and many skills of agents. Segmenting customers is a marketing task, while agent segmentation is human-resource-management. The need for type-skill matching suggests new types of operational challenges. For example, at the real-time level, one should manage the so-called *agent selection* and *call selection* problems, choosing to which free agent should an arriving call be routed, if any, and which waiting call should be attended by an agent who becomes idle, if any. In addition, multi type/skill environments significantly complicate the staffing and scheduling problems discussed above. SBR in the ED regime is relatively tractable [2,13]. However, QED SBR is the subject of intense research [1]. Readers are referred to [7] and [22] for more details.

Human Behavior

One of the most challenging aspects in the modeling of call centers is the incorporation of human factors, for both customers and agents. This opens up a vast agenda for multi-disciplinary research, involving psychology, marketing, operations research and statistics. Below we present two relevant examples from our studies on call centers.

Short Service Times

Figure 1 shows the empirical distribution of service times in a call center of an Israeli bank during July, 1999. We observe a peak of very short service times: more than 7% of the calls were shorter than 10 seconds. These short calls were due to certain agents who were taking "rest breaks" by hanging up on customers. At the end of October, the problem was discovered and corrected. Figure 2 reflects the data of December, 1999: no peak is observed and, moreover, the lognormal distribution provides an excellent approximation to the empirical data.

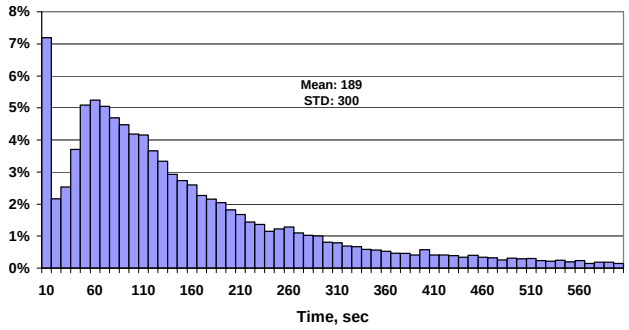

Figure 1. Histogram of service times in an Israeli call center. July 1999.

The problem of agents "abandoning" their calls can arise when short service durations (or many calls per shift) are a prime performance objective. The problem becomes immediately apparent from a histogram in Figure 1, based on call-by-call data. However, it can be hardly discovered through the prevalent standard of reporting only half-hour averages.

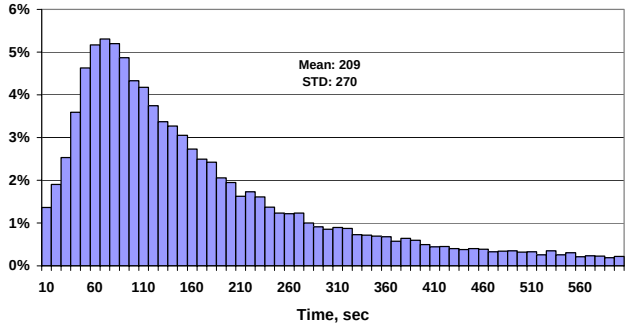

Figure 2. Histogram of service times in an Israeli call center. December 1999.

Psychology of Customer's Impatience

Figure 3 presents empirical hazard rates of patience times for Regular and Priority customers in an Israeli call center. The value of the hazard rate at time t is proportional to the likelihood to abandon during a short time interval after t seconds of wait, given that the customer already waited t seconds. (See [7] for detailed explanations.)

Figure 3 provides one with two important observations. First, priority customers turn out to be more patient than regular customers. This could be the reflection of a more urgent need on the part of priority customers; or could be an evidence of their higher level of trust that they will be served soon after arrival. Second, both functions have peaks of abandonment around 10-15 and 60 seconds, which turns out to reflect two announcements to customers: upon joining the queue and for those who have waited one minute, respectively. The announcements inform customers on their relative position in the tele-queue. This phenomenon gives rise to important questions. Do, in fact, announcements encourage abandonment, which could be in contrast to their original goal? Do they, on the other hand, provide customers with an opportunity to take a rational decision concerning abandonment which could decrease frustration and, probably, overall abandonment? (In principle, announcements could imply larger immediate abandonment but smaller abandonment during the periods between announcements.)

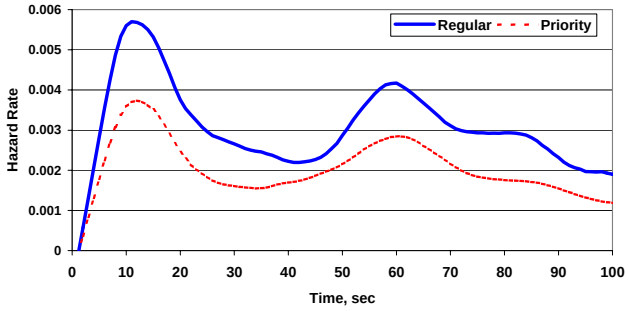

Figure 3. Empirical hazard rates of patience time in an Israeli call center.

Integration of theoretical, field and laboratory studies is needed in order to answer these questions, as well as many similar ones. See [16], for an example of a psychological study that is based on laboratory experiments.

Operational Models and Customer Relationship (Revenue) Management (CRM)

CRM automatic systems promise to enable companies to better track and understand the service experience of their customers, and then analyze its effect on the long-term relationship with the company (e.g. purchasing behavior, amenability to cross-selling). The interaction of our operational models with CRM could, hopefully, manifest itself in the ability to answer questions such as: "How a change in the service process (e.g. adding an agent for answering calls) affects company revenues?"

THE "SERVICE ENGINEERING" COURSE AT THE TECHNION

Many of the issues discussed above are taught, or at least addressed, at the Technion's "Service Engineering" (ServEng) course [19]. The course has been taught for over ten years at the Faculty of Industrial Engineering and Management (IE&M). It started as a seminar for graduate students and has gradually developed into the present undergraduate compulsory course, taught almost each semester and attended by over 120 students yearly. Its site [19] contains course materials (lecture notes/slides, recitations, homework), related research papers, slides of seminars, software and databases.

Teaching Goals

Although the service industry generates more than 70% of the GNP of many developed countries, prior to ServEng IE&M students had been exposed mainly to methods and techniques inspired by manufacturing applications. (This situation is likely to prevail among IE departments.) ServEng aims at filling this gap by providing students with appropriate models and tools for design, operation and analysis of service systems. Examples from various service sectors are presented at lectures and recitations, with the call center industry being the central application area.

Course Syllabus

The course has four general parts: 1. *Prerequisites*: measurements and models; 2. *Building Blocks*: demand, services, (im)patience; 3. *Models*: deterministic (Fluid) and stochastic - mainly queueing models, both conventional (Markovian) and approximations; 4. *Applications*: design, ED/QD/QED workforce-management and skills-based routing. The course's teaching philosophy was inspired by the book Hall [10], which serves as a recommended textbook.

M*easurements*, at the level of individual service transactions, are prerequisites for design, analysis and management of service systems. After a ServEng Introduction, we survey transactional measurement systems in face-to-face, telephone, internet and transportation systems. These measurements immediately give rise to deterministic (fluid/flow) models of a service station, which capture average behavior and enables relatively simply yet far-reaching analysis – for example, capacity (bottleneck) analysis. Then we proceed with an introduction to *Modeling*, using Dynamic Stochastic PERT models as a modeling framework that captures operational congestion, due to resource constraints and synchronization gaps.

The next segment is dedicated to three building blocks of a basic service-model. First we study *service demand*, emphasizing the importance of reliable forecasting techniques. (For example, arrivals of incoming calls to a call center are typically Poisson or Poisson-related.) Then we analyze the *service process*, describing models for service-durations. (For example, service durations in call centers "are" log-normally distributed [4].) We end with *customers' impatience* and its manifestation – the abandonment phenomena, which is important in call centers and other services (e.g. Internet and even Emergency Rooms).

The building blocks are now fused into basic queueing models where customers are i.i.d. and servers are i.i.d. A central role is played by Markovian Queues, emphasizing the applicability of the Erlang-A queue [15]. Then we discuss design principles (pooling to exploit economies of scale) and present operational workforce management techniques (staffing and scheduling), including staffing in the QED and ED operational regimes. We conclude the course with models that acknowledge customers differentiation (priorities) and servers heterogeneity/skills (SBR). An optional last lecture surveys queueing networks as models of multi-stage service systems.

Data-Based Teaching

ServEng students are trained with real-data and software. Early generations of the course used one-month tellers' data from a bank in Israel [14], in support of recitations and homework. Later, we added one-year call center data from another small bank [4]. The tellers' data has been since used in recitations while the telephone data in homework. DATA-MOCCA [21] currently serves in examples, lecture presentations and few homework assignments. As mentioned, we are in the process of making DATA-MOCCA publicly accessible and, then, we shall be able to incorporate it much more actively in the course.

The main software tool that students use is 4CallCenters [6]. This package, based on [8], allows them to solve staffing problems, using various queueing techniques that are inspired by call centers but are applicable more broadly (for example, to nurse staffing).

Our Service Engineering course is an ongoing R&D process. We already mentioned the incorporation of DATA-MOCCA. We are also planning to enrich near-future versions of the course with examples and techniques from health care and hospital operations management.

CONCLUSION

In this contribution, we surveyed possible applications of SSME in call centers and described the Technion's Service Engineering course. We believe that only such integration of data-based research, teaching and practice can provide the service industry with the necessary engineering tools as well as qualified specialists that are capable and trained to apply these tools.

We emphasize the need for multi-disciplinary approach to the Service Engineering problems [7]. For example, in order to understand and exploit the phenomenon of customers' abandonment in call centers, as described above, one should use Statistics and Operations Research to measure and model impatience, Psychology to understand and interpret customers' behavior, and Marketing to assess

the economical impact of abandonment. We hope that such cooperation between academic and industry researchers, from various branches of science, will provide solutions to the numerous challenges that arise in the Service Industry.

REFERENCES

1. Atar R. (2005) A diffusion model of scheduling control in queueing system with many servers. Annals of Applied Probability, 15(1B), 820-852.
2. Bassamboo A., Harrison J.M. and Zeevi A. (2006) Design and control of a large call center: Asymptotic analysis of an LP-based method. Operations Research, 54, 419-435.
3. Borst S., Mandelbaum A., and Reiman M. (2004). Dimensioning large call centers. Operations Research, 52(1), 17-34.
4. Brown L.D., Gans N., Mandelbaum A., Sakov A., Shen H., Zeltyn S. and Zhao L. (2002). Statistical analysis of a telephone call center: A queueing science perspective. Journal of the American Statistical Association (JASA), 100(469), 36-50.
5. Feldman Z., Mandelbaum A., Massey W. and Whitt W. (2005) Staffing of time-varying queues to achieve time-stable performance. Submitted to Management Science. Available at http://iew3.technion.ac.il/serveng/References/references.
6. 4CallCenters Software (2005). Available at http://iew3.technion.ac.il/serveng/4CallCenters/Downloads.htm.
7. Gans N., Koole G. and Mandelbaum A. (2003). Telephone call centers: a tutorial and literature review. Invited review paper, Manufacturing and Service Operations Management, 5(2), 79-141.
8. Garnett O., Mandelbaum A. and Reiman M. (2002). Designing a telephone call-center with impatient customers. Manufacturing and Service Operations Management, 4, 208-227.
9. Halfin S. and Whitt W. (1981). Heavy-traffic limits for queues with many exponential servers. Operations Research, 29, 567-588.
10. Hall R.W. (1991) Queueing Methods for Services and Manufacturing, Prentice-Hall.
11. IBM Research site. Service Sciences, Management and Engineering, http://www.research.ibm.com/ssme/
12. Mandelbaum A. (2006). Call Centers. Research Bibliography with Abstracts. Version 7. Available at http://iew3.technion.ac.il/serveng/References/references.
13. Mandelbaum A. and Stolyar A. (2004) Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. Operations Research, 52(6), 836-855.
14. Mandelbaum A. and Zeltyn S. (1998) Estimating characteristics of queueing networks using transactional data. Queueing Systems: Theory and Applications (QUESTA), 29, 75-127.
15. Mandelbaum A. and Zeltyn S. (2005) Service engineering in action: the Palm/Erlang-A queue, with applications to call centers. Invited chapter to *IAO* book project. Available at http://iew3.technion.ac.il/serveng/References/references.
16. Munichor N. and Rafaeli A. (2006) Numbers or apologies? Customer reactions to tele-waiting time fillers. To appear in the Journal of Applied Psychology. Available at http://iew3.technion.ac.il/Home/Users/anatr/JAP-Tele-wait-FINAL.pdf.
17. National Science Foundation. Service Enterprise Engineering (SEE) program. Available at http://nsf.gov/funding/pgm_summ.jsp?pims_id=13343&org=NSF&more=Y.
18. Palm C. (1957). Research on telephone traffic carried by full availability groups. Tele, Vol. 1, 107 pp.
19. "Service Engineering" course website, Technion, http://iew3.technion.ac.il/serveng.
20. Service research at the Fraunhofer Institute for Industrial Engineering. Available at http://www.management.iao.fhg.de/English/Overview.pdf.
21. Trofimov V., Feigin P., Mandelbaum A. and Ishay E. (2005) DATA-MOCCA: Data Model for Call Center Analysis. Technical Report, Technion. Available at http://iew3.technion.ac.il/serveng/References/references.
22. Wallace R.B. and Whitt W. (2005) A staffing algorithm for call centers with skill-based routing. Manufacturing and Service Operation Management, 7, 276-294.
23. Whitt W. (2005). Engineering solution of a basic call-center model. Management Science, 51(2), 221-235.