Queueing Models of Call Centers
An Introduction

Ger Koole[1] & Avishai Mandelbaum[2]
[1]Vrije Universiteit, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
[2]Industrial Engineering and Management, Technion, Haifa 32000, Israel

October 2001

Abstract

This is a survey of some academic research on telephone call centers. The surveyed research has its origin in, or is related to, queueing theory. Indeed, the "queueing-view" of call centers is both natural and useful. Accordingly, queueing models have served as prevalent standard support tools for call center management. However, the modern call center is a complex socio-technical system. It thus enjoys central features that challenge existing queueing theory to its limits, and beyond.

This paper (and its future updates) can be downloaded from www.cs.vu.nl/obp/callcenters and ie.technion.ac.il/serveng.

Acknowledgements

G.K. would like to thank Sandjai Bhulai and Geert Jan Franx for their useful comments on the very first version of this paper, and an anonymous referee (of a different paper) for pointing out some sources of which he was not aware.

Some of the writing was done while A.M. was visiting Vrije Universiteit - the hospitality of G.K. and the institutional support are greatly appreciated. A.M. thanks Sergey Zeltyn for his direct and indirect contribution to the present project: Sergey helped in the preparation of the figures and tables, and he is the co-producer of the material from ie.technion.ac.il/serveng which was used here. Thanks are also due to Sergey and Anat Sakov for their approval of importing pieces of [51]. Finally, the research of A.M. was partially supported by the ISF (Israeli Science Foundation) grant 388/99-02, by the Technion funds for the promotion of research and sponsored research, and by Whartons' Financial Institutions Center.

Contents

1 Introduction

Call centers, or their contemporary successors contact centers, are the preferred and prevalent way for many companies to communicate with their customers. The call center industry is thus vast, and rapidly expanding in terms of both workforce and economic scope. For example, it is estimated that 3% of the U.S. and U.K. workforce is involved with call centers, the call center industry enjoys a annual growth rate of 20% and, overall, more than half of the business transactions are conducted over the phone. (See `callcenternews.com/resources/statistics.shtml` for a collection of call center statistics.)

Within our service-driven economy, telephone services are unparalleled in scope, service quality and operational efficiency. Indeed, in a large best-practice call center, many hundreds of agents could cater to many thousands of phone callers per hour; agents utilization levels could *average* between 90% to 95%; no customer encounters a busy signal and, in fact, about half of the customers are answered *immediately*; the waiting time of those delayed is measured in seconds, and the fraction that abandon while waiting varies from the negligible to mere 1-2% (e.g., see Figures 4, 5 and 7). The design of such an operation, and the management of its performance, surely must be based on sound scientific principles. This is manifested by a growing body of academic multi-disciplinary research, devoted to call centers, and ranging from Mathematics and Statistics, through Operations Research, Industrial Engineering, Information Technology and Human Resource Management, all the way to Psychology and Sociology. (The bibliography [47] covers over 200 research papers.) *Our goal here is to survey part of this literature, specifically that which is based on mathematical queueing models and which potentially supports Operations Research and Management.* Along the way, we take the opportunity to point out some challenges, both theoretical and practical, that give rise to worthy (at least in our opinion) open research directions.

There already exist several academic surveys on call centers. (There are numerous survey papers in the business literature, which are not addressed here.) We are aware of five such articles: Pinedo et al. [56], giving basics of call center management, including some analytical models; Anupindi & Smythe [7], which describes the technology that enables current and plausibly future call centers; Grossman et al. [36] and Mehrotra [53], which are both short overviews of some OR challenges in call center research and practice; and finally Anton [6], who provides a managerial survey of the past, present and future of customer access (contact) centers.

One can possibly view our survey as a supplement to [56, 53], aimed at academic researchers that seek an entry to the subject, as well as to queueing theorists and practitioners who develop call center applications. The surveys [6, 7] add some interesting background information. Additional articles that we recommend as part of a quantitative introduction to call centers are Brigandi et al. [20], Mandelbaum, Sakov and Zeltyn [51], Evenson et al. [25] and Duxbury et al. [23]. The first one [20] presents a long-term effort that demonstrates the payoffs in call center modeling. The second [51], parts of which have been adapted to the present text, provides a statistical description of call center operations. The rest [25, 23] are managerial surveys of performance drivers and state-of-the-art. One could also add the book Cleveland and Mayben [22] as a practitioners' overview.

While the focus of our survey is mathematical queueing models of call centers, no model will actually be formulated here, but rather a multitude of models will be referred to, and only some described. Our conception, and hope, is that readers will be stimulated enough to pursue the details through their sources, namely the original research articles that are surveyed here.

The working environment of a call center

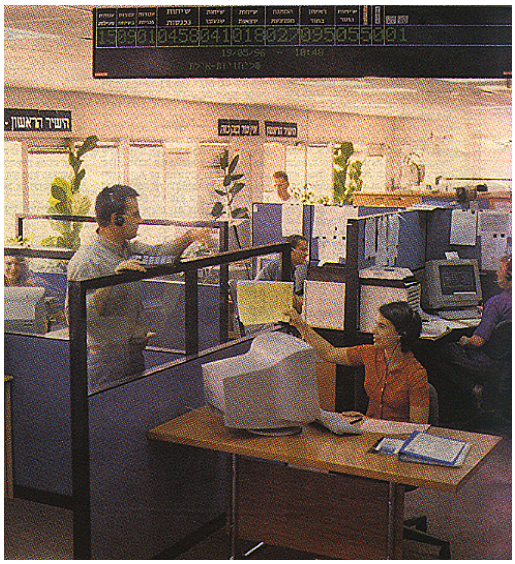

Figure 1: The working environment of a call center

1.1 What is a call center?

A *call center* constitutes a set of resources (typically personnel, computers and telecommunication equipment), which enable the delivery of services via the telephone. The working environment of a large call center (Figure 1) could be envisioned as an endless room with numerous open-space cubicles, in which people with earphones sit in front of computer terminals, providing tele-services to unseen customers. Most call centers also support Interactive Voice Response (IVR) units, also called Voice Response Units (VRU's), which are the industrial versions of answering machines, including the possibilities of interactions. But more generally, a current trend is the extension of the call center into a *contact center*. The latter is a call center in which the traditional telephone service is enhanced by some additional multi-media customer-contact channels, commonly VRU, e.mail, fax, Internet or chat (in that order of prevalence).

Most major companies have reengineered their communication with customers via one or more call centers, either internally-managed or outsourced. The trend towards contact centers has been stimulated by the societal hype surrounding the Internet, by customer demand for channel variety, and by acknowledged potential for efficiency gains (requests for e.mail and fax services can be "stored" for later response and, when standardized and well-managed, they can be made significantly cheaper that telephone services). *Our survey deals almost exclusively with pure telephone services.* Indeed, to the best of our knowledge, no analytical model has yet been dedicated to truely multi-media contact centers, though a promising framework (skill-based routing) and a few models that accommodate

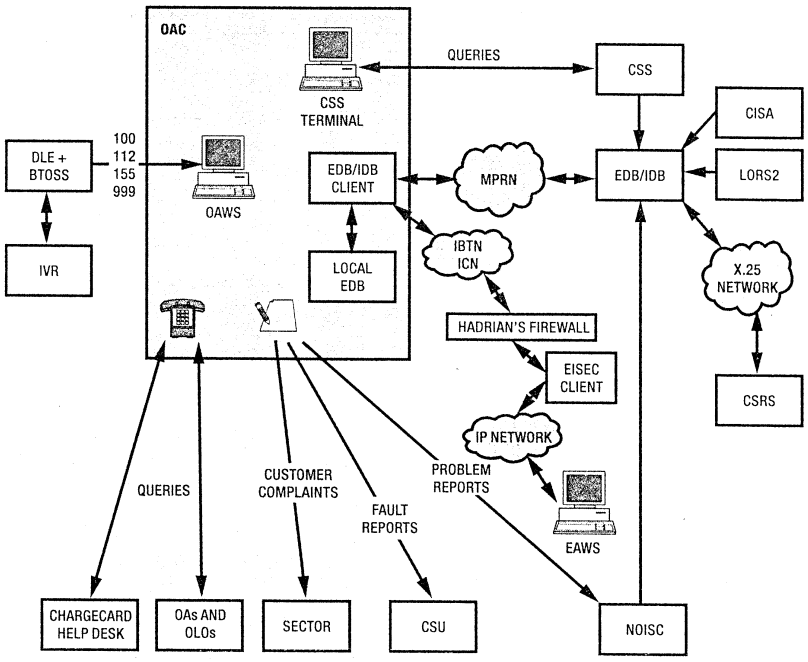

Figure 2: Call Center Technology (Operator Assistance in British Telecom UK [23])

IVR's, e.mails and their blending with telephone services, will be described later in Section 3.

1.2 Technology

The large-scale emergence of call centers, noticeably during the last decade, has been enabled by technological advances in the area of Information and Communication Technology (ICT). First came PABX's (Private Automatic Branch Exchanges, or simply PBX), which are the telephone exchanges within companies. A PABX connects, via trunks (telephone lines), the public telephone network to telephones within the call centers. These, in turn, are staffed by telephone *agents*, often called CSR's for Cusomer Service Representatives, or simply "rep's" for short. Intermediary between the PABX and the agents is the ACD (Automatic Call Distribution) switch, whose role is to distribute calls among idle qualified agents. A secondary responsibility of the ACD is the archival collection of operational data, which is of prime importance as far as call center research is concerned (see Subsection 2.2 below for an elaboration). While there exists a vast telecommunications literature on the physics of telephone-traffic and the hardware (technology) of call centers, *our survey focuses on the service contact between customers and agents*, sometimes referred to as the service's "moment of truth".

Advances in information technology have contributed as importantly as telecommunication to the accelerated evolution of call centers. (See Figure 2 for a concrete example.) To wit, rather than search

for a paper file in a central archive, that renders impossible an immediate or even fast handling of a task related to that file, nowadays an agent can access, almost instantaneously, the needed file in the company's data base. A new trends in ICT is the access of customer files in an automatic way. The relevant technology is CTI (Computer Telephony Integration), which does exactly what its name suggests. In fact, this can go further. Consider, for example, a customer who seeks technical support from a telephone help-desk. That customer can be often automatically identified by the PABX, using ANI (Automatic Number Identification). This triggers the CTI to search for the customer's history file; information from the file then *pops up* on the agent's computer screen, detailing all potentially relevant support for the present transaction, as well as pointers for likely responses to the support request. Having identified the customer's need, this could all culminate in an almost instantaneous automatic e.mail or fax that resolves the customer's problem. In a business setting, CTI and ANI are used to identify, for example, cross- or up-selling opportunities and, hence, routing of the call to an appropriately skilled agent.

1.3 The world of call centers

Call centers can be categorized along many dimensions: functionality (help desk, emergency, telemarketing, information providers, etc.), size (from a few to several thousands of agent seats), geography (single- vs. multi-location), agents charateristics (low-skilled vs. highly-trained, single- vs. multi-skilled), and more. A central characteristic of a call center is whether it handles *inbound* vs. *outbound* traffic. (Synonyms for inbound/outbound are incoming/outgoing.) *Our focus here is on inbound call centers*, with some attention given to mixed operations that blend in- and out-going calls. An example of such blending is when agents are utilizing their idle time to call customers that left IVR requests to be contacted, or customers that abandoned (and had been identified by ANI) to check on their wishes. Pure outbound call centers are typically used for advertisement or surveys - they will be only briefly described (and contrasted with pure inbound and mixed operations) in Subsection 3.5.

Modern call/contact centers however are challenged with multitude types of calls, coming in over different communication channels (telephone, internet, fax, e.mail., chat, mobile devices, ...); agents have the skill to handle one or more types of calls (e.g., they can provide technical support for several products in several languages by telephone, e.mail or chat). Furthermore, the organizational architecture of the modern call center varies from the very flat, where essentially all agents are exposed to external calls, to the multi-layered, where a layer represents say a level of expertise and customers could potentially be transferred through several layers until being served to satisfaction. Further yet, a call center could in fact be the virtual embodiment of few-to-many geographically dispersed call centers (from the very large, connected over several continents - for example, mid-West U.S.A. with Ireland and India - to the very small, constituting individual agents that work from their homes in their spare time).

1.4 Management and quality of service

There exists a large body of literature on the management of call centers, both in the academia (Section VII in [47] contains close to 50 references) and even more so in the trade literature. The traditional view of the call center has been that of a necessary evil - something that a well-respected company must maintain, and pour money into - indeed too much of it. However, and as already indicated, call centers have evolved into the main contact frontier of businesses with their customers and, as such, present highly promising business opportunities. (One talks about the transformation of call centers

from being mere operational "cost centers" to strategically-significant "profit centers".) Call center managers are thus faced with the conflicting goals of providing consistent high service quality, through a turbulent fast expanding service channel, to an often huge customer base - a truly fertile ground for Operations Research models and analysis, with a special role reserved for queueing theory and queueing science.

Typically, call center goals are formulated as the provision of service at a given quality, subject to a specified budget (more on this momentarily). While Service Quality is a very complicated notion, to which numerous articles and books have been devoted [32, 12, 27], a highly simplified approach suffices for our purposes. We measure service quality along two dimensions: qualitative (psychological) and quantitative (operational). The former relates to the way in which service is provided and perceived (am I satisfied with the answer, is the agent friendly, etc.; for example, [66]). The latter relates more to service accessibility (how long did I have to wait for an answer, was I forced into calling back, etc.). Models in support of the qualitative aspects of service quality are typically empirical, originating in the Social Sciences or Marketing (see Sections III, IV and VIII in [47]). *Models in support of quantitative management are typically analytical, and here we focus on the subset of such models that originates in Operations Research in general and Queueing Theory in particular.*

Common practice is that upper management decides on the desired service level and then call center managers are called on to defend their budget. Similarly, costs can be associated with service levels (eg. toll-free services pay out-of-pocket for their customers' waiting), and the goal is to minimize total costs. These two approaches are articulated in Borst, Mandelbaum and Reiman [15]. It occurs, however, that profit can be linked directly to each individual call, for example in sales/mail-order companies. Then a direct trade-off can be made between service level and costs so as to maximizes overall profit. Two papers in which this is done are Andrews & Parsons [5] and Akşin & Harker [2]. *In what follows we concentrate on the service level vs. cost (efficiency) trade-off.* The fact that salaries account for 60–70% of the total operating costs of a call center *justifies our looking mostly at personnel costs.* This is also the approach adopted by *workforce management tools*, that are used on a large scale in call centers. By concentrating on personnel, one presumes that other resources (such as ICT) are not bottlenecks (see however the work of Akşin and Harker [1, 2]).

1.5 Performance measures

Operational service level is typically quantified in terms of some congestion or performance measures. Our experience, backed up by [27], suggests a *focus on abandonment, waiting and/or retrials*, which underscores the natural fit between queueing models and call centers (Subsection 1.7).

Abandonment is measured by the fraction of customers that abandon the queue prior to being served (either out of all customers, or out of those actually delayed, or sometimes out of those waiting over some small threshold, say 5 seconds). *Waiting* is measured by either its average (ASA = Average Speed of Answer), or by some percentile of the waiting-time distribution. An industry standard for telephone services seems to be the 80/20 rule, under which at least 80% of the customers must wait no more than 20 seconds. (We are aware of no rationalized justification for these parameter values; moreover, different standards are applied to emails, faxes and regular mail.) Finally, *retrials* can be quantified by the fraction of customers whose request is satisfied on first attempt, or by the average number of visits needed to resolve a single problem.

Performance measures are of course intercorrelated - see [69] for the remarkably linear relation between the fraction of abandoning customers and average waiting time. They could also convey more information that actually meets the eye. For example, in contrast to waiting statistics which

are objective, abandonment and retrial measures are *subjective* in that they incorporates customers' view on whether the offered service is worth its wait (abandonment) or returning to (retrials). As another example, it turns out that one can quantify customers' patience in terms of the ratio between the fraction abandoned to the fraction served - indeed, it is shown in [51] that this ratio can be also interpreted as that between the average time that customers are *willing* to wait to the average time that they *expect* to wait.

The practice of service levels must be handled with care. For example, a low fraction of abandoning customers need not indicate a high level of service but possibly an urgent need for it. Low levels of waiting could correlate highly with short service times which, in turn, could result in many retrials - in other words, service times and delays are effectively long, being accumulated over several visits. It is also unclear whether service level must hold for every time interval, or on average over the whole day, or perhaps for an arbitrary customer. The usual situation is that the service level must hold for every time interval. (Koole and van der Sluis [44] emphasize the advantages of looking at an overall service level. For an example of rigorous considerations of these issues, involving convex analysis of the relevant performance measures, see Part 1 of Homework 11, in `ie.technion.ac.il/serveng`.)

For performance measures to be useful, they must be archived at a proper resolution and observed at the appropriate frequency. Ideally, one would like to store, for each individual transaction at the call center, its operational and business characteristics. This raw data can then be mined for exploratory purposes, or aggregated into performance measures for management use. For example, Figures 4 and 7 exhibit the prevailing standard, under which operational data is averaged over half-hour intervals. (See Section 2 for details.) Such an averaging, however, is insufficient for deeper needs, as amply demonstrated in [51].

1.6 A scientific approach to management

In the practice of call center management, a quantitative approach often amounts to merely monitoring, explaining or intervening with quantitative changes in performance. The call center manager tracks performance indicators and *reacts* when they reach unacceptable levels; for example, too many customers are waiting or too many agents are idle. These reactions are typically based on subjectively-biased experiences, and a decision is doomed "poor" or "wrong" if the resulting performance turns out worse than expected.

In a more scientific approach, management is pro-active rather than reactive - for example, ensuring that waiting is scarce rather than adding agents when waiting becomes excessive. Here quantitative models - analytical or simulation - turn out useful for developing rules-of-thumb and intuition, or practically supporting design and control. For example, the "what-if" scenarios in the Introduction to [15] demonstrate, via a simple analytical model, that call centers are typically extremely sensitive to changes in underlying parameters; this is closely related to the square-root principle for staffing, which is a rule-of-thumb that is presented in Subsection 3.1.1 below. Models have in fact become integral parts of the widely used workforce scheduling tools; but such uses rarely go beyond the rudimentary M/M/s (Erlang-C) queue, let alone the more sophisticated models that are surveyed in Section 3.

Analytical models are commonly contrasted with simulation (Section VIII in [47]),which has been growing in popularity. This is partly because of an improved user-friendliness in simulation tools, partly in view of the scarcity of mathematical skills required for the alternatives to simulation, but perhaps mostly due to the widening gap between the complexity of the modern call center relative to the analytical models available to accommodate this complexity.

We shall not delve here on the virtues and vices of analytical vs. simulation models. Our contention

is that, ideally, one should blend the two: analytical models for insight and calibration, simulation also for fine tuning. In fact, our experience strongly suggests that, having analytical models in one's arsenal, even limited in scope, improves dramatically one's use of simulation. These analytical models are often, as it turns out, related to *queueing theory* and their state-of-the-art, in the context of call centers, is described in Section 3.

1.7 Queueing Theory and Science

Queues in service operations are often the arena where customers, service providers (servers, or agents) and managers establish contact, in order to jointly create the service experience. Process-wise, queues play in services much the same role as inventories in manufacturing. But in addition, "human queues" express preferences, complain, abandon and even spread around negative impressions. Thus, *customers* treat the queueing experience as a window to the service-providing party, through which their judgement of it is shaped for better or worse. *Managers* can use queues as indicators (queues are the means, not the goals) for control and improvement opportunities. Indeed, queues provide unbiased quantifiable measures (these are not abundant in services), in terms of which performance is relatively easy to monitor and goals are naturally formulated.

Research in quantitative call center management is concerned with the development of scientifically-based design principles and tools (often culminating in software), that support and balance service quality and efficiency, from the likely conflicting perspectives of customers, servers, managers, and often also society. Queueing models constitute a natural convenient nurturing ground for the development of such principles and tools [31, 15]. However, the existing supporting (Queueing) theory has been somewhat lacking, as will now be explained.

The bulk of what is called Queueing Theory, consists of research papers that formulate and analyze queueing models with a realistic flavor. Most papers are knowledge-driven, where "solutions in search of a problem" are developed. Other papers are problem-driven, but most do not go far enough in the direction of a practical solution. Only some articles develop theory that is either rooted in or actually settles a real-world problem, and scarcely few carry the work as far as validating the model or the solution [34, 38]. In concert with this state of affairs, not much is available of what could be called *Queueing Science*, or perhaps the Science of Congestion, which should supplement traditional queueing theory with empirically-based models [69], observations [51] and experiments [59, 46]. In call centers, and more generally service networks, such "Science" is lagging behind that in telecommunications, computers, transportation and manufacturing. Key reasons for the gap seem to be the difficulty of measuring service operations (see Section 2), combined with the need to incorporate human factors (which are notoriously difficult to quantify) - see Subsections 2.3.3 and 3.2 for a discussion of human patience while waiting in *tele-queues*.

1.8 Call centers as queueing systems

Call centers can be viewed, naturally and usefully, as queueing systems. This comes clearly out of Figure 3, which is an operational scheme of a simple call center. (See Subsection 3.1 for an elaboration.)

In a queueing model of a call center, the customers are callers, servers (resources) are telephone agents (operators) or communication equipment, and tele-queues consist of callers that await service by a system resource. The simplest and most-widely used such model is the $M/M/s$ queue, also known in call center circles as Erlang C. For most applications, however, Erlang C is an over-simplification: for example, it assumes out busy signals, customers impatience and services spanned over multiple visits.

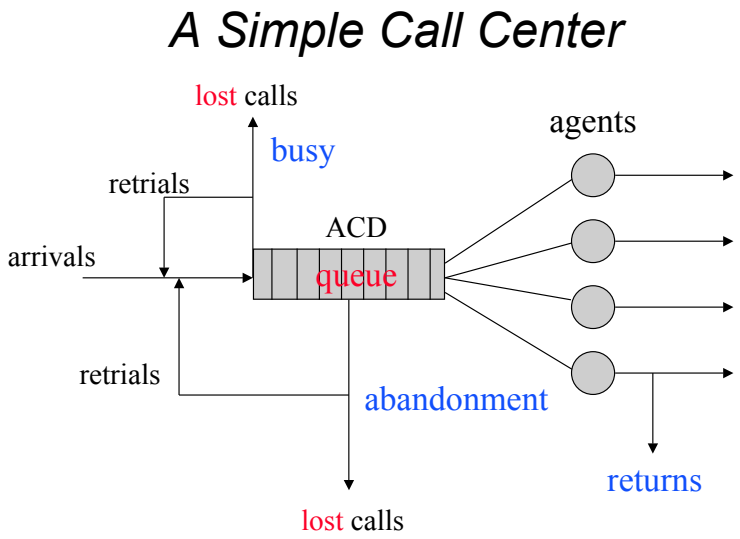

Figure 3: Operational Scheme of a Simple Call Center.

These features are captured in Figure 3, which depicts a single finite-queue with abandonment [31] and retrials [63, 38]. But the modern call center is often a much more complicated queueing *network*: even the mere incorporation of an IVR, prior to joining the agents' tele-queue, already creates two stations in tandem [19], not to mention having multiple teams of specialized or cross-trained agents [30, 13], that are geographically dispersed over multiple interconnected call centers [43], and who are faced with time-varying loads [50] of calls by multi-type customers [8, 2].

1.9 Keeping up-to-date

A fairly complete list of academic publications on call centers has been compiled in [47]. There are over 200 publications, arranged chronologically within subjects, each with its title and authors, source, full abstract and keywords. Section I in [47], entitled "Operations Research, Operations Management", covers analytical queueing models of call centers. The material in [47] has been updated through September 2001, and attempts will be made to keep it up-to-date. Indeed, authors of papers related to call centers are encouraged to send their work to the second author (see the cover page of [47] for guidelines).

Given the speed at which call center technology and research are evolving, advances are perhaps best followed through the Internet - either via sites of researchers active in the area, or industry sites - see Section XI of [47] for a list of 15 web sites. As an example, the first author maintains a web site with call center research done in his research group. It can be found at `www.cs.vu.nl/obp/callcenters`. The site of the second author (`ie.technion.ac.il/serveng`) includes material from a course entitled *Service Engineering*, the inspiration and data for which came out of call centers that are conceptualized as queueing systems. Alternatively, one could activate any standard search engine with the keywords "call center" (for example Google (`www.google.com`)) - this could become the start of a possibly long but potentially very interesting journey.

1.10 The overall picture

Any generalization or classification has the disadvantage of ignoring certain relations. Moreover, there are a multitude of ways to survey a broad subject, each having its advantages and disadvantages. Yet a coherent presentation must strike a compromise on some sort of classification which, in our case, is as follows:

A modeling study must incorporates, at some point, the collection and analysis of relevant data. We thus start our survey with a description of call center data. Directly measurable quantities, such as arrival rates and service time distributions, are referred to as *primitives* - these are typically the required inputs (parameters) for any performance model.

Next, we move on to models that are used to calculate and optimize call center performance over short time intervals (in practice 60-, 30- and less frequently 15-minute intervals). Such models are used for two distinct goals: first, they assist in the operational control of call centers, in the sense that for example call routing decisions are based on or even taken by these models; second, they provide input for the staffing algorithms in the form of required numbers and skills of agents, over all relevant time intervals.

Staffing models are the subject of our last section. At present, they constitute the most used mathematical models of call centers, often embedded within so-called *workforce management systems*. Our survey follows the practice of workforce management, which is typically exercised hierarchically: from the strategic, through tactical to the operational and online-control levels [29]. The queueing models of Section 3 are most appropriate for the online-control level For the other levels, mathematical programming models are the ones used. To the best of our understanding, there is a scarcity (both in theory and practice) of strategic models that support business plans and integrate with lower levels of decision (Xu [68]).

To sum up, call centers offer numerous opportunities for scientific and engineering research, that will enhance the understanding of their operations and support their management. Such research must be based on analytical models, and validated with real data (as in Green and Kolesar [34]), to be of practical value. In the sequel we survey the state-of-the art of this research. As will be unraveled, a very wide gap does exist between the available and what is still needed, but a considerable effort has already been undertaken.

2 Data analysis and forecasting

Any modeling study of call centers must necessarily start with a careful data analysis. For example, the simplest Erlang C queueing model of a call center requires the estimation of calling rate and mean service (holding) times. Moreover, the performance of call centers in peak hours is extremely sensitive to changes in its underlying parameters. (See Figure 7, and the discussion in Subsection 3.2.) It follows that an extremely accurate estimation/forecasting of parameters is a prerequisite for a consistent service level and an efficient operation.

Section II in [47] lists only 16 papers on the statistics and forecasting of call center data. Given the data-intensive hi-tech environment of modern call centers, combined with the importance of accurate estimation, it is surprising, perhaps astonishing, that so little research is available and so much is yet needed. (Compare this state-of-affairs with that of Internet and telecommunication - here, only few year ago, a fundamental change in the research agenda was forced on by data analysis, which revealed new phenomenon, for example heavy-tails and long-range-dependence.)

There is a vast literature on statistical inference and forecasting, but surprisingly little has been

devoted to stochastic processes and much less to queueing models in general and call centers in particular (see Section II in [47] for some exceptions). Indeed, the practice of statistics and time series in the world of call centers is still at its infancy, and serious research is required to bring it to par with its needs.

The scarcity of statistical research of call centers renders our survey effort relatively easy. We start with a brief discussion on model types: empirical, explanatory and theoretical. Then we move on to describing existing sources for call center data. We conclude with some models for primitives and traffic that are based on these data: first, common models for call arrivals and service/holding times, followed by less common models for some aspects of customers' behavior, specifically (im)patience and retrials. We restrict attention to individual call centers for lack of any empirical studies on networks.

2.1 Empirical, explanatory and theoretical models

Parts of the present section are adapted from the report Mandelbaum, Sakov & Zeltyn [51]. This is an in-depth empirical analysis, mostly descriptive, of call center operational data, gathered at a small Israeli bank over the 12 months of 1999. (Both the report and the data are downloadable from `ie.technion.ac.il/serveng`.) Descriptive analysis culminates in empirical *descriptive models*, the simplest of which are tables or histograms of parameters and performance; for example, an end product is a histogram of service duration by service type, or of customers' patience by customer type or of waiting times for those ultimately served. It is to be contrasted with statistical analysis, that seeks to develop mathematical or *theoretical models*; for example, identifying that the arrival process is a Poisson process or a mixture of such [42]; or that service duration is exponentially distributed, with parameters that depend on time-of-day and are correlated with patience. (Such statistical analysis that accompanies [51] is now in progress, jointly with Larry Brown, Noah Gans and Linda Zhao of the Wharton Business School.)

Intermediary between empirical and theoretical models are *explanatory models*, for example regression and time series. These go beyond say histograms by identifying and capturing relations in terms of explanatory variables, but they go short of theoretical models in that there is no attempt to explain these relations. For a clarification, consider three possible models for the arrival process of calls to a call center during a given day (Section 4 in [51]): a descriptive model could be simply a deterministic fluid-like model of the time-varying calling rate [51], based on averaging out stochastic variability in previous similar days; a theoretical model could be a time-inhomogeneous Poisson process [41], expressing the fact that arrivals are completely random; an explanatory model could be a histogram for the total number daily arrivals, by the hour, based on an ARIMA-related time-series [4] that captures trends due to holidays and advertisement campaigns.

Queueing models constitute mathematical relations among building blocks, for example arrivals and services, which we refer to here as *primitives*. Queueing analysis of a given model starts with assumptions on its primitives and culminates in properties of its performance measures. Validation of the model then amounts to a comparison of its primitives and performance measures - typically theoretical, against their analogs in a given call center - mostly empirical.

While models of primitives and performance measures could be empirical, theoretical or explanatory, the latter type will not be surveyed here. One reason is that explanatory models for primitives have been rarely used in queueing theory. A further justification is that Queueing Science typically validates theoretical primitives and performance measures against their empirical analogs. For example, theoretical analysis of the $G/G/s$ queue gives rise to Kingman's law of congestion: in heavy traffic, the waiting time of delayed customers is close to being exponentially distributed. Empirical

analysis, as in [51], can then validate Kingman's law against the reality in one or many call centers operating in heavy traffic.

2.2 Call center data

We distinguish between three types of call center data: operational, marketing, and psychological. *Operational* data is typically collected by the Automatic Call Distributor (ACD), which is part of the telephony-switch infrastructure (typically hardware-, but recently more and more software-based). *Marketing* or *Business* data is gathered by the Computer Telephony Integration/Information (CTI) software, that connects the telephony-switch with company data-bases, typically customer profiles and business histories. Finally, *psychological* data is deduced from surveys of customers, agents or managers. It records subjective perceptions of service level and working environment, and will not be discussed here further.

Existing performance models are based on operational ACD data. The ultimate goal, however, is to integrate data from the three sources mentioned above, which is essential if one is to understand and quantify the role of (operational) service-quality as a driver for business success. For example, in a sales call center, customers' (im)patience (psychological) affects both staffing levels (operational) and the likelihood of a purchase (marketing). But there is ways to go. First, "dialogues" between ACDs and CTI's are non-existing (the two typically originate in separate vendors). Moreover, our experience has been that both types of data are very difficult to access: ACD data for technical reasons and CTI data due to confidentiality concerns. (Interestingly, this state of affairs seems to be different with Internet services.)

2.2.1 Operational data (ACD)

Most modern call centers are equipped with an ACD: this is the switch that routes calls to agents, while tracing and capturing the history of each call as it flows through the call center. ACD data include each call's arrival time, waiting time in the tele-queue, and service duration. (A related software tool goes under the name of Customer Relations Management (CRM) - it also records individual service transactions, but more in terms of work content and customer value rather than operational characteristics.)

ACD data is typically used through aggregated reports. These consist of counts and averages over 15/30/60 minutes periods at the lowest level, and daily/weekly/yearly periods at higher levels (Figures 5 and 7). In such reports one can find, among other things, the total number of calls served or abandoned during the given period, average waiting times, the agents' utilization levels, etc. The call-log files, with individual call histories, is commonly *erased* after being aggregated. The reasons, we believe, are at least the following. First is the desire to save storage space, which nowadays is economically unfounded: a whole month worth of data, from a large call center, would fit onto a single compact disc. But more importantly is the lack of understanding on what can be done with individual transaction data (the increasing popularity of CRM helps here) and sometimes also the lack of capabilities for deciphering vast data warehouses (here Data Mining appears to come to some rescue.)

2.2.2 Marketing data (CTI)

The other main type of data is marketing (business) data. It is typically collected by CTI software (middleware), that integrates telephone data, specifically the caller ID, with computer data bases that

include the caller's profile and business history. (Some associate CRM tools with this kind of data, rather than with ACD data). Having made the integration, the CTI software pops up a relevant description of the customer on the agent's terminal screen. This description includes, for example, the history of the previous calls and, if relevant, dollar-figures for past sales and future tele-marketing targets. (It is less relevant, for example, in Help Desks, which are Technical Support Centers; here history would include, for example, past complaints and repairs.)

2.2.3 Additional data sources

Two additional sources of data are important to acknowledge. First, some companies record individual calls for legal needs (e.g., online insurance) or training reasons. While potentially useful, there is of yet no simple machinary to translate this data into, say, a spreadsheet. A second source are subjective survey data, which could include both operational and marketing data. While serving a useful rough benchmarking role (see, for example, www.e-interactions.com), such data should be handled with care. It definitely can *not* serve as a substitute, or even a proxy, for the ACD and CTI data discussed above.

2.2.4 On the use of data

The data that is needed for a modeling study depends on the goal of that study. Call centers are dynamic environments, where agent groups and routing tables are frequently changing. This makes historical data useful only if carefully used. In addition, a proliferation of suppliers of hardware and software, and the many custom-made management information systems used to retrieve data, make most data analysis projects of call centers rather unique undertakings. (Though experience does help: for example, the project of the small call center in [51] is now redone at Wharton with a large U.S. bank, controlling four networked call centers - this would have been impossible as a first project!)

We believe that a call center should keep at least several months worth of *transactional data*, namely data at the individual call level, together with utilization profiles of individual agents. The data should be complete, showing for each call its full trace (as in [51]), and for each agent lists of periods of availability for service, breaks and their reasons, etc. Aggregate data should always be kept and archived, with the level of aggregation (per interval, per day) increasing (if necessary) as the data gets older. It is also useful to keep management data, such as attained service levels. (Incidentally, computing these service levels from the data often results in a discrepancy from the service level obtained directly from the ACD. Finding out the causes for these differences can be a very interesting exercise.)

In summary, as modeling efforts of call centers become commonplace, which undoubtedly is to come, the practice of gathering and analyzing *individual* calls, customers and agents will gain popularity as well - both separately and integratively over the operational, marketing and psychological data sources.

2.3 Models of primitives

As argued above, queueing models are created from primitives. For example, the most basic operational model of a call center is the M/M/s queue with parameters λ, μ and s: the primitives are the arrival process (assumed Poisson at a constant rate λ), the service times (assumed exponentially distributed with mean μ^{-1}), and the agents (s of them); there are additional implicit assumptions, such as independence among the primitives, FCFS service disciplines and more. In order to apply the

$M/M/s$ model, one must estimate its parameters based on historical data. But a more fundamental question is whether the underlying $M/M/s$ assumptions are at all valid - are arrivals Poisson? are service times exponential? and if not, what are alternative *useful* models for arrivals and services? For our purposes, usefulness means that they can serve as primitives for analytical models - a definition that depends, of course, on modeling scope and capabilities of the model builder. This defines the main research agenda for Queueing Science, in the context of call centers: to identify, validate and archive (publish) useful models for primitives and performance measures of call centers.

2.3.1 Call arrivals

The arrival process records the epochs that calls arrive to the call center. It can be described at different levels of detail, and from various points of view (as done in Subsection 2.1). Arrivals to call centers are typically random. For our purposes, *randomness* can be explained as follows: there are many potential, statistically identical callers to the call center; there is a very small yet non-negligible probability for each of them calling at any given minute, say, and they decide on whether to call independently of each other. Under such circumstances, theory dictates that the arrival process fits well a *Poisson process*. If more customers are likely to call say at 10:30 am than at 1:00 pm, one gets a time-inhomogeneous Poisson process. Common call-center practice is to assume constant arrival rates over, say, individual hours or half hours. Such an approximation, by a piecewise-constant arrival rate function allows one to use standard steady-state models. This is reasonable if steady state is achieved relatively fast, in particular predictable variability does not change abruptly.

There are scenarios where the Poisson assumptions are violated. A simple example is when callers react to an external event, such as a telephone number shown in a TV commercial, which can be modeled by adding a Poissonian number of arrivals at a *predictable* point in time. (This is still referred to a Poisson point process, which "enjoys" discontinuity in its cumulative arrival rate.) Other examples are when busy-signals or retrials occur frequently, or when call centers are interconnected via overflow protocols, for example via centralized load-balancing. In analogy to Internet traffic, it is conceivable that phenomena such as long-range dependence, or heavy tails of the interarrival times, would emerge, but there has been no empirical support of that.

Overall, traffic models are very important to call centers as they are used to forecast future traffic. Evidently, without good forecasts it is impossible to schedule personnel efficiently. As traffic volume depends on many factors, statistical techniques using explanatory variables are used to represent the day and the time, and all kinds of influences related to holidays, for example. However, call centers do not always have sufficient historical data to base forecasts on, and certain factors such as weather conditions cannot be predicted. A possible solution to this is offered in Jongbloed and Koole [42], where a method is developed to derive *intervals* for arrival rates rather than point estimates. Gordon & Fowler [33] also pay attention to this problem, but in less detail.

Many companies notice the influence of internal and external events (such as advertisement campaigns, or the release of new product versions) on the traffic load. They also experience a relation between the number of users of a products and the traffic related to a product (and whether a customer is a new or a long-time user, etc.). For these reasons certain centers prefer to base their forecasts on the numbers of customers and acticipated events, and just use the historical data to estimate the behavior of individual customers. As far as we know, no academic studies have been undertaken that explore this practice.

2.3.2 Service duration

Most applications of queueing theory to call centers assume *exponentially* distributed service durations as their default. The main reason is the lack of empirical evidence to the contrary, which leads one to favor convenience and apply models that are analytically tractable (for which there are readily available formulae). And indeed, models with exponential service times are amenable to analysis, especially when combined with the natural assumption that arrival processes "are" Poisson (see Section 3 for a discussion of queueing models).

In some studies, the exponentiality assumption was indeed validated as an acceptable fit. Representatives are Kort [46], who summarizes models of the Bell System Public Switched Telephone Network, developed in the 70's and 80's; and Harris et al. [38], that analyzes IRS call centers. Our experience has confirmed these findings for human services that are homogeneous and unpaced (not only telephone services), but this is definitely not the case with mixtures of such, as will be discussed momentarily. In many service time formulae or their approximations service times manifest themselves merely through their means and standard deviation. Consequently, for practical purposes, if means and standard deviations are close to each other, then one can assume exponentiality of service times. Define the squared-coefficient-of-variation C by $C^2 = E^2/\sigma^2$, where E is the average service time, and σ its standard deviation. Having $C < 1$ can be accommodated by Erlang distributions; $C > 1$ arises with hyperexpontials; and mixtures of Erlangs (as in [63]) are in fact dense among all service time distributions. This subfamily of phase-type distributions is very convenient numerically but less so theoretically, having ample parameters. To this end, parametric models are desired, which has lead us to the important issue of fitting a *parametric* statistical family to service times.

Two such families arose in practice: Erlang or more generally Gamma, and the lognormal distribution [14]. Both families are explored in Chlebus [21], who analyzes holding time distributions in cellular communication systems. Other confirmations for the lognormal fit are provided by the service times in [51] (where the fit has been found truly remarkable), and in the analysis of the data from the call center of a Dutch bank. The excellent fit is not only for the overall service time, but also when restricted to service types, individual agents, etc. It follows that the natural log ($\log_e$) of service time is *normally* distributed. The implications of these findings are presently being explored. One example is the analysis of covariates that affect service time, by simply applying standard regression techniques to log(service time).

Often time, non-standard service duration arise in practice. Consider the following very partial list. First, various ingredients of the call center have associated service times, for example IVR's, but little is known about their distributions. Management practice could effect dramatically service duration - for example, agents could hang up on customers (Figure 17 and Table 40 in [51]), say in order to adhere to a distorted incentive scheme that rewards mostly the *number* of served calls at day-end. Technology also has been effecting service duration. We already mentioned CTI; a simpler example is when greetings and farewell of telephone services are recorded for automation, which reduces both service duration and its stochastic variability. Next, the heterogeneity of servers could be such that a separation of services into types becomes a must for appropriate modeling, and a virtue for efficiency (Whitt [67]). And finally [63], "operators may initially work faster during periods of overload to work off the customer queue, but may tire and work slower than usual if the heavy load is sustained and if no relief is provided".

2.3.3 Abandonment and retrials

One of the most challenging aspects in developing queueing models for call centers is the incorporation of human factors, for both customers and agents, in a practical manner. This opens up a vast multi-disciplinary research agenda which here we only touch on. Specifically, we address the operational service experience of a customer, once joining the *tele-queue*: most attention is given to customers' (im)patience, with little also devoted to the tendency to retry (redial). One should mention that there has been a lot of work on human factors while waiting in queues, which originates in psychology and marketing. (The articles in Section III of [47], entitled "Consumer Psychology", would have ample leads.) However, most of this work has to do with *visible* queues, such as in a bank or a clinic, while telephone queues are phantom, and hence the waiting experience could be vastly different.

Human factors could be tricky to quantify, measure and model. Consider, for example, human (im)patience while waiting for service at the phone. It surely depends on the communication channel (telephone, internet, IVR), or on customer type (who are more patient - regular of VIP customers?). Is patience an absolute or relative quantity? For example, if a customer is willing to wait 10 minutes on the phone, does it indicate ample patience? Or does it indicate more of a need? Or perhaps it depends on prior expectations, namely willing to wait 10 minutes indicates ample patience if prior expectation is for 1 minute's wait, but little patience if one expects 1 hour's wait.

In most work, redial behavior is quantified in terms of some perseverance function that gives the probability of an nth attempt, given a survival beyond the $(n-1)$th attempt. A basic description of patience is the distribution of the time beyond which a customer would not be willing to wait; an equivalent description is the hazard rate, which provides a natural dynamic depiction of patience, as it evolves while waiting (more on this momentarily).

Models for human factors during telephone services were first developed in the 40's by Palm [54]) and later, during the 70's and 80's, in France by Roberts [59]) and in the U.S.A., reported in Kort [46]. All this work analyzed the behavior of telephone subscribers, as they attempt to establish telephone connections. For example, "customer may abandon a call attempt in any of three stages of call setup: while waiting for a dial tone, during dialing, or while waiting for network response after dialing is complete." [46].

Palm [54] develops explanatory models for human patience. Roberts [59] presents descriptive models, based on experimental observations, for redial and abandonment behavior. Kort describes models, mostly parametric, for customer opinion and behavior models. To the best of our knowledge, the only analogous study of human factors in call centers, for customers waiting to be served by an agent, is [51] who develop descriptive models for customer patience. The rest of the present subsection will be devoted to a survey of all these works.

As mentioned above, Palm [54] was the first to recognize the need for incorporating (im)patience in performance analysis. Palm modeled impatience (while waiting for a telephone connection) as a distribution of the time beyond which a customer would not be willing to wait. To quantify this function, Palm introduced an *inconvenience* function of time $I(t)$, $t \geq 0$, the derivative of which he called *irritation*. As a plausible form for irritation, Palm proposed $dI(t) = c \cdot t^{\lambda} dt$ and closed the circle by axiomatizing that irritation is proportional to the hazard rate of customers patience. In other words, the distribution of patience is assumed *Weibull*.

Palm's findings are remarkably consistent with those reported in Kort [46] where, based on abandonment behavior in laboratory testing, Weibull was proposed as the distribution of patience while waiting for a dial tone. From direct measurements of patience, Palm concluded that λ is close to unity, possibly slightly less. Kort, on the other hand, has $\lambda = 1.23$. Kort has in fact two additional

models of patience: abandonment time while dialing, described in terms of a shifted exponential; and abandonment time prior to network response (after dialing), where a mixture of two lognormal distribution fitted well.

Descriptive models (histograms) of patience, while waiting for a network connection, were also developed in Roberts [59]. A parametric model for the data in [59] was later described in Baccelli and Hebuterne [10], where Erlang with 3 phases turned out a reasonable fit. Roberts also estimated the distribution of the time that a customer would have to wait for service, *given infinite patience*. Assuming customers' familiarity with the system, based on prior service experience, this is also the distribution of the time that a customer *expects* to wait [52, 69]. One deduces an improved two-dimensional assessment of human patience, where one compares the time that a customer is willing to wait against the time the customer expects to wait. (See Subsection 6.2.1 in [51] for more details).

Mandelbaum, Sakov & Zeltyn [51] contains an empirical study of abandonment in a call center (see also Subsection 6.3). Descriptive models of patience are developed, which requires tools from statistical survival analysis in order to accommodate *censored data*: indeed, observations of the patience of those who got served (the time that they would be willing to wait for service) are censored by their actual waiting time. (See the Appendix in [69] for details.) The need for censoring was already recognized by Roberts [59], who circumvented it in an own-developed method. (Interestingly, no censoring was required in [54, 46], as patience data there was uncensored.)

A comparison of [51] with [54, 59, 46] is not straightforward for several reasons. Indeed, all deal with (im)patience, but in different circumstances. Thus, the natural time units for [54, 59, 46] are seconds, while patience in a call center is typically measured in minutes (an average of about 10 minutes in [51]). In addition, the call center in [51] provides, during waiting at the phone, some form of information on the position of customers in queue. Such information turns out to cause noticeable spikes in the hazard rate of patience, exactly at those (predictable) times when announcements are made. This demonstrates that the information encourages abandonment, perhaps in contrast to original intentions where information was intended to reduce uncertainty in order to prevent abandonment.

3 Performance models

The essence of *operations management* in a call center is the matching of service requests (demand) with resources (supply). The fundamental tradeoff is between service quality vs. operational efficiency. *Performance analysis* supports this tradeoff by calculating attained service level and resource occupancy/utilization as functions of traffic load and available resources. We start with describing the simplest such models and then expand to capture main characteristics of today's highly complex contact centers. Examples include patience and abandonment [54, 31], retrials [38], time-varying operations [41, 50], skill-base-routing [30, 16, 45], human factors [69, 8], and networked call centers [43].

There exist no single analytical model that accommodates all, or even most, aspects of the modern call center. But this need not be a vice: relatively simple models that capture only key characteristics, unilaterally, typically suffice for management support (insight, design, control, etc).

3.1 Single-type customers and single-skill agents

A schematic operational model of a *simple* call center is depicted in Figure 3. The connotation is that of the old-times switch-board, either those operated by telephone companies or as part of individual organizations, where telephone operators were connecting incoming calls physically to the proper

extension/line. (Old papers on telephone services, as the classical Erlang [24] and Palm [54], were in fact modeling such switch-boards.) Modern technology has now replaced these human operators by the ACD, that routes customers calls to idle agents. What renders the operation depicted above, as well as its model, "simple" is that there is a single type of calls that can be handled by all agents (statistically identical customers and servers).

The simplest and most used performance model is the stationary $M/M/s$ queue. It describes a single-type single-skill call center with s agents, operating over a short enough time-period so that calls arrive at a constant rate, yet randomly (Poisson); staffing level and service rates are also taken constant. The assumed stationarity could be problematic if the system does not relax fast enough, for example due to events such as an advertisement campaign or a mew-product release. The model assumes out busy signals, abandonment, retrials and time-varying conditions. All of these, except for busy signals, are accommodated by the fluid and diffusion approximations of Mandelbaum, Massey, Reiman, Reider and Stolyar ([49, 50]. But the state-of-the art of these approximations is not yet ripe for serious applications.

The reason for using the $M/M/s$ queue is of course the fact that there exist closed form expressions for most of its performance measures. However, $M/M/s$ predictions could turn out highly inaccurate because, as articulated in Subsection 2.3 and recalled above, reality often "violates" its underlying assumptions, and these violations are not straightforward to model. For example, non-exponential service times leads one to the $M/G/s$ queue which, in stark contrast to $M/M/s$, is analytically intractable. One must then resort to approximations, out of which it turns out that service time affects performance through its coefficient-of-variation $C = E/\sigma$). Specifically,

$$E[\text{Wait for M/G/s}] \approx E[\text{Wait for } M/M/s] \times \frac{(1+C^2)}{2}.$$

Thus, performance deteriorates (improves) as stochastic variability in service times increases (decreases). The above approximation was advocated by Sze [63], where it is attributed to Lee and Longton from 1959. Sze was in fact motivated by a traffic mix problem, where service duration was "definitely not exponential" but rather a mixture of two types of services, each well fitted with an Erlang distribution. (Service time was therefore phase-type.)

When modeling call centers, the useful approximations are typically those in heavy-traffic, namely high agents' utilization levels at peak hours. Consider again the $M/G/s$ queue. For small to moderate number of agents s, Kingman's classical result asserts that Waiting Time is approximately exponential, with mean as given above. Large s, on the other hand, gives rise to a different asymptotic behavior. This was first discovered by Halfin and Whitt [37] for the $M/M/s$ queue, and recently extended to $M/PH/s$ by Puhalski and Reiman [57]. We now discuss these issues within the context of two key challenges for call center management: agent *staffing* in Subsection 3.1.1 and site *pooling* in Subsection 3.1.2.

3.1.1 Square-root safety staffing

The *square-root safety-staffing principle*, introduced formally by Borst, Mandelbaum and Reiman [15] but having existed long before, recommends a number of servers s given by

$$s \;=\; R + \Delta \;=\; R + \beta\sqrt{R}\,, \quad -\infty < \beta < \infty\,,$$

where $R = \frac{\lambda}{\mu}$ is the *offered load* (λ=arrival rate, μ=service rate) and β represents *service grade*. The actual value of β depends on the particular model and performance criterion used, but the form of s

is extremely robust and accurate. As an example, for the $M/M/s$ queue analyzed in [15], β could be taken a positive function of the ratio between hourly staffing and delay costs, Δ is the safety staffing, and performance measures are approximately given by the following formulae:

$$\mathrm{P}\{\mathrm{Wait} > 0\} \approx P(\beta) = \left[1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right]^{-1} ,$$

$$\mathrm{E}[\mathrm{Wait}] \approx \mathrm{E}(S) \cdot \frac{\mathrm{P}\{\mathrm{Wait} > 0\}}{\beta\sqrt{R}} ,$$

$$\mathrm{P}\{\mathrm{Wait} > T \cdot \mathrm{E}(S)\} \approx \mathrm{P}\{\mathrm{Wait} > 0\} \cdot e^{-T\Delta} .$$

Here Φ and ϕ are, respectively, the distribution and density functions of the standard normal distribution (mean=0, variance = 1). It is shown in [15] that the square-root principle is essentially asymptotically optimal for large heavily-loaded call centers ($\lambda \uparrow \infty$, $s \uparrow \infty$), and it prescribes operation in the rationalized (Halfin-Whitt) regime (more on this in Subsection 3.1.2).

The square-root principle is applicable beyond $M/M/s$ (Erlang C). Garnett et al. [31] verify it for the $M/M/s$ model with abandonment (Subsection 3.2) - here β can take also negative values, since abandonment guarantee stability at all staffing levels; for time-varying models, as in Jennings et al. [41], β varies with time; and Borst and Seri [16] use it for skill-based routing. Finally, Puhalskii and Reiman [57] support the principle for the $M/G/s$ queue, given service times that are square integrable. (Extensions to heavy-tailed service times would plausibly give rise to safety staffing with power of R other than half.)

In all the extensions of [15], only the *form* $s = R+\beta\sqrt{R}$ was verified, theoretically or experimentally, but the determination of the exact value of β, based on economic considerations, is still an important open research problem. The square-root principle embodies another operational principle of utmost importance for call centers - economies of scale (EOS) - which we turn to.

3.1.2 Operational regimes, pooling and economies of scale

Consider Figure 4, that summarizes the performance of a large U.S. mail-catalogue retailer.

Focus on the peak period of 10:00-11:00: 765 customers customers called; service time is about 3.75 minutes on average with an after-call-work of 30 seconds and auxiliary work to the order of 5% of the time; ASA is about 1 seconds and only 1 call abandoned (after 1 second - which seems more like a "typo"). But there were about 95 agents handling calls, resulting in about 65% utilization - clearly a *quality-driven* operation.

At the other extreme there are *efficiency-driven* call centers: with a similar offered work as above, ASA could reach many minutes and agents are utilized very close to 100% of their time.

Within the quality-driven regime, almost all customers are served immediately upon calling. At the efficiency-driven regime, on the other hand, essentially all customers are delayed in queue. However, as explained in [15] and elaborated on momentarily, well-managed large call centers operate within a *rationalized* regime, where quality and efficiency are balanced in the face of scale economies. This is the case in Figure 5, summarizing the performance of 12 call centers, operated by a large U.S. health insurance company: one observes a *daily* average of 2.8% abandonment (out of those called), 31 second ASA, 318 seconds AHT (Average Handling Time, namely service duration), with 91% agents' utilization (and over 95% in a couple of the call centers).

A characteristic of the rationalized regime is a fraction of delayed customers that is neither close to zero (quality-driven) nor to unity (efficiency-driven). The second author obtained more refined data

Copy of Summary Interval - Order PK

Date: 7/7/97
Split/Skill: Order PK

Time	Avg Speed Ans	Avg Aban Time	ACD Calls	Avg ACD Time	Avg ACW Time	Aban Calls	% ACD Time	% Ans Calls	Avg Pos Staff	Calls Per Pos	%Serv Lev	%Aux Time	%ACW Time	%ACD Time
Totals	:00:02	:00:28	10456	:03:47	:00:25	46	53	98	70	149		8		
12:00 AM*	:00:00	:00:00	26	:04:31	:00:02	1	76	51	7	4	51	2	16	61
12:30 AM*	:00:03	:04:10	14	:07:27	:00:33	1	89	52	5	3	48	1	26	63
1:00 AM*	:00:00		9	:04:54	:11:29	0	91	90	1	7	90	0	26	65
5:30 AM*			0			0	0		0	0		33	0	0
6:00 AM*	:00:00		12	:03:21	:00:19	0	21	100	7	2	100	9	2	19
6:30 AM*	:00:00		27	:02:51	:00:20	0	32	100	14	2	100	5	3	29
7:00 AM*	:00:00		62	:03:34	:00:15	0	38	100	21	3	100	13	4	34
7:30 AM*	:00:00		93	:03:11	:00:34	0	36	100	30	3	100	7	4	32
8:00 AM*	:00:00		120	:03:37	:00:40	0	39	100	47	3	100	8	6	33
8:30 AM*	:00:00		193	:03:04	:00:14	0	44	100	61	3	100	10	7	37
9:00 AM*	:00:01		293	:03:25	:00:25	0	54	99	75	4	97	9	7	47
9:30 AM*	:00:02	:00:06	381	:03:45	:00:22	2	60	97	91	4	93	8	8	52
10:00 AM*	:00:02	:00:01	416	:03:49	:00:26	1	63	97	94	4	96	5	8	55
10:30 AM*	:00:00		349	:03:35	:00:33	0	52	99	96	4	99	6	8	44
11:00 AM*	:00:00		352	:03:50	:00:27	0	51	100	102	3	100	7	6	45
11:30 AM*	:00:00		349	:03:44	:00:18	0	49	100	97	4	100	8	5	45
12:00 PM*	:00:01		354	:03:59	:00:18	0	52	95	95	4	95	8	5	47
12:30 PM*	:00:00		336	:03:38	:00:21	0	52	99	97	3	99	9	6	46
1:00 PM*	:00:00		347	:03:53	:00:32	0	51	99	98	4	99	11	8	44
1:30 PM*	:00:00		368	:03:52	:00:14	0	56	99	99	4	99	11	7	50
2:00 PM*	:00:01		393	:03:55	:00:17	0	51	100	106	4	100	10	5	46
2:30 PM*	:00:00		403	:03:58	:00:13	0	54	100	112	4	100	10	4	50
3:00 PM*	:00:00	:00:04	410	:04:02	:00:16	1	57	98	110	4	98	8	5	51
3:30 PM*	:00:00		347	:03:59	:00:14	0	50	100	100	3	100	7	5	45
4:00 PM*	:00:00		382	:03:48	:01:37	0	54	100	98	4	100	6	7	47
4:30 PM*	:00:00		379	:03:41	:00:19	0	55	99	97	4	99	8	5	50
5:00 PM*	:00:00		411	:03:53	:00:19	0	53	100	109	4	100	9	5	48
5:30 PM*	:00:01		387	:03:58	:00:19	0	58	99	96	4	99	10	6	51
6:00 PM*	:00:01	:00:21	371	:03:28	:00:25	1	53	98	91	4	98	9	6	47
6:30 PM*	:00:00		260	:03:26	:00:13	0	41	100	90	3	100	8	4	37
7:00 PM*	:00:00		269	:03:24	:00:17	0	42	100	78	3	100	9	5	38

Page 1 of 2

Figure 4: A quality-driven call center.

Command Center Intraday Report

Date		Updated Through: All Day									
06/13 - Tue		Recvd	Answ	Abn %	ASA	AHT	Occ %	On Prod%	On Prod FTE	Sch Open FTE	Sch Avail %
Total:		**129,960**	**126,321**	**2.8%**	**31**	**318**	**90.9%**	**88.4%**	**1531.7**	**1585.0**	**96.6%**
INQ	Charlotte	20,577	19,860	3.5%	30	307	95.1%	85.4%	222.7	234.6	95.0%
INQ	Columbus MCSC	7,973	7,773	2.5%	36	314	94.9%	89.8%	89.2	94.5	94.4%
INQ	Phoenix	17,102	16,757	2.0%	31	298	92.7%	91.8%	187.3	194.8	96.2%
INQ	Scranton	1,257	1,254	0.2%	6	515	78.6%	28.9%	28.5	35.1	81.2%
INQ	Tampa	9,174	8,859	3.4%	42	366	91.5%	93.6%	123.1	125.9	97.8%
CEN	Bourbonnais	6,070	5,937	2.2%	33	362	86.7%	90.2%	86.0	88.4	97.3%
CEN	Bristol	10,667	10,505	1.5%	25	355	95.1%	93.1%	136.3	139.6	97.6%
CEN	Columbus Claims	5,258	5,153	2.0%	27	293	86.7%	89.8%	60.5	62.2	97.3%
STH	Atlanta	7,514	7,338	2.3%	40	318	82.1%	89.5%	98.6	99.8	98.8%
STH	Sherman	19,669	18,833	4.3%	46	252	93.8%	90.6%	175.5	174.9	100.4%
STH	Wilmington	10,422	9,888	5.1%	21	285	89.9%	92.1%	108.7	114.6	94.8%
WST	Visalia	14,277	14,164	0.8%	10	382	87.2%	85.0%	215.2	220.6	97.6%

Figure 5: Performance of 12 call centers in the rationalized regime.

from the above-mentioned health insurance company, from which it was calculated that, overall, only about 40% of the customers were delayed while the other 60% accessed an agent immediately without any delay.

The rationalized regime was first identified in practice by Sze [63], from which we loosely quote the following: "The problems faced in the Bell System operator service differ from queueing models in the literature in several ways: 1. Server team sizes during the day are large, often 100-300 operators. 2. The target occupancies are high, but are not in the heavy traffic range. Approximations are available for heavy and light traffic systems, but our region of interest falls between the two. Typically, 90-95% of the operators are occupied during busy periods, but because of the large number of servers, only about half of the customers are delayed." Theory that supports the rationalized regime was first developed by Halfin and Whitt [37].

Thus large call centers operate in a regime that seems to circumvent the traditional tradeoff between service-level and resource-efficiency - EOS is the enabler. We demonstrate this through an example where we quantify the operational benefits from *pooling* several call centers. This example is adapted from Homework 11 in `ie.technion.ac.il/serveng`. It demonstrates that our three operational regimes (efficiency-driven, quality-driven and rationalized) can be naturally associated with different staffing guidelines, and that the regimes vary significantly with respect to the EOS that pooling yields.

Consider the pooling of m statistically identical call centers into a single operation. Each call center has the same λ and μ; the arrival rate to the pooled call center is $m \times \lambda$, and its μ is unaltered. One asks: what should be the staffing levels after pooling, so as to sustain service level.
Three scenarios are explored:
- *Scenario 1 (Efficiency-driven staffing):* ASA = E[Wait|Wait>0] is sustained.
- *Scenario 2 (Quality-driven staffing):* Agents' occupancy level is sustained.
- *Scenario 3 (Rationalized staffing):* Congestion = delay probability is sustained.
(The terminology hints at the connection with our operational regimes.)

Define the *Total Service Factor* (TSF) by

$$\text{TSF} \;=\; \text{P}\{\text{Wait} > T \cdot \text{E}(S) \mid \text{Wait} > 0\} \;.$$

(This differs from the standard definition, which is TSF = P{Wait $> T$}.) Similarly, let

$$\text{ASA} \;=\; \text{E}\left[\frac{\text{Wait}}{\text{E}(S)} \;\middle|\; \text{Wait} > 0\right] \;.$$

(The standard is ASA=E[Wait], but our approach is natural and, moreover, it gives rise to simpler mathematical formulae for performance.) The main operational and performance characteristics of the call center are summarized in Figure 6, which we now briefly explain. (The framed entries are those sustained after pooling; the function $P(\beta$ was introduced for the square-root staffing principle.)
- Scenario 1 implies a decrease of the service grade β to $\beta/\sqrt{m}$, and an increase of the delay probability from $P(\beta)$ to $P(\beta/\sqrt{m})$ (which could be significant even for small m's). Note, however, that ASA and TSF are unchanged. We observe convergence, as $m \uparrow \infty$, to an *efficiency-driven* regime where servers are close to 100% utilized and essentially all customers are delayed.
- Scenario 2 exhibits overall improvement of service-level: ASA decreases to ASA/m, TSF decreases to $(\text{TSF})^m$ and the delay probability decreases from $P(\beta)$ to $P(\beta\sqrt{m})$. This is our *quality-driven* regime since, as $m \uparrow \infty$, essentially all customers are served immediately upon arrival.
- Scenario 3 implies the same service-grade and probability of wait as in the base-case. ASA decreases to ASA/$\sqrt{m}$, and TSF decreases to $(\text{TSF})^{\sqrt{m}}$. This is the *rationalized* regime: it is both efficiency-

Economies of Scale

Base case: M/M/N with parameters λ, μ, N

Scenario: $\lambda \to m\lambda$ $(R \to mR)$

	Base Case	Efficiency-driven	Quality-driven	Rationalized
Offered load	$R = \frac{\lambda}{\mu}$	mR	mR	mR
Safety staffing	Δ	Δ	$m\Delta$	$\sqrt{m}\Delta$
Number of agents	$N = R + \Delta$	$mR + \Delta$	$mR + m\Delta$	$mR + \sqrt{m}\Delta$
Service grade	$\beta = \frac{\Delta}{\sqrt{R}}$	$\frac{\beta}{\sqrt{m}}$	$\beta\sqrt{m}$	$\boxed{\beta}$
Erlang-C = P{Wait>0}	$P(\beta)$	$P\left(\frac{\beta}{\sqrt{m}}\right) \uparrow 1$	$P(\beta\sqrt{m}) \downarrow 0$	$\boxed{P(\beta)}$
Occupancy	$\rho = \frac{R}{R+\Delta}$	$\frac{R}{R+\frac{\Delta}{m}} \uparrow 1$	$\boxed{\rho = \frac{R}{R+\Delta}}$	$\frac{R}{R+\frac{\Delta}{\sqrt{m}}} \uparrow 1$
$\text{ASA} = \text{E}\left[\frac{\text{Wait}}{\text{E}(S)} \middle\vert \text{Wait} > 0\right]$	$\frac{1}{\Delta}$	$\boxed{\frac{1}{\Delta} = \text{ASA}}$	$\frac{1}{m\Delta} = \frac{\text{ASA}}{m}$	$\frac{1}{\sqrt{m}\Delta} = \frac{\text{ASA}}{\sqrt{m}}$
$\text{TSF} = \text{P}\left\{\frac{\text{Wait}}{\text{E}(S)} > T \middle\vert \text{Wait} > 0\right\}$	$e^{-T\Delta}$	$\boxed{e^{-T\Delta} = \text{TSF}}$	$e^{-mT\Delta} = (\text{TSF})^m$	$e^{-\sqrt{m}T\Delta} = (\text{TSF})^{\sqrt{m}}$

Figure 6: Erlang C in the efficiency, quality and rationalized regimes.

driven (occupancy increases to 100%) and quality-driven (a significant fraction, namely $1 - P(\beta)$, of the customers are served *immediately*.)

Note that, agents' utilization increases in both Scenarios 1 and 3: pooling allows for higher productivity as well as improved service-level - a clear manifestation of EOS.

3.2 Busy signals and abandonment

Each caller within a call center occupies a trunk-line. When all the lines are occupied, a calling customer gets a busy signal. Thus, a manager could eliminate *all* delays by dimensioning the number of lines to be equal to the number of agents. in which case $M/M/s/s$, or Erlang-B ("B" for Blocking) becomes the "right" model. But then there would typically be ample busy-signals. Moreover, prevailing practice goes in fact the other way: it is to dimension amlple lines so that a busy signal becomes a rare event. But then customers are forced into long delays. This is costly for the call center (think 1-800 costs) and possibly also for the customers - they might well prefer a busy-signal over an information-less delay, and hence they abandon the tele-queue before being served.

The busy-signal vs. delay vs. abandonment trade off has not yet been formally and fully analyzed, to the best of our knowledge. A simulation study of $M/M/s/B$ is presented by Feinberg [26], where B stands for the overall number of lines ($B \geq s$); it is argued that only 10% lines in excess of agents provides good performance: more lines would give rise to too much waiting and fewer to too many busy signals. A more appropriate framework would be the $M/M/s/B + G$ queue, where $+G$ indicates arbitrarily distributed patience (following the notation and results of [10]). An analytically tractable model is the $M/M/s/B + M$, in which patience is assumed exponential. (For mathematical details see Riordan [58], pages 109–112, and [31].) Procedures for estimating the mean patience, as an input

6/13/00 - Tue

Charlotte - Center

Time	Recvd	Answ	Abn %	ASA	AHT	Occ %	On Prod%	On Prod FTE	Sch Open FTE	Sch Avail %
0	20,577	19,860	3.5%	30	307	95.1%	85.4%	222.7	234.6	95.0%
8:00	332	308	7.2%	27	302	87.1%	79.5%	59.3	66.9	88.5%
8:30	653	615	5.8%	58	293	96.1%	81.1%	104.1	111.7	93.2%
9:00	866	796	8.1%	63	308	97.1%	84.7%	140.4	145.3	96.6%
9:30	1,152	1,138	1.2%	28	303	90.8%	81.6%	211.1	221.3	95.4%
10:00	1,330	1,286	3.3%	22	307	98.4%	84.3%	223.1	229.0	97.4%
10:30	1,364	1,338	1.9%	33	296	99.0%	84.1%	222.5	227.9	97.6%
11:00	1,380	1,280	7.2%	34	306	98.2%	84.0%	222.0	223.9	99.2%
11:30	1,272	1,247	2.0%	44	298	94.6%	82.8%	218.0	233.2	93.5%
12:00	1,179	1,177	0.2%	1	306	91.6%	88.6%	218.3	222.5	98.1%
12:30	1,174	1,160	1.2%	10	302	95.5%	93.6%	203.8	209.8	97.1%
13:00	1,018	999	1.9%	9	314	95.4%	91.2%	182.9	187.0	97.8%
13:30	1,061	961	9.4%	67	306	100.0%	88.9%	163.4	182.5	89.5%
14:00	1,173	1,082	7.8%	78	313	99.5%	85.7%	188.9	213.0	88.7%
14:30	1,212	1,179	2.7%	23	304	96.6%	86.0%	206.1	220.9	93.3%
15:00	1,137	1,122	1.3%	15	320	96.9%	83.5%	205.8	222.1	92.7%
15:30	1,169	1,137	2.7%	17	311	97.1%	84.6%	202.2	207.0	97.7%
16:00	1,107	1,059	4.3%	46	315	99.2%	79.4%	187.1	192.9	97.0%
16:30	914	892	2.4%	22	307	95.2%	81.8%	160.0	172.3	92.8%
17:00	615	615	0.0%	2	328	83.0%	93.6%	135.0	146.2	92.3%
17:30	420	420	0.0%	0	328	73.8%	95.4%	103.5	116.1	89.2%
18:00	49	49	0.0%	14	180	84.2%	89.1%	5.8	1.4	416.2%

Figure 7: Performance of a large call center in the rationalized regime.

parameter to performance analysis, are given in [31, 51]. Alternatively, mean patience could be used as a tuning parameter, where its value is determined to establish a fit between practice and theory - this will be the approach taken in the following example.

3.2.1 Performance sensitivity in heavy traffic, via Erlang A

In heavy traffic, even a small fraction of busy-signals or abandonment could have a dramatic effect on performance, and hence must be accounted for. This will now be demonstrated via the $M/M/s+M$ model [54, 10, 31], which adds an abandonment feature to $M/M/s$ (Erlang C): specifically, one models customers' patience as exponentially distributed, independently of everything else; customers abandon if their patience expires before they reach an agent. We shall refer to the $M/M/s+M$ queue as Erlang A, "A" for Abandonment, and for the fact that this model interpolates between Erlang B and Erlang C.

A model for a call center with busy-signas should be $M/M/s/B+M$, to account for the existence of B lines. Performance analysis of the $M/M/s/B+M$ queue has been implemented at www.4callcenters.com. This web-site includes two tools, iProfler and Charisma, to support workforce management of inbound call centers. iProfiler is available for online experimentation, free of charge. It will now be used to demonstrate performance sensitivity of a large call center in heavy traffic. In this example, there were sufficiently many lines so that the busy signal phenomenon was negligible. We thus use Erlang A.

Consider Figure 7, which summarizes the daily operation of the Charlotte call center from Figure 5. Note the significant differences in performance over the busy half-hour periods while, on the other hand, the numbers of calling customers, as well as AHT and the number of agents working ("on production") do not seem to vary that significantly. Let us understand these performance differences. For example, during the period 10:30-11:00, the absence of only 5 agents (out of the 223 working) would likely result in almost doubling of both ASA and the fraction abandoning. We arrived at this projection by choosing the average of customers' patience (30 minutes) so that the predicted theoretical performance was close to the observed one. Interestingly and significantly, a model in which average patience is 30 minutes differs dramatically from a model which does not acknowledge abandonment ("infinite patience"): with our parameters, the latter would give rise to an unstable system (agents are required to be busy "more than 100%" of their time); stability could nevertheless be achieved by adding only 2 agents (225 all together), but in this case ASA would get close to 7 minutes - an order of magnitude error in predicting performance if one ignores abandonment (that is, if one uses Erlang C instead of Erlang A).

3.2.2 More models

The prevalent analytical models for performance analysis are sometimes Erlang B and mostly Erlang C: the first is typically inappropriate for not acknowledging waiting; the second lacks central features, notably customer abandonment [31] and heterogeneity [30]. While heterogeneity could requires a leap in modeling capabilities, the Erlang A model is ripe for applications (www.4callcenters.com), and we strongly recommend it as *the standard* to replace the prevalent Erlang C model.

Brandt, Brandt, Spahl, & Weber [19] consider a call center with a finite number of lines, exponential patience and, prior to waiting, an IVR message of constant-duration. The model is thus a two-dimensional network, allowing for only approximations. Brandt & Brandt [18] solve the system with generally distributed patience (times to abandonment) and a finite number of lines. Also Brandt & Brandt [17] study a system with generally distributed patience and a secondary "call back" queue; again, this gives rise to approximations of a two-dimensional network.

Mandelbaum & Shimkin [52] take another perspective: they assume that rational customers compare their expected *remaining* waiting time with their subjective value of service. They provide evidence why rational callers should abandon at some time while being queued. Finally, Zohar, Mandelbaum & Shimkin [69] provide numerical evidence for the thesis of rational adaptive customers and present a new model for abandonment (simpler and more practical than that in [52]). For a discussion on service levels, including abandonment, we recommend Cleveland & Mayben [22].

Reality is even more complicated than described above, as demonstrated by the following reasoning. Decisions on agent staffing must take into account customer patience; the latter, in turn, is influenced by the waiting experience which, circularly, depends on staffing levels. An appropriate framework, therefore, is that of an equilibrium (Game Theory), arrived at through customer self-optimizing and learning. This is the perspective of Mandelbaum & Shimkin [52] and Zohar, Mandelbaum & Shimkin [69], which constitutes merely a first step. In [52], abandonment arises as an equilibrium behavior of rational customers who optimaly compare their expected *remaining* waiting time with their subjective value of service. In [69], the model of [52] is simplified, which enables some support for adaptive behavior (learning) of customers. For a discussion on service-levels that includes abandonment, we recommend Cleveland & Mayben [22].

In Akşin & Harker [1], a model is considered where computer resources are assumed the bottlenecks, and hence they are explicitly modeled. Here all agents compete, in a processor sharing manner, for the

same computer resource. This leads to certain counterintuitive phenomena: for example, performance levels could decrease as the number of agents increase. (In fact, [1] analyses a multi-skill environment.)

3.3 Performance over multiple intervals and overload

To make the translation to intra-day performance, and thus to inhomogeneous Poisson arrivals, (weighted) sums of interval performances are taken, where for each interval another call arrival rate is taken. Green & Kolesar [35] call this the *pointwise stationary approximation*. An alternative idea would be to take the average arrival rate, and use this as input for a performance model. This can give extremely bad results, even if the occupancy is constant; see Green & Kolesar [34, 35].

Insight on the asymptotic behavior of the minimal occupancy as the load tends to ∞ can be obtained from Borst, Mandelbaum & Reiman [15] - see above, where we introduced the square-root staffing principle.

Standard modeling applications for call centers use stationary performance measures for each interval, say of 30 minutes duration. This works in general pretty well. But exceptions arise with abrupt significant changes in arrival rate, particularly when overload occurs during one or more intervals. Then a backlog is built up, and nonstationarity has to be accounted for. As already mentioned, such a behavior could arise from an external event, such as advertising a telephone number on TV, or when the call center opens in the middle of the day. Such abrupt overloads can be modeled with the help of fluid models, as in Mandelbaum, Massey, Reiman, & Rider [49]. These results are extended in Mandelbaum, Massey, Reiman, Rider, & Stolyar[50]. Unfortunately these fluid approximations work less well in underload situations, as has been argued in Altman, Jiménez, & Koole [3]. A numerical way to include nonstationary behavior is described in Fu, Marcus & Wang [28]. Jennings, Mandelbaum, Massey & Whitt [41] propose staffing guidelines, which were developed heuristically and gave rise to a time-varying square-root staffing principle.

3.4 Skill-based routing: on-line and off-line

The operational characteristics of multi-type/channel multi-skill contact centers could get very complicated [30]. Simply conceptualize a call center of say a large European company, which provides technical support in all major European languages for a broad product line. Nevertheless, and out of necessity, most call centers are multi-type multi-skill operations, and hence practice is here awaiting theoretical research for guidelines.

If each skill has dedicated agents, then of course the call center can be regarded as several independent single-skill call centers operating in parallel. But then one does not exploit the economies of scale, due to resource flexibility, of a large call center with multi-skill agents. At the other extreme, complete flexibility where all agents can do all tasks (for example, be able to support all products in all languages) is typically unrealistic. Thus a compromise must be struck where a subset of tasks, which we refer to as a *skill*, can be performed by a subgroup of agents - namely a *skill group*. Skills of different skill groups could overlap, which enables the benefits from economies of scale without the need to train all agents at all skills.

The operational challenges are then both off- and on-line. One should determine off-line the overall number of agents required of each skill, which are to be part of the company's permanent or temporary pool of agents; and out of these, how many and who should occupy a given shift. On-line, one should determine for an idling agent which caller to attend to first; and for an arriving call, who will be the agent to cater to it. In this section we survery on-line problems. The off-line issues are related more to human resource management and are discussed in Section 4 on staffing and workforce management.

Skill-based routing refers to the *on-line* strategy that matches callers and agents. It is nowadays part of any advanced ACD, often provided as a list of options that managers can choose from, but without any guidelines to accompany them. We now survey some related available research. For more information, readers are referred to the short literature survey in [30] and the OR and Simulation sections in [47].

Garnett & Mandelbaum [30] constitute an introduction to skill-based routing and its operational complexities. Via simulation, it is demonstrated there that advantages can be considerable, already for simple scenarios. Perry and Nilsson [55] provide a useful brief introduction to both theory and practice.

A common way of implementing skill-based routing is by specifying two selection rules: agent selection - how does an arriving call select an idle agent, if there is one; and call selection - how does an idle agent select a waiting call, if there is one. Here are some details. Agents are first divided into groups such that all agents within the group share the same skills. In general, several groups could have the same skill. The PABX/ACD contains, for each skill, an ordered list of agent groups containing that skill. An arriving call for a certain skill is then assigned to the first group in the list that has an agent available. When no agent with the right skill is available, then the call is assigned to the first agent with the skill that becomes available. If an available agent can handle each one of several waiting calls, then some priority rule is employed in order to determine which call to handle first. As far as we know, this common protocol has not been analyzed analytically.

If one leaves out the possibility that a call finds all agents occupied, then a flow of calls of a certain type from one agent group to the next group occurs only if all agents are occupied, i.e., it is overflow. These are notoriously hard to analyze, see [40], because the overflow process is not Poisson. The performance of this type of an overflow queueing network in the context of call centers is studied in Koole & Talim [45].

It is also possible to program a PABX in such a way that a call is assigned to a group only if there is at least a certain threshold number of agents available for service. Thus agents are reserved idle for future high-priority calls while low-priority calls are presently waiting to be served. This becomes useful if a group has skills of varying importance, and it is advisable to reserve several agents free for the most important call types.

Although the above protocol is commonplace, it is certainly not optimal. E.g., it can occur that the last agent with skill A is occupied by a call of skill B, while there are multiple agents available with skills B and C. This effect cannot be avoided by changing the routing lists, due to the random behavior of the system. In fact, to reach optimal routing, one has to take the number of available agents in all groups into account. This way the routing becomes completely dynamic. The standard way to solve this type of problems is by Dynamic Programming. Unfortunately, it is impossible to apply standard Dynamic Programming to identify the optimal assignment, neither theoretically (the problem as of now seems too hard) nor practically, due to the so-called *curse of dimensionality* [11]: the number of possible configurations is exponential in the number of agent groups, making it numerically infeasible to apply standard algorithms from Markov decision theory. One way to overcome the problem's complexity is to consider simple structures and specific strategies. For example, [55] consider a two-channel system, where waiting customer are assigned an *aging factor*, proportional to their waiting time. Then customers with the largest aging factor is chosen for service. Alternatively, one could analyze provably-reasonable approximations, for example Borst & Seri [16]. Both [55] and [16] consider the on-line routing problem as well as the of-line staffing problem - namely, how many agents are to be available for answering calls so as to maintain an acceptable grade of service. ([16] actually applies the square-root staffing principle.)

3.5 Call blending and multi-media

Different multi-media services require differing response times. Specifically, telephone services should be responded to within seconds or minutes and, once started, should not be interrupted; e.mail and fax, on the other hand, can be "stored" towards response within hours or days, and can definitely be preempted by telephone calls, and then resumed; chat services are somewhere in between. In Mandelbaum, Massey, & Reiman [48], a mathematical asymptotic framework of Markovian Service Networks is developed, where multi-type customers are served according to preemptive-resume priority disciplines. The primitives of a Markovian service network are time-varying, abandoment and retrials are accomodated, and the asymptotics is in the rationalized (Halfin-Whitt) regime. The framework of [48] is thus applicable for performance analysis of large multi-media call centers - as indeed was done in [49, 50]. Note however that the framework can not accommodate non-preemptive priority disciplines or finite buffers (busy signals).

We now continue with models that include IVR and e.mail. Brandt and Brandt [17], already mentioned in the context of abandonment, propose a (birth-and-death) queueing model for a call center with impatient callers and an integrated IVR: callers that are patient enough, and which have been waiting online beyond a given threshold, are then transferred to ("stored in") an IVR-queue; the latter is served later, as soon as no customers are waiting online, and the number of idle agents exceeds another threshold. Armony and Maglaras [8] establish the asymptotic optimality in equilibrium of such a threshold strategy, when customers act rationally. By this we mean that customers who are not served immediately optimize among balking, abandoning, or opting for a return call (or a later e.mail) if they assess their anticipated delay as exceeding its worth. The equilibrium formulation is inspired (but differs from) [52, 69]; the asymptotics is taken in the rationalized (Halfin-Whitt) regime.

If we mix traffic from multiple channels, then additional questions arise. Historically, these questions first arose in the context of mixing inbound and outbound traffic, but they are also applicable to multi-media traffic. The solution is called *call blending*, where agents are made to switch between inbound and outbound traffic, depending on the traffic loads of inbound traffic. A mathematical model for call blending is presented and solved in Bhulai & Koole [13].

Pure outbound Call centers are becoming more prevalent, mainly in surveys and tele-marketing. They use devices called *predictive dialers* that automatically call up customers, according to a prepared list. In order to reduce idleness of the most expensive call center resource, its agents, it often happens that the PABX calls the next customer on the list while, in fact, there are no agents available to take the call. Thus, the central problem is balancing between agent productivity (is there always a customer right away?) and customer dissatisfaction (no agent is idle while a customer picks up the phone), in a manner that is consistent with the company-specific relative importance of these two goals. For more information on predictive dialers, see Samuelson [62].

3.6 Geographically dispersed call centers

Another subject, growing in importance, is that of multiple geographically dispersed call centers. By interconnecting them properly (dynamic load balancing), performance can get close to that of a single virtual call center, thus exploiting fully the economies of scale. This is the case in Figure 5, the header of which reads "Command Center Intraday Report": and indeed, load balancing is exercised from a single Command Center that overseas the 12 call centers represented in the table. An ACD that distributes calls to several call centers is often referred to as a network-ACD.

Servi & Humair [61] analyze the problem of setting routing probabilities, but more can be gained if routing is completely dynamic. Kogan et al. [43] compares two basic strategies for a network-ACD:

a centralized FIFO vs. a distributed strategy that routes an arriving call to the call center with least expected delay. Both strategies require information-exchange over the network. While FIFO is, of course, much more taxing, it could nevertheless be still inferior, given certain delays in switching calls between centers. This paper provides references to previous works on the subject, by the same group at AT&T.

4 Workforce management

Up to now, we have mainly discussed mathematical models for determining minimal occupancy levels. Next we shall see how this fits into the overall picture of workforce management as it is practiced in call centers. A well written practical book about workforce management, that also touches on some mathematical issues, is Cleveland & Mayben [22].

4.1 Decisions at the tactical and strategic levels

Long-term strategic issues are related to the way in which a call center operates and to the way in which its human resources are managed. Questions in the first category include the deployment of a voice response unit (VRU) to answer simple questions thus saving costly human intervention, whether or not to have different skill groups, etc. Questions in the second category deal for example with the type of agents contracts used: part-time vs. full-time, a fixed or flexible number of working hours, etc. It is least expensive to have full-time non-flexible contracts because training part-time agents is relatively expensive (the costs per working hour is higher), and because flexibility has its price as well. Whether employing full-time agents is desirable from a scheduling point of view depends on the organization and traffic load of the call center. The organization determines routing possibilities, the way skills are dealt with, the possibilities of call blending, etc. Then, using a performance model, it should be determined what mix of contract types is best for the call center. It is interesting to note that the operational performance model can play an important role in the strategic decisions of the call center. A model for hiring and training decisions can be found in Gans & Zhou [29].

The importance of good staffing decisions is illustrated by the following reasoning. Staffing policies should of course take into account agents' turnovers (often very high), training (could be costly) and on-the-job learning (possibly slow); the latter factors depend on job-enrichment and career paths (Human Resource Management), which are enabled by the technology of skill-based-routing (Operations Research) that, again circularly, affect staffing. Similarly, operationally-driven staffing could lead to exaggerated agents utilization that, in turn, promote burnout and hence turnovers. Avoiding such a negative spiral by having good working conditions and inviting incentives is typically less expensive than trying to recover out of one.

At the tactical level there are issues related to the skills that agents acquire. Until recently, career-growth opportunities within call centers were unusual, resulting in low motivation (hence low service level) and fast burnout (hence high turn-over). Some even refer to the call centers as the *modern-age sweat shops*. But in today's call centers agents could start as specialists with a single skill, and then get up the ranking by acquiring more skills and being granted more responsibilities. The result is the possibility for a full working career, from the novice to the manager of a large multi-national operation. But then the challenge is to design and implement appropriate hiring and training plans which ensure that, at all times, each call center has at its disposal the right number of agents with the right skills. This would require a multi-disciplinary effort, with an important part being played by performance models and analysis.

4.2 Operational decisions

Operational decisions vary from those taken a few weeks in advance of the day concerned until those controls that take effect immediately. In most call centers there is a planner that is responsible for agents rosters. This person starts every week or every couple of weeks with preparing a forecast for the specified period. Based on this forecast, occupancy levels are determined, and together with agent and management input (concerning days off, meeting, etc.), a roster is determined. This process is very often supported by a *workforce management tool*. Such a WFM tool integrates a forecasting algorithm, a performance model that allows for the computation of staffing levels, and (mostly) mathematical programming models that have as output agent rosters. See the next subsection for a discussion of staffing models. These WFM tools are decision support systems in the true sense of the word, in that they are often very user friendly, and that they allow for ample communication between the system and the user. E.g., search Google for "call center workforce management" for a list of vendors.

Establishing the agent roster is not yet the end of story: changes will take place to this roster until the scheduled day itself, based on changing forecasts and internal and external events. When the roster is executed then a supervisor is responsible for the service levels and the productivity. He or she monitors the service levels, changes agents of group, organizes call blending (if this is not done automatically), etc.

During the day, data is fed back into the workforce management tool. Based on this feedback, forecasts are updated or new ones created, and the process repeats itself.

4.3 Staffing models

Consider first a single-skill call center. One simplifies the problem of determining the working hours of each agent by splitting it into two steps: first the shifts are determined, and then the agents are assigned to shifts. Different approaches for the first step are used. E.g., a heuristic approach is advocated in Henderson & Berry [39], while Segal [60] uses linear programming. Other aspects such as break placements are also studied in detail (see Aykin [9] and references therein). An overview of the area (not necessary dealing with call centers) can be found in the introduction to Thompson [64]. A paper that describes a case study in which agents had to be assigned to shifts (the second step) is Thompson [65].

Translating minimal interval occupancies to agent rosters is a task that occurs also in contexts other than call centers and hence it has been well studied. See [60] and [39]. These papers use approaches based on integer programming.

There are also some papers that focus on break placement (see Aykin [9] and references therein). However, in practice there are many additional constraints that make it necessary to use other optimization techniques. Therefore we see in practice techniques such as constraint satisfaction and techniques based on local search employed in software. In the literature there are no papers that deal with this assignment problem in its full generality.

The model needed becomes even more complex if one realizes that splitting the problem in two (making shifts first, and then assigning shifts to agents) can turn out to be significantly suboptimal: the availability of agents can greatly influence the types of shifts required. This calls for a method in which shift determination and shift assignment are integrated. Heuristic methods allow for this. Koole & van der Sluis [44] go a step further: they argue that shift determination (and assignment) should be integrated with the determination of the desired occupancy level. Indeed, requiring that service levels are satisfied for every time interval together with minimal shift lengths generally result in overstaffing during some intervals. But overstaffing in certain intervals could be compensated by

understaffing in other intervals. This requires combined performance and staffing models that work with a single service level for the entire day.

The situation in a multi-skill call center is even more complicated. Many different agent combinations might be possible for fulfilling service requirements. For example, in a two-skill call center, it might be that during a certain interval 5 skill-1 and 10 skill-2 agents are necessary. But 13 agents with both skills might also suffice, or an in-between configuration, with both single skill and double skill agents, might do. Many different agent combinations potentially solve the occupancy model, even if shift duration are fixed. This makes necessary the integration of the performance model and the occupancy model . Creating rosters becomes an extremely difficult task, for which no methods are available in the research literature.

Another type of problems arises if we consider multiple channels. The central difference, from a scheduling point of view, is that processing capacity for the low-service level channels can be assigned to different intervals. E.g., work that comes in through faxes or regular mail can be handled by agents at times when there is over-capacity. As such, contact centers enjoy an advantage over call centers in that some of the offered load can be shifted between intervals (inventoried). This can help reduce costs but, at the same time, it adds complexity to scheduling.

5 Conclusions

The world of call centers is a challenging fertile area for the applications of queueing models. Simple models are already incorporated into workforce management tools and are widely used. However, current explosion in scale and scope necessitate the use of more complex models, and many of these are yet to be formulated, analyzed and implemented. Only then shall we gain the full benefits of the modern contact center.

References

[1] O.Z. Akşin and P.T. Harker. Analysis of a processor shared loss system. *Management Science*, 47:324–336, 2001. 1.4, 3.2.2, 3.2.2

[2] O.Z. Akşin and P.T. Harker. Capacity sizing in the presence of a common shared resource: Dimensioning an inbound call center. Working paper, 2001. 1.4, 1.4, 1.8

[3] E. Altman, T. Jiménez, and G.M. Koole. On the comparison of queueing systems with their fluid limits. *Probability in the Engineering and Informational Sciences*, 15:165–178, 2001. 3.3

[4] B. Andrews and S.M. Cunningham. L.L. Bean improves call-center forecasting. *Interfaces*, 25(6):1–13, 1995. 2.1

[5] B. Andrews and H. Parsons. Establishing telephone-agent staffing levels through economic optimization. *Interfaces*, 23(2):14–20, 1993. 1.4

[6] J. Anton. The past, present and future of customer access centers. *International Journal of Service Industry Management*, 11:120–130, 2000. 1, 1

[7] R. Anupindi and B.T. Smythe. Call centers and rapid technology change. Teaching Note. Submitted, 1997. 1, 1

[8] M. Armony and C. Maglaras. Customer contact centers with multiple service channels. Working paper, 2001. 1.8, 3, 3.5

[9] T. Aykin. Optimal shift scheduling with multiple break windows. *Management Science*, 42:591–602, 1996. 4.3, 4.3

[10] F. Baccelli and G. Hebuterne. On queues with impatient customers. In *Performance '81*, pages 159–179. North-Holland, 1981. 2.3.3, 3.2, 3.2.1

[11] R. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961. 3.4

[12] L. Bennington, J. Commane, and P. Conn. Customer satisfaction and call centers: an Australian study. *International Journal of Service Industry Management*, 11:162–173, 2000. 1.4

[13] S. Bhulai and G.M. Koole. A queueing model for call blending in call centers. In *Proceedings of the 39th IEEE CDC*, pages 1421–1426. IEEE Control Society, 2000. 1.8, 3.5

[14] V.A. Bolotin. Telephone circuit holding time distributions. In J. Labetoulle and J.W. Roberts, editors, *Proceedings of the 14th International Teletraffic Conference*, pages 125–134, 1994. 2.3.2

[15] S.C. Borst, A. Mandelbaum, and M.I. Reiman. Dimensioning large call centers. Working paper, 2000. 1.4, 1.6, 1.7, 3.1.1, 3.1.1, 3.1.1, 3.1.1, 3.1.2, 3.3

[16] S.C. Borst and P. Seri. Robust algorithms for sharing agents with multiple skills. Working paper, 2000. 3, 3.1.1, 3.4, 3.4, 3.4

[17] A. Brandt and M. Brandt. On a two-queue priority system with impatience and its application to a call center. *Methodology and Computing in Applied Probability*, 1:191–210, 1999. 3.2.2, 3.5

[18] A. Brandt and M. Brandt. On the $M(n)/M(n)/s$ queue with impatient calls. *Performance Evaluation*, 35:1–18, 1999. 3.2.2

[19] A. Brandt, M. Brandt, G. Spahl, and D. Weber. Modelling and optimization of call distribution systems. In V. Ramaswami and P.E. Wirth, editors, *Proceedings of the 15th International Teletraffic Conference*, pages 133–144. Elsevier Science, 1997. 1.8, 3.2.2

[20] A.J. Brigandi, D.R. Dargon, M.J. Sheehan, and T. Spencer III. At&t's call processing simulator (caps) operational design for inbound call centers. *Interfaces*, 24(1):6–28, 1994. 1, 1

[21] E. Chlebus. Empirical validation of call holding time distribution in cellular communications systems. In *Proceedings of the 15th International Teletraffic Conference*, 1997. 2.3.2

[22] B. Cleveland and J. Mayben. *Call Center Management on Fast Forward*. Call Center Press, 1997. 1, 3.2.2, 3.2.2, 4

[23] D. Duxbury, R. Backhouse, M. Head, G. Lloyd, and J. Pilkington. Call centres in BT UK customer service. *British Telecommunications Engineering*, 18:165–173, 1999. 1, 1, 2

[24] A.K. Erlang. Solutions of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Electroteknikeren*, 13:5–13, 1917. In Danish. 3.1

[25] A. Evenson, P.T. Harker, and F.X. Frei. Effective call center management: Evidence from financial services. Working paper 99–25–B, Wharton Financial Institutions Center, 1998. 1, 1

[26] M.A. Feinberg. Performance characteristics of automated call distribution systems. In *GLOBECOM '90*, pages 415–419. IEEE, 1990. 3.2

[27] R.A. Feinberg, I.-S. Kim, L. Hokama, K. de Ruyter, and C. Keen. Operational determinants of caller satisfaction in the call center. *International Journal of Service Industry Management*, 11:131–141, 2000. 1.4, 1.5

[28] M.C. Fu, S.I. Marcus, and I-J. Wang. Monotone optimal policies for a transient queueing staffing problem. *Operations Research*, 48:327–331, 2000. 3.3

[29] N. Gans and Y.-P. Zhou. Managing learning and turnover in employee staffing. Working paper 99-39, Wharton Financial Institutions Center, 1999. 1.10, 4.1

[30] O. Garnett and A. Mandelbaum. An introduction to skills-based routing and its operational complexities. Teaching note. 1.8, 3, 3.2.2, 3.4, 3.4, 3.4

[31] O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. Working paper. 1.7, 1.8, 3, 3.1.1, 3.2, 3.2, 3.2.1, 3.2.2

[32] A. Gilmore and L. Moreland. Call centres: How can service quality be managed? *Irish Marketing Review*, 13:3–11, 2000. 1.4

[33] J.J. Gordon and M.S. Fowler. Accurate force and answer consistency algorithms for operator services. In J. Labetoulle and J.W. Roberts, editors, *Proceedings of the 14th International Teletraffic Conference*, pages 339–348, 1994. 2.3.1

[34] L. Green and P. Kolesar. Testing the validity of a queueing model of police patrol. *Management Science*, 37:84–97, 1989. 1.7, 1.10, 3.3

[35] L. Green and P. Kolesar. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science*, 37:84–97, 1991. 3.3, 3.3

[36] T.A. Grossman, D.A. Samuelson, S.L. Oh, and T.R. Rohleder. Call centers. In S.I. Gass and C.M. Harris, editors, *Encyclopedia of Operations Research and Management Science*. 2nd edition, 1999. To appear. 1

[37] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29:567–587, 1981. 3.1, 3.1.2

[38] C.M. Harris, K.L. Hoffman, and P.B. Saunders. Modeling the irs telephone taxpayer information system. *Operations Research*, 35:504–523, 1987. 1.7, 1.8, 2.3.2, 3

[39] W.B. Henderson and W.L. Berry. Heuristic methods for telephone operator shift scheduling: An experimental analysis. *Management Science*, 22:1372–1380, 1976. 4.3, 4.3

[40] A. Hordijk and A. Ridder. Stochastic inequalities for an overflow model. *Journal of Applied Probability*, 24:696–708, 1987. 3.4

[41] O.B. Jennings, A. Mandelbaum, W.A. Massey, and W. Whitt. Server staffing to meet time-varying demand. *Management Science*, 42:1383–1394, 1996. 2.1, 3, 3.1.1, 3.3

[42] G. Jongbloed and G.M. Koole. Managing uncertainty in call centers using Poisson mixtures. Technical Report 2000-3, Department of Stochastics, Vrije Universiteit Amsterdam, 2000. Electronically available as www.cs.vu.nl/˜koole/papers/S2000-3.ps. 2.1, 2.3.1

[43] Y. Kogan, Y. Levy, and R.A. Milito. Call routing to distributed queues: Is FIFO really better than MED? *Telecommunication Systems*, 7:299–312, 1997. 1.8, 3, 3.6

[44] G.M. Koole and H.J. van der Sluis. An optimal local search procedure for manpower scheduling in call centers. Technical Report WS-501, Vrije Universiteit Amsterdam, 1998. Electronically available at www.cs.vu.nl/obp/callcenters. 1.5, 4.3

[45] G.M. Koole and J. Talim. Exponential approximation of multi-skill call centers architecture. In *Proceedings of QNETs 2000*, pages 23/1–10, 2000. 3, 3.4

[46] B.W. Kort. Models and methods for evaluating customer acceptance of telephone connections. *IEEE*, pages 706–714, 1983. 1.7, 2.3.2, 2.3.3, 2.3.3, 2.3.3, 2.3.3, 2.3.3, 2.3.3

[47] A. Mandelbaum. Call centers (centres): Research bibliography with abstracts. Electronically available as ie.technion.ac.il/∼serveng/References/ccbib.pdf, 2001. 1, 1.4, 1.4, 1.6, 1.9, 1.9, 1.9, 1.9, 1.9, 2, 2, 2.3.3, 3.4

[48] A. Mandelbaum, W.A. Massey, and M.I. Reiman. Strong approximations for Markovian service networks. *Queueing Systems*, 30:149–201, 1998. 3.5, 3.5

[49] A. Mandelbaum, W.A. Massey, M.I. Reiman, and R. Rider. Time varying multiserver queues with abandonments and retrials. In P. Key and D. Smith, editors, *Proceedings of the 16th International Teletraffic Conference*, 1999. 3.1, 3.3, 3.5

[50] A. Mandelbaum, W.A. Massey, M.I. Reiman, R. Rider, and A. Stolyar. Queue lengths and waiting times for multiserver queues with abandonment and retrials. Working paper, 2000. 1.8, 3, 3.1, 3.3, 3.5

[51] A. Mandelbaum, A. Sakov, and S. Zeltyn. Empirical analysis of a call center. Working paper, 2000. (document), 1, 1, 1.5, 1.5, 1.7, 2.1, 2.1, 2.1, 2.1, 2.1, 2.2.4, 2.2.4, 2.3.2, 2.3.2, 2.3.3, 2.3.3, 2.3.3, 2.3.3, 2.3.3, 2.3.3, 3.2

[52] A. Mandelbaum and N. Shimkin. A model for rational abandonments from invisible queues. *Queueing Systems*, 36:141–173, 2000. 2.3.3, 3.2.2, 3.2.2, 3.2.2, 3.2.2, 3.2.2, 3.5

[53] V. Mehrotra. Ringing up big business. *OR/MS Today*, pages 18–24, August 1997. 1, 1

[54] C. Palm. Methods of judging the annoyance caused by congestion. *Tele*, 4:189–208, 1953. 2.3.3, 2.3.3, 2.3.3, 2.3.3, 2.3.3, 2.3.3, 3, 3.1, 3.2.1

[55] M. Perry and A. Nilsson. Performance modeling of automatic call distributors: Assignable grade of service staffing. In *XIV International Switching Symposium*, pages 294–298, 1992. 3.4, 3.4, 3.4

[56] M. Pinedo, S. Seshadri, and J.G. Shanthikumar. Call centers in financial services: Strategies, technologies, and operations. In E.L. Melnick, P. Nayyar, M.L. Pinedo, and S. Seshadri, editors, *Creating Value in Financial Services: Strategies, Operations and Technologies*. Kluwer, 1999. 1, 1

[57] A.A. Puhalskii and M.I. Reiman. The multiclass GI/PH/N queue in the Halfin-Whitt regime. *Advances in Applied Probability*, 32:564–595, 2000. 3.1, 3.1.1

[58] J. Riordan. *Stochastic Service Systems*. Wiley, 1961. 3.2

[59] J.W. Roberts. Recent observations of subscriber behavior. In *Proceedings of the 9th International Teletraffic Conference*, 1979. 1.7, 2.3.3, 2.3.3, 2.3.3, 2.3.3, 2.3.3, 2.3.3, 2.3.3

[60] M. Segal. The operator-scheduling problem: A network-flow approach. *Operations Research*, 24:808–823, 1974. 4.3, 4.3

[61] L. Servi and S. Humair. Optimizing Bernoulli routing policies for balancing loads on call centers and minimizing transmission costs. *Journal of Optimization Theory and Applications*, 100:623–659, 1999. 3.6

[62] D.A. Sumuelson. Predictive dialing for outbound telephone call centers. *Interfaces*, 29(5):66–81, 1999. 3.5

[63] D.Y. Sze. A queueing model for telephone operator staffing. *Operations Research*, 32:229–249, 1984. 1.8, 2.3.2, 2.3.2, 3.1, 3.1.2

[64] G.M. Thompson. Improved implicit optimal modeling of the labor shift scheduling problem. *Management Science*, 41:595–607, 1995. 4.3

[65] G.M. Thompson. Assigning telephone operators to shifts at New Brunswick Telephone Company. *Interfaces*, 27(4):1–11, 1997. 4.3

[66] G. Tom, M. Burns, and Y. Zeng. Your life on hold: The effect of telephone waiting time on customer perception. *Journal of Direct Marketing*, 11:25–31, 1997. 1.4

[67] W. Whitt. Improving service by informing customers about anticipated delays. *Management Science*, 45:192–207, 1999. 2.3.2

[68] W. Xu. Long range planning for call centers at FedEx. *The Journal of Business Forecasting Methods & Systems*, 18:7–11, Winter 1999/2000. 1.10

[69] E. Zohar, A. Mandelbaum, and N. Shimkin. Adaptive behavior of impatient customers in tele-queues: Theory and emperical support. working paper, 2000. 1.5, 1.7, 2.3.3, 2.3.3, 3, 3.2.2, 3.2.2, 3.2.2, 3.5