Telephone Call Centers: a Tutorial and Literature Review

Noah Gans[*] Ger Koole[†] Avishai Mandelbaum[‡]

[*]The Wharton School, University of Pennsylvania,Philadelphia, PA 19104, U.S.A.
gans@wharton.upenn.edu http://opim.wharton.upenn.edu/∼gans

[†]Vrije Universiteit, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
koole@cs.vu.nl http://www.cs.vu.nl/∼koole

[‡]Industrial Engineering and Management, Technion, Haifa 32000, Israel
avim@tx.technion.ac.il http://ie.technion.ac.il/Home/Users/avim.phtml

September 2, 2002

Abstract

Telephone call centers are an integral part of many businesses, and their economic role is significant and growing. They are also fascinating socio-technical systems in which the behavior of customers and employees is closely intertwined with physical performance measures. In these environments traditional operational models are of great value – and at the same time fundamentally limited – in their ability to characterize system performance.

We characterize the state of research on telephone call centers. We begin with a tutorial on how call centers function and proceed to survey academic research devoted to the management of their operations. We then outline important problems that have not been addressed and identify promising directions for future research.

Acknowledgments

The authors thank Lee Schwarz, Wallace Hopp and the editorial board of *M&SOM* for initiating this project. Thanks are also due to L. Brown, A. Sakov, H. Shen, S. Zeltyn and L. Zhao for their approval of importing pieces of [33, 106].

Financial support was provided by NSF Grants SBR-9733739 and DMI-0223304 (N.G. and A.M.), the Wharton Financial Institutions Center (N.G. and A.M.), ISF Grants (A.M., jointly with N. Shimkin and R. Atar), and the Technion funds for the promotion of research and sponsored research (A.M.)

Some data originated with member companies of the Call Center Forum at Wharton, to whom we are grateful. Some material was adapted from the Service Engineering site prepared by A.M. and S. Zeltyn (http://ie.technion.ac.il/serveng/).

Parts of the manuscript were written while A.M. was visiting the Vrije Universiteit and the Wharton School – the hospitality of the hosting institutions is greatly appreciated.

Key Words: telephone call center, contact center, tele-services, tele-queues, capacity management, staffing, hiring, workforce management systems, ACD reports, queueing, Erlang C, Erlang B, Erlang A, QED regime, time-varying queues, call routing, skills-based routing, forecasting, data mining.

Contents

1 Introduction

Call centers and their contemporary successors, contact centers, have become a preferred and prevalent means for companies to communicate with their customers. Most organizations with customer contact – private companies, as well as government and emergency services – have reengineered their infrastructure to include from one to many call centers, either internally managed or outsourced. For many companies, such as airlines, hotels, retail banks, and credit card companies, call centers provide a primary link between customer and service provider.

The call center industry is thus vast and rapidly expanding, in terms of both workforce and economic scope. For example, a recent analyst's report estimates the number of agents working in U.S. call centers to have been 1.55 million in 1999 - more than 1.4% of private-sector employment - and to be growing at a rate of more that 8% per year [43, 140]. In 1998, AT&T reported that on an average business day about 40% of the more than 260 million calls on its network were toll-free [1]. One presumes that the great majority of these 104 million daily – "1-800" calls terminated at a telephone call center.

The quality and operational efficiency of these telephone services are also extraordinary. In a large, best-practice call center, many hundreds of agents can cater to many thousands of phone callers per hour; agent utilization levels can *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 a mere 1-2%.

Given this performance, one would imagine that the design and management of call center operations are based on sound scientific principles. But in fact, the software that is often used to support them makes use of only the simplest analytical models. These models have performed an important role in the management of call centers, but they leave much to be desired. More sophisticated approaches are needed to accurately describe the reality of call-center operations, and they can improve call-center performance significantly.

More broadly, the continued growth in both the economic importance and complexity of call centers has prompted increasingly deep investigation of their operations. This is manifested by a growing body of academic work devoted to call centers, research ranging in discipline from Mathematics and Statistics, through Operations Research, Industrial Engineering, Information Technology and Human Resource Management, all the way to Psychology and Sociology (see Mandelbaum [97]).

In this article, we provide a tutorial on call centers, as well as a review of the academic literature that is related to the management of their operations. We first outline important operational problems, and our focus is on mathematical models which potentially support their management. Our literature review thus addresses primarily *analytical* models that support *capacity management*.

Analytical models can be contrasted with simulation techniques, which have been growing in popularity (see Section VIII in [97]). This growth has occurred partly because of improved user-friendliness of simulation tools and partly in view of the scarcity of mathematical skills required for the analytical alternatives. Perhaps it is mostly due to the widening gap between the complexity of the modern call center and the analytical models available to accommodate this complexity.

We will not dwell here on the virtues and vices of analytical versus simulation models. Our contention is that, ideally, one should blend the two: analytical models for insight and calibration,

simulation 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.

There are two related reasons for our focus on capacity management. First, in most call centers capacity costs in general, and human resource costs in particular, account for 60–70% of operating expenses. Thus, from a cost perspective capacity management is critical. Second, the majority of research to date has addressed capacity management. This no doubt reflects the traditional emphasis of operations management (OR / IE) research, and it also is due to researchers' sensitivity to the economic importance of capacity costs.

Nevertheless, these traditional operational models do not capture a number of critical aspects of call-center performance, and we also discuss what we believe to be important determinants that have not been adequately addressed. These topics include a better understanding of the role played by human factors, as well as the better use of new technologies, such as networking and "skills-based routing" tools. Indeed, these behavioral and technological issues are closely intertwined, and we believe that the ability to address these problems will often require multi-disciplinary research.

1.1 Additional Resources

Our article is a complement to a number of research resources that already exist. In particular, Mandelbaum [97] provides a comprehensive bibliography of call-center-related work. It includes references and abstracts that cover well over 250 research papers. Indeed, 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 through industry sites. For a list of web sites, see Section XI of [97].

There also exists a number of academic review articles of which we are aware: Pinedo et al. [119] provides the basics of call center management, including some analytical models; Anupindi and Smythe [10] describes the technology that enables current and plausibly future call centers; Grossman et al. [67] and Mehrotra [114] are both short overviews of some OR challenges in call center research and practice; Anton [9] provides a managerial survey of the past, present and future of customer contact centers; and Koole and Mandelbaum [89] is more narrowly focused on queueing models. One may view our survey as a supplement to these articles, one that is aimed at academic researchers that seek an entry to the subject, as well as at practitioners who develop call-center applications.

Additional articles that we recommend as part of a quantitative introduction to call centers include the following. Buffa et al. [36] is an early, comprehensive treatment of the hierarchical framework used by call centers to manage capacity. The series of four articles by Andrews et al. [6, 7, 8, 122] constitutes an interesting record of this group's work with the call center of the catalogue retailer, L.L. Bean. Similarly, Brigandi et al. [31] present work by AT&T that demonstrates the monetary value of call center modelling. Mandelbaum, Sakov and Zeltyn [106], parts of which have been adapted to the present text, provides a thorough descriptive analysis of operational data from a call center, and Brown et al. [33] is its complementary statistical analysis. Evenson et al. [48] and Duxbury et al. [44] discuss performance drivers and the state of the art. Finally, Cleveland and Mayben [42] is a well-written overview by and for practitioners (though we take exceptions to some of its views, notably its treatment of customer abandonment [59] and its capacity-sizing recommendations [133]).

1.2 Paper Structure and Reading Guide

The headings within the Table of Contents provide some detail on the material covered in the various sections. Here we offer a complementary overview of the paper's structure. We also provide a guide for readers with specific interests.

The paper is structured as follows. Section 2 provides a tutorial background on call centers and how they function. Section 3 presents a "base case" example – of a homogeneous population of customers being served by a homogeneous group of agents – to outline a hierarchy of capacity-management problems that is central to call-center operations. In Section 4 we then review research that considers the setting of the base-case problems introduced in Section 2. Topics include queueing-theoretic staffing models, mathematical-programming models used for scheduling agents, and longer-term models for hiring and training. Section 5 describes advances in call routing, multimedia, and networking technology that significantly broaden the operational problems discussed in Sections 3 and 4. It then reviews current research, as well as a number of open questions that are motivated by the new capabilities. In Section 6, we turn to an important and underdeveloped area for call-center research: data analysis and forecasting for call-center operations. Section 7 describes what we currently believe to be the future of call-center research: new approaches and broad questions that, as of now, are largely unexplored. Finally, Section 8 offers a brief conclusion.

Thus, the paper offers a systematic introduction to call center operations and research. In reading it from beginning to end, one should be able to develop a fairly complete understanding of how call centers operate, as well as what insights academic research provides into their operational problems.

Those with an interest in a particular area of call-center operations or research may wish to focus on specific sections of the paper, however. After reading or skimming through Section 2, the sections they should concentrate on will vary. What follows is a list of potential topic choices.

- Queueing performance models for multiple-server systems: Sections 3.2, 4.1–4.4, 4.7, and 7.
- Queueing control models for multiple-server, multi-class systems: Section 5 and 7.
- Human resources problems associated with personnel scheduling, hiring and training: Sections 3.2, 3.5, 4.5–4.6, 4.7.2, and 7.
- Service quality, and customer and agent behavior: Sections 2.5, 3.5, 6.3.2 –6.3.4.
- Statistical analysis of call-center data: Sections 2.3, 3.3, 6, and 7.

For readers with special interests, we also provide forward pointers at the ends of relevant sections.

Note that Sections 2.2–2.3 introduce and define commonly used call-center names and acronyms that we use throughout the paper. A summary of the abbreviations and their definitions can be found in Appendix A.

2 Overview of Call-Center Operations

This section provides a tutorial on call-center operations. In §2.1, we provide background on the scope of call-center operations. Then in §2.2 we describe how call centers work, and we define

common call-center nomenclature. Next, §2.3 describes how call centers commonly monitor their operations and measure their operating performance. In §2.4 we highlight the relationship between call centers and queueing systems. Then §2.5 discusses measures of service quality commonly used in call centers.

2.1 Background

At its core, 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) can 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 phantom customers.

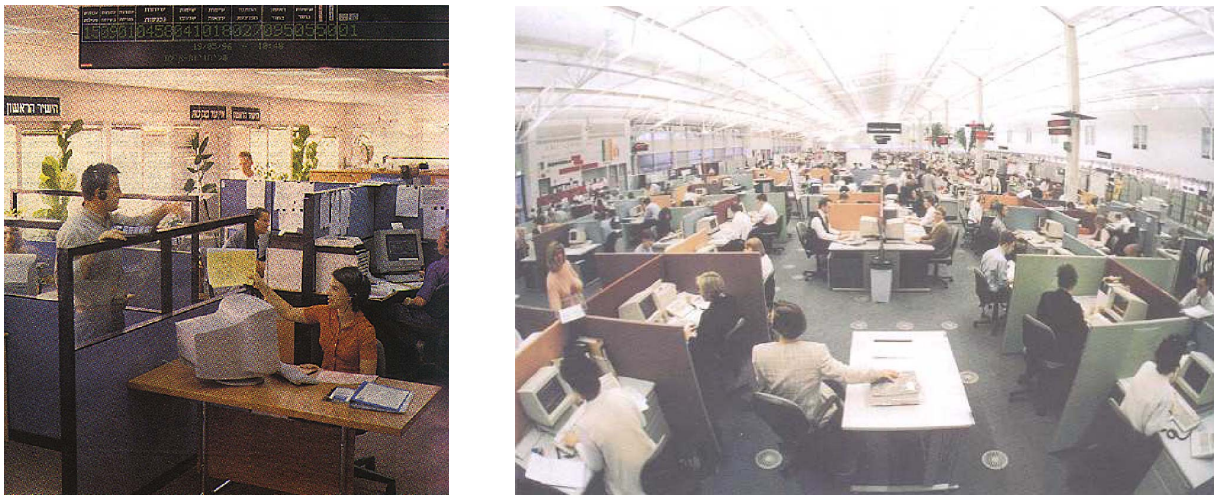


Figure 1: The Working Environment of a Call Center (right image of First Direct [93])

Call centers can be categorized along many dimensions. The functions that they provide are highly varied: from customer service, help desk, and emergency response services, to tele-marketing and order taking. They vary greatly in size and geographic dispersion, from small sites with a few agents that take local calls – for example at a medical practice – to large national or international centers in which hundreds or thousands of agents may be on the phone at any time.

Furthermore, the latest telecommunications and information technology allow a call center to be the *virtual* embodiment of a few or many geographically dispersed operations. These range from small groups of very large centers that are connected over several continents – for example, in the U.S.A., Ireland, and India – to large collections of individual agents that work from their homes.

The organization of work may also vary dramatically across call centers. When the skill-level required to handle calls is low, a center may cross-train every employee to handle every type of call, and calls may be handled first come, first-served (FCFS). In settings that require more highly-skilled work, each agent may be trained to handle only a subset of the types of calls that the center serves, and "skills-based routing" may be used to route calls to appropriate agents. In turn, the organizational structure may vary from the very flat – in which essentially all agents are exposed

to external calls – to the multi-layered – in which a layer represents a level of expertise – and customers may be transferred through several layers before being served to satisfaction.

A central characteristic of a call center is whether it handles *inbound* or *outbound* traffic. Inbound call centers handle *incoming* calls that are initiated by outside callers calling *in* to a center. Typically, these types of centers provide customer support, help desk services, reservation and sales support for airlines and hotels, and order-taking functions for catalog and web-based merchants. Outbound call centers handle *outgoing* calls, calls that are initiated from within a center. These types of operations have been traditionally associated with tele-marketing and survey businesses. A recent development in some inbound centers is to initiate outbound calls to high-value customers who have abandoned their calls before being served.

Our focus in this article is on inbound call centers, with some attention given to mixed operations that blend incoming and outgoing calls. In fact, we are aware of almost no academic work devoted to pure outbound operations, the exception being Samuelson [126]. Within inbound centers, the agents that handle calls are often referred to as *customer service representatives* (CSRs) or "reps" for short. (Appendix A summarizes the call-center acronyms used in this review, and it displays the page numbers on which they are defined.)

In addition to providing the services of CSRs, many inbound call centers use *interactive voice response* (IVR) units, also called *voice response units* (VRUs). These specialized computers allow customers to communicate their needs and to "self-serve." Customers interacting with an IVR use their telephone key pads or voices to provide information, such as account numbers or indications of the type of service desired. (In fact, the latest generation of speech-recognition technology allows IVRs to interpret complex user commands.) In response, the IVR uses a synthesized voice to report information, such as bank balances or departure times of planes. IVRs can also be used to direct the center's computers to provide simple services, such as the transfer of funds among bank accounts. For example, in many banking call centers, roughly 80% of customer calls are fully self-served using an IVR. (Interestingly, the process by which customers who wish to speak to a CSR identify themselves, using an IVR, can average 30 seconds, even though subsequent queueing delays often reach more no more than a few seconds.)

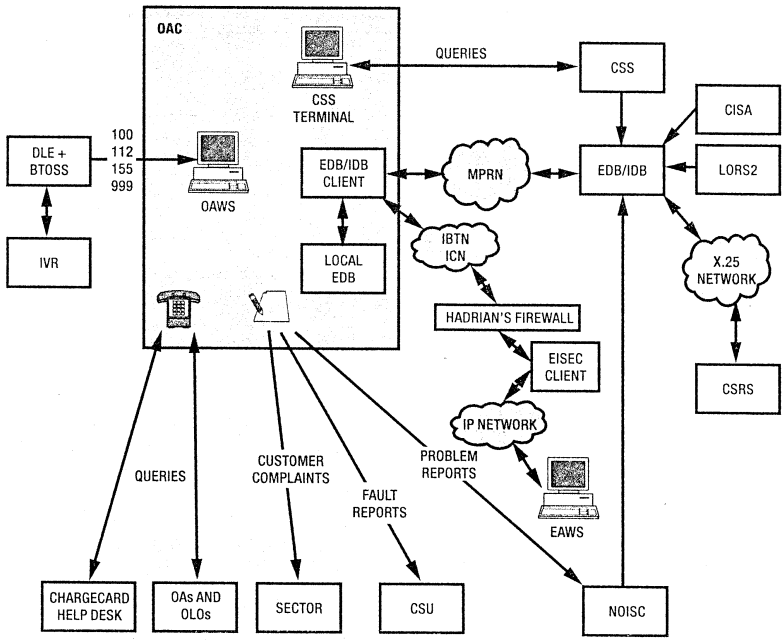
A current trend is the extension of the call center into a *contact center*. The latter is a call center in which agents and IVRs are complemented by services in other media, such as email, fax, web pages, or chat (in that order of prevalence). The trend toward contact centers has been stimulated by societal hype surrounding the internet and by customer demand for channel variety, as well as by the potential for efficiency gains. In particular, requests for email and fax services can be "stored" for later response and it is possible that, when standardized and well managed, they can be made significantly less costly than telephone services.

Our survey deals almost exclusively with pure telephone services. To the best of our knowledge, no analytical model has yet been dedicated to truly multi-media contact centers, though a promising framework (skills-based routing) and a few models that accommodate IVRs, emails and their blending with telephone services, will be described in Section 5.

2.2 How an Inbound Call is Handled

The large-scale emergence of call centers has been enabled by technological advances in information and communications systems. To describe these technologies, and to illustrate how they function,

we will walk the reader through an example of the process by which a call center serves an incoming call. Figure 2 provides an actual schematic diagram of the equipment involved.



BTOSS: BT OPERATOR SERVICES SUBSYSTEM
CISA: CHARGING INFORMATION SUPPLY AND AUDIT
CSRS: CASHLESS SERVICES REPLACEMENT SYSTEM
CSS CUSTOMER SERVICE SYSTEM
CSU: CUSTOMER SERVICES UNIT
DLE: DIGITAL LOCAL EXCHANGE
EAWS: EXTERNAL ASSISTANCE WORKSTATION
EISEC: ENHANCED INFORMATION SYSTEM FOR EMERGENCY CALLS
EDB: EXTERNAL DATABASE

IBTN: INTERNAL BACKBONE TRANSMISSION NETWORK
ICN: INTERNAL CORPORATE NETWORK
IDB: INTERNATIONAL DATABASE
IP: INTERNET PROTOCOL
IVR: INTERACTIVE VOICE RESPONSE
LORS2: OTHER LICENSED OPERATOR RECEIPT SYSTEM 2
NOISC: NATIONAL OPERATOR INFORMATION SYSTEMS CENTRE
OA: OPERATOR ASSISTANCE
OAC: OPERATOR ASSISTANCE CENTRE
OAWS: OPERATOR ASSISTANCE WORKSTATION

Figure 2: Call Center Technology (Operator Assistance in British Telecom UK [44])

Consider customers in the U.S. who wish to buy a ticket from a large airline using the telephone. They begin the process of buying the ticket by calling a toll-free "800" number. The long-distance or *public service telephone network* (PSTN) company that provides the 800 service to the airline knows two vital pieces of information about each call: the number from which the call originates, often called the *automatic number identification* (ANI) number; and the number being dialed, named the call's *dialed number identification service* (DNIS) number. The PSTN provider uses the ANI and DNIS to connect callers with the center.

The airline's call center has its own, privately-owned switch, called a *private automatic branch exchange* (PABX or PBX), and the caller's DNIS locates the PABX on the PSTN's network. If the airline has more than one call center on the network – both reachable via the same 800 number – then a combination of the ANI, which gives the caller's location, and the DNIS may be used to route the call. For example, a caller from Atlanta may be routed to a Dallas call center, while another caller from Chicago – who calls the same 800 number – may be routed to a center in North Dakota. Conversely, more than one DNIS may be routed to the same PABX. For example, the airline may maintain different 800 numbers for domestic and international reservations and have both types of call terminate at the same PABX.

The PABX is connected to the PSTN through a number of telephone lines, often called *trunk*

lines, that the airline owns. If there are one or more trunk lines free, then the call will be connected to the PABX. Otherwise, the caller will receive a busy signal. Once the call is connected it may be served in a number of phases.

At first, calls may be connected through the PABX to an IVR that queries customers on their needs. For example, in the case of the airline, callers may be told to "press one" if they wish to find flight status information. If this is the case, then through continued interaction with the IVR customers may complete service without needing to speak with an agent.

Customers may also communicate a need or desire to speak with a CSR, and in this case calls are handed from the IVR to an *automatic call distributor* (ACD). An ACD is a specialized switch, one that is designed to route calls, connected via the PABX, to individual CSRs within the call center. Modern ACDs are highly sophisticated, and they can be programmed to route calls based on many criteria.

Some of the routing criteria may reflect callers' status. For example, an airline may wish to specially route calls from Spanish-speaking customers. This identification can happen in a number of ways: though the DNIS, because a special 1-800 number is reserved for Spanish-speaking customers; through the ANI, which allows the call-center's computer system to identified the originating phone number as that of a Spanish-speaking customer; or through interaction with the IVR, which allows callers who press "3" to identify themselves as Spanish speakers.

The capabilities of agents may also be used in the routing of calls. For example, when agents at our example airline's call center begin working, they log into the center's ACD. Their login IDs are then used to retrieve records that describe whether they are qualified to handle domestic and/or international reservations, as well as whether or not they are proficient in Spanish.

Given its status, as well as that of the CSRs that are currently idle and available to take a call, the incoming call may be routed to the "best" available agent. If no suitable agent is free to take the call, the ACD may keep the call "on hold" and the customer waits until such an agent is available. While the decision of whether and to whom to route the call may be programmed in advance, the rules that are needed to solve this "skills-based routing" problem can turn out to be very complex.

Customers that are put on hold are typically exposed to music, commercials, or other information. (A welcome, evolving trend is to provide delayed customers with predictions of their anticipated wait.) Delayed customers may judge that the service they seek is not "worth" the wait, become impatient, and hang up before they are served. In this case, they are said to *abandon* the queue or to *renege*. Customers that do not abandon are eventually connected to a CSR.

Once connected with a customer, agents can speak on the telephone while, at the same time, they work via a PC or terminal with a corporate information system. In the case of our example airline, agents may discuss flight reservations with customers as they (simultaneously) query and enter data into the company's reservation system. In large companies, such as airlines and retail banks, the information system is typically *not* dedicated to the call center. Rather, many call centers, as well as other company branches, may share access to a centralized corporate information system.

Computer-telephone integration (CTI) "middleware" can be used to more closely integrate the telephone and information systems. For instance, CTI is the means by which a call's ANI is used to identify a caller and route a call: it takes the ANI and uses it to query a customer database in the company's information systems; if there exists a customer in the database with the same

ANI, then routing information from that customer's record is returned. In our airline example, the routing information would be the customer's preferred language.

Similarly, CTI can be used to automatically display a caller's customer record on a CSR's workstation screen. By eliminating the need for the CSR to ask the caller for an account number and to enter the number into the information system, this so-called "screen pop" saves the CSR time and reduces the call's duration. If applied uniformly, it can also reduce variability among service times, thus improving the standardization of call handling procedures.

In more sophisticated settings, CTI is used to integrate a special information system, called a *customer relationship management* (CRM) system, into the call center's operations. CRM systems track customers' records and allow them to be used in operating decisions. For example, a CRM system may record customer preferences, such as the desire for an aisle seat on an airplane, and allow CSRs (or IVRs) to automatically deliver more customized service. A CRM system may also enable a screen pop to include the history of the customer's previous calls and, if relevant, dollar-figures of past sales the customer has generated. It may even suggest cross-selling or up-selling opportunities, or it may be used to route the incoming call to an agent with special cross-selling skills.

Once a call begins service, it can follow a number of paths. In the simplest case, the CSR handles the caller's request, and the caller hangs up. Even here, the service need not end; instead, the CSR may spend some time on *wrap-up* activities, such as an updating of the customer's history file or the processing of an order that the customer has requested. It may also be the case that the CSR cannot completely serve the customer and the call must be transferred to another CSR. Sometimes there are several such hand-offs.

Finally, the service need not end with the call. Callers who are blocked or abandon the queue may try to call again, in which case they become *retrials*. Caller who speak with CSRs but are unable to resolve their problems may also call again, in which case they becomes *returns*. Satisfactory service can also lead to returns.

2.3 Data Generation and Reporting

As it operates, a large call center generates vast amounts of data. Its IVR(s) and ACD are special-purpose computers that use data to mediate the flow of calls. Each time one of these switches takes an action, it records the call's identification number, the action taken, the elapsed time since the previous action, as well as other pieces of information. As a call winds its way through a call center, a large number of these records may be generated.

From these records, a detailed history (trace) of each call that enters the system can be reconstructed: when it arrived; who was the caller; what actions the caller took in the IVR and how long each action took; whether and how long the caller waited in queue; whether and for how long a CSR served the call; who was the CSR. If the call center uses CTI, then additional data from the company's information systems may be included in the record: what the call was about; the types of actions taken by a CSR; related account information.

Call centers have not typically stored or analyzed records of individual calls, however. This may be due, in part, to the historically high cost of maintaining adequately large databases – a large call center generates many gigabytes of call-by-call data each month – but clearly these quantities of data are no longer prohibitively expensive to store. It is also likely due to the fact that the

software used to manage call centers – itself developed at a time when data storage was expensive – often uses only simple models which require limited, summary statistics. Finally, we believe that it is due to lack of understanding of how and why more detailed analyses should be carried out. (Section 6 describes current work that analyzes call-by-call data. Sections 6 and 7 argue for the long-term value of this type of work.)

Instead, call centers most often summarize call-by-call data from the ACD (and related systems) as averages that are calculated over short time intervals, most often 30 minutes in length. Figure 3 displays twenty one half-hours' worth of data from such a report.

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%	39.1%	5.8	1.4	416.2%

Figure 3: Example Half-Hour Summary Report From an ACD (from [89])

These ACD data are used both for planning purposes and to measure system performance. They are carefully and continuously watched by call-center managers. They will also be central to the discussion that continues through much of this article. Therefore, it is worth describing the columns of the report in some detail.

The first four columns indicate the starting time of the half-hour interval, as well as counts of calls arriving to the ACD: [Recvd], sometimes called *offered*, is the total number of calls arriving that half hour; [Answ], sometimes called *handled*, the number of arriving calls that were actually answered by a CSR; and [Abn %], the percentage of arriving calls that abandoned before being served (equals (1-[Answ]/[Recvd])$\times$ 100%). Note that the number of calls offered to the ACD may be much smaller than the total number of calls arriving to the center. First, [Recvd] does not account for busy signals, which occur at the level of the PSTN and PABX. Furthermore, as already mentioned, in some industries it is not unusual for 80% of the calls arriving to a call center to be

"self service" and to terminate in the IVR.

[Abn %] is an important measure of system congestion. The next column reports another one: [ASA] is the *average speed of answer*, the amount of time [Answ] calls spend "on hold" waiting to speak to a CSR. (Because ASA does not include the time that abandoned calls spend waiting, a reasonably full picture of congestion requires, at a minimum, both ASA and Abn % statistics.) Call centers sometimes report additional measures of the delay in queue. For example the *service level*, also called the *telephone service factor* (TSF), is the fraction of calls whose delay fell below a prespecified "service-level" target. Typically the target is 20 or 30 seconds. Some call centers also report the delay of the call that waited on hold the longest during the half hour.

To interpret the remaining statistics in Figure 3, it is helpful to define the following three states of CSRs who are logged into the ACD: 1) *active*, namely handling a call; 2) sitting idle, *available* to handle a call; and 3) not actively handling a call but not idle, *unavailable* to take calls. Over the course of each half-hour reporting interval, the ACD tracks the time that each CSR that is logged into the system spends in each of these states, and it aggregates [total active], [total available], and [total unavailable] time (across all logged-in CSRs) to calculate the figure's statistics.

The next column in Figure 3 reports the [AHT], the *average handle time* per call, another name for average service time (equals [total active] ÷ [Answ]). In some reports this total is broken down into component parts: *'talk' time*, the average amount of time a CSR spends talking to the customer during a call; *'hold' time*, the average time a CSR puts a customer "on hold" during a call, once service has begun; and *'wrap' time*, the average amount of time a CSR spends completing service, after the caller has hung up.

The remaining columns detail the productivity of the call center's CSRs. [On Prod FTE] is the average number of *full time equivalent* (FTE) CSRs that were active or available during the half hour (equals ([total active]+[total available]) ÷ 30 minutes). [Occ %], the system *occupancy*, is a measure of system utilization that excludes the time that CSRs were unavailable to serve calls (equals [total active] ÷ ([total active] + [total available]) × 100%). [On Prod %] is the fraction of time that logged-in CSRs were actively handling or able to handle calls (equals ([total active]+[total available]) ÷ ([total active]+[total available]+[total unavailable]) × 100%). [Sch Open FTE] is the number of FTE CSRs that had been *scheduled* to be logged in during the half hour; it is the planned version of [On Prod FTE]. Finally, [Sch Avail %] relates the actual time spent logged-in to the original plan (equals [On Prod FTE] ÷ [Sch Open FTE] × 100%).

Thus, the report records three sources of loss in CSR productivity. The first is idle time that is presumably induced by naturally occurring stochastic variability in arrival and service times and is captured by (100%–[Occ %]). The second is the fraction of time that CSRs were originally scheduled to be available to take calls but were not, which is calculated as (100%–[On Prod %]). This percentage can be tracked against an operating standard that the call center maintains to make sure that CSRs are not spending "too much" logged-in time unavailable. Similarly, the third source, (100%–[Sch Avail%]), allows call-center managers to track the fraction of time CSRs are not logged in, perhaps away from their work stations taking unplanned breaks. The latter two measures are often monitored to diagnose perceived disciplinary problems: CSRs' lack of *compliance* with their assigned schedules.

Note that the occupancies in Figure 3 are quite high, 97-100% during much of the day. This does not mean, however, that every CSR spends 97-100% of his or her work day speaking with customers. For example, suppose the arrival rate of calls to a center is a constant 2,850 per half

hour over an 8-hour day. The [AHT] of a call is one minute, so the call center expects 2,850 minutes of calls to be served in every half hour, or 95 FTE CSRs worth of calls (95 CSRs $\times$ 30 minutes per CSR = 2,850 CSR minutes each half hour.) The center does not allow CSRs to be unavailable, and in every half hour it makes sure that 100 CSRs are taking calls, so that [Sched Open FTE] = [On Prod FTE] = 100. Therefore, [Occ %] = 95% and [On Prod %] = [Sch Avail %] = 100% in every half hour. The call center has 200 CSRs on staff, however, and each CSR is scheduled to spend only *half* of the day on the phone. Indeed, as we will see in Section 3, CSRs are typically given breaks and off-phone work that lower their overall, daily utilization to more sustainable levels.

It is also worth noting that, although the statistics described above are averaged over all agents working, many can be archived also at the individual-agent level. This practice is useful for monitoring individual compliance, and it can be used as a part of incentive compensative schemes.

While the specifics of ACD reports may vary from one site to the next, the reports almost always (as far as we have seen) contain statistics on the four categories of data shown in Figure 3: numbers of arrivals and abandonment, average service times, CSR utilization, and the distribution of delay in queue. This is hardly surprising – it simply reflects the fact that call centers can be viewed, naturally and usefully, as queueing systems.

Readers who are primarily interested in statistical analysis of call-center data can now proceed to Section 3.3.

2.4 Call Centers as Queueing Systems

Figure 4 is an operational scheme of a simple call center. In it, the relationship between call centers and queueing systems is clearly seen.

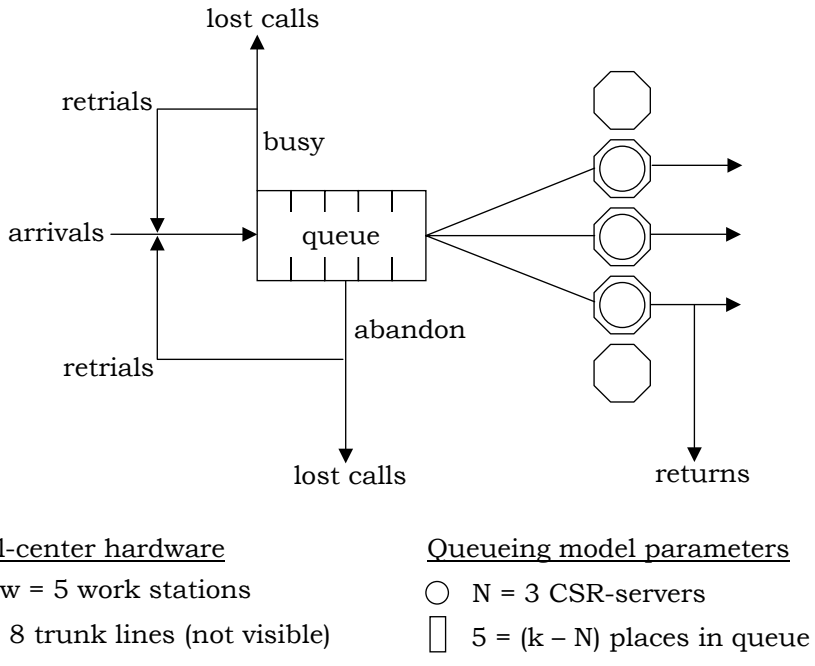


Figure 4: Operational Scheme of a Simple Call Center (adapted from [89])

The call center depicted in the figure has the following setup. A set of k trunk lines connects calls to the center. There are $w \leq k$ work stations, often referred to as seats, at which a group of $N \leq w$ agents serve incoming calls. An arriving call that finds all k trunk lines occupied receives a *busy* signal and is blocked from entering the system. Otherwise it is connected to the call center and occupies one of the free lines. If fewer than N agents are busy, the call is put immediately into service. If it finds more than N but fewer than k calls in the system, the arriving call waits in *queue* for an agent to become available. Customers who become impatient hang up, or *abandon*, before being served. For the callers that wait and are ultimately helped by a CSR, the service discipline is first-come, first-served.

Once a call exits the system it releases the resources it used – trunk line, work station, agent – and these resources again become available to arriving calls. A fraction of calls that do not receive service become *retrials* that attempt to reenter service. The remaining blocked and abandoned calls are *lost*. Finally, served customers may also *return* to the system. Returns may be for additional services that generate new revenue, and as such may be regarded as good, or they may be in response to problems with the original service, in which case they may be viewed as bad.

Thus, the number of trunk lines k acts as an upper bound on the number of calls that can be *in the system*, either waiting or being served, at one time. Similarly, the number of CSRs taking calls, $N \leq w$, provides an upper bound on the number of calls that can be *in service* simultaneously. Over the course of the day, call-center managers can (and do) dynamically change the number of working CSRs to track the load of arriving calls.

Less frequently, if equipped with the proper technology, managers also vary the number of active trunk lines k. For example, a smaller k in peak hours reduces abandonment rates and waiting (as well as the associated "1-800" costs, to be discussed later); this advantage can be traded off against the increase in busy signals.

For any fixed N, one can construct an associated queueing model in which callers are customers, the N CSRs are servers, and the queue consists of callers that await service by CSRs. When N changes, $(k - N)$, the number of spaces in queue, changes as well. As in Figure 3, model primitives for this system would include statistics for the arrival, abandonment, and service processes. Fundamental model outputs would include the long-run fraction of customers abandoning, the steady-state distribution of delay in queue, and the long-run fraction of time that servers are busy.

In fact, these types of queueing models are used extensively in the management of call centers. The simplest and most-widely used model is that of an $M/M/N$ queue, also known in call-center circles as Erlang C, which we later describe in more detail. For many applications, however, the model is an over-simplification. Just looking at Figure 4, one sees that the Erlang C model ignores busy signals, customer impatience, and services that span multiple visits.

In practice, the service process sketched above is often much more complicated. For example, the incorporation of an IVR, with which customers interact prior to joining the agents' queue, creates two stations in tandem: an IVR followed by CSRs. The inclusion of a centralized information system adds a resource whose capacity is shared by the set of active CSRs, as well as by others whom may not even be in the call center. The picture becomes far more complex if one considers multiple teams of specialized or cross-trained agents that are geographically dispersed over several, interconnected call centers, and who are faced with time-varying loads of calls from multiple types of customers.

2.5 Service Quality

Service quality is a complex and important topic that is closely related to the understanding of CSR and customer behavior, and we return to these subjects in Section 7. Here, we briefly review three notions of service quality that are most commonly tracked and managed by call centers.

The first view of quality regards the *accessibility* of agents. Typical questions are, "How long did customers have to wait to speak to an agent? How many abandoned the queue before being served?" This type of quality is measured via ACD (and related) reports, described above, and queueing models are used to manage it. In this article, we concern ourselves with problems associated with capacity management, and our emphasis will be on measures of accessibility.

The second view is of the *effectiveness* of service encounters, and it parallels the notion of *rework* in the manufacturing literature. The question here is "Did the service encounter completely resolve the customer's problem, or was additional work required?" Among call centers in the U.S., a call without rework is sometimes referred to as "one and done." This type of quality is typically measured by sampling inspection; agent calls are listened to at random – either live or on tape – and they are judged as requiring rework or not. To our knowledge, there do not exist wide-spread, formalized schemes for managing service effectiveness.

The last type of quality that is consistently monitored is that of the content of the CSRs' *interactions* with customers. Typical questions concern the CSR's *input* to the encounter and include, "Did the CSR use the customer's name? Did s/he speak to the customer with a 'smile' in his or her voice? Did the CSR manage the flow of the conversation in the prescribed manner?" As with the question of "one and done," answers to these questions are found by listening to a random sample of each CSR's calls. Sometimes the *output* of interactions is tracked, and the question "Was the customer satisfied?" is asked. Customer satisfaction data are typically collected via surveys.

Of course, the notion of the quality of the customers' experiences extends beyond their interaction with CSRs. For example, it critically includes the time spent waiting on hold, in queue.

In particular, we note that the nature of the time customers spend waiting on hold, in a *tele-queue*, is fundamentally different than that spent in a *physical queue*, at a bank or a supermarket checkout line, for example. Customers do not see others waiting and need not be aware of their "progress" if the call center does not provide the information. As Cleveland and Mayben [42] point out, customers that join a physical queue may start out unhappy – when they see the length of the queue which they have joined – and become progressively happier as they move up in line. (For experimental evidence of this effect, see Carmon and Kahneman [37].) In contrast, customers that join a tele-queue may be optimistic initially – because they do not realize how long they will be on hold – and become progressively more irritated as they wait. Indeed, call centers that inform on-hold callers of their expected delays can be thought of as trying to make the tele-queueing experience more like that of a physical queue.

Readers who are primarily interested in issues associated with service quality and customer and agent behavior can now continue with Section 3.5.

3 A Base Example: Homogeneous Customers and Agents

In this section we use a baseline example to describe the standard operational models that are used to manage capacity. We begin in §3.1 with background on capacity management in call centers. Then in §3.2 we define a hierarchy of capacity management problems, as well as the analytical models that are often used to solve them: queueing performance models for low-level staffing decisions; mathematical programming models for intermediate-level personnel scheduling; and long-term planning models for hiring and training. In §3.3–3.4 we describe standard practices in call-center forecasting. Finally, §3.5 offers a qualitative discussion of longer-term problems in system design.

3.1 Background on Capacity Management

Higher utilization rates imply longer delays in queue, and in managing capacity, call centers trade off resource utilization with accessibility. This trade-off is central to the day-to-day operations of call centers and to the *workforce management* (WFM) software tools that are used to support them. It is also the concern of much of the research that is discussed in later sections.

In some cases, revenues or costs can be directly associated with system performance. One can then seek to maximize expected profits or to minimize expected costs. For example, call centers that use toll-free services pay out-of-pocket for the time their customers spend waiting, and these "1-800" costs grow roughly linearly with the average number in queue: a call center that is open 24 hours a day, 7 days a week, and averages 40 calls in queue will pay about $1 million per year in queueing expenses (when the cost per minute per call is $0.05). Similarly, order-taking businesses can sometimes estimate the opportunity cost of lost sales due to blocking (busy signals) or abandonment. For example, see Andrews and Parsons [8] and Akşin and Harker [4].

More typically, however, call center goals are formulated as the provision of a given level of accessibility, subject to a specified budget constraint. Common practice is that upper management decides on the desired service level and then call center managers are called on to defend their budget. (See Borst et al. [25] for a discussion of the constraint satisfaction and cost minimization approaches.)

Furthermore, call center managers' view of system capacity most often focuses on agents. CSR salaries typically account for 60–70% of the total operating costs, and managers presume that other resources, such as information systems, are not bottlenecks. While centers often do maintain extra hardware capacity, such as workstations, Akşin and Harker [3, 4] show that planning models that do not account for other bottlenecks, when they exist, could be a problem.

We next introduce a "base case" example that reflects the capacity planning approach used by most call centers. We note that the example does not represent the state of the art or, for that matter, best practice. It does, however, give a sense of the state of common practice. Furthermore, the description of the example – and its inherent problems and limitations – will provide a framework by which we will organize our subsequent discussion of call-center research.

The subsection is divided into three parts. We begin by describing, from the bottom up, a hierarchy of capacity-planning problems (introduced formally already in Buffa et al. [36]). We then describe forecasting and estimation procedures which are commonly used to determine inputs to the capacity-planning process. Finally, we sketch how the elements are put together within the

context of the call center's day-to-day operations.

3.2 Capacity Planning Hierarchy

Consider the call center whose statistics are reported in Figure 3. One sees that the pattern of arrivals and service times the center experiences is changing over the course of the day. Offered calls (per half hour) peak from 11:00–11:30am, dip over lunch, and then peak again, from 2:30–3:00pm. Average handle times also appear to change significantly from one half hour to the next.

Indeed, in most call centers, the arrival rate and mix of calls entering the system vary over time. Over short periods of time, minute-by-minute for example, the is significant *stochastic* variability in the number of arriving calls. Over longer periods of time – the course of the day, the days of the week or month, the months of the year – there also can be *predictable* variability, seasonal patterns that arriving calls follow. (See Figure 5.)

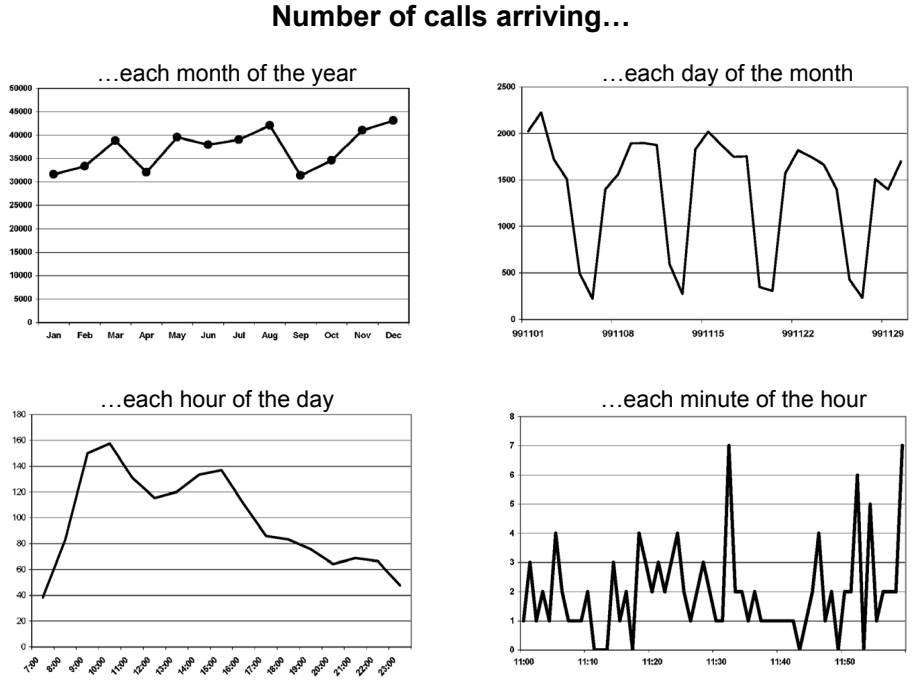


Figure 5: A Hierarchical View of Arrival Rates (adapted from [106], following [36])

Because service-capacity cannot be inventoried, managers vary the number of available CSRs to track the predictable variations in the arrival rates of calls. In this manner, they attempt to meet demand for service at a low cost, yet with an acceptable delay. In turn, capacity-planning naturally takes place from the bottom, up: *queueing* models determine how many CSRs must be available to serve calls over a given half-hour or hour; *scheduling* models determine when during the week or month each CSR will work; *hiring* models determine the number of CSRs to hire and train each month or quarter of the year.

At the lowest level of the hierarchy, the arrival times of individual calls are not predictable (lower right panel of Figure 5). Here, common practice uses the $M/M/N$ (Erlang C) queueing

model to estimate *stationary* system performance of short – half-hour or hour – intervals. In doing so, the call center implicitly assumes constant arrival and service rates, as well as a system which achieves a steady state quickly within each interval. Furthermore, the arrival process is assumed to be Poisson, service times are assumed to be exponentially distributed and independent of each other (as well as everything else in the system), and the service discipline is assumed to be first-come, first-served. Blocking, abandonment, and retrials are ignored.

Given these assumptions, the Erlang C formula (see (3) below) allows for straightforward calculation of the stationary distribution of the delay of a call arriving to the system. This and other steady state performance measures are used to make the capacity-accessibility trade-off.

The calculations begin as follows. Let λ_i be the arrival rate for 30-minute interval i. Similarly, let $\mathsf{E}[S_i]$ and $\mu_i = \mathsf{E}[S_i]^{-1}$ be the expected service time and service rate for the interval. Then define

$$R_i \triangleq \lambda_i/\mu_i = \lambda_i \mathsf{E}[S_i] \tag{1}$$

to be the *offered load* and

$$\rho_i \triangleq \lambda_i/(N\mu_i) = R_i/N \tag{2}$$

to be the associated average system utilization (also called "traffic intensity"). Note that R_i, often dubbed the number of offered *Erlangs*, is a unit-less quantity. That is, over half hour i, an average of R_i units of service time is offered to the call center per unit of time, and CSRs are busy an average of $\rho_i \times 100\%$ of the time.

Given the Erlang C's no-blocking and no-abandonment assumptions, *at least* R_i CSRs are required to work for a half hour to serve this expected load. Furthermore, N must be strictly greater than R_i, equivalently $\rho_i < 1$, for the system to have a steady-state. In this case, the Erlang C formula

$$\mathrm{C}(N, R_i) \triangleq 1 - \left(\sum_{m=0}^{N-1} {R_i}^m/m!\right) \Bigg/ \left(\sum_{m=0}^{N-1} {R_i}^m/m! + \left(\frac{{R_i}^N}{N!}\right)\left(\frac{1}{1-R_i/N}\right)\right) \tag{3}$$

defines the steady-state probability that all N CSRs are busy.

The application of the "Poisson arrivals see time averages" (PASTA) [150] property then allows us to obtain our first measure of system accessibility, the fraction of arriving customers that must wait to be served:

$$\mathsf{P}\{\text{Wait} > 0\} = \mathrm{C}(N, R_i)\,. \tag{4}$$

In turn, given the event that an arriving customer must wait, the conditional delay in queue is exponentially distributed with mean $(N\mu_i - \lambda_i)^{-1}$, and additional steady-state measures of accessibility are straightforward to calculate:

$$\begin{aligned} \text{ASA} \triangleq \mathsf{E}[\text{Wait}] &= \mathsf{P}\{\text{Wait} > 0\} \cdot \mathsf{E}[\text{Wait} \,|\, \text{Wait} > 0] \\ &= \mathrm{C}(N, R_i) \cdot \left(\frac{1}{N}\right)\left(\frac{1}{\mu_i}\right)\left(\frac{1}{1-\rho_i}\right), \end{aligned} \tag{5}$$

the average waiting time before being served, and

$$\begin{aligned} \text{TSF} \triangleq \mathsf{P}\{\text{Wait} \le T\} &= 1 - \mathsf{P}\{\text{Wait} > 0\} \cdot \mathsf{P}\{\text{Wait} > T \,|\, \text{Wait} > 0\} \\ &= 1 - \mathrm{C}(N, R_i) \cdot e^{-N\mu_i(1-\rho_i)T}, \end{aligned} \tag{6}$$

the fraction of customers that wait no more than T units of time, for some T that defines the desired telephone service factor. All three stationary measures are monotone in N: $\mathsf{P}\{\text{Wait} > 0\}$ decreasing, ASA decreasing, and TSF increasing.

Figure 6 depicts the empirical relationship between ASA and system occupancy, ρ, at a relatively small call center, analyzed in Brown et al. [33]. The "cloud" of points in the figure's left panel plots the result for each of 3,867 hourly intervals that the call center was open during 1999. The right panel highlights the relationship between ASA and ρ by furthering averaging the occupancies and ASAs of adjacent points. The data plotted in Figure 6 clearly parallel the theoretical relationship defined by (5).

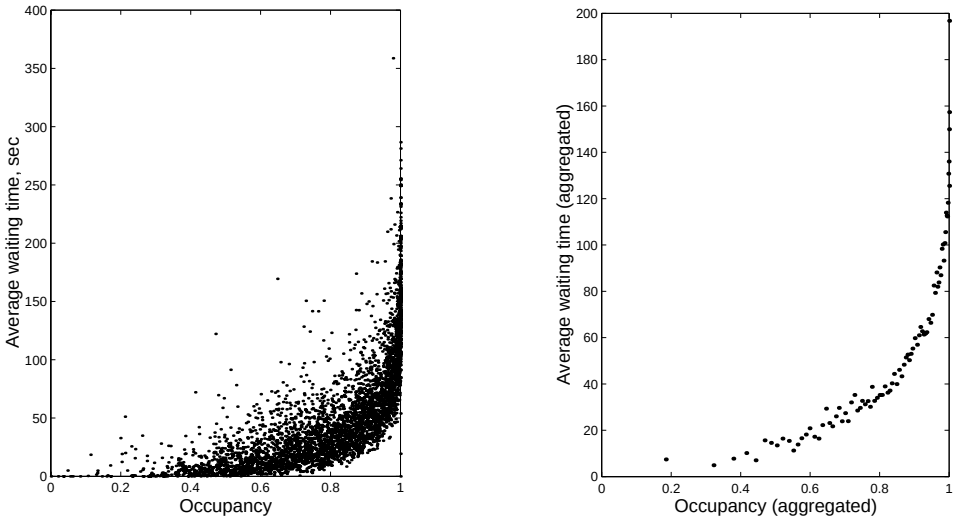


Figure 6: Congestion Curves Based on Raw and Aggregate Data (from [33])

In is interesting to note that $\mathsf{P}\{\text{Wait} > 0\}$ is a fundamental measure of accessibility, from which ASA and TSF are derived, and it also plays an important part in asymptotic characterizations of accessibility. (See Section 4.) Yet it is almost never tracked by call center management.

Rather, call centers typically choose ASA or TSF as the standard used for determining staffing levels. For example, a call center might define ASA* to be an upper bound on the acceptable average delay of arriving calls. Then the monotonicity of ASA with respect to N is used to find the minimum number of agents required to meet the service-level standard:

$$N_i \quad = \quad \min\{N \leq w \mid \text{ASA} \ \leq \ \text{ASA}^*\}\,. \tag{7}$$

Over relatively long time intervals, variations in arrival rates become more predictable. For example, the fluctuations shown in the lower left and upper right panels of Figure 5 are fairly typical patterns of arrivals over the course of the day and month. Common practice assumes that these fluctuations are completely predictable.

Point forecasts for system parameters are then inputs to the next level up in the planning hierarchy, staff scheduling. More specifically, each half hour interval's forecasted λ_i and μ_i give rise to a target staffing level for the period, N_i. For a call center that is open twenty four hours a day, seven days a week, repeated use of the Erlang C model will produce 1,440 N_i's in a 30-day month. The vector of N_i's becomes the input to the scheduling model.

We distinguish between two elements of the scheduling process, shifts and schedules. A *shift* denotes a set of half-hour intervals during which a CSR works over the course of the day. A *schedule* is a set of daily shifts to which an employee is assigned over the course of a week or month. Both shifts and schedules are often restricted by union rules or other legal requirements and can be quite complex. For example, a feasible shift may start on the half-hour, last 9 hours, including an hour total of break time. One half hour of this break must be devoted to lunch, which must begin sometime between two and three hours after the shift begins, and the other to a morning or afternoon pause. A feasible schedule may require an employee to work five, 9-hour shifts each week of the month, on Sunday, Monday, Tuesday, Friday, and Saturday. Another may require a CSR to work a different set of shifts each week of the month.

Now, suppose there is a collection of $j = 1, \ldots, J$, feasible schedules to which employees may be assigned and that the monthly cost of assigning an agent to schedule j equals c_j. This cost includes wage differentials and overtime costs that are driven by schedule assignments; it need not include regular wage and benefit costs that do not change with the schedule. Then determination of an optimal set of schedules can be described as the solution to an integer program (IP). Given $i = 1, \ldots, I$, half-hour intervals during the planning horizon we let

$$a_{ij} = \begin{cases} 1, & \text{if an agent working according to schedule } j \text{ is available to take calls during interval } i; \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Figure 7 shows the complete A-matrix for schedules that cover one 11-hour day (for simplicity, rather than 30-days). To enhance readability, only the matrix's ones are shown, not the zeros. Each of the 22 rows represents a different 1/2-hour interval, and each of the 10 columns represents a different schedule to which employees may be assigned. Inspection reveals that the first 5 columns all have the same structure; the only difference among them is the time that an employee assigned to the schedule would start. Similarly, the second set of 5 columns share the same structure. Every one of the 10 schedules has an employee take calls for 7 hours of a 9 hour day.

Letting the decision variables x_j, $j = 1, \ldots, J$, represent the numbers of agents assigned to the various schedules, one solves

$$\min\{c'x \mid Ax \geq \vec{N};\ x \geq 0;\ x \text{ integer}\} \tag{9}$$

to find a least-cost set of schedules. That is, the optimal solution to (9) defines the number of CSRs to assign to each monthly schedule, j, subject to the lower bounds on available CSRs imposed by the service-level constraint. This formulation can become quite large – with thousands of time slots (rows) and feasible schedules (columns) – in which case it cannot be solved to optimality. For call centers in which $\sum_i N_i$ is large, the rounded (up) solution of a linear programming relaxation may perform well, however (see Mandelbaum and Ruszczynski [105]).

In practice, the formulation of the scheduling problem may differ somewhat from (9). One alternative is to impose an aggregate service-level constraint for a longer period of time, such as a day, rather than one for each half-hour or hour (see Koole and van der Sluis [90]). Another is to minimize the deviation between the recommended staffing levels, N_i, and the actual staffing levels obtained from the assigned schedules (see Buffa et al. [36]). Both of these alternatives reduce overall staffing levels, in effect by relaxing the service-level constraints.

Furthermore, a solution to (9) defines only how many agents are assigned to the various schedules, not necessarily which person works on what schedule. For large call centers, the final assignment of employees to schedules is an even more complex problem for which even feasible solutions

time		j = 1	2	3	4	5	6	7	8	9	10
8:00-8:29am	i = 1	1					1				
8:30-8:59	2	1	1				1	1			
9:00-9:29	3	1	1	1			1	1	1		
9:30-9:59	4		1	1	1		1	1	1	1	
10:00-10:29	5	1		1	1	1		1	1	1	1
10:30-10:59	6	1	1		1	1	1		1	1	1
11:00-11:29	7	1	1	1		1	1	1		1	1
11:30-11:59	8	1	1	1	1		1	1	1		1
12:00-12:29pm	9		1	1	1	1	1	1	1	1	
12:30-12:59	10			1	1	1		1	1	1	1
1:00-1:29	11	1			1	1			1	1	1
1:30-1:59	12	1	1			1	1			1	1
2:00-2:29	13	1	1	1			1	1			1
2:30-2:59	14	1	1	1	1		1	1	1		
3:00-3:29	15		1	1	1	1		1	1	1	
3:30-3:59	16	1		1	1	1	1		1	1	1
4:00-4:29	17	1	1		1	1	1	1		1	1
4:30-4:59	18	1	1	1		1	1	1	1		1
5:00-5:29	19		1	1	1			1	1	1	
5:30-5:59	20			1	1	1			1	1	1
6:00-6:29	21				1	1				1	1
6:30-6:59	22					1					1

Figure 7: An Example A-Matrix for the Scheduling Problem

are difficult to construct. Here, heuristics are often used. One common method ranks employees by job tenure or seniority and allows higher-ranking employees to choose their schedules first.

Figure 8 shows how the number of busy CSRs tracks the arrival of work at a fairly large (virtual) call center, under study by Brown et al. [34]. Note that, although WFM systems typically schedule CSRs to start working every 15 or 30 minutes, the figure shows the number of busy CSRs closely tracking the offered load in the morning. This may be due to one of two factors, or perhaps a combination of the two: either additional CSRs log into the ACD every few minutes in the morning, as the arrival rate grows; or there exist additional (underutilized) CSRs who are available to take calls, but do not show up in the chart, because they never take calls. Note also the system overcapacity during the peak of the day, an interval over which the center operates at about 80% utilization. We believe that this relatively low occupancy is due to a skills-based routing scheme in which specialized CSRs are prohibited from taking regular calls and are, hence, underutilized.

The solution to (9) also defines the total number of employees required to be assigned to monthly schedules, $\mathbf{1}'x$. Typically this number is then "grossed up" – say by a factor $\alpha \in (0, 1)$ – to account for unplanned breaks, time spent training and in meetings, absenteeism and other factors that reduce employees' productive capacity. For example, statistics, such as [On Prod %] and [Sch Avail %] in Figure 3, can be used in the estimation of α. Thus, the number of agents needed in month t becomes $n_t = \mathbf{1}'x/\alpha$.

At the top of the planning hierarchy, a long-term hiring problem is solved to ensure that monthly staffing requirements are met. The horizon for the hiring problem, $\mathcal{T}$, may be on the order of six months to one year.

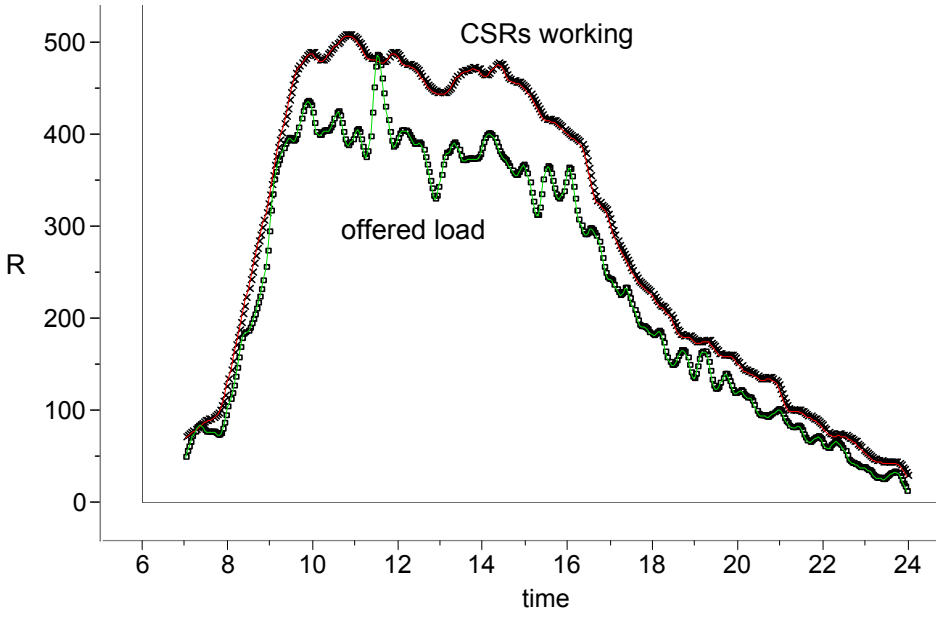


Figure 8: The Numbers of CSRs Working Tracks the Offered Load (from [34])

The gross numbers of employees needed each month over the planning horizon, $\{n_t : t = 1, \ldots, \mathcal{T}\}$, are found by solving the scheduling problem (9) (and the underlying staffing problems, (7)) for each month t. Other input data for the hiring problem include the following: an estimate of the monthly turnover rate, β; and an estimate of the lead time, τ, that is required to recruit and train a new employee, once the decision to hire has been made.

It is worth noting that these latter two factors can be significant. For example in many centers employee turnover exceeds 50% per year; hiring and training lead times can be two or three months, or more.

Given these data, a simple method of addressing the long-term problem that we have seen is to myopically hire enough new employees so that, by the time they are trained, the projected number of employees on hand meets or exceeds the projected requirements. More formally, suppose y_t employees are on hand at the start of month t and that, due to previous months' hiring decisions, y_j employees will start working in months $j = t+1, \ldots, t+\tau-1$. Then the number, z_t, to hire in month t is

$$z_t \quad = \quad \left(n_{t+\tau} \; - \; \sum_{j=t}^{t+\tau-1} y_j (1-\beta)^{\tau-(j-t)} \right)^+ , \tag{10}$$

so that $y_{t+\tau} = z_t + \sum_{j=t}^{t+\tau-1} y_j \geq n_{t+\tau}$. Here, the terms within the summation account for the after-turnover numbers of employees on hand at $t+\tau$, before the hiring decision at t is made. By hiring the difference between $n_{t+\tau}$ and that total, (10) assumes that no turnover occurs among employees during recruitment and training.

Readers who are interested in queueing performance models can now proceed to Section 4. Those primarily interested in problems associated with personnel scheduling, hiring, and training can now continue with Section 3.5.

3.3 Forecasting

The hierarchy of capacity planning models, described above, requires the following inputs: arrival rates λ_i, service rates μ_i, productivity factor α, turnover rate β, and lead time τ. Much of the data required to build estimates for these parameters come from ACD reports, such as that shown in Figure 3. For example, the [Rcvd] and [AHT] columns of the report state actual arrival rates and average service times for each half-hour of the day.

The sources of the other data vary. WFM systems sometimes track productivity figures for employees, such as Figure 3's [On Prod %], through the ACD. Employee turnover rates, hiring lead times, and training requirements are (clearly) not captured by ACD systems, however. These data are collected from employee records by the call center's human resources (HR) department.

Arrival rates are often forecast on a "top-down" basis. The process begins by aggregating the reported number of calls arriving each half hour into monthly totals, such as those found in the upper-left of Figure 5. These totals are the historical basis of forecasts that are to be built on a combination of simple time-series methods, such as exponential smoothing, and managerial opinion regarding what will happen to the business that the call center supports. (For an early book on exponential smoothing see Brown [32]; for a recent one see Makridakis et al. [96].) The result is a month-by-month forecast of call volumes.

Once these top-level forecasts are set, the monthly totals are then allocated by day-of-week and day-of-month, as well as by time of day. (See the upper-right and lower-left of Figure 5.) For example, it may be assumed that 20% of July's calls are handled in the first week of the month and Mondays account for 27% of each week's total volume. Similarly, each half-hour may be allocated a fixed percentage of a day's total call volume.

Common call-center practice is then to assume constant arrival rates over individual half hours or 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 quickly, in particular when the event rate ($\lambda + N\mu$ in an M/M/N queue) is large when compared to the duration of the interval, and when predictable factors that drive the rates are relatively stable over the interval.

In addition to using day-of-week and day-of-month allocations, managers may flag certain days as special and increase or decrease anticipated call volumes accordingly. For example, suppose July 4th falls on Tuesday. Then the anticipated volume for the 4th may be adjusted down, below normal. Conversely, the volume for the 5th may be adjusted upward, in anticipation of customers who put their calls off from the 4th to the 5th. Again, these adjustments are made using a combination of data-analysis and experience-based judgment.

In theory, the half-hourly ACD records of average service times could also be used to generate detailed forecasts of μ_i's. In practice, however, many call centers do not forecast service times or other parameters in detail. Instead, grand averages for historical service rates, productivity rates, and turnover rates are calculated.

For capacity-planning purposes, the parameters μ, α, and β are often assumed to be objects of managerial control, and how they are set is the result of negotiation. For example, upper management may assign call-center managers an objective of reducing employee turnover rates by 3% or of reducing average handle times by 5 seconds.

Readers who are primarily interested in statistical analysis of call-center data can now proceed to Section 6.

3.4 The Forecasting and Planning Cycle

In most call centers there is a planner that is responsible for agent rosters. Every week or every few weeks, this person begins preparing a forecast for the specified period. Based on this forecast, required numbers of CSRs are determined and, together with agent and management input (concerning days off, meetings, etc.), a roster is determined. This process is very often supported by WFM (workforce management) software, whose core function is to forecast arrival rates and average service times and then solve (7) and (9). These WFM packages allow call center planners and managers to refine and redefine their operating plans.

The establishment of an initial agent roster is not yet the end of story, however. Forecasts are continually updated and changes are made to roster until the scheduled day itself. When the roster is executed, a supervisor is responsible for service levels and CSR productivity. He or she monitors abandonment rates and waiting times and changes agents' deployments, based on real-time operating conditions. During the day, data are fed back into the workforce management tool, forecasts are updated, and the process repeats itself.

3.5 Longer-Term Issues of System Design

Beyond workforce management, more strategic decisions concern the design of the service process and system. Often, HR planning and the use of technology are tied together through service process design.

In the case of a single call center with universal (flexible) agents, these issues can be easily illustrated. For example, such a call center may attempt to reduce HR costs by having more calls resolved in its IVR. In this case, additional IVR resources may need to be purchased, and the IVRs must be programmed to handle the newly added service. At the same time, the expected change in CSR load must be estimated: the fraction of customers that decides to self-serve using the IVR causes arrival rates to decline; the elimination of these calls from the original mix causes the average service time to change as well. The changes then flow through the staffing models described above, and an estimated reduction in CSR head count may be made. Thus, investment in the IVR is traded off against HR savings.

Newer technology expands the possibilities for call-center design, and it also makes the task of evaluating and implementing the options more complex. For example, consider how IVRs and skills-based routing makes the use of part-time CSRs become economically attractive. In general, "part-timers" are valuable because they may work only during the daily peak in arriving calls, thereby reducing the number of full-time agents that are (paid but) not well utilized at other times of the day. CSR training is expensive, however, and turnover among part-time employees is high. To make cost-effective use of the part-timers, their training may need to be reduced. This implies that they will be able to serve only a subset of the calls handled by the center. To identify which incoming calls can be handled by part-time CSRs, the IVR is programmed so that customers identify the type of service they desire. To make sure that only simple calls are routed to part-time CSRs, the center invests in skills-based routing.

Interestingly, while skills-based routing allows for more efficient use of CSR resources, many call centers also see the technology as a means for reducing employee turnover. The idea expands on the use of part-time workers described above. A set of skills is designed to act as a career path; as agents learn new skills they move up the ladder. This, in turn, is hoped to improve employee motivation and morale and to reduce job burnout and turnover.

Readers who are interested in problems associated with personnel scheduling, hiring, and training can now proceed to Section 3.5. Those interested in service quality, as well as customer and CSR behavior, can now continue with Sections 6.3.2–6.3.4.

4 Research within the Base-Example Framework

In this section we review research that bears directly on the capacity-planning problems described in Section 2. As such, it reflects the state-of-the art within a narrow context: a single type of call is handled by a homogeneous pool of CSRs at a single location. (We consider models of multiple call-types, CSR skills, and locations in Section 5.) Even so, this special case provides a challenging set of problems, and its results offer essential insights into the nature of capacity management in all call centers, simple and complex.

In §4.1–4.4 we cover queueing models used to determine short-term staffing requirements. Then §4.5 reviews research devoted to the problem of scheduling CSRs. Next, §4.6 addresses models for long-term hiring and training. Finally §4.7 discusses open problems in each of the three areas of research.

4.1 Heavy-Traffic Limits for Erlang C

The Erlang C model, described in §3.2, has been widely adopted primarily because of its ease of use. In particular, there exist simple expressions such as (4)-(6) for most performance measures of interest. At the same time, the model has notable limitations.

Although the Erlang C formula is easily implemented, it is not easy to obtain insight from its answers. For example, to find an approximate answer to questions such as "how many additional agents do I need if the arrival rate doubles?" we have to perform a calculation. An approximation of the Erlang C formula that gives structural insight into this type of question would be of use to better understand economies of scale in call-center operations.

Erlang-C-based predictions can also turn out highly inaccurate because of violations of underlying assumptions, and these violations are not straightforward to model. For example, non-exponential service times lead one to the M/G/N queue which, in stark contrast to the M/M/N system, is analytically intractable.

Thus, approximations are useful both to aid insight and to extend modelling scope, and when modelling call centers, the most useful approximations are typically those for heavy-traffic regimes – those in which agent utilization is high. The heavy-traffic assumption naturally reflects the highly utilized nature of large call-center operations, particularly the peak-hour conditions that drive overall system scale.

Consider the M/G/N queue. For small to moderate numbers of agents, N, Kingman's classical

"Law of Congestion" asserts that delay in queue is approximately exponential, with mean as given by

$$\mathsf{E}[\text{Wait for M/G/}N] \approx \mathsf{E}[\text{Wait for M/M/}N] \ \times \ \frac{1+{c_s}^2}{2} \tag{11}$$

(see Whitt [143]). Here $c_s = \sigma(S)/\mathsf{E}[S]$ denotes the coefficient-of-variation of the service time, a unit-less quantity that naturally quantifies stochastic variability. Furthermore, the heavy-traffic regime assumed by Kingman [85] – and, more broadly, traditional heavy-traffic analyses – implies that essentially all customers experience some delay before being served. (For recent texts on heavy traffic, see Chen and Yao [41] and Whitt [146].)

Then, given $C(N,R) \approx 1$, (11) becomes

$$\mathsf{E}[\text{Wait for M/G/}N] \ \approx \ \left(\frac{1}{N}\right) \ \mathsf{E}[S] \ \left(\frac{1}{1-\rho}\right) \ \left(\frac{1+{c_s}^2}{2}\right) \tag{12}$$

From (12) we clearly see that the effect on congestion of both utilization, ρ, and stochastic variability, c_s, is non-linear, in fact increasing convex. Indeed, even small increases in load (utilization), ρ, can have an overwhelming, negative effect on highly utilized systems. Performance also deteriorates with longer and more variable service times, $\mathsf{E}[S]$ and ${c_s}^2$, and it improves with increased parallelism, N.

4.1.1 Square-Root Safety Staffing

The use of Kingman's Law for call centers was advocated by Sze [137], where it was attributed to Lee and Longton [95]. Sze was motivated by a traffic mix problem in a call center with the following characteristics. We loosely quote from [137]: "The problems faced in the Bell System's 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. While approximations are available for heavy and light traffic systems, 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."

Sze [137] tests a number of asymptotic approximations for M/G/N systems and, interestingly, favors the approximation (11). This approximation, in particular, identifies exponential service times with any other service time for which $c_s = 1$. But, as will be seen later in Figure 14, this identification can be inaccurate in the case of many highly utilized servers. (Perhaps the conclusion in [137] is due to testing only phase-type service-time distributions, which allows (11) to be a reasonable approximation.)

Indeed, for many call centers, N is in the tens or hundreds, rather than ones, and larger N gives rise to an asymptotic regime that differs from that of Kingman's Law in that significantly many customers do not wait and service quality is carefully balanced with server efficiency. For this reason, we call it a *Quality and Efficiency Driven* (QED) operational regime.

The QED regime for the M/M/N queue was first analyzed by Halfin and Whitt [69]. Formally, in this regime a service rate μ is fixed, as well as a target value $\alpha \in (0,1)$ for $\mathsf{P}\{\text{Wait} > 0\}$. Thus, it is defined as one in which some, but not all, customers wait for service. Then scaling $\lambda \uparrow \infty$ and $N \uparrow \infty$, Halfin and Whitt demonstrate

$$\mathsf{P}\{\text{Wait} > 0\} \to \alpha \iff \sqrt{N}(1-\rho_N) \to \beta \tag{13}$$

for some fixed *service grade* $\beta \in (0, \infty)$, so that $\rho_N = \frac{\lambda}{N\mu} \uparrow 1$. They then derive the following asymptotic expression for the Erlang-C formula:

$$\mathsf{P}\{\text{Wait} > 0\} \approx \mathsf{P}(\beta) = \left[1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right]^{-1} , \tag{14}$$

where $\alpha = \mathsf{P}(\beta)$ in (13). Here Φ and ϕ are, respectively, the distribution and density functions of the standard normal distribution (mean=0, variance=1).

For a fixed service grade, β, (13) suggests a *square-root safety-staffing principle* that recommends the number of servers N to be

$$N = R + \Delta = R + \beta\sqrt{R} , \quad 0 < \beta < \infty , \tag{15}$$

where, again, $R = \lambda/\mu$ is the offered load. The quantity $\Delta = \beta\sqrt{R}$ is "safety staffing" against stochastic variability. Note that $\beta \leq 0$ implies a utilization of 100% or more, hence an unstable system. As β increases, so does the level of safety staffing. In turn, $\mathsf{P}\{\beta\} \approx \mathsf{P}\{\text{Wait} > 0\}$ decreases with β.

Recalling that $\{\text{Wait}|\text{Wait} > 0\}$ is exponentially distributed with mean $(N\mu - \lambda)^{-1}$, one deduces from expressions (5),(6) and (15) that square-root safety-staffing with $\Delta = \beta\sqrt{R}$ obtains

$$\mathsf{E}[\text{Wait}] = \mathsf{P}\{\text{Wait} > 0\} \cdot \mathsf{E}[\text{Wait}|\text{Wait} > 0] \approx \mathsf{P}\{\text{Wait} > 0\} \cdot \frac{\mathsf{E}[S]}{\Delta} , \tag{16}$$

as well as the following simple expression for the distribution of delay:

$$\mathsf{P}\{\text{Wait} > T\} \approx \mathsf{P}\{\text{Wait} > 0\} \cdot e^{-(T/\mathsf{E}[S])\Delta} . \tag{17}$$

While Halfin and Whitt's formal analysis did not appear until the early 1980's, "folk" versions of this square-root law have long been recognized. Erlang [47] himself described the square-root relationship as early as 1924, and he reports that square-root rules had been in use at the Copenhagen Telephone Company since 1913.

Related, infinite-server heuristics that generate square-root staffing rules have also been long recognized (see Whitt [142], and the references in Borst et al. [25]). In infinite server systems, the number of busy CSRs found by an arriving call has a Poisson distribution, and the heuristic assumes that in large finite systems, this number is *nearly* Poisson if delays are not prevalent. In turn, a Poisson random variable with mean R is approximately a normally distributed random variable with mean R and standard deviation $\sqrt{R}$. Then, given a target delay probability of α, one chooses β in (15) such that

$$\alpha = 1 - \Phi(\beta) \equiv \bar{\Phi}(\beta).$$

This is justified by

$$\mathsf{P}\{\text{Wait} > 0\} = \mathsf{P}\{\text{Number of busy servers} > N\} \approx \mathsf{P}\{R + Z\sqrt{R} > R + \beta\sqrt{R}\} = \bar{\Phi}(\beta). \tag{18}$$

Here Z denotes a standard normal random variable, and the PASTA property ensures that $\mathsf{P}\{\text{Wait} > 0\} = \mathsf{P}\{\text{Number of busy servers} > N\}$. For small $\mathsf{P}\{\text{Wait} > 0\}$, $\bar{\Phi}^{-1}(\beta) \approx P^{-1}(\beta)$, and the heuristic's recommendation essentially matches that of Halfin and Whitt.

Borst, Mandelbaum, and Reiman [25] prove that, for a variety of natural delay-cost functions, staffing based on the square-root principle is, in fact, (asymptotically) optimal for large, heavily-loaded systems. That is, the paper shows that to minimize cost, it is optimal to operate in the

QED regime. The same conclusion applies when minimizing staffing levels subject to constraint on performance measures, which is more common in practice.

Square-root safety staffing turns out to be exceptionally accurate and robust: it is tested in [25] over all regimes, from very light to very heavy traffic, and it rarely deviates by more than a *single* server from the exactly optimal staffing level. The introduction to [25] offers further details through a set of staffing scenarios.

Borst et al. [25] also derives an explicit means of determining the optimal β, a problem which they term "dimensioning." Figure 9 graphs the optimal β for the case in which delay costs and staffing costs are both linear functions of time. In this case, let r denote the ratio of delay cost per hour to CSR cost per hour. Then, the optimal β can be seen to be growing exceptionally slowly with r:

$$\beta(r) \approx \begin{cases} \sqrt{r/\left(1 + r(\sqrt{\pi/2} - 1)\right)} & 0 \le r < 10\,, \\ \sqrt{2\ln\left(r/\sqrt{2\pi}\right) - \ln\left(2\ln\left(r/\sqrt{2\pi}\right)\right)} & 10 \le r < \infty\,. \end{cases} \tag{19}$$

From the left chart one sees that for an r of 10, which reflects delay costs that are 10 times that of staffing costs, the optimal β is about 1.68. For a call center with offered load of $R = 400$, this implies that safety-staffing should equal about 34 ($1.68 \times \sqrt{400} = 33.6$) and the call center then operates at 92.2% ($400 \div 434$) utilization.

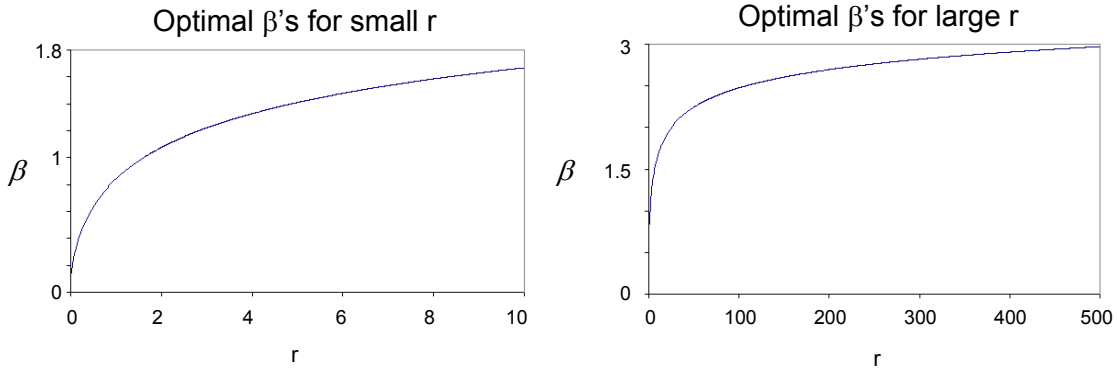


Figure 9: Optimal β for Linear Waiting and Staffing Costs (from [25])

4.1.2 Operational Regimes, Pooling and Economies of Scale

The square-root safety-staffing principle leads to additional insights concerning the nature of economies of scale in the M/M/N queueing systems. In particular, the analysis in Borst et al. [25] gives rise to three asymptotic cases, each of which displays different economies of scale.

In the first case, waiting costs of customers dominate the cost of capacity, and the optimal staffing policy uses an asymptotically fixed utilization rate. Staffing levels grow linearly with the offered load, and there are no economies of scale. In a large system, the vast majority of callers are served without delay. This is dubbed a (service) *quality-driven* regime.

An example of such a system is shown in Figure 10, which summarizes the performance of a large U.S. catalogue retailer. Focus on the peak period of 10:00am-11:00am: 765 customers called;

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		280	: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

Figure 10: A Quality-Driven Call Center (from [89])

service time is about 3.75 minutes on average, with after-call-work of 30 seconds and auxiliary work requiring roughly 5% of CSRs' time; ASA is about 1 second 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.

Another, common example of a quality-driven regime is in the operation of an IVR. Here, capacity is relatively inexpensive when compared to the cost of CSR assistance. To encourage customer self-service, companies ensure that capacity is ample enough that callers virtually *never* encounter congestion.

At the other extreme, staffing costs dominate the imputed cost of customer delay. In this case, the number of excess CSRs – beyond the number required to handle the offered load (R) – is asymptotically fixed in the optimal regime. Thus, as the offered load increases, utilization quickly approaches 100% (at a rate that equals the number of servers).

This is called an *efficiency-driven regime*. Examples are found in email response and many help-desk operations (that offer "free" service to customers who have recently purchased hardware or software). In these systems, essentially all customers are delayed in queue, ASA is on the order of an expected service time, and agents are utilized very close to 100% of the time.

In between these two extremes, are call centers that fall within the *quality and efficiency-driven* (QED) regime, in which quality and efficiency are carefully balanced. As they grow, these

centers display *both* the economies of scale shown in efficiency-driven systems and the high accessibility that is the characteristic of quality-driven operations.

This is the case in Figure 11, which summarizes the performance of 12 call centers, operated by a large U.S. health insurance company: one observes a *daily* average of 31-second ASA, 318-second AHT, with 91% agent utilization, in fact over 95% in a couple of the call centers. (Note also that 2.8% of calling customers abandoned. Customer impatience, however, is beyond the explanatory scope of Erlang C, and we address it in Subsection 4.2.2.)

Command Center Intraday Report

Date 06/13 - Tue		Updated Through: All Day									
		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 11: Performance of 12 Call Centers in the QED Regime (from [89])

Recall from (13) that the QED regime is characterized by a fraction of delayed customers that is neither close to zero (quality-driven) nor to unity (efficiency-driven). Indeed, more refined data from the above-mentioned health insurance company show that, overall, only about 40% of the customers were delayed, while the other 60% accessed an agent immediately, without any delay. Thus, the call-center characteristics described by Sze [137] identify the QED regime.

Economies of scale are the enabler that allows the QED regime to circumvent the traditional tradeoff between service level and resource efficiency. To sharpen this insight, we consider the following problem that is commonly addressed by call-center managers: the pooling of geographically dispersed call centers. This pooling may be achieved either physically – by closing some operations and expanding others – or "virtually" – through the use of networking technology that allows calls to be routed to various sites. For this problem we can compare how the different regimes affect the economies of scale enabled through pooling.

As a first step, we use (16)–(17) to define the following analogues to (5)–(6):

$$\widetilde{\text{ASA}} \;=\; \mathsf{E}\left[\frac{\text{Wait}}{\mathsf{E}[S]}\;\middle|\;\text{Wait}>0\right] \;\approx\; \frac{\mathsf{E}[S]}{\Delta}\,, \tag{20}$$

and

$$\widetilde{\text{TSF}} \;=\; \mathsf{P}\left\{\frac{\text{Wait}}{\mathsf{E}[S]}>T\;\middle|\;\text{Wait}>0\right\} \;\approx\; e^{-T\Delta}\,. \tag{21}$$

Note that these definitions modify the standard versions of ASA and TSF in two ways: they are conditioned on the event that delay is nonzero, and waiting time is measured in units of expected

service duration, $\mathsf{E}[S]$. This gives rise to simple expressions that are straightforward to compare across regimes.

We observe that in each of the three regimes, a single measure of system performance is fixed, which then determines the other performance measures:

- in the *efficiency-driven* regime, excess capacity Δ and, in turn, $\widetilde{\text{ASA}}$ and $\widetilde{\text{TSF}}$ are fixed;
- in the *quality-driven* regime, system utilization $\rho = \frac{R}{R+\Delta}$ is held constant; and
- in the *QED* regime, the service grade β and, in turn, $\mathsf{P}\{\beta\} \approx \mathsf{P}\{\text{Wait} > 0\}$ are fixed.

The above scalings have been formalized in Whitt [145].

Now 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. Figure 12 summarizes the results. Note that, within each column, the boxed entries highlight the performance measures that are fixed under that regime's scaling.

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	QED
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	$\rho = \frac{R}{R+\Delta}$	$\frac{R}{R+\frac{\Delta}{m}} \uparrow 1$	$\boxed{\rho = \frac{R}{R+\Delta}}$	$\frac{R}{R+\frac{\Delta}{\sqrt{m}}} \uparrow 1$
$\widetilde{\text{ASA}} = \text{E}\left[\frac{\text{Wait}}{\text{E}(S)} \middle\vert \text{Wait} > 0\right]$	$\frac{1}{\Delta}$	$\boxed{\frac{1}{\Delta} = \widetilde{\text{ASA}}}$	$\frac{1}{m\Delta} = \frac{\widetilde{\text{ASA}}}{m}$	$\frac{1}{\sqrt{m}\Delta} = \frac{\widetilde{\text{ASA}}}{\sqrt{m}}$
$\widetilde{\text{TSF}} = \text{P}\left\{\frac{\text{Wait}}{\text{E}(S)} > T \middle\vert \text{Wait} > 0\right\}$	$e^{-T\Delta}$	$\boxed{e^{-T\Delta} = \widetilde{\text{TSF}}}$	$e^{-mT\Delta} = (\widetilde{\text{TSF}})^m$	$e^{-\sqrt{m}T\Delta} = (\widetilde{\text{TSF}})^{\sqrt{m}}$

Figure 12: Erlang C in the Efficiency, Quality and QED Regimes (from [89])

Under efficiency-driven staffing, the service grade decreases from β to $\beta/\sqrt{m}$, and the delay probability increases from $P(\beta)$ to $P(\beta/\sqrt{m})$ (which can be significant even for small m's). Note, however, that $\widetilde{\text{ASA}}$ and $\widetilde{\text{TSF}}$ are unchanged. As $m \uparrow \infty$, we observe fast convergence to a system in which servers are 100% utilized – so that the system behaves as a single server that processes m times more quickly – and essentially all customers are delayed.

For the quality-driven system, there is a significant overall improvement of the service-level: $\widetilde{\text{ASA}}$ decreases to $\widetilde{\text{ASA}}/m$, $\widetilde{\text{TSF}}$ decreases to $(\widetilde{\text{TSF}})^m$ and the delay probability decreases from $P(\beta)$ to $P(\beta\sqrt{m})$. As $m \uparrow \infty$, essentially all customers are served immediately upon arrival.

Finally, in the QED regime, the service grade and probability of wait remain constant (by definition). In contrast, $\widetilde{\text{ASA}}$ decreases to $\widetilde{\text{ASA}}/\sqrt{m}$, and $\widetilde{\text{TSF}}$ decreases to $(\widetilde{\text{TSF}})^{\sqrt{m}}$. Note that it is both efficiency driven (occupancy increases to 100%) and quality-driven (a significant fraction, namely $1 - P(\beta)$, of the customers is served *immediately*).

4.2 Busy Signals and Abandonment

The Erlang C model provides an exceedingly simple means of trading off capacity and accessibility. In turn, its heavy-traffic limits provide insight into these tradeoffs that deepen our understanding of economies of scale in call centers and how they should be managed. There are, however, significant limitations to the Erlang C model.

In particular, recall that arriving calls have three ways in which they may exit the system: a call that finds all k trunk lines occupied encounters a busy signal and is blocked; a caller that becomes impatient may abandon the queue before being served; and a caller that waits for a CSR is served and then leaves. The Erlang C model ignores the effect of the first two of these three.

4.2.1 Busy Signals: Erlang B

A call center can eliminate *all* delays by setting the number of lines to be equal to the number of agents. In this case the so-called Erlang B formula (think "B" for blocking) characterizes the blocking (busy-signal) probability for the associated M/M/N/N system. There are no queues, and accessibility is measured solely in terms of the fraction of customers that encounter a busy signal. A serendipity is the well-known *insensitivity* of the blocking probability with respect to the service-time distribution. This accommodates general, rather than exponential, service-time distributions (hence M/G/N/N, rather than M/M/N/N).

In the QED regime, the Erlang B system displays square-root results that are analogues to those for the Erlang C system. Again, Erlang [47] reports that the Copenhagen Telephone Company had made use of this relationship as early as 1913, but a formal analysis appears to have been first carried by Jagerman [81]. He shows that, for large M/G/N/N systems with $N = R + \beta\sqrt{R}$, $-\infty < \beta < \infty$, the blocking probability is of order $1/\sqrt{N}$: $\sqrt{N} \cdot \mathsf{P}\{\text{All trunks are busy}\} \to \phi(\beta)/\Phi(\beta)$. Thus, even in the absence of the ability to queue, accessibility remains high in the QED regime.

For $\beta > 0$, the fraction of callers that is blocked in an Erlang B system is small. Furthermore, (14) shows that, under the same conditions ($\beta > 0$), the fraction is small enough that it would not overwhelm the system if allowed to queue. Of course, the Erlang C system's infinite space in queue (number of trunk lines) is not practically attainable.

In between the Erlang B and Erlang C systems, one trades off blocking with delay: the former decreases with the available space in queue, while the latter increases. How much queue-space should be allowed? Feinberg [51] performs a simulation study of an M/M/N/k system which systematically varies $k \geq N$. The paper argues that a mere 10% excess of lines over agents suffices for good performance: more lines would give rise to too much waiting; fewer, too many busy signals.

It turns out that another square-root principle emerges here, given the offered load and N grow as in the QED regime. Letting $k = N + b\sqrt{N}$ in such an M/M/N/k queue gives rise to a 'double' QED regime in which blocking is of the same order as in Jagerman [81], order $1/\sqrt{N}$, but with a smaller constant. The square-root principle for queue-dimensioning was addressed in [104], and the resulting steady-state distribution is formally characterized in Massey and Wallace [111]. Whitt [147] develops process limits for systems that also include abandonment, as well as service times that are mixtures of an exponential distribution and a point mass at zero. In turn, Whitt [148] uses these exact results as the basis for a diffusion approximation of G/GI/n/k systems.

One might think that queue size is unimportant in call centers, that waiting customers are only logical entities in a phantom queue. As was already mentioned, however, queue size determines overall "1-800" delay costs, which can be significant (millions of dollars). Furthermore, although the above discussion motivates the tradeoff between busy signals and delays, it fails to acknowledge the most prevalent outcome of excessive congestion – the build-up of impatience that culminates in a customer abandoning the tele-queue.

4.2.2 Abandonment: Erlang A

A model that incorporates both busy signals and abandonment is the so-called M/M/N/k+G queue. In this model, patience is defined as the maximal amount of time that the customer is willing to wait for service; if not served within this time, he or she abandons the tele-queue. The "+G" notation indicates that patience is generally distributed, *i.i.d.* over customers and independently of everything else. Baccelli and Hebuterne [17] were motivated by telephone services to analyze the performance of the M/M/N/k + G system. Brandt and Brandt [29] extend the results to cover more general birth-and-death processes in which arrival and service rates may vary from state to state. (It is both remarkable and useful that *general* patience is amenable to exact analysis.)

A special case is the M/M/N/k + M queue, in which patience is assumed to be exponentially distributed. A performance analysis "engine" for the M/M/N/k + M queue is publicly available at `www.4callcenters.com`. (The web site includes two tools, iProfiler and Charisma, to support workforce management of call centers. iProfiler is available for online use, free of charge.) For mathematical details, see Palm [116] and Riordan [123] (pages 109–112), as well as the more recent Garnett et al. [59], which specifically addresses call centers.

Prevailing practice is to install an ample number of lines, enough so that a busy signal becomes a rare event. In this case, one has an M/M/N+M system, which we shall refer to as "Erlang A" ("A" for Abandonment, and for the fact that this model interpolates between Erlang B and Erlang C).

In analogy to the heavy-traffic analysis of Erlang C models [25, 69], Garnett et al. [59] develop three operational regimes for an Erlang A system: efficiency-driven, quality-driven and QED. As before, the regimes are characterized by their delay probability: close to 1, close to 0, and within $(0, 1)$, respectively. And, as before, the QED regime, with $N \approx R + \beta\sqrt{R}$ agents, is robust enough to cover the full operational spectrum. Here, however, the service grade, β, can take both positive

and negative values, since abandonment stabilizes the system at all staffing levels. The operational characteristics of the QED regime are appealing, and they can be summarized as follows: server idleness, ASA and, most importantly, the fraction abandoning are all of order $(1/\sqrt{N})$. (The introduction of [59] is recommended for an informal elaboration.) Figure 3 summarizes the daily performance of a call center that operates, most of the time, in the QED regime.

We now use the Erlang A model to demonstrate mathematically how, in heavily loaded call centers, customer abandonment behavior significantly affects system performance. Consider Figure 3, which in fact details the daily operation of the Charlotte call center that is listed in Figure 11. Note that across busy half hours – for example, from 10:00am-11:30am – the number of agents working ("on production") does not vary significantly. At the same time, changes in the offered load, the numbers of arriving calls and AHT's, are matched by changes in both the ASA and abandonment rate.

To get a sense of how abandonment affects performance, we fit the Erlang A model within individual half-hour intervals. From Figure 3 we naïvely use the count of arriving calls and the AHT as estimates of λ and μ. (We will discuss estimation in more detail in §6.) We round (up) "On Production FTE" to arrive at an approximate N. Then given three out of four parameters for the M/M/N+M, model, we search for the abandonment rate that (roughly) generates the ASA and abandonment percentage observed during the half hour.

For example, during the period from 10:30am-11:00am, the procedure yields an estimated mean time to abandonment of 30 minutes. Given this estimate, the absence of only 5 agents (out of the 223 working) would likely result in almost a *doubling* of both the ASA and the fraction abandoning.

Interestingly and significantly, a model in which average patience is 30 minutes differs dramatically from a model which does not acknowledge abandonment ("infinite patience"). For the half-hour 10:30am-11:00am, the latter would give rise to an unstable system in which 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 be close to 7 minutes – an order of magnitude error in the predicted performance if one ignores abandonment.

Thus, in heavy traffic, even a small fraction of calls that abandons the queue (or is blocked) can have a dramatic effect on system performance, and it should be accounted for when determining minimum staffing levels. For this reason we recommend the use of Erlang A as *the standard* to replace the prevalent Erlang C model.

Indeed, a common complaint one hears from call-center managers is that workforce management systems consistently recommend overstaffing. While some managers develop an intuitive sense of how to adjust staffing levels down, a better approach is to model abandonment in the first place.

4.3 Time-Varying Arrival Rates

As is clear from Figures 5 and 8, the arrival rate of calls can change significantly – and predictably – throughout the day. The "Erlang" models (C, B, and A), however, assume that arrival rates (and other system parameters) are constant. Hence, in practice the models are typically used only for shorter intervals of time, such as 30 minutes, for which the arrival rate is (hopefully) fairly constant. The instantaneous arrival rate, $\lambda(t)$, is averaged over the desired interval, T, to calculate an interval average, $\lambda = \frac{1}{T}\int_0^T \lambda(t)\,dt$. The average arrival rate is then used to calculate (stationary) performance measures, such as ASA or $\mathsf{P}\{\text{Wait} > 0\}$, for the interval.

Of course, stationary models often cannot adequately capture the performance of highly time-varying systems. Furthermore, the use of stationary performance measures implicitly assumes that the time required for the system to relax is small when compared to the interval for which the measure is used. This should be the case for systems in which the event rate, $\lambda + N\mu$, is large when compared with the length of the interval, T, typically 30 minutes. But exceptions arise, again, with abrupt changes in the arrival (or, for that matter, service) rate, or when overload occurs during one or more intervals. In this case a backlog builds up, and nonstationarity must be accounted for.

One method of accommodating time-varying parameters is numerical. Yoo [151] and Ingolfsson et al. [80] investigate exact methods, numerically solving the Chapman-Kolmogorov forward equations for $\mathrm{M}_t/\mathrm{M}/N_t$ systems to calculate the associated transient system behavior. Yoo [151] and Ingolfsson and Cabral [79] approximate continuously varying parameters with small, discrete intervals and use the so-called randomization method (see Grassmann [61]) to explicitly calculate the change in system occupancy from one small interval to the next.

Another natural means of accommodating changes in the arrival rate is to reduce the interval over which a stationary measure is applied. In the limit, one heuristically applies a stationary Erlang formula for instantaneous $\lambda(t)$, and as $\lambda(t)$ changes, one calculates a continuously varying measure of accessibility, such as $\mathsf{P}_t\{\text{Wait} > 0\}$. In turn, one averages the point-wise estimates to approximate the average performance for the interval: $\frac{1}{T}\int_0^T \mathsf{P}_t\{\text{Wait} > 0\}\,dt$. This is the essence of the *point-wise stationary approximation* (PSA) of Green and Kolesar [63]. But the PSA does not explicitly consider nonstationary behavior that may be induced by abrupt changes in the arrival rate, and it appears to perform less well in these cases [63]. For example, without accounting for abandonment, the PSA does not allow for instances t in which $\lambda(t) > N\mu$, and such short-term overloads can (in fact can be designed to) occur.

Mandelbaum and Massey [98] analyze a single-server queue with time-varying arrival and service rates. (Formally, the $M_t/M_t/1$ queue.) A notable qualitative outcome is that a time-varying queue alternates among several phases, the major ones being under-loaded (or sub-critical), over-loaded (or super-critical) and critically-loaded. Roughly speaking, steady-state analysis applies to under-loaded regimes, and over-loaded regimes are well approximated by fluid-models. Critically-loaded regimes exhibit, in some sense, the null-recurrence behavior of a Markov chain and are rather subtly described. The phase transitions of [98] should also apply to *multi-server queues* in efficiency-driven operations, but this is yet to be verified.

Interestingly, allowing the number of servers to increase can simplify things. This is well illustrated in Mandelbaum et al. [101, 102, 103], which model a call center with abandonment and retrials. In these papers, all parameters can vary with time. Then, given fast transitions through periods of critical loading, fluid and diffusion limits are derived for the queue-length and waiting-time processes.

Indeed, the fluid models provide excellent fits to the transient behavior of systems that, otherwise, are far beyond the capabilities of exact analysis. Moreover, the models are intuitive and easy to set up. For example, primitives in [101] are as follows: the arrival rate, λ_t; a number n_t of servers, each processing at rate μ_t; the abandonment rate θ_t; and the initial queue length $Q(0)$. Then the first-order differential equation,

$$\frac{dQ(t)}{dt} \;=\; \lambda_t - \mu_t \min(Q(t), n_t) - \theta_t (Q(t) - n_t)^+ \,,$$

defines the system occupancy $Q(t)$ as it evolves. A queue with abandonment and retrials is similarly modelled with two intuitive, autonomous first-order differential equations.

Figure 13 compares the fluid approximation of a system with retrials to actual (simulated) system performance. In the example, the arrival rate is $\lambda_t = 10$ per hour at all times except 9–11am, when it is increased to 110 per hour. All the parameters are constant. In particular $u_t = 1$ per hour and $n_t = 10$ at all times. The figure' s circles represent the average number in queue from the simulation of a Markovian system with the given parameters. The solid line that runs through the circles is the *theoretical* queue length calculated from the fluid model – truly a remarkable fit.

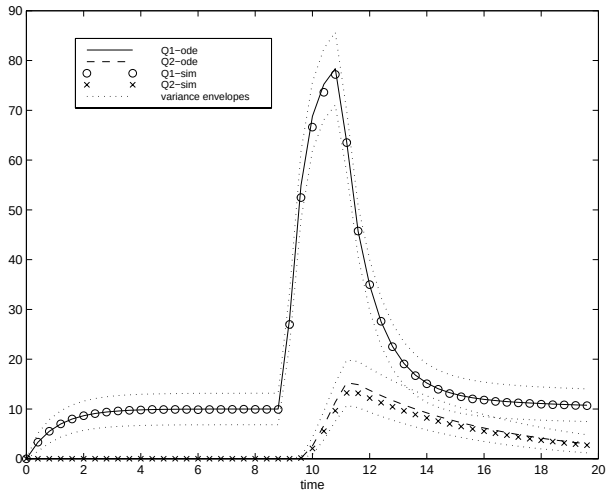


Figure 13: Fluid Model of the Transient Behavior of an Overloaded Queue (from [101])

Jennings et al. [83] extend the square-root staffing principle to account for intertemporal effects, allowing the number of servers to change in response to a time-varying offered load. The scheme is based on infinite-server approximations, as in Section 4.1.1, and uses results for the $M_t/G/\infty$ queue, developed by Eick et al. [45, 46]. In particular, [83] assumes the number of busy agents at t is Poisson-distributed with mean $\mathsf{E}[\lambda(t - S_e)]\mathsf{E}[S]$. ($S_e$ is the equilibrium distribution of the service time: $\mathsf{P}\{S_e \leq t\} = \mu \int_0^t \mathsf{P}\{S > \xi\}d\xi$.) Thus, busy servers *lag* arriving calls by an equilibrium service-time distribution.

The staffing heuristic then uses a time-varying version of (15) with $R_t = \mathsf{E}[\lambda(t - S_e)]\mathsf{E}[S]$, to determine staffing levels as $\lambda(t)$ varies. In these time-varying models, the appropriate β also varies with t, and the paper uses the heuristic procedure in (18) for its calculation. Numerical tests in [83] show that the procedure performs well under a wide variety of conditions. And indeed, current work by Mandelbaum et al. [100] formally justifies the underlying asymptotic scheme.

Infinite-server approximations to time-varying queues have been further analyzed in Massey and Whitt [113]. The paper develops an asymptotic characterization of the time lag between the maximum arrival rate and the maximum number of busy servers, as the arrival changes ever more slowly. The asymptotic analysis leads to a refined, modified offered load (MOL) heuristic that performs well in numerical tests.

The idea that the number of busy servers (peak congestion) lags the arrival of calls (peak load) has also been used to improve the performance of other approximations. Green and Kolesar [64] show that a "lagged" version of the PSA performs better than a non-lagged version when estimating peak-hour congestion. Similarly, Green et al. [65] show that using a lagged arrival rate often significantly improves the performance of the original, Erlang-C-based staffing models, themselves.

Finally, Ingolfsson et al. [78] have recently evaluated the accuracy and computational requirements of a number of approaches, including the exact calculation of the Chapman-Komogorov forward equations, the method of randomization [61], the infinite-server approximations of [45, 46], and the MOL approximation of [113]. The results show that the method of randomization generally produces results that are close to exact with about one third of computational time. Among the quicker but more approximate methods, the MOL tends to outperform the infinite-server approximation. We believe the latter will continue to prove useful, however, because of its particularly simple and intuitive nature.

4.4 Uncertain Arrival Rates

While the models described above explicitly represent uncertainty in interarrival times, the overall arrival rate is assumed to be known. As Section 6.3.1 describes in more detail, however, this is typically not the case. Rather, the arrival rate is predicted from historical data and is not known with certainty.

It can be risky to ignore arrival-rate uncertainty, and for call centers that operate in the QED regime, such as those in Figure 11, the danger is particularly acute. For example, if a call center plans to operate at 95% utilization and the arrival rate turns out to be 5% higher than planned, either actual system utilization climbs to 99.75% – and waiting times explode – or customer abandonment far exceeds planned-for levels. Given this potential for difficulty, it is natural to increase safety staffing above the level required by the expected arrival rate. Surprisingly, however, there is little work devoted to an exploration of how much.

Specifically, suppose Λ is the random arrival rate and let $f(\Lambda)$ be an arrival-rate dependent measure of system performance. Typically these measures are nonlinear functions of Λ. For example, given fixed N and μ, Figure 6 shows the highly convex relationship between ASA and λ. Common practice is to staff so that $f(\mathsf{E}[\Lambda])$ attains the desired level of performance, but the nonlinear nature of $f(\cdot)$ implies that actual performance, $\mathsf{E}[f(\Lambda)]$, will not match that of the plan (and typically will be worse).

To account for this nonlinearity, Chen and Henderson [40] develop simple upper and lower bounds on $\mathsf{E}[f(\Lambda)] - f(\mathsf{E}[\Lambda])$. These bounds can then be used as an aid in selecting bounds for staffing levels. Ross [125] numerically tests the following heuristic, proposed by Grassmann [62]: add the variance of Λ to the offered load when determining safety staffing, so that $N \approx R + z\sqrt{R + \mathrm{var}(\Lambda)}$. Here, z is a number of standard deviations derived from an infinite-server approximation, such as (18). In [125] the heuristic is shown to consistently underestimate the number of CSRs needed to attain a desired service level, and [125] suggests and tests modified versions of the heuristic that correct for the bias.

Jongbloed and Koole [84] offer two approaches for setting staffing levels: the first assumes there exists a fixed staffing level to be determined, and the second that a separate pool of flexible workers can be called in if needed. More specifically, suppose a call center wishes to set an 80-20 TSF: $\mathsf{P}\{\text{Wait} \leq 20 \text{ seconds}\} = 80\%$. Given a fixed λ and μ, the call center would simply choose a staffing level, N, to meet the target. If Λ is random and N and μ are fixed, however, then larger realizations of Λ imply lower $\mathsf{P}\{\text{Wait} \leq 20 \text{ seconds}\}$'s. Therefore, a fixed N will only achieve the desired TSF a fraction of the time. The paper's first approach chooses a target probability, α, with which the call center should meet or exceed the 80-20 TSF. Then given this α, it calculates a λ_α such that $\mathsf{P}\{\Lambda \leq \lambda_\alpha\} = \alpha$, as well as a staffing level, N_α, so that the 80-20 TSF is met for

λ_α. N_α becomes the fixed staffing level. The second approach sets target levels for two pools of employees: full time CSRs, who work no matter what the realization of Λ; and flexible CSRs, who can be called in if the realization of Λ is "too" large. In this case, the call center chooses two target probabilities, $\alpha_1 < \alpha_2$, and two associated arrival rates, $\lambda_1 < \lambda_2$, so that $\mathsf{P}\{\Lambda \leq \lambda_i\} = \alpha_i$, $i = 1, 2$. In turn, the number of full-time CSRs, N_1, is chosen so the 80-20 TSF is met for λ_1. Similarly, N_2 is set so that the 80-20 TSF is met for λ_2, and the number of call-in CSRs becomes $(N_2 - N_1)$.

The above methods can be naturally combined with square-root safety staffing. Suppose one seeks to guarantee a service grade β (or one of its analogues) with a certain probability, α. For the first approach, let $R_\alpha = \lambda_\alpha/\mu$ be the associated upper bound on the offered load, with λ_α defined as before. Then to ensure a service grade of β with a probability of α one staffs $N_\alpha = R_\alpha + \beta\sqrt{R_\alpha}$ CSRs. For the second, let $R_1 = \lambda_1/\mu$ and $R_2 = \lambda_2/\mu$, so that $N_1 = R_1 + \beta\sqrt{R_1}$ and $N_2 = R_2 + \beta\sqrt{R_2}$ are the associated staffing levels.

Readers who are primarily interested in queueing performance models can now proceed to Section 4.7.

4.5 Staff Scheduling and Rostering

In the example presented in Section 3, a two-stage process is used to transform half-hourly staffing requirements into schedules for individual CSRs. First a minimum-cost set of schedules is found. Then, individual agents are assigned to various schedules. We now discuss articles devoted to these problems.

Scheduling Problems: The staff scheduling problem (9) is quite general and is tied to call centers only through the fact that queueing formulae are used to determine the underlying half-hourly staffing requirements, N_i. In fact, the IP (9) has a history that dates back to Danzig in the 1950's. (For brief histories see [16, 138].) It seeks a minimum cost (positive, integral) linear combination of the columns of A that "covers" the requirements defined on the right-hand-side, $\vec{N}$.

For this reason (9) is known as a *set-covering* formulation. Although simple to state, set-covering formulations can be difficult to solve for problems with many rows (time slots) and columns (feasible schedules). Our understanding is that, in practice, these scheduling problems are not solved to optimality. Rather, suboptimal solutions are arrived at via simulated annealing and other heuristic methods.

In particular, it is the presence of breaks in the middle of employee shifts that cause trouble. If every employee were to work consecutive half-hour periods, without breaks, then every column of the matrix A would have contiguous ones. In this case, the matrix is called totally unimodular (TU). In turn, it is well known that for TU A, the optimal solution of the linear program (LP) relaxation of (9) is integral (for example, see Nemhauser and Wolsey [115]). With the introduction of breaks, however, A is no longer TU, and the LP solution need not be integral.

Given that breaks are the source of computational difficulty, a natural heuristic approach is to tackle the problem in two stages. In the first step, shifts without breaks – often called "tours" – are defined, and an LP is run to find a minimum cost set of tours that cover the staffing requirements. In the second stage, breaks are heuristically placed within tours and additional tours are added as needed. An example of this procedure can be found in Segal [128], which uses a network-flow formulation in the first stage of the problem.

An alternative is to restrict the size of the problem by exogenously limiting the numbers of columns of A that may be considered. This is the approach taken by Henderson and Berry [75] who first heuristically select a "good" subset of columns and then use rounding to make LP relaxations feasible.

Another approach is to avoid the set-covering formulation (9) all together and look for alternatives that may be easier to solve (see Thompson [138]). For example, Aykin [16] and Brusco and Jacobs [35] define two sets of variables: the first defines the start-time and duration of shifts; and the second defines possible break times. Additional constraints are then used to define which break times are feasible for what shifts. For more on these alternative formulations, see the papers and their references.

Even with these alternative approaches, the solution of scheduling IP's can be computationally burdensome. When schedules use short intervals – for example, 15 or 30-minute periods – and last for many days, the number of rows (time periods) grows into the hundreds, and the number of columns (feasible shifts) into the thousands [138]. Again, in practice one finds solutions via "global search" heuristics, such as simulated annealing and genetic algorithms, as well as local search techniques.

Finally, it is worth noting that the scheduling problem may become easier, rather than harder, for larger call centers. In particular, even though there may be many thousands of feasible schedules, the optimal solution of the LP relaxation (of the scheduling IP) will only include a small fraction of them. (Some fraction of the rows will have binding constraints, and each binding constraint will correspond to a schedule.) Given this fixed set of feasible schedules, larger numbers of CSRs make the simple rounding (up) of the LP relaxation more attractive. (See Mandelbaum and Ruszczynski [105].) For example, a 1000-CSR call center that uses only 50 feasible schedules would add between 0 to 50 extra CSRs due to rounding. That's 0%–5% above the (infeasible) LP lower bound.

Joint Staffing and Scheduling: One can also consider the underlying staffing (queueing) problem together with the scheduling problem. For example, any feasible solution to the IP (9) will meet the minimum staff level N_i in all periods. Furthermore, scheduling constraints make it likely that a solution will strictly exceed N_i in some – perhaps many – periods. In these periods the call center will strictly exceed the service-level constraint ASA*.

An alternative formulation of (9) relaxes the interval-by-interval service-level restriction and, instead, seeks a minimum cost set of schedules, subject to an aggregate service-level constraint: $\sum_i f_i \, \mathsf{E}_i[\text{Wait}] \leq \text{ASA}^*$, where $f_i = \lambda_i / \sum_j \lambda_j$. A significant complication in this formulation, however, is the fact that the scheduling problem is no longer linear. Nevertheless, Koole and van der Sluis [90] demonstrate that, when the A-matrix (8) is TU, a greedy procedure exists for finding an optimal sets of schedules.

This joint approach to staffing and scheduling also allows for more straightforward inclusion of time-varying arrival rates. Yoo [151], Ingolfsson et al. [80], and Ingolfsson and Cabral [79] all take this approach. In these papers, a higher-level scheduling routine proposes potential schedules for CSRs, and a lower-level service-level evaluator then explicitly calculates (transient) system performance for the resulting $\text{M}_t/\text{M}/N_t$ system. The methods of generating high-level schedules vary: [151] tests greedy heuristics and dynamic programming (DP); [80], genetic algorithms; and [79], an IP cutting-plane heuristic. (The DP algorithm presented in [151] relies on the stochastic submodularity of the occupancy of the $\text{G}/\text{M}/N$ system with respect to initial queue size and staffing

level. This property is formally demonstrated in Fu et al. [53].) Finally, Atlason et al. [15] use a scheme in which sample schedules from a high-level IP are evaluated for feasibility via discrete-event simulation. The approach assumes that the service level is concave with respect to the staffing level, and it uses this property as the basis for the systematic introduction of cutting planes in the top-level IP. We note that the example problems solved in all of these papers are not large, and the computational feasibility of (at least the current versions of) the approach remains unclear.

Assignment of CSRs to Schedules: Once a set of schedules is determined, an assignment problem must be solved to match agents and schedules. Call centers with few workers can perform the assignment manually, but operations with 100's or 1000's of CSRs require supporting software. Often call centers use a process called "shift bidding" to make the assignment. In shift bidding each employee first states preferences for various schedules. Then employees are ranked, typically according to seniority, and they are assigned to schedules according to their ranking. Thompson [139] describes a case study in which a decision support system is used to automate the process.

We are also aware of a call center that posts schedules on a web site and allows employees (who are typically students) to choose shifts on a FCFS basis. In this operation, most shifts are taken within 10 minutes of posting, which turns the "bidding" process into essentially a lottery. As a consequence, some CSRs can go for months without working a shift, a reality that can persist only when agents tolerate volatile schedules.

4.6 Long-Term Hiring and Training

Like the scheduling problem described above, the longer-term problem of deciding how many employees to hire and train is not necessarily tied to call centers, and a number of disciplines have addressed problems related to hiring and training. Broadly speaking, this research has focused on two fundamental problems with the "solution" to the hiring problem defined by (10).

The first set of work introduces an element of control into (10). That is, rather than myopically hiring a minimal number of employees in each period, this work considers minimizing staffing costs over some time horizon and looks for optimal hiring rules. Furthermore, the hiring systems modelled in this stream are much more complex and include attributes such as: capacity improvements that come with experience; multiple stages of learning and training; and the ability to hire "experienced" employees.

The bulk of the literature that explores these control issues uses a mathematical programming approach derived from *aggregate planning*. Perhaps the first is the seminal work of Holt, Modigliani, Muth and Simon [77]. Also well known are a monograph by Grinold and Marshall [66] and a volume on planning models edited by Charnes, Cooper and Niehaus [38]. Akşin [2] uses dual prices from these types of mathematical programs to account for the "appreciation" and "depreciation" in employee value that comes with experience.

The second set of work recognizes that employee advancement and turnover occur at random, when considered at the aggregate level. It thus models the evolution of an operation's population of employees as a stochastic process, mostly Markov Chains. In contrast to the research based on mathematical programming, this work has not typically considered issues of control. Batholomew, Forbes and McLean [19] provide a comprehensive summary of research in this area.

More recent work has sought to combine elements of stochastic modelling and control. Bordoloi and Matsuo [24] derives steady-state performance measures for a heuristic class of linear control

rules. Gans and Zhou [56] develop a DP model of long term hiring that admits a more general class of controls. In [56], the lower-level scheduling problem (9) is explicitly modelled as the core of the DP's one-period cost function, and optimal hiring policies are characterized as analogues to "order-up-to" policies in the inventory literature. In these policies the current numbers of employees determine an optimal number of new employees to target having on hand. More specifically, if the current number of new employees falls below a threshold, then the difference between the current and the threshold number of employees is hired; if the current number exceeds the threshold, no hiring is done.

4.7 Open Questions

The research that we have reviewed thus far has addressed capacity management in the simple setting of a single call center with a single type of calls. As the preceding sections should make clear, a great deal of progress has been made in understanding how best to manage this base case. Nevertheless, even here there remain significant open questions. Section 4.7.1 describes problems related to lower-level queueing models, and Section 4.7.2 outlines questions related to higher-level scheduling and training.

4.7.1 Simple Multi-Server Queues in the QED Regime

Dimensioning: The $N = R + \beta\sqrt{R}$ *form* of the square-root safety-staffing principle, along with its insight into economies of scale, has been well established. For the Erlang C system, Borst et al. [25], also provide a complete analysis of the dimensioning problem of how to determine the economically optimal value of β. For all other models, however, work remains to be done.

For various other "Erlang" models, the stationary behavior in the asymptotic QED regime has been characterized. In particular, performance analysis for the Erlang B system can be found in Jagerman [81], and that for the Erlang A system in Fleming et al. [52] and Garnett et al. [59]. Current work by Mandelbaum, Massey and Rider [100] analyzes the time-varying system developed in Jennings et al. [83], work which requires deep mathematics (for example, excursion theory). In each of these models, as well as all those surveyed in the sequel, the optimal β is yet to be determined.

Time-Varying Conditions: Recent work on fluid and diffusions approximations, such as that in Mandelbaum et al. [101, 102, 103], offer a framework for incorporating both abandonment and nonstationarity, as well as retrial behavior. This approach should also work well for modelling the abrupt overloads that arise from predictable events, such as rushes of calls that are generated by TV advertising. Fluid approximations may work less well in underloaded situations, however, as argued in Altman, Jiménez, and Koole [5].

The employment of square-root safety-staffing to a time-varying system, as done heuristically in [83], gives rise to a QED operation. As Borst et al. [25] suggest, the optimal dimensioning of such a system should also lead to this same regime. The justification of the heuristics and the optimization of staffing levels are mathematically challenging problems, however. (See [100, 103].)

General Service and Interarrival Times: Asymptotic analysis in the QED regime also should be extended to cover non-exponentially distributed service times. This is important practically, as service times in call centers can well be non-exponential [23, 33, 137], and they can affect performance

in subtle ways [107].

Exact analysis of the performance of systems with general service times is challenging theoretically, however. For general service times, a state-descriptor of the queue must account for the state of each server and, in the QED limits, the number of servers increases indefinitely. Puhalskii and Reiman [121] prove weak convergence for (the queue and virtual wait) processes of the GI/PH/N system to a complex, multidimensional diffusion process, but not its steady state. Whitt [148] characterizes the steady-state behavior of a one-dimensional diffusion approximation for G/GI/N/k systems.

Jelenkovic et al. [82] analyze the GI/D/N system. The fact that service times are deterministic allows for an especially simple analysis. The reason is that waiting times in GI/D/N systems are the same under both FCFS service and the cyclical assignment of servers (ideal load balancing), and the latter allows each server to be analyzed individually. For example, the M/D/N queue is equivalent to an E_N/D/1 system, which has Erlang(N) inter-arrival times (the sum of N i.i.d. exponentials). Taking QED limits of a GI/D/N queue thus amounts to applying the central limit theorem to the inter-arrival times of a single-server queue.

Simulation experiments of M/GI/N queues in the QED regime are conducted in Mandelbaum and Schwartz [107]. As Figure 14 shows, deterministic service times cut ASA in half, when compared to the analogous M/M/N system. This is consistent with conventional heavy traffic (11), as explained in [82].

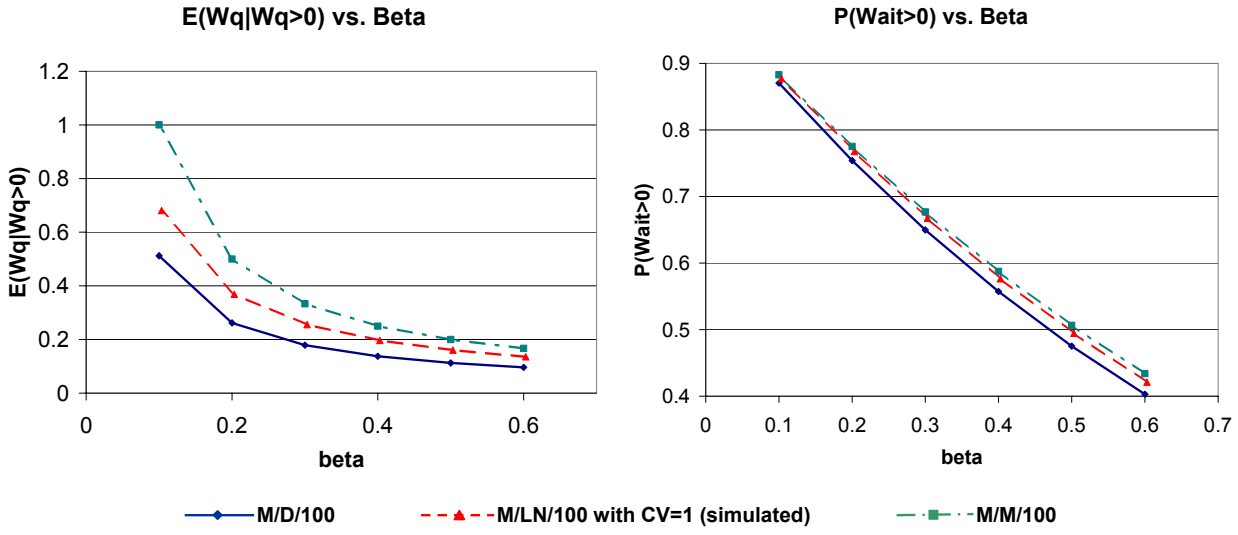


Figure 14: Effect of Service-Time Distribution on $\widetilde{ASA}$ and P{Wait > 0} (from [107])

Surprisingly, lognormally distributed service times, with mean and coefficient of variation both equal to one, also *reduce* congestion. This finding runs directly counter to (11), which predicts the same ASA as in an M/M/N system (or an even a worse ASA, taking the heavier tails of the lognormal distribution into account).

The paper [107] takes the notion of heavy tails to the extreme, considering an M/GI/100 system with two-valued service times: $1-\epsilon$, with a high probability, and 100 with a small probability, so

that both the mean and variance of the service-duration equal 1. Conventional heavy traffic results would identify the system's performance with that of an M/M/100 queue, but simulations confirm that in the QED regime it actually behaves like an M/D/100 system. The many servers in this case render the effect of the service tail negligible.

In contrast to service times, general inter-arrival times present no extra difficulty in the QED regime, this as long as they are not too "heavy tailed." (Heavy enough of a tail seems to necessitate a change of scale, to something like $N \approx R + \beta R^H$, where $H > 1/2$ depends on the weight of the tail [146].) Indeed, Halfin and Whitt [69] already analyze the GI/M/N queue.

Thus, the QED regime and its accompanying square-root principal appear to be widely applicable. For many models, however, complete understanding and supporting analysis still remain an important open avenue for research.

Readers who are primarily interested in queueing performance models can now proceed to Section 7. Those who have an interest in queueing control models can now continue with Section 5.

4.7.2 Staffing and Hiring models

How best to set staffing levels, given uncertain arrival rates, is a problem of great importance. As already noted, for call centers that operate in the QED regime, it is no doubt critical. We believe that a complete treatment of the problem needs to model customer abandonment, however.

Specifically, suppose there is an arrival-rate forecast, Λ, and that μ and N are fixed. Then in models which do not account for abandonment, such as the Erlang C, a realization of $\Lambda > N\mu$ makes the system unstable: the queue length explodes, and positive increasing (cost) measures of the backlog make no sense – they are infinite. Models that include customer abandonment have no such problem, however. Even when $\Lambda > N\mu$, abandonment stabilizes the system, stationary queue-length and delay distributions exist, and one can meaningfully trade off capacity against measures of system accessibility, such as abandonment rate and delay.

A second problem of practical importance is the *joint* determination of staffing levels and schedules. Previous research suggests that, in comparison to the standard procedure – which first sets half-hourly staffing requirements, using (7), and then uses the scheduling IP (9) to determine CSR schedules – the joint approach offers both cost and service-level benefits. Work remains to be done, however, to extend this approach for larger systems. In particular, for higher call volumes, analytical approximations, such as [83, 113], should be of use in the lower-level problem of evaluating service-levels.

Similarly, the practice of *first* determining CSR schedules, and *then* assigning particular agents to schedules, can potentially be improved upon. In reality, not all CSRs are necessarily available for every feasible schedule, and the number of agents available for various schedules can greatly influence the types of schedules required.

Of course, in theory one would find even better solutions by integrating staffing, shift determination, and assignment problems together. While the integrated problem is likely to be too complex to solve to optimality, heuristic approaches may be of value. Even if a combined approach to the three is impractical, work is required to understand which two of the three – staffing and scheduling, or scheduling and assignment – should be combined.

Finally, the analysis of hiring models should be extended in at least two directions. First, in many call centers, additional training and promotions are not automatically granted to all CSRs; rather, managers control the numbers of CSRs that advance from one skill-level to the next. While the ability to control the numbers of advancing CSRs is modelled in some deterministic analyses, it is not accounted for in stochastic analyses, such as that of Gans and Zhou [56]. Second, while the DP approach to stochastic hiring models does allow for the characterization of optimal policies, it is nevertheless computationally burdensome. Alternative methods of searching for optimal "hire-up-to" numbers need to be developed.

Readers who are interested in higher-level human resources problems can now proceed to Section 7.

5 Routing, Multimedia, and Networks

The research reviewed in Section 4 addresses a highly simplified setting, one in which a single type of call arrives to a call center at a single location that handles only inbound calls. In fact, advances in call center technology expand the possibilities and make capacity management decisions more complex in all of these aspects: skills-based routing technology allows for distinctions to be made among many types of calls and many skills of servers; the growth of email, chat, and web-based services expands call centers into multi-media "contact" centers; and networking technology allows for multiple locations to be linked into larger, "virtual" call centers.

In this section we describe the new capabilities, and we review a growing body of research that addresses it. We start in §5.1–§5.1.1 by providing a qualitative discussion of skills-based routing and associated capacity-planning problems. Then in §5.1.2–5.1.4 we review queueing research that applies to the control of these systems, as well as insights into staffing that some of the models provide. Next, §5.2 provides a qualitative description of problems relating to multimedia and a brief review of recent models in this area. Finally, §5.3 describes new networking technology.

5.1 Skills-Based Routing

Consider a call center of a large European company which provides technical support for a product in all major European languages. There are several approaches for staffing the operation.

One strategy is to establish a single pool of agents, each of whom is cross-trained to offer service in all of the languages the center supports. This type of operation is straightforward to manage; it may be viewed as a classical center that handles a single type of call. Labor cost for the required multi-lingual agents may be exorbitant, however. If the number of languages supported is high, it may even be impossible to find qualified people.

At the other end of the spectrum, one may staff a separate pool of agents for each language. In this case, the call center can be regarded as several smaller, independent call centers operating in parallel. The advantage is that one need not hire multilingual agents. From §4.1.2 we know that the cost is a loss of economies of scale that a single, large pool of agents would provide.

In between, one might partition languages into separate subgroups – for example, French, Italian, and Spanish; Dutch, German, and English – and staff these intermediate groups in parallel. Still, this excludes the possibility of hiring agents that do not speak an entire subset, and it does

not make full use of agents who speak languages beyond a subset.

Another, more flexible alternative is to use *skills-based routing*. In this scheme, each agent is acknowledged as speaking an individual subset of the languages offered; agents identify the languages for which they qualify when they log into the system. Arriving calls are then identified by language – either through the number called by the customer (DNIS) or through the customer's interaction with the IVR – and the ACD is programmed to route calls only to qualified agents.

Of course, the applications of skills-based routing reach far beyond the language domain. For example, it can be used to route customers with different types of questions to agents with various sets of expertise or to route high-value customers to more highly trained agents. Whenever one defines many types of calls, and agents are heterogenous in the types of calls they (can and) may handle, skills-based routing becomes a necessity.

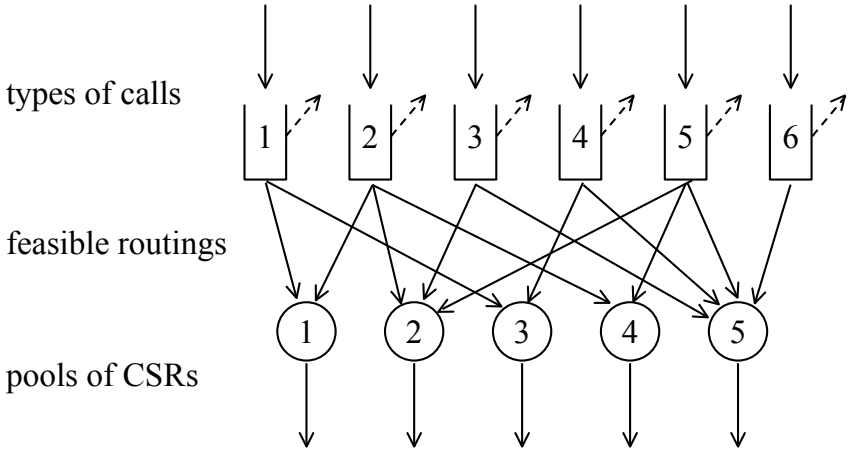


Figure 15: An Example of Skills-Based Routing

Figure 15 is an example of a system with an elaborate skills-based routing structure. In it 6 *types* of calls are routed to 5 *pools* of agents. All agents in a pool can handle the same set of call-types; we equivalently say that the agents within a pool have the same *skills*. Arrows between call types and agent pools describe the various pools' skills. Dashed arrows at the sides of queues represent customer abandonment. (We note that the nomenclature for skills-based routing has not yet become standardized. For example, one ACD manufacturer refers to customer types as "skills" and to agents with the same skills as having the same "skill-set.")

It is important to note that, in addition to call content and agent training, call types and agent skills may be defined according to any of a vast set of attributes. Examples include operational attributes, such as the forecasted duration of service, and economic attributes, such as how much an agent is compensated.

The majority of advanced ACDs now offer some capability to do skills-based routing, provided as a menu of options from which managers can choose. But our experience is that skills-based routing capabilities are scarcely offered with guidelines on how best to use them. Indeed, the technology has raced ahead of managers' and academics' understanding of how it may best be used, and the characterization of effective strategies for skills-based routing poses challenging, unanswered questions at all levels of the capacity-planning hierarchy described in §3.

5.1.1 Capacity Planning under Skills-Based Routing

At the lowest level, new call-routing problems emerge. When an agent becomes free and one or more calls (for which the agent is qualified) is waiting to be served, one must choose which call to attend to first, if any. For an arriving call that finds one or more appropriately-skilled agents free, one must decide to which agent the call should be routed, if any. Often these are respectively dubbed *call selection* and *agent selection* problems.

These call-routing problems feed requirements for minimal staffing levels that are a multi-skilled analog to the single-skilled call center's solution of (7). In the single-skilled center, there is one type of agent, and the minimum number of agents required for interval i is a scalar, N_i. In the multi-skilled call center there are many pools of agents. In turn, the "minimum" number of agents required is a vector in which each element represents the number of agents in a particular pool.

Note that there is an extra layer of complexity, however. There is typically more than one "minimal" vector of agents that feasibly fulfill the center's service requirements. In a call center that serves two types of calls, for example, a certain time interval may require 5 CSRs that can handle type-1 calls and 10 CSRs that can handle type 2; alternatively, 13 CSRs that are cross-trained to handle both call-types might also suffice. (The characterization of feasible staffing induces a partial order of staffing vectors.)

Furthermore, the means of determining whether a given fixed vector of agents is or is not feasible is through the application of a call-routing scheme. Thus, the "on-line" control problem of routing calls, and the "off-line" problems of determining half-hour staffing levels per skill – as well as of designing customer types and server skills – are closely intertwined. (In fact, the determination of system stability – whether or not there exists *some* routing scheme that is capable of keeping up with arriving work – is not trivial, though it can be arrived at via the solution of an LP [11, 18, 55].)

The upper levels of the capacity planning hierarchy also become much more complex. In the intermediate scheduling integer-program (9), each N_i, again, becomes a set of minimal staffing vectors for interval i. A solution to the program is feasible if, in each i, the numbers of agents on hand exceeds at least one of the interval's minimal vectors. Finally, the top-level problem expands from the consideration of how many employees to hire each period to how many to hire *and train*. Furthermore, in any period, training may be applied to transform one class of existing employee into another.

While skills-based routing affects the entire chain of capacity management activities, research into how best to solve skills-based routing problems is just beginning. Some of the work on long-term hiring described in §3.2 is formulated with multiple skills in mind, but we are not aware of any work that directly addresses the problem. The state of research into the intermediate scheduling problem appears to be even less developed; we are not aware of any work on staff scheduling that explicitly addresses the multi-skilled problem we have described.

The state of research for the low-level call routing and staffing problems is somewhat more advanced. There exists work that starts to address the problems in a preliminary fashion. In the next subsections, we review this work and offer a view of future research needs and opportunities.

5.1.2 Call Routing and Staffing

A standard approach for determining effective routing policies in Markovian queueing systems is via DP. The system state represents the servers' profiles (what types of calls are currently in service by which server skill) and the queue profile (how many of each type of call is in queue). System controls are rules for routing waiting calls to idle servers at event epochs, namely the times at which a call arrives to the system or completes service. Effective policies satisfy service-level constraints with fewer (rather than more) CSRs, perhaps also minimizing 1-800 delay costs. In fact, effective routing policies are likely to be dynamic in that call-routing decisions critically depend on the current system state.

The identification of effective routing policies through DP is often impractical, however. For a large call center with many types of calls, the dimensionality of the state space is large, making the derivation of structural properties of effective policies difficult, if not impossible. Similarly, the size of the state space explodes, so that the application of standard DP techniques to numerically find effective controls also becomes infeasible. Rather than tackling the realistic problem, as is, research to date has attempted to reduce complexity in roughly three ways: topology simplification, control simplification and asymptotic analysis.

Topology Simplification: The first means of reduction is to consider simple special network topologies, such as those shown in Figure 16. These configurations represent building blocks for more complex systems. For example, in a "V" design a single pool of agents handles two (or more) types of calls. In a "W" design, two pools of agents cater to three types of calls: pool 1 serves types 1 and 2; and pool 2 serves types 2 and 3.

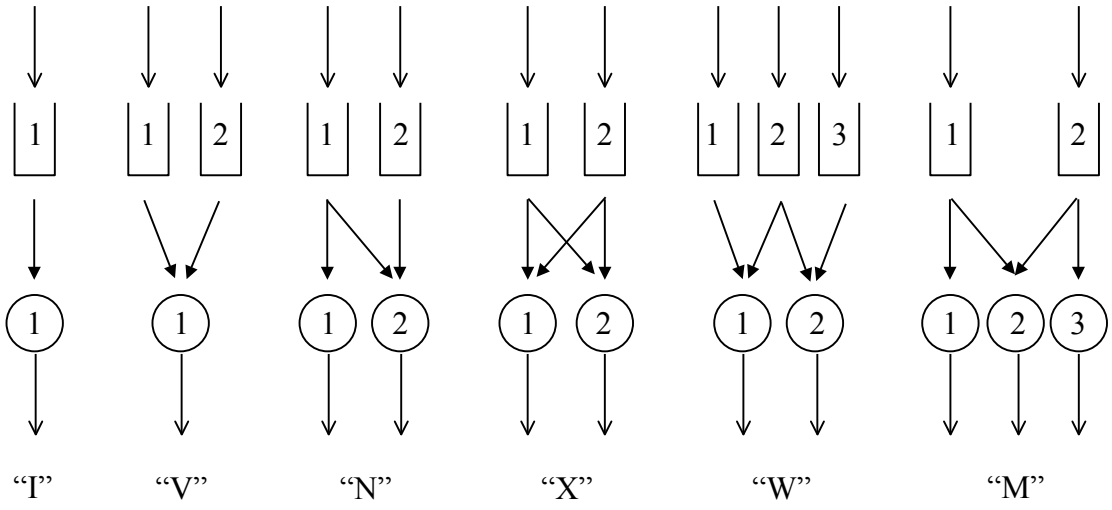


Figure 16: Some Canonical Designs for Skills-Based Routing (from [58])

The "X" design, in which two types of calls can be served by either of two pools of agents, represents full flexibility. It also reflects the fact that skill groups may be defined on a relative, rather than absolute, basis. For example, an X-design arises when CSR pool 1 is assigned call type 1 as a "primary skill," CSR pool 2 is assigned call type 2 as primary, and both pools have the other type of call assigned as secondary. A pool takes "secondary skill" calls only when deemed necessary: say, only if it has idle CSRs *and* the other pool is congested. In this case, skills-

based routing captures the fact that different call-type-to-pool assignments have differing (perhaps implicit) costs or rewards.

It is also important to note that the same network topology can be used quite differently, given various levels of traffic and routing schemes. For example, an "N" design can be used when type-1 customers are VIP but there are not enough specialized pool-1 CSRs to serve them. In this case, pool-2 CSRs can contribute to maintaining an adequate service level for type 1's. Conversely, the same N design can be used when type-2 customers are VIP and pool-2 capacity is in excess. Here, acceptable resource efficiency can be maintained by routing type-1 calls to idle pool-2 CSRs.

Garnett and Mandelbaum [58] is an introductory teaching note that lays out the canonical structures shown in Figure 16. The paper also uses simulation to demonstrate how various routing policies can effect dramatic differences in system performance.

Perry and Nilsson [118] consider a V-design in which two classes of calls are served by a single pool of crossed-trained agents. The model represents a single group of operators that provides directory assistance, as well as serving toll/assist calls. Bhulai and Koole [21] and Gans and Zhou [57] also model variants of a V-design. Stanford and Grassmann [134], who model a call center with monolingual and bilingual CSRs, as well as Shumsky [131], analyze an N-design.

Control Simplification: A second method of simplifying skills-based routing is to consider more easily characterizable, heuristic methods of capacity sizing and routing control. Stanford and Grassmann [134] and Shumsky [131] both consider fixed, static priority policies: the former use matrix-geometric methods for performance analysis and staffing; and the latter an approximate analysis. Perry and Nilsson [118] use a scheme, first analyzed by Kleinrock [86, 87], that assigns to each waiting customer an *aging factor* that grows proportionally to its waiting time. The call selection problem is solved by serving the call with the greatest attained age. The results are used to determine both the number of agents and the aging factors needed to yield specified expected waiting times.

Borst and Seri [27] apply more complex heuristics for both the call-routing and the staffing-level problems. For a fixed set of agents, they propose a dynamic scheme in which the number of calls of each class that actually has been served is compared to the number that, nominally, should have been served under a long-run average allocation scheme. The farther "behind" the actual number of services, the higher the resulting priority. (Alternatively, in a call center that has service-level agreements with various clients, the scheme penalizes those who 'misbehave', namely those who request more service than what had been agreed upon.) The paper also determines bounds for the number of agents required to offer a given level of service: it applies the square-root staffing principle to identify a lower bound, and it uses results concerning the achievable performance of multi-server systems (Federgruen and Groenevelt [50]) to produce an upper bound.

Finally, consider the following protocol, which is prevalent in practice. Agents are first divided into pools, such that all agents within a pool can serve the same types of calls. Call types are assigned fixed priorities and, when an agent becomes available, the call-selection problem is solved by assigning the highest priority call that the agent is qualified to handle. Similarly, for each type of call, there exists an ordered list of qualified agent pools, and the agent-selection problem is solved by assigning arriving calls to the first pool in the list that has an agent available.

We note that, among the criteria used for ranking pools, there exists a notion of dominance. Specifically, suppose the call types that one pool handles are a superset of the those that are handled by another pool. Then the first pool dominates the second pool. In turn, suppose an arriving call

can be served by either of these two pools, and it finds both with idle CSRs. Then the call should *not* be routed to the dominant pool. Rather, it should be routed to the dominated pool, thereby reserving the more flexible CSRs.

To the best of our knowledge, there does not yet exist a general analysis of these fixed-priority routing schemes. Nevertheless, there are two sets of results that bear mentioning. First, for the W design Stolyar [136] has shown that, for any static priority scheme, there exist conditions for which: 1) the static scheme is unstable; 2) yet another routing scheme makes the system stable. Second, the static method of agent selection used for arriving calls makes them "overflow" from one pool of CSRs to the next. Such overflow problems are notoriously hard to analyze, because the inter-overflow process is not Poisson. An approximate analysis of performance of this type of overflow behavior is performed in Koole and Talim [91].

Asymptotic analysis: The third approach for simplifying skills-based routing is asymptotic analysis. The asymptotic regime is heavy traffic, and two such regimes have been considered. The first is the efficiency-driven regime of conventional heavy-traffic, originally proposed by Kingman [85]. The second is the QED regime. In the next two subsections, we describe research that analyzes skills-based routing in each of the two regimes.

5.1.3 Skills-Based Routing in the Efficiency-Driven Regime

For an efficiency-driven operation, one lets the agents' utilization approach 100% in a way that, in the limit, *all* customers are delayed in queue. (The agent selection problem then becomes irrelevant.) As one takes limits, the number of agents either remains fixed, in which case the backlog of waiting calls grows without bound, or it is allowed to increase while controlling ASA, but at a rate slow enough so that the fraction delayed still approaches 100%.

In these conventional heavy-traffic conditions, the results of Gans and van Ryzin [55] imply that, even though there may be many ways in which arriving calls can be assigned to various pools of CSRs, one need only consider a small number of possible assignments when minimizing the system backlog. Harrison and Lopez [72] further characterize the nature of type-skill matchings (or minimal skill-overlaps) that make such small sets of assignments most efficient. In particular, they identify that, in heavy traffic, efficient sets of assignments enable *complete resource pooling* (CRP), a condition in which the set of CSRs act as a (pooled) single, virtual "super" server. Note that all of the designs of Figure 16, except for "X", satisfy this CRP condition; eliminating any of the four arrows in "X" would do as well.

This pooling condition is the corner-stone for analysis of efficiency-driven operations, and it seems likely to be relevant in the QED regime. Section 5 of Stolyar [135], which characterizes optimal policies for a more general *resource pooling* condition, compactly lays out the relevant literature. Section 3 and the beginning of Section 5 in Williams [149] have the full story.

Harrison and Lopez [72] is the first skills-based-routing model that resembles the reality of a call center. In [55] both the model, which lists only call-center-wide processing rates, and the measure of system congestion, the minimum time required to work off the entire backlog of waiting calls, are aggregate. In contrast, in [72] the assignment of individual calls to CSRs is explicitly modelled, and occupancy costs are defined as growing linearly with the backlog of each type of call. Both [55, 72] consider discrete-review policies that process sets of calls in large batches, however. This class of policies is reasonable for emails, for example, but it is clearly inappropriate for inbound

calls.

In contrast, for the N-design of Figure 16, Bell and Williams [20] prove the asymptotic optimality of threshold controls. More specifically, they assume linear occupancy costs and that type-1 customers are VIP. They then establish that, whenever the length of the type-1 queue exceeds a critical threshold, type-1 calls should get priority over type-2 calls at CSR pool 2. Williams [149] conjectures that dynamic, threshold-based policies are also asymptotically optimal for the model in [72]. It is important to note, however, that the calculation of the conjectured thresholds requires prior processing (the solution of linear programs) that intimately depends on model parameters and topology.

An alternative to thresholds is provided by index controls: each queue is assigned an index, that depends only that queue's state; the queue chosen for service is then the one with the highest index. A striking example is van Mieghem's [141] analysis of the V-design with a single-server, which proves the asymptotic optimality of a simple Generalized cμ (Gcμ) rule for waiting costs that are convex increasing. By equipping each agent with its own index for call selection, Mandelbaum and Stolyar [109] verify that these Gcμ rules remain asymptotically optimal in the context of skills-based routing.

To elaborate, consider a general skills-based design in which type-i calls are served by pool-j agents at rate μ_{ij}. (Here μ_{ij} is the reciprocal of an average service time, and $\mu_{ij} = 0$ if j's cannot serve i's). Delay costs are quantified in terms of type-dependent increasing convex functions: $C_i(w)$ is the cost incurred by an i customer that waits in queue w units of time, before being served. Then each server j that becomes idle at time t adheres to the following Gcμ rule: choose to serve the longest-waiting i^* customer for which

$$i^* \in \arg\max_i C_i'(W_i(t))\,\mu_{ij}. \tag{22}$$

Here C_i' is the derivative of C_i, and $W_i(t)$ is the longest waiting time (that of the head-of-the-line customer) in queue i at time t. In [109] it is proved that, under complete resource pooling (as in [72, 149]), and for costs with $C_i(0) = C_i'(0) = 0$, the above parsimonious Gcμ rule is asymptotically optimal in heavy traffic. Qualitatively speaking, the result demonstrates that an exceedingly simple call-selection index performs well for system that are efficiency-driven – even within complex routing designs. (In these circumstances, agent-selection arises infrequently enough to be handled arbitrarily.)

We note that quadratic costs recover the aging factor of [86, 87, 118] that is introduced in the previous subsection. The assumption $C_i'(0) = 0$ rules out linear costs, but it is conjectured in [109] that these can be accommodated by carefully choosing aging factors that vary with system parameters.

More importantly, the natures of threshold and GCμ controls differ fundamentally. The former require careful prior calculations (of thresholds) and management by exception: type-2 calls get priority at CSR pool 2 until the number of waiting type-1 calls "crosses" an "emergency" boundary. In contrast, GCμ rules are simple and robust (surprisingly enough, they do not depend even on arrival rates), but they are based on a continuous reevaluation of the state-dependent indices.

To summarize, conventional heavy traffic analysis had yielded strikingly simple classes of policies that should perform well in efficiency-driven environments. This regime is appropriate for slower-turnaround work, such as emails or faxes, that may be processed after some delay. It is not appropriate for work that must be performed in the quality or QED regimes, however.

5.1.4 Skills-Based Routing in the QED Regime

With the exception of Borst and Seri's [27] heuristic use of square-root laws, research to date on skills-based routing in the QED regime has been limited to variants of the V-design. Given a V-design, the QED regime is straightforward to characterize as simply maintaining square-root safety staffing. Formally, let λ_i and μ_i denote the arrival and service rates for type i customers, respectively. The offered load is then $R = \sum_i \frac{\lambda_i}{\mu_i}$, and square-root safety staffing, as before, prescribes a number of agents $N \approx R + \beta\sqrt{R}$. The service grade β is arbitrary if there is abandonment and positive otherwise.

Two papers [14, 73] analyze general V-designs in the QED regime (that it, under square-root safety staffing). Being inspired by call centers, both allow for multiple types of impatient customers.

Harrison and Zeevi [73] assume a Markovian model with preemption, work-conservation (no idling when there are customers waiting), and linear costs that are discounted over an infinite horizon. They identify with the QED limit a diffusion control problem which they solve. The solution yields control policies for the original problem that are conjectured to be asymptotically optimal. A two-type example is then solved numerically to concretize the results.

In Atar et al. [14], the cost per unit of time can be a non-linear function of the number of customers waiting to be served in each class, the number actually being served, the abandonment rate, the delay experienced by customers, the number of idling servers, as well as certain combinations thereof. Service times and impatience clocks are exponentially distributed, but interarrival times are renewals. In the QED limit, the queueing scheduling problem converges to a diffusion control problem. Its solution is used to construct controls that are then proved asymptotically optimal for the original queueing system.

The analysis yields both qualitative and quantitative insights. For example, it implies that, for natural cost structures, the advantage of preemption is negligible in the QED regime. Similarly the benefit of violating work-conservation appears to be negligible. Thus, when solving the "call-selection" problem in the QED regime, the ability to idle CSRs need not be considered.

Finally, Armony and Maglaras [12, 13] consider a more specific V-design in which a single pool of CSRs serves two classes of customers: one that opts to queue for service, and another that elects to be called back by the call center at a later time. (The two classes endogenously arise from a model of choice behavior.) In [12] arriving customers have static, equilibrium estimates of the distribution of delay for immediate service, and in [13] customers are given a state-dependent estimate of delay. In both cases, the authors demonstrate that a threshold control policy – that gives priority to waiting customers if and only if the queue of "call-back" customers falls below a fixed level – asymptotically minimizes the expected delay for immediate service (among all non-idling policies). Furthermore, in [13] they demonstrate that systems that use dynamic estimates of delay outperform those that relay on static estimates. Both papers also provide staffing guidelines that follow from square-root laws.

It is worth noting that the asymptotic analysis of [14, 73] has so far shed little 'qualitative light' on the structure of optimal controls for skills-based routing in the QED regime. This differs fundamentally from the threshold controls of [12, 13] and the index controls of the efficiency-driven regime [109, 141]. QED complexity stems from the absence of complete resource pooling, and the fact that the agent selection problem plays an important role – indeed, the QED characteristic is that a significant fraction of the customers find idle servers upon arrival.

In summary, this is a newly-emerging and important area of research. There is much to achieve, both on the staffing and control fronts.

5.2 Call Blending and Multi-Media

The integration of telephony and data-processing infrastructure has allowed call centers to expand their range and provide additional services. These so-called "contact centers" have the ability to handle a wide range of media. Common examples include email and electronic faxes. Other, less common examples include *callbacks*, in which customers signal – via a company's web site or IVR – that they wish to be called back by the center, and *chat*, in which agents communicate in (more or less) real time over the internet with customers, using text.

In fact, the latest generation of systems has the potential to route any type of electronically mediated work. For example, consider a claims-processing center for a large U.S. based health insurance company. The information system with which CSRs interact, as they handle customer requests, is the company's claims-processing and adjudication system. The telephone and claims systems are integrated via CTI, and the company has the ability to route both phone calls and "screens" of claims-processing work to idle CSRs.

In one sense, multimedia may be thought of as an example of skills-based routing. The various types of work – call, email, claims-processing screen – parallel various call types. Each agent's skills define the types of media s/he is capable of handling.

At other levels, however, differences among media are deeper than differences among calls: one important difference is the natural time scales at which the various media must be responded to; another is the discipline required for service; yet another is the mental 'setup' required to switch among media.

Typically, telephone calls are to be served within seconds or minutes and, once started, should not be interrupted. Responses to email and fax requests, on the other hand, can be delayed for hours or perhaps days, and they may be preempted and resumed. Response times for chat services fall somewhere in between, and CSRs that handle chat may serve several customers at once.

Differences in timing naturally lead one to consider priority schemes in which telephone calls receive high priority and emails and faxes, low. In addition, limits in the structure of shifts and schedules often prompt the solution to the staffing problem (9) to include periods of over-capacity (i s.t. $\sum_j a_{ij}x_j > N_i$). During these intervals, agents that might otherwise be idle can become productive by handling low-priority work. Historically, the problem first arose in the context of mixing inbound and outbound calls, a process commonly known as *call blending*.

Thus, contact centers enjoy an advantage over call centers in that slower response-time traffic can be shifted between intervals (inventoried), and this can help to increase CSR utilization and reduce operating costs. At the same time, blending of traffic of various priorities, subject to medium-specific service-level constraints, creates problems that are akin to those of skills-based routing. Bhulai and Koole [21] and Gans and Zhou [57] address the lower-level routing problem: given a fixed set of CSRs, how to maximize the throughput of low-priority traffic, for example email, subject to a service-level constraint on incoming calls. They demonstrate that variants of threshold reservation policies, which reserve agents for inbound calls, are effective. Brandt and Brandt [29] assume the use of these threshold policies and analyze analogous systems with customers that have generally distributed patience. To the best of our knowledge, extensions of higher-level scheduling

and hiring problems have not been tackled in this context.

We are aware of three papers that model the use of IVRs and callbacks. Brandt, Brandt, Spahl, and Weber [30] consider a call center with a finite number of lines, customers with exponential patience and, prior to waiting, an IVR message of constant-duration. The system is modelled as a two-dimensional network, and approximations to steady state performance measures are derived. Brandt and Brandt [28] analyzes a birth-and-death model of a system in which callers who have waited beyond a given threshold are transferred to a callback queue. The callback queue is served only when there are no "live" callers waiting and the number of idle agents exceeds some threshold. This again gives rise to approximate measures of performance of a two-dimensional network. Finally, Armony and Maglaras [12, 13] model a system in which customers who are not immediately served may choose to wait in queue for service, to wait for a later callback, or to balk if they deem both waits too long.

Finally, Mandelbaum, Massey, and Reiman [99] develop a framework for the analysis of Markovian service networks. In these systems, multiple types of customers are served according to preemptive-resume priority disciplines. The primitives include time-varying abandonment and retrial intensities, and the asymptotics are in the QED regime. This framework is thus applicable for performance analysis of large multi-media call centers. While the framework does not yet accommodate non-preemptive priority disciplines or finite buffers (busy-signals), both features appear to be within theoretical grasp: given the results of Atar et al. [14], the gains from preemption promise to be negligible in the QED regime; and current work by Massey and Wallace [111] addresses finite-buffer systems.

5.3 Networking

Telephone networking technology allows companies to link geographically dispersed call centers, and through careful management their performance can approach that of a single "virtual" call center. In this manner companies can better exploit potential economies of scale. For example, this is the case in Figure 11, the header of which reads "Command Center Intraday Report." Here, load balancing is exercised from a single Command Center that overseas the 12 call centers represented in the table.

There currently exist several methods for networking call centers, the main variants of which are summarized in Figure 17. There exist true *network ACD* systems that have the ability to hold calls centrally and route individual calls to call centers as agents become free. There also exist less elaborate *load balancing* schemes in which calls are routed from a central switch to call centers with traditional ACDs. Here, calls queue locally, at the individual centers. In *overflow systems*, calls that are queued at one call center may be laterally transferred to another, and combinations of the load-balancing and overflow schemes may also be used.

In addition to these main variants there may be many other hybrids. One with which we are familiar allows calls to queue simultaneously at more than one center. In this so-called "interflow" scheme, each call first arrives to an individual center. If it is predicted to be served within a fixed time limit, say 15 seconds, then the call queues only at that center. If the expected delay is greater than the threshold, however, then the call simultaneously queues at the original site, as well as any other site whose expected delay falls below the expected-delay threshold (possibly none).

Figure 18 shows one hour's worth of data from a network of four call centers (nodes) that use

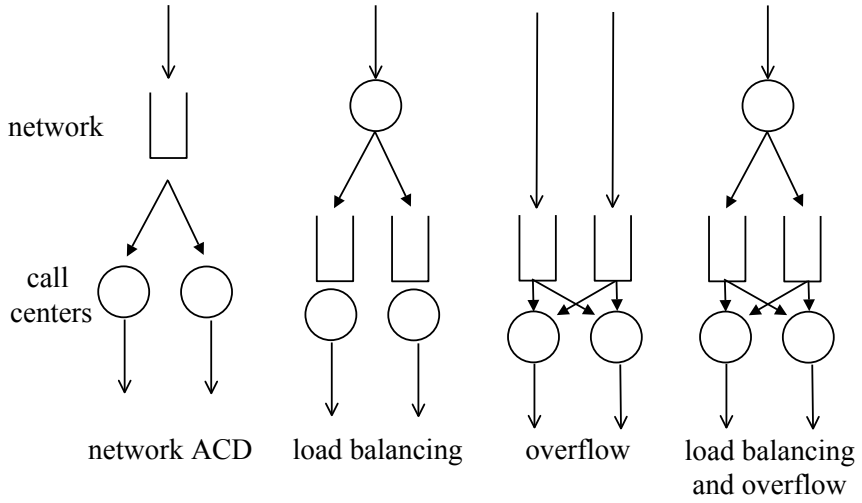


Figure 17: Common Methods of Networking Call Centers

this interflow scheme. Numbers with asterisks indicate the volumes of calls arriving to each of the nodes, as well as each call's final disposition: served or abandoned. Numbers on the arrows between nodes indicate the numbers of calls that were served at a location that differed from the originating node. From the figure, one sees that nodes 2 and 4 were not interconnected. One also sees that of the 2,092 calls arriving to node 1, for example, less than 1.4% abandoned, and almost 34% were ultimately served at locations 2 and 3.

While there is a fairly extensive literature on load balancing, little of it appears to be directly applicable to these systems. Servi and Humair [129] analyze the problem of setting static routing probabilities in load balancing systems. More can be gained if routing is dynamic, and Kogan et al. [88] compare two basic strategies for the dynamic routing of calls: the first is a network-ACD that queues calls centrally and routes calls to individual CSRs as they become available at various sites; and the second is a dynamic load-balancing scheme that immediately routes an arriving call to the site with the least expected delay, at which point the call queues. The paper demonstrates numerically that, to the extent that a FCFS system imposes a delay in switching calls to centers, it may be inferior to the less elaborate dynamic load-balancing method. Numerical tests in Borst et al. [26] show that similar dynamic load-balancing schemes perform well, so that "a multiple site system approaches quite closely the performance of a single virtual facility" [27].

Thus, the growing trend of networking call centers has barely been investigated. In particular, both dynamic load balancing and overflow protocols can – in theory – lead the arrival processes at individual call centers to not be Poisson. Questions of how the assumptions are violated, by how much, and how this impacts capacity management have yet to be addressed systematically. Because the effect of complex networking protocols are difficult to predict, empirical analysis that captures actual behavior would help to deepen academics' and managers' understanding of the benefits and drawbacks of the various networking schemes.

Readers who are primarily interested in problems associated with queueing performance or control can now proceed to Section 7.

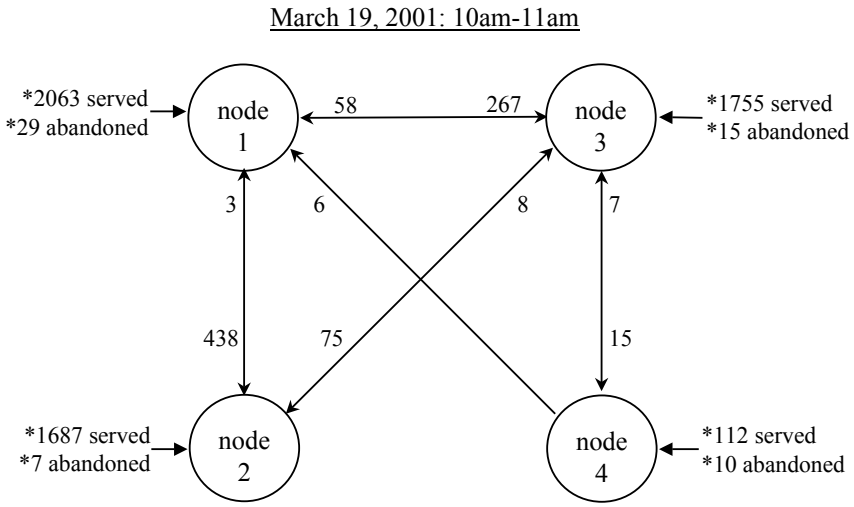


Figure 18: Example Flow of Calls within a Network of Four Call Centers (from [34])

6 Data Analysis and Forecasting

The modelling and control of call centers must necessarily start with careful data analysis. For example, the simple Erlang C queueing model described in Section 3 requires the estimation of a calling rate (λ_i) and a mean service time (μ_i^{-1}) for each half-hour interval. Moreover, as Figure 3 and the accompanying discussion in Section 4.2 indicate, the performance of call centers in peak hours can be extremely sensitive to changes in these underlying parameters.

It follows that accurate estimation and forecasting of parameters are prerequisites for a consistent service level and an efficient operation. Furthermore, given the computer-mediated, data-intensive environment of modern call centers, one might imagine that highly developed estimation and forecasting methods would exist.

But in fact, though there is a vast literature on statistical inference and forecasting, surprisingly little has been devoted to stochastic processes, and much less to queueing models in general and call centers in particular. For example, Section II in [97] lists only 17 papers on the statistics and forecasting of call center data. Indeed, the practice of statistics and time series analysis is still in its infancy in the world of call centers, and serious research efforts are required to bring it up to par with prevalent needs.

The scarcity of statistical research of call centers renders this part of our survey more speculative. In §6.1–6.2 we describe general categories of call-center data, as well as various statistical approaches that are useful for their analysis. Then in §6.3 we review data analysis and statistical research that is related to call-center operations: the analysis of system primitives, such as arrival rates, service times, and abandonment behavior; and that of system performance measures, such as waiting times and aggregate customer abandonment rates. We conclude the section in §6.4 by stating our views concerning data analysis and forecasting work that needs to be done.

6.1 Types of Call Center Data

Recall from §2.2 the description of how a call is handled. This process generates a great deal of data, which we divide into four categories: operational, marketing, human resources, and psychological.

Operational data reflect the physical process by which calls are handled. These data are typically collected by pieces of the telephone infrastructure such as IVRs and ACDs. They can be usefully organized in two, complementary fashions.

Operational *customer* data provide listings of every call handled by a site or network of call centers. Each record includes time-stamps for when the call arrived, when it entered service or abandoned, when it ended service, as well as other identifiers, such as who was the CSR and at which location the call was served.

Operational *agent* data provide a moment-by-moment history of the time each logged-in agent spent in various system states: available to take calls, handling a call, performing wrap-up work, and assorted unavailable states. These data allow one to deduce the numbers of agents working at any time. Often these records include identifiers of the calls being served and (with difficulty) can be matched to the operational customer data described above, for joint analysis.

Marketing or *business* data are gathered by a company's corporate information system. They may include records of the transactions that took place over the customer's entire history with the company, through call centers as well as through other channels. They may also capture information concerning the customer's current status at the business.

In theory, operational and marketing data can be seamlessly integrated via CTI software, which connects the telephone infrastructure with a company's customer databases. That is, given the existence of CTI, one might expect companies to record and analyze a full view of what happens to each call as it enters the system: marketing data concerning what happened during the service, together with operational data concerning how and when the service happened. In practice, however, the use of CTI appears, thus far, to be limited to facilitating the service process through "screen pops" which save CSRs time, not to the joint reporting of call data. Incompatibility between data storage schemes of (older) ACD and (newer) CTI systems may be the problem that prevents this integration from taking place.

Human resources data record the history and profile of agents. Typical data include information concerning employees' tenure at the company, what training they have received and when, and what types of call they are capable of handling. With one frequent exception, these data generally reside within the records of a company's human resources department. The exception is that of "skills" data, which define the types of calls that agents can handle. This information is needed, by the ACD (or those that manage it), to support skills-based routing.

Finally, *psychological* data are collected from surveys of customers, agents or managers. They record subjective perceptions of the service level and working environment.

Two additional sources of data are important to acknowledge. First, some companies record individual calls for legal needs (e.g., brokerage and insurance businesses) or training reasons. While potentially useful, we are not aware of any simple machinery that can extract these data for analysis (say, into a spreadsheet). A second source is subjective surveys in which call center managers report statistics that summarize their operations. These surveys can include both operational and marketing data, such as arrival and utilization rates, average handle times, and the average

dollar value of a transaction. While they may facilitate rough benchmarking (see, for example, www.benchmarkportal.com), these data should be handled with care. By their nature, they are biased and should *not* serve as a substitute, or even a proxy, for the operational and marketing data discussed above.

6.2 Types of Data Analysis

As in any statistical work, the analysis of call-center data can take a number of forms. We briefly make two sets of distinctions.

Descriptive, Explanatory and Theoretical Analysis: We first distinguish among descriptive, explanatory, and theoretical analysis. Each mode is important, and we briefly describe the three in turn.[1]

Descriptive models organize and summarize the data being analyzed. The simplest of these are tables or histograms of parameters and performance. An example is a histogram of service duration by service type, or of customers' patience by customer type, or of waiting times for those ultimately served.

These can be contrasted with *theoretical models* that seek to test whether or not the phenomenon being observed conforms to various mathematical or statistical theories. Examples include the identification of an arrival process as a Poisson process or of service durations as being exponentially distributed.

In between descriptive and theoretical models fall *explanatory models*. These are often created in the context of regression and time series analysis. Explanatory models go beyond, say, histograms by identifying and capturing relationships in terms of explanatory variables. For example, average service times of calls may be systematically higher from 11am to 3pm and lower at other periods. At the same time, these models fall short of theoretical models in that there is no attempt to develop or test a formal, mathematical theory to explain the relationships.

Queueing models constitute theoretical models which mathematically define relationships among building blocks, for example arrivals and services, which we refer to here as *primitives*. Queueing analysis of a given model starts with assumptions concerning its primitives and culminates in properties of performance measures, such as the distribution of delay in queue or the abandonment rate. 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.

For example, theoretical analysis of the G/G/N queue gives rise to Kingman's law of congestion: in heavy traffic, the waiting time of delayed customers is close to being exponentially distributed with expected delay defined as in (11). Empirical analysis of call centers operating in heavy traffic can then validate or refute Kingman's law, as in [33] (see Figure 6).

Estimation versus Prediction: We also distinguish between two closely related, but different, statistical tasks: estimation and prediction. *Estimation* concerns the use of existing (historical) data to make inferences about the parameter values of a statistical model. *Prediction* concerns the use of the estimated parameters to forecast the behavior of a sample outside of the original data

[1] Parts of the present section are adapted from the report Mandelbaum, Sakov and Zeltyn [106]. This is an in-depth empirical analysis, mostly descriptive, of call center operational data, gathered at a small Israeli bank over the twelve months of 1999. Both the report and the data are downloadable from ie.technion.ac.il/serveng/.

set (used to make the estimate). Predictions are "noisier" than estimates, because, in addition to uncertainty concerning the estimated parameters, they contain additional sources of potential errors.

As an example, consider a simple model in which the arrival rate to a call center (each day from 9:00am–9:30am) is a linear function of the number of customers receiving a promotional mailing. That is

$$\lambda_i = \alpha + \beta x_i + \varepsilon_i \,, \tag{23}$$

where λ_i is the arrival rate, x_i is the number of mailings, α and β are unknown constants, and the ε_i are *i.i.d.* normally-distributed noise terms with mean zero. Given n sample points (x_i, λ_i), one may use regression techniques (such as least squares) to produce parameter estimates $\hat{\alpha}$ and $\hat{\beta}$. There is uncertainty, however, regarding how closely these estimates match the true α and β. That is $\hat{\alpha}$ and $\hat{\beta}$ are random variables that are functions of the n *i.i.d.* samples, and given our estimated function

$$\hat{\lambda}_i = \hat{\alpha} + \hat{\beta} x_i \,, \tag{24}$$

the associated estimation error is distributed as

$$\lambda_i - \hat{\lambda}_i \;=\; (\alpha - \hat{\alpha}) + (\beta - \hat{\beta}) x_i \,.$$

Now suppose we are told the number of mailings that customers will receive on day $n+1$, and we are asked to predict what λ_{n+1} will be. Then we use $\hat{\lambda}_{n+1}$ to predict the $(n+1)$st arrival rate, and from (23)–(24) we see that the prediction error is distributed as

$$\lambda_{n+1} - \hat{\lambda}_{n+1} \;=\; (\alpha - \hat{\alpha}) + (\beta - \hat{\beta}) x_{n+1} + \varepsilon_{n+1} \,.$$

In particular, the ε_{n+1} term makes the prediction error larger than the estimation error that arises from the use of $\hat{\alpha}$ and $\hat{\beta}$. (For more on estimation and prediction see, for example, Bickel and Doksum [22], as well as [33].)

6.3 Models for Operational Parameters

First we review work devoted to primitives: arrivals, service times, abandonment (patience) and retrials. Then we address the validation of performance measures.

6.3.1 Call Arrivals

The arrival process records the epochs at which calls arrive to the center. Among the queueing primitives of call centers, it has been studied most extensively. For example, consider the following three models for the arrival process of calls to a call center, during a given day.

Descriptive models, such as histograms of interarrival times, reflect short-term patterns of randomness in arrivals. Conversely, more aggregate descriptions of the arrival process, such as those developed in Mandelbaum, Sakov and Zeltyn [106], may average out stochastic variability over several (similar) days to develop deterministic, fluid-like models that reflect predictable variability in the calling rate. For example, the lower-right panel of Figure 5 reveals the stochastic nature of arrivals in the short term, and the remaining panels highlight predictable variability.

Explanatory models can be used to forecast future arrival rates of calls to a call center, a crucial first step in the process of scheduling personnel. Arrival rates depend on many factors – day of week or month, time of day, holidays, etc. – and statistical techniques using explanatory variables can be used to represent them. For example, Andrews and Cunningham [6] develop autoregressive integrated moving average (ARIMA) models that estimate the number of daily calls for orders (buying merchandise) and inquiries (e.g. checking order status) at L.L. Bean. In addition to day of week, covariates include the presence of holidays, catalogue mailings, as well as forecasts for orders that are independently produced by the company's marketing department.

Classical theoretical models posit that arrivals form a Poisson process. It is well known that such a process results from the following behavior: there exist 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, so that the average number of calls arriving within a minute is moderate; and callers decide whether or not to call independently of each other. When the average numbers of arrivals change over the time of day then one obtains a time-inhomogeneous Poisson process.

Suppose the same "seasonal" cycle of arrival rates – such as the daily pattern shown in the lower-left panel of Figure 5 – repeats itself. Then a common method of estimating the arrival rate is to break the cycle into smaller intervals, collect several cycles' worth of samples for each subinterval, and use each sample mean as an estimate of that subinterval's arrival rate. Henderson [74] performs an asymptotic analysis of this scheme, one in which the length of the subinterval (appropriately) shrinks as the number of cycles' worth of data grows. The analysis shows that, when the underlying arrival process is time-inhomogeneous Poisson and cyclical, the limiting sample rate function is a consistent estimator of the original arrival-rate function.

Massey et al. [112] approximate a general time-inhomogeneous Poisson process as one with a piecewise linear rate function. That is, over intervals $(0, T_1], \ldots, (T_{i-1}, T_i], \ldots, (T_{n-1}, T_n]$, the arrival process is Poisson with rate $\lambda_i(t) = a_i + b_i \cdot t$. Using simulated arrivals, they compare ordinary least squares (OLS), weighted least squares (WLS), and maximum likelihood (ML) methods of estimating (a_i, b_i). They find that, given an arrival rate that is a linear Poisson process, all three methods perform well, as long as $a_i > 0$. For $a_i \approx 0$, however, WLS and ML (which are asymptotically equivalent) perform better. They also offer standard tests for validating the linear Poisson model.

Of course, forecasts of arrival rates are not exact. Call centers may not always have sufficient historical data upon which reliable estimates can be based. Furthermore, unpredictable factors, such as weather conditions, also make future arrival rates uncertain. As Section 4.4 describes, uncertainty in the arrival rates can be dangerous to ignore, particulary for highly utilized call centers operating in the QED regime. Therefore, it is desirable to develop distributional (rather than point) forecasts.

Jongbloed and Koole [84] analyze the numbers of arrivals to a Dutch call center by time of day. For each time interval (e.g. 10:00–10:30am) they collect several days' worth of data, and they show that the data do not appear to be *i.i.d.* samples of Poisson random variables: while the mean and variance of a Poisson distribution are the same, the samples' variances are much larger than their means. Thus, given only time-of-day information, the process turns out to be doubly Poisson: the rate itself is random.

The paper [84] goes on to develop parametric (Gamma-Poisson) and nonparametric (ML) methods of estimating a distribution for the (unknown) arrival rate of a Poisson processes. (The Gamma distribution is a natural prior for the arrival rate, since it is a conjugate prior for the Poisson dis-

tribution.) Gordon and Fowler [60] also address this problem, but in less detail.

Brown et al. [33] analyze the arrival process of calls to a relatively small call center in Israel. They confirm that arrivals (by call type) do correspond to a Poisson process in which the arrival intensity varies. As in [84], the arrival process appears to be doubly Poisson, however. Indeed, when numbers of arrivals are stratified by time of day and day of week, the samples for specific time-day pairs (e.g. Monday from 10-11am) do not appear to be *i.i.d.* samples of Poisson random variables.

The paper [33] also develops a method for generating prediction (as opposed to estimation) confidence intervals of a non-stationary arrival rate. The prediction model includes an autoregressive term – each day's aggregate arrival rate is conditioned on that of the previous period – and the introduction of the autoregressive element noticeably reduces prediction error. Thus, the arrival rates are (positively) serially correlated

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

6.3.2 Service Duration

The service duration is typically defined as the time an agent spends handling a call. This may include time speaking with the customer, time during which the customer is "on hold" and the agent is processing the customer's request, as well as time after the caller hangs up that the agent continues processing the request.

Work on the service phase has almost exclusively concentrated on description and validation of theoretical models for service duration. Thus, there exists published work that presents data analysis, such as histograms, as well as tests for "goodness of fit" with certain parametric families of distributions, but little in the way of explanatory work.

Furthermore, there does not exist an extensive theory for the distributional form that service durations should follow. Mixtures of Erlang distributions are, in fact, dense among all distributions, and this subfamily of phase-type distributions is convenient numerically. (For example, see Latouche and Ramaswami [94].) But having ample parameters, they are less convenient theoretically. To this end, lower-dimensional parametric models are desired.

The most frequently used parametric model of service is that of *exponentially* distributed durations. In practice, the main "theoretical" justification for its use has been analytical tractability, along with a lack of empirical evidence to the contrary. Nevertheless, some studies have compared empirical distributions of service durations to exponential distributions and found an acceptable fit. One example is Kort [92], which summarizes models of the Bell System Public Switched Telephone Network, developed in the 70's and 80's. Another is Harris et al. [71], which analyzes IRS call centers. Our experience has frequently confirmed these findings for human services that are

homogeneous and unpaced (not only telephone services).

In addition to the exponential distribution, two other parametric statistical families have been found to arise in applications: Erlang, or more generally Gamma, distributions and the lognormal distribution. Both families are explored in Chlebus [39], who analyzes holding time distributions in cellular communication systems. Other confirmations for the lognormal fit are provided by the service times in Bolotin [23] and in Mandelbaum, Sakov and Zeltyn [106], where the fit has been found to be truly remarkable (right panel of Figure 19). It has also shown up in an (unpublished) analysis of data from the call center of a Dutch bank. The excellent fit of lognormal arises not only for overall service times. It is also found when service times are stratified by service types, by individual agents, and so on. Furthermore, Bolotin [23] suggests a theoretical (psychological) argument for the appropriateness of the lognormal form of service durations.

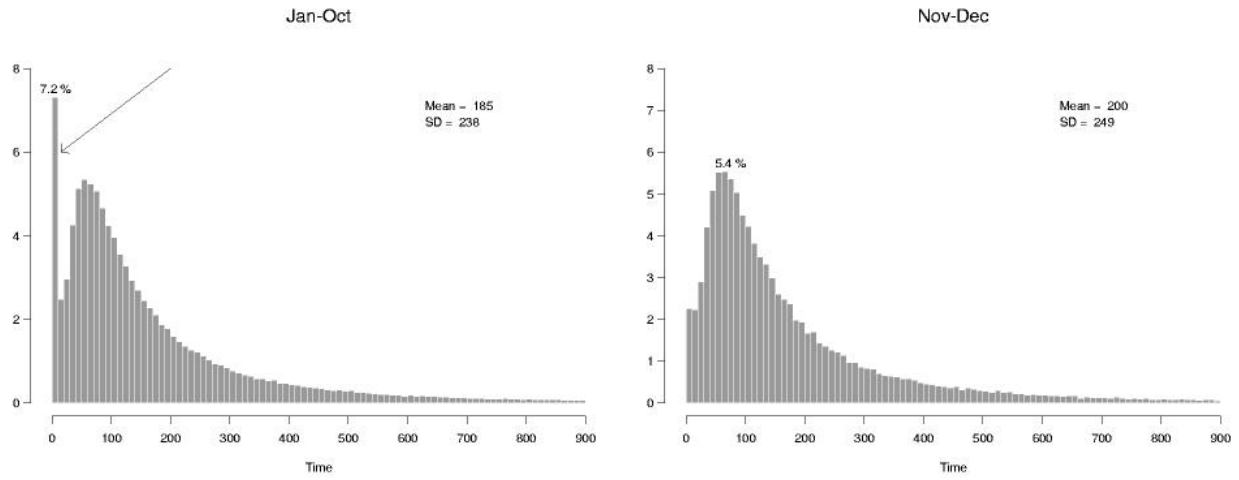


Figure 19: Service-Time Distributions from a Call Center (from [106])

A comparison of the two histograms in Figure 19 is worth making. Observe that the left panel, which is the empirical distribution of call times from January to October of 1999, has a spike near 0 seconds (the origin): about 7% of the calls were shorter that 10 seconds. These very short calls were due to certain agents who were taking small "rest breaks" by hanging up on customers. In contrast, the right panel, which reflects only November and December data, shows no such spike. At the end of October, the problem was discovered and corrected.

The problem of agents 'abandoning' their calls arises when short service durations (or many calls per shift) are highlighted as a prime performance objective. The problem becomes immediately apparent from call-by-call data. It cannot be discovered, however, through the prevalent standard of reporting only half-hour averages.

6.3.3 Abandonment and Retrials

A final set of primitives in operational queueing models concerns the behavior of the customers who are calling to be served. From Figure 4 we recall that these include abandonment, retrials and returns. Among published papers, most attention has been given to patience and abandonment, with little devoted to the tendency to retry, namely redial, and even less to returns.

The Impatience Function: A basic description of patience is the cumulative distribution function, say F, of the time beyond which a customer would not be willing to wait. An equivalent description is its corresponding hazard rate function. Assuming that F has a density $f = F'$, its hazard rate function is given by $h(t) = f(t)/(1 - F(t))$, $t \geq 0$. Intuitively, $h(t)dt$ is the probability that a customer, who has already survived waiting for t units of time, will abandon within the next dt units, namely during $(t, t + dt]$. Thus, the hazard rate $h(t)$ provides a natural dynamic depiction of (im)patience, as it evolves while waiting. We will refer to this as the *impatience function*, or simply *impatience*.

Figure 20 plots two impatience functions. It has appeared in [33, 106], and it has several noteworthy features. First, note that the impatience of "priority" customers lies strictly below that of "regular" customers, so that the high priority customers emerge as (stochastically) *more* patient (less impatient) than regular customers. This could be a reflection of a more urgent need, on the part of priority customers, to speak with an agent. Second, the impatience functions of both types of customers are *not* monotone and have secondary peaks at about 60 seconds. The peak, as it happens, reflects an announcement to customers who have waited 60 seconds, informing them of their relative place in the tele-queue (but not on their anticipated waiting time). As can be seen, the information here encourages abandonment. This could be in contrast to its original goal, namely *preventing* abandonment by reducing the uncertainty about waiting times. (For a theoretical exploration of the benefits of inducing abandonment, see Whitt [144].)

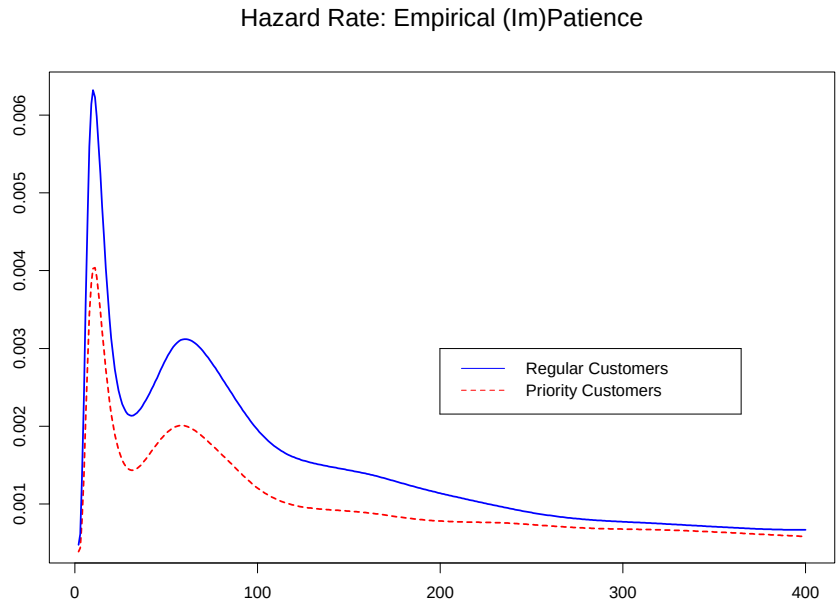


Figure 20: Impatience Functions of Regular and Priority Customers (follows [33, 106])

Models of Impatience: The first model of impatience was developed by Palm [116] in the 1940's. Palm analyzes the waiting process for a telephone connection, trying to estimate the distribution of patience. To his end, he introduces an *inconvenience* function of time $I(t)$, $t \geq 0$, the derivative of which he calls *irritation*. As a plausible form for irritation, Palm proposes

$$dI(t) = c \cdot t^{\lambda} dt, \quad t \geq 0,$$

and he closes the circle by axiomatizing that irritation is proportional to our impatience function (hazard rate). This reasoning implies that the distribution of patience (the time a customer is willing to wait for service) is *Weibull*. The special case of exponentially distributed patience, as in Erlang A, corresponds to $\lambda = 0$, which is irritation (or impatience) that is constant over time.

In calibrating λ, Palm [116] considers three circumstances, two of which have close present-day analogues. Short delays, during which a subscriber waits to be connected to a circuit, are similar to today's call-center tele-queueing delays. From direct measurements of patience, Palm concludes that, in this case, λ is close to one, possibly slightly less. Medium delays, after which an operator calls back a subscriber with a requested connection, are similar to those generated by present-day "call back" systems [12, 13]. These induce exponentially distributed impatience, with $\lambda = 0$.

Palm's findings are remarkably consistent with those reported by Kort in the 1980's. Based on abandonment behavior in laboratory testing in the U.S.A., Kort [92] proposes Weibull as the distribution of patience while waiting for a dial tone and deduces $\lambda = 1.23$. Kort also has two additional models of patience: time to abandonment while dialing, described in terms of a shifted exponential; and time to abandonment prior to network response (after dialing), where a mixture of two lognormal distributions fits well.

Roberts [124] presents descriptive models for redial and abandonment behavior, based on experimental observations in France. A parametric model for the data in Roberts [124] is derived by Baccelli and Hebuterne [17], who find that an Erlang distribution with 3 phases provides a reasonable fit.

Analogous studies of abandonment behavior in call centers exist in three related papers. The first two, by Mandelbaum, Sakov, and Zeltyn [106] and Brown et al. [33], develop descriptive models for customer patience. The third, by Zohar et al. [152], is motivated by the following linear relationship between the abandonment fraction and expected delay; it holds for patience that is exponentially distributed, say with parameter θ:

$$\mathsf{P}\{\text{Abandon}\} = \theta \cdot \mathsf{E}[\text{Wait}]. \tag{25}$$

(To verify the relationship, start with the flow conservation equation, $\lambda \cdot \mathsf{P}\{\text{Abandon}\} = \theta \cdot \mathsf{E}[\text{Queue-Length}]$, in which both sides represent the effective abandonment rate. Then substitute Little's Law, $\mathsf{E}[\text{Queue-Length}] = \lambda \cdot \mathsf{E}[\text{Wait}]$, and cancel out the λ.) Empirical support for this relationship is given in Figure 21, which is based on the same data set as Figure 20.

Notably though, the impatience in Figure 20 is hardly exponentially distributed, a prerequisite for (25) to hold. This fact suggests that the linear relationship in Figure 21 is to be expected under broader circumstances. And indeed, suppose that the distribution of patience has a density function whose value at the origin, say $r(0)$, is positive. Mandelbaum and Schwartz [110] show that, in the quality-driven and QED regimes,

$$\mathsf{P}\{\text{Abandon}\} \approx r(0) \cdot \mathsf{E}[\text{Wait}], \tag{26}$$

which is an asymptotic generalization of (25). In fact, the distribution of patience becomes critical only at its origin because, in the quality-driven and QED regimes, customers wait little, if at all. In the efficiency-driven regime, however, the asymptotic behavior differs.

A comparison of the recent [33, 106, 152] with the older [92, 117, 124] is not straightforward for several reasons. While all address impatience, they do so in different circumstances. For example, the natural time unit for [92, 117, 124] is seconds, while patience in a call center is typically

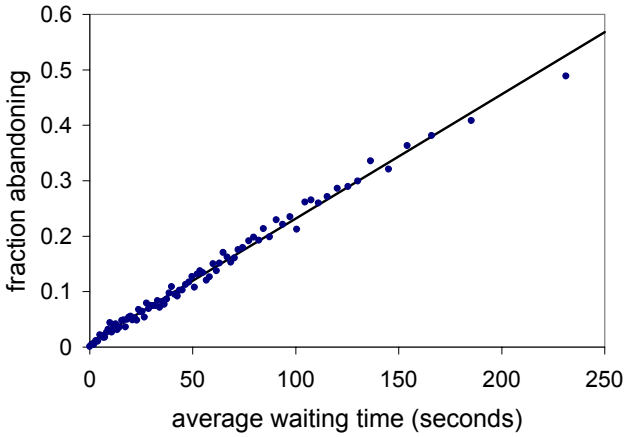


Figure 21: Empirical Relationship Between Abandonment and Delay (from [33])

measured in minutes, an average of about 10 minutes in [33, 106, 152]. Furthermore, the customers in [33, 106, 152] hear a 60-second announcement that reflects their status in queue, information which was not available to the participants in [92, 117, 124].

Censored Sampling: Data associated with patience and abandonment are typically censored, since the patience of those who get served before they abandon is not fully observed. Consequently, the need to account for censoring is an important topic in much of the work cited above. Both Palm [117] and Kort [92] avoid the problem by sampling from (unfortunate!) subjects who called a non-working service and waited until they were exhausted. Mandelbaum, Sakov, and Zeltyn [106] account for censoring by using standard tools from the statistical theory of survival analysis, especially the Kaplan-Meier estimator of a distribution function. (See the Appendix in [152] for details.) Roberts [124] handles censoring in a self-developed method, which is essentially equivalent to the Kaplan-Meier approach.

Redials and Revisits: Recalling Figure 4, there are three modes of return to a call center: redials after encountering a busy signal, redials after abandoning and revisits after service. The well developed theory of retrial queues is mostly devoted to the first. (For example, see Falin and Templeton [49].) This is the least relevant to the practice of call centers, however. Indeed, managers have typically resolved the tradeoff between busy signals and abandonment in favor of the latter, by operating with an ample number of trunk lines. As a result, immediate returns are mostly redials after abandonment, and delayed returns are revisits after service.

In most work, redials are quantified in terms of some perseverance function that gives the probability of an nth attempt, given a survival beyond the $(n-1)$th attempt. Andrews and Parsons [8] report (but do not actually show) that redials to an L.L. Bean call center are infrequent enough and spread out enough over time that they do not alter the Poisson nature of arrivals. (See the discussion at the end of §6.3.1.) Hoffman and Harris [76] develop a method of jointly estimating arrival rates (before retrials) and redial rates from ACD data.

There is no work on revisits after service, to the best of our knowledge. This does not diminish from their prime importance, however. For example, revisits to a retail call center for subsequent purchases may indicate satisfactory service, which is the opposite for revisits to technical support – and both constitute central performance measures of their operation.

6.3.4 System Performance

Given assumptions on primitives – arrivals, service times, abandonment, retrials – and the relationships among them, queueing models derive measures of system performance such as the distribution of the number of calls in queue, the distribution of delay in queue faced by a typical arriving customer and the overall abandonment rate.

Roberts [124] estimates the *virtual wait*, the distribution of the time that a customer would have to wait for service, *given infinite patience*. (These plots are consistent with Kingman's exponential law of congestion (11).) If one assumes that customers are familiar with the system, based on prior service experience, the actual delay would also coincide with the distribution of the time that a customer expects to wait, as explained in [108, 130, 152].

Indeed, one may think of two dimensions of customer patience: the first is the time that a customer is *willing* to wait, and the second, the time the customer *expects* to wait. Mandelbaum et al. [106] divide the expectation of the former by the latter to develop a "patience index." They find, for example, that customers with internet questions are willing to wait 528 seconds, on average, but their expected (virtual) wait is roughly half that, and their patience index is 1.98. In contrast, customers calling to make stock trades are willing to wait even longer (678 seconds), but they are served much more quickly, and their patience index is much higher (4.74). Thus, the index reveals that stock-trading customers are *relatively* more patient, while customers with internet questions are relatively less. Furthermore, this view of relative patience would not be captured by either expected wait or willingness to wait, alone.

Brown et al. [33] demonstrate that, given exponentially distributed times to abandonment and virtual waits, the patience index for a set of customers should equal the empirically-observed ratio

$$\frac{\#\text{ customers served}}{\#\text{ customers abandoned}}\ . \tag{27}$$

They then analyze the data from [106] comparing a patience index to the ratio above. To avoid problems of censoring, the patience index is calculated using first quartiles, rather than means. Nevertheless, the two measures of patience – theoretical patience index and empirical abandonment ratio – are shown to have a strong linear relationship. (This, despite the fact that, for these data, patience is not exponentially distributed.) Thus, the results suggest that simple counts of services and abandonments can be used to estimate customer patience.

The call center analyzed in [33, 106] has 15-20% abandonment rates. Brown et al. [33] also use these data to demonstrate that measures of system performance predicted by the Erlang A model fit the center's actual performance well. Again a good fit emerges, despite the fact that neither service times nor patience are exponentially distributed, as assumed in the Erlang A model. Our experience suggests that such robustness, perhaps surprisingly, is not uncommon. This is clearly a phenomenon worthy of further research.

Performance measures are, of course, correlated. An example is the remarkably linear relationship between the fraction of abandoning customers and average waiting time, theoretically justified in (25) and displayed in Figure 21 [110]. This implies that one need measure only one of the two statistics; the other can be arrived at through inference. More subtle is the fact that, in these data, arrival rates, service times, and delay in queue all tend to peak at the same times of day [33]. This correlation could be the result to any of several phenomena: that peak hours are the most convenient hours for customers with longer service times to call; that CSRs "pace" themselves

during busy periods by slowing down; that during periods with longer delays, customers with the more pressing problems – hence longer service duration – are more likely to not abandon. Analysis in [33] suggests that the first hypothesis is the one most likely to be true.

Readers who are primarily interested in issues associated with service quality and customer and CSR behavior can now proceed to Section 7.

6.4 Future Work in Data Analysis and Forecasting

There has been recent progress in the analysis of call-center data. Call-by-call data from a small number of sites have been obtained and analyzed, and these limited results have proven to be fascinating. In some cases, such as the characterization of the arrival process and of the delay of arriving calls to the system, conventional assumptions and models of system performance have been upheld. In others, such as the characterization of the service-time distribution and of customer patience, the data have revealed fundamental, new views of the nature of the service process. Of course, these limited studies are only the beginning, and the effort to collect and analyze call-center data can and should be expanded in every dimension.

Perhaps the most pressing practical need is for improvements in the forecasting of arrival rates. For highly utilized call centers, more accurate, distributional forecasts are essential. While there exists some research that develops methods for estimating and predicting arrival rates [6, 33, 84, 112], there is surely room for additional improvement to be made. However, further development of models for estimation and prediction will depend, in part, on access to richer data sets. We believe that much of the randomness of Poisson arrival *rates* may be explained by covariates that are not captured in currently available data.

More broadly, there is need for the development of a wider range of descriptive models. While a characterization of arrival rates, abandonment from queue, and service times are essential for the management of call centers, they constitute only a part of the complete picture of what goes on. For example, there exist (self) service times and abandonment (commonly called "opt-out") behavior that arise from customer use of IVRs. Neither of these phenomena is likely to be the same as its CSR analogue. Similarly, sojourn times and abandonment from web-based services have not been examined in multi-media centers.

Parallel, descriptive studies are also needed to validate or refute the robustness of initial findings. For example, lognormal service times have been reported in two call centers, both of which are part of retail financial services companies. Perhaps the service-time distribution of catalogue retailers or help-desk operations have different characteristics. Similarly, one would like to test Figure 20's finding that the waiting-time messages customers hear while tele-queueing promote, rather than discourage, abandonment.

It would also be interesting to put Palm [116], Roberts [124], Kort [92] and Mandelbaum et al. [106] in perspective. These studies provide empirical and exploratory models for (im)patience on the phone in Sweden in the 40's, France in the late 70's, the U.S. in the early 80's, and Israel in the late 90's. A systematic comparison of patience across countries, for current phone services, should be a worthy, interesting undertaking.

There is the opportunity to further develop and extend the scope of explanatory models. Indeed, given the high levels of system utilization in the QED regime, a small percentage error in the forecast of the offered load can lead to significant, unanticipated changes in system performance.

In particular, the state of the art in forecasting call volumes is still rudimentary. Similarly, the fact that service times are lognormally distributed enables the use of standard parametric techniques to understand the effect of covariates on the (normally distributed) natural log of service times [33].

In well-run QED call centers, only a small fraction of the customers abandon (around 1-3%), hence about 97% of the (millions of) observations are censored. Based on such figures, one can hardly expect any reasonable estimate of the whole patience distribution, non-parametrically at least. Fortunately, however, the analysis of [110], as represented in (26), suggests that only the behavior of impatience near the origin is of relevance, and this is observable and analyzable.

Indeed, call-center data are challenging the state-of-the-art of statistics, and new statistical techniques seem to be needed to support their analysis. Two examples are the accurate non-parametric estimation of hazard rates, with corresponding confidence intervals, and the survival analysis of tens of thousands, or even millions, of observations, possibly correlated and highly censored.

Finally, in the following section we will also argue that a broader goal should be, in fact, the analysis of *integrated* operational, marketing, human resources, and psychological data. That is, the analysis of these integrated data is essential if one is to understand and quantify the role of operational service quality as a driver for business success.

7 Future Directions in Call-Center Research

The work described above constitutes only a beginning. Below, we describe what we believe to be some natural next steps in the evolution of call center research. In all cases, both empirical and theoretical work are needed to develop the depth of understanding required.

7.1 A Broader View of the Service Process

The service process at most call centers takes place over multiple stages. Aside from a few exceptions, however, the work we have described thus far has been geared to a simple, single stage of service; there currently exists little analysis directed at the broader service process.

For example, many large call centers use IVRs, but an understanding of how they function is far from complete. There exist some theoretical papers that explicitly or implicitly model the use of IVRs in certain circumstances. Srinivasan and Talim [133] analyze a two-station network model of an IVR, possibly followed by CSR service. They show that industry practice – which first uses an Erlang B model to fix the number of trunk lines, then an Erlang C model to set the number of CSRs – can lead to problematic recommendations, such as having more agents than trunks. (Cleveland and Mayben [42] describe this industry practice.) Other examples of models that explicitly include the use of IVRs include Brandt, Brandt, Spahl, and Weber [30], Brandt and Brandt [28], and Armony and Maglaras [12, 13]. The only empirical treatment of IVR service-time distributions of which we are aware, however, is the brief description provided in Mandelbaum et al. [106].

Furthermore, IVR technology is rapidly evolving. The current generation of speech recognition and artificial intelligence (AI) technologies have increased the range, as well as the speed, of IVR-supported self-service. These advances are likely to make IVR-based service increasingly important,

and study is required to understand how the technologies can best improve the service process.

The service that CSRs provide to customers may also be thought of as occurring over multiple stages. Calls may be "escalated" from a front-line CSR to a supervisor or problem specialist. In some operations, such as insurance claims, service requests commonly require several phone calls to be resolved. It is also often the case that, having satisfied the customer's service request, a CSR has the opportunity to "cross-sell" another product or service.

Furthermore, customer transitions among IVR, CSR, and other "nodes" of a call-center's internal network can exhibit strong interdependencies. For example the time customers spend interacting with an IVR is fundamentally related to the time required for agents to serve them. Indeed, many businesses encourage customers to use IVRs as a means of self service, with the specific hope of reducing the time spent with a CSR.

Other technologies can also affect the nature of service durations in a systematic fashion. As noted before, the "screen pops" enabled by CTI can reduce and standardize service times. Similarly, the automatic greetings and farewells used by telephone operators reduce both the service duration and its stochastic variability.

Because of the shared nature of many call-center resources, customers affect each others' service times. For example, Akşin and Harker [3] develop a theoretical model in which the corporate information system is a bottleneck. Customers' service durations with agents are driven by agent queries to the shared system, and as congestion increases, service durations increase as well. Simulation studies in the paper demonstrate how this can lead to counterintuitive phenomena: for example, performance levels may decrease as the number of agents increase. As far as we know, there exists no empirical work that systematically investigates the claim.

7.2 An Exploration of Intertemporal Effects

Just as the process by which a given call is served may be complex, there also exist interdependencies that exist over time. We believe that a fuller description and modelling of these interdependencies are essential for a complete understanding of call-center behavior (as well as of service-delivery systems more generally).

At the most basic level, primitives used in queueing models systematically vary over time. For arrival rates, which exhibit regular, seasonal patterns, this fact is well accepted. It also appears to be true for service rates though, in this case, the documentation is far less complete and the reasons far less clear.

There are many sources of systematic variation in service times that should be better described and analyzed. For example, Gustafson [68] documents "learning-curve" effects: as they gain experience, CSRs become systematically faster on the job. Similarly, Sze [137] describes a phenomenon (sometimes called "shift fatigue") in which "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."

Again, there exists limited work that address some of these effects. The data analysis of Brown et al. [33] (described in §6.3.4) shows that service times at one call center are longer during periods with higher arrival rates. The hiring model in Gans and Zhou [56] accommodates learning curve effects, should they exist. In both cases, however, our understanding of what drives the duration

of service times and how these effects should be modelled is only rudimentary.

Similarly, over the course of the day or week, customer patience may vary because customers have learned to expect more or less congestion during certain intervals. Again, Zohar et al. [152] describe and model this type of behavior, but work must be done to understand this effect at a more fundamental level.

7.3 A Better Understanding of Customer and CSR Behavior

Most of the operational primitives of call-center queueing models – arrivals, services, abandonment, retrials – are functions of human behavior. For example, the need to view intertemporal changes in these primitives is intimately related to the fact that peoples' behavior changes according to circumstances and over time.

Indeed, 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 agenda for multi-disciplinary research.

We note that there exists a substantial body of research, which originates in psychology and marketing, that studies people's behavior while waiting in queues. The articles in Section III of Mandelbaum [97], entitled "Consumer Psychology", have ample leads. Most of this work concerns physical queues, such as those found in a bank or a clinic. As we noted in Section 2.4, however, tele-queues are phantom, hence the experience of waiting in them is likely to differ significantly from that in physical queues.

The subtlety of human factors can make their effect on tele-queues difficult to properly quantify, measure and model. Consider, for example, human impatience while waiting for a tele-service. It surely depends on the communication channel – telephone, internet, IVR – and on the type of customer. But who is more patient, a regular customer or a VIP? VIP customers may become impatient and abandon the tele-queue more quickly (Figure 20). At the same time, the actual waiting time performance may be better, and hence their relative patience greater (Section 6.3.4).

Furthermore, CSRs' and customers' experiences are linked. For example, Schneider et al. [127] show that the employees' perceptions of working conditions and customers' perceptions of service quality affect each other over time. (For more on this interaction, see also the references within [127].)

Customer and CSR Behavior as Equilibrium Phenomena: Both customers and employees have the ability to adjust their expectations, based on experience, and to adjust behavior, based on expectations. In call centers, this adjustment can also be seen in response to management practice, which sometimes dramatically affects behavior. For example, an incentive scheme that rewards agents for maintaining a low AHT can lead CSRs to hang up on customers. (The empirical distribution displayed on the left side of Figure 19 reflects a similar pattern, though in this case agents were taking small "rest breaks" by hanging up on customers.) Similarly, announcements made to customers waiting "on hold" can lead them to abandon the tele-queue, as noted in §6.3.3.

The CSR and customer phenomena described above suggest that an appropriate framework for including human behavior is that of a game-theoretic or economic equilibrium, arrived at through learning and self-optimization. This is the perspective of a number of recent papers. Shumsky and Pinker [132] consider settings in which front-line CSRs act as "gatekeepers" who may attempt to

solve customers' problems or send them onto a specialist, and they use principal-agent models to analyze the impact of incentive compensation schemes.

Mandelbaum and Shimkin [108] develop a theoretical equilibrium-analysis of rational customers who compare their expected *remaining* waiting time with a subjective value they ascribe to service. This is equivalent to assuming a linear cost structure, and it implies that additional factors, such as the likelihood of "never" being served, are required to motivate a waiting customer to abandon the queue. Such a motivation is not needed in Shimkin and Mandelbaum [130], who show that when waiting costs are nonlinear, an analogous equilibrium can be achieved. Zohar et al. [152] provide empirical evidence for the thesis of rational, adaptive customers, and they present a simpler and more analytically tractable form of the original, linear model. Armony and Maglaras [12, 13] and Whitt [145] both develop similar notions of abandonment and congestion as equilibrium phenomena.

Multiple Levels of Equilibria: Furthermore, the notion of an equilibrium exists simultaneously at a number of levels. In real time, system congestion interacts with customers' patience to engender balking and abandonment behavior, and these behaviors help to stabilize the system. Over longer periods, balking and abandonment behavior during one interval can be seen to lead to retrials during another. Thus, one might think of arrival rates, service times, patience, and the propensity to redial at any point in time as being jointly dependent on queueing system performance over the course of the week.

That is, system performance may more properly be modelled as depending on arrival-rate, service-time, patience, and retrial *functions* that vary systematically over the week (see Mandelbaum et al. [103]). The form of the functions represents a fixed point, arrived at through customer and CSR experience and adjustment over many weeks.

The work cited above represents a promising start. Still, little is known about the processes that underly waiting and service behaviors. Just as the analysis of abandonment data have guided Zohar, Mandelbaum and Shimkin [152] to simplify the theoretical model developed in [108], a deeper economic and psychological understanding of abandonment and service behaviors will enable continuing improvements in their theoretical characterization.

Finally, the hierarchy of equilibria described above may be extended one level further. Just as customers' expectations concerning queueing delays may be learned through experience, their repeated contacts with a company – through its call centers and through other communications channels – may lead them to form broader expectations concerning the value the company provides. These expectations would then lead customers to alter their buying patterns, for example, for better or worse.

7.4 A Call for Multi-Disciplinary Research

The notion that service quality affects customer buying behavior requires one to consider elements of quality beyond the operational measures on which we have concentrated thus far. For example, as outlined in §2.5, an adequate treatment of service quality is likely to include notions of the effectiveness of service encounters, as well as judgments concerning the content of CSRs' interactions with customers. It is also likely to include financial, as well as operational, measures of success.

For instance, it may be that calls which reduce future rework or improve the likelihood of future purchases are judged more "effective," and effective calls require longer service durations. Conversely, customer abandonment that results from system congestion can reduce the likelihood

of future purchases, and (given a fixed set of servers) shorter service times reduce abandonment. Then, given the current state of the system – the overall level of congestion, which specific customers are waiting in queue, and which are being served by what specific CSRs – one would like to know whether it is best to have more effective (longer) or quicker calls.

Thus, a broader view of service quality may affect managers' and academics' notions of what gets optimized during the service process. Research that supports this view is just emerging, however. For example, Pinker and Shumsky [120] posit that "learning curve" behavior drives the quality of service offered by CSRs, and they consider how learning and quality are affected by CSR specialization. Gans [54] and Hall and Porteus [70] offer models of service quality affecting customer churn. These models attempt to connect operational, human resources, and marketing decisions, but they are highly stylized. To become practical, the analysis must capture more detail of the service process.

In fact, CRM systems (promise to) enable companies to better track and understand how each service experience affects a customer's long-term buying behavior. Skills-based routing systems provide a natural complement, in that they allow a company to exercise much finer control over who serves that customer, as well as how and when. These systems only provide the necessary infrastructure, however. Research is required to understand exactly how customers respond to service and, in turn, exactly how their service should be controlled.

The research required to support this scheme falls across several disciplines. In broad terms, expertise in marketing, data mining, and statistics are required to segregate customers into classes and to develop those classes' cost or index functions, as in the Gcμ rule of (22). Knowledge of human resources is required to design agent skill-sets, as well as to develop effective hiring and training plans. Operations-based research is needed to understand how best to route customers and their work to CSRs. Information systems expertise is required to ensure that the underlying CRM and routing systems are capable of performing the required functions.

Furthermore, because customers, CSRs and systems jointly interact, much of the required research is inherently multi-disciplinary. We highlight a few of the many elements of call-center design and management that would benefit from such an approach:

- In skills-based routing, operational decisions determine the duration of time that calls wait on hold, as well as the nature of the CSR who serves the call. These, in turn, affect the caller's experience, hence short and long-term behavior. Thus, a solution to the problem requires an integrated view of both operational and marketing issues.

- Skills-based routing decisions also affect customer abandonment from queue, and impatience is, fundamentally, a psychological process. Similarly, the customer's perception of service depends on his or her interaction with a CSR. Thus, operating policies should also be informed by a proper understanding of the psychology of individuals and of social interactions.

- The numbers of different CSRs, as well as the types of "skill sets" that they have, affect how the weekly scheduling and real-time routing problems can be solved. Thus, these HR problems of organizational design and management are linked to marketing outcomes through operational, call-routing controls. They also have an operational element, themselves.

- Incentive schemes complement skill sets and job-ladders in the design of CSRs' work. Tools from microeconomics, such as principal-agent models, can provide insight into possible or likely outcomes of proposed system designs.

- Advances in automation technologies, such as speech recognition and AI, affect the design of IVRs, web-based interaction, and call-scripting – as well as how they are integrated. These changes will have a direct effect on the time required to complete tasks. More subtly, and as importantly, they will also affect system performance through their impact on customer satisfaction and behavior.

- The design of information flows also affects both CSR and customer behavior and, in turn, system performance. A simple example for CSRs is the use of flashing panels to provide real-time feedback on the length of the queue. An analogous example for customers is the communication of information concerning expected delay in queue.

- The need for statistical tools arises everywhere in the analysis of call center operations. Examples include: the forecasting of arrival rates and service times; the characterization of the hazard-rate of abandonment (impatience function); and the validation or refutation of queueing-theoretic performance models.

- Data mining and statistical analysis will also be essential in developing the link between operating decisions and their marketing consequences. For example, they should be used to determine which customers have high (potential) value and should receive better service.

8 Conclusion

Telephone call centers are an economically important new form of operation. They employ a growing fraction of the work force and mediate a significant volume of trade in developed economies.

While tools from operations management and operations research have proved to be essential for their management, a number of problems related to call centers' most basic operational characteristics have yet to be thoroughly tackled. In particular, the forecasting of arrival rates, the characterization of customer and agent behavior, and the analysis of the time-varying nature of these systems need to be more fully developed, and they represent challenges for academics and managers alike.

Furthermore, a number of new opportunities also exist for extending call center capabilities. Skills-based routing, networking, and speech-recognition are examples of promising technologies for which an understanding is just beginning to be developed. A broad range of multi-disciplinary work is needed to help them fully realize their potential.

We believe that this research is exciting because it will also have impact beyond call centers themselves. Indeed, the service sector represents 70% or more of most developed economies, and this fraction continues to grow. In many parts of the sector, operational, marketing and human resource issues are also tightly intertwined. Thus, the research frameworks and insights that are derived from multi-disciplinary call-center research are certain to apply more broadly.

A Glossary of Call-Center Acronyms

Acronym	Description	Definition
ACD	automatic call distributor	p. 7
ANI	automatic number identification	p. 6
ASA	average speed of answer	p. 10
CRM	customer relationship management	p. 8
CSR	customer service representative	p. 5
CTI	computer-telephony integration	p. 7
DNIS	dialed number identification service	p. 6
IVR	interactive voice response unit (also called VRU)	p. 5
PABX	private automatic branch exchange (also called PBX)	p. 6
PBX	private automatic branch exchange (also called PABX)	p. 6
PSTN	public switched telephone network	p. 6
TSF	telephone service factor (also called the 'service level')	p. 10
VRU	interactive voice response unit (also called IVR)	p. 5
WFM	workforce management	p. 14

References

[1] AT&T. As reported on callcenternews.com/resources/statistics.shtml. 1

[2] O.Z. Akşin Accounting for appreciating human assets in a service sector firm. Working Paper, INSEAD, 2002. 4.6

[3] O.Z. Akşin and P.T. Harker. Analysis of a processor shared loss system. *Management Science*, 47:324–336, 2001. 3.1, 7.1

[4] O.Z. Akşin and P.T. Harker. Capacity sizing in the presence of a common shared resource: Dimensioning an inbound call center. To appear in *European Journal of Operational Research*, 2003. 3.1

[5] 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. 4.7.1

[6] B. Andrews and S.M. Cunningham. L.L. Bean improves call-center forecasting. *Interfaces*, 25(6):1–13, 1995. 1.1, 6.3.1, 6.4

[7] B. Andrews and H. Parsons. L.L. Bean chooses a telephone agent scheduling system. *Interfaces*, 19:1–9, 1989. 1.1

[8] B. Andrews and H. Parsons. Establishing telephone-agent staffing levels through economic optimization. *Interfaces*, 23(2):14–20, 1993. 1.1, 3.1, 6.3.3

[9] J. Anton. The past, present and future of customer access centers. *International Journal of Service Industry Management*, 11:120–130, 2000. 1.1

[10] R. Anupindi and B.T. Smythe. Call centers and rapid technology change. Teaching Note. Submitted, 1997. 1.1

[11] M. Armony and N. Bambos. Queueing networks with interacting service resources. *Proceedings of the 37th Allerton Conference*, 42–51, 1999. 5.1.1

[12] M. Armony and C. Maglaras. On customer contact centers with a call-back option: customer decisions, routing rules, and system design. Working Paper, Columbia University, 2001. 5.1.4, 5.2, 6.3.3, 7.1, 7.3

[13] M. Armony and C. Maglaras. Contact centers with a call-back option and real-time delay information. Working Paper, Columbia University, 2002. 5.1.4, 5.2, 6.3.3, 7.1, 7.3

[14] R. Atar, A. Mandelbaum, and M. Reiman. Scheduling a multi-class queue with many i.i.d. servers: asymptotic optimality in heavy traffic. Working Paper, Technion, 2002. 5.1.4, 5.2

[15] J. Atlason, M.A. Epelman, and S.G. Henderson. Call center staffing with siumlation and cutting plane methods. Working Paper, University of Michigan, 2002. 4.5

[16] T. Aykin. Optimal shift scheduling with multiple break windows. *Management Science*, 42:591–602, 1996. 4.5

[17] F. Baccelli and G. Hebuterne. On queues with impatient customers. In *Performance '81*, 159–179. North-Holland, 1981. 4.2.2, 6.3.3

[18] N. Bambos and J. Walrand. Scheduling and stability aspects of a general class of parallel processing systems. *Advances in Applied Probability*, 25:176–202, 1993. 5.1.1

[19] D.J. Bartholomew, A.F. Forbes, and S.I. McClean. *Statistical Techniques for Manpower Planning*. John Wiley and Sons, 2nd edition, 1991. 4.6

[20] S.L. Bell and R.J. Williams. Dynamic scheduling of a system with two parallel servers in heavy traffic with complete resource pooling: asymptotic optimality of a continuous review threshold policy. *Annals of Applied Probability*, 11:608–649, 2001. 5.1.3

[21] S. Bhulai and G.M. Koole. A queueing model for call blending in call centers. In *Proceedings of the 39th IEEE CDC*, 1421–1426, 2000. 5.1.2, 5.2

[22] P.J. Bickel and K.A. Docksum. *Mathematical Statistics: Basic Ideas and Selected Topics*, vol. I, 2nd ed. Prentice-Hall, 2001. 1

[23] V.A. Bolotin. Telephone circuit holding time distributions. In *Proceedings of the 14th International Teletraffic Conference*, 125–134, 1994. 4.7.1, 6.3.2

[24] S.K. Bordoloi and H. Matsuo. Human resource planning in knowledge-intensive operations. *European Journal of Operational Research*, 130:169–189, 2001. 4.6

[25] S.C. Borst, A. Mandelbaum, and M.I. Reiman. Dimensioning large call centers. Working Paper, Technion, 2000. 3.1, 4.1.1, 4.1.1, 9, 4.1.2, 4.2.2, 4.7.1

[26] S.C. Borst, A.D. Flockhart, M.I. Reiman, and J.B. Seery. DEFINITY queue to best: multisite routing simulations. Compas Document ID 53921, Bell Laboratories, Lucent Technologies, 1996. 5.3

[27] S.C. Borst and P. Seri. Robust algorithms for sharing agents with multiple skills. Working Paper, 2000. 5.1.2, 5.1.4, 5.3

[28] 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. 5.2, 7.1

[29] A. Brandt and M. Brandt. On the $M(n)/M(n)/s$ queue with impatient calls. *Performance Evaluation*, 35:1–18, 1999. 4.2.2, 5.2

[30] A. Brandt, M. Brandt, G. Spahl, and D. Weber. Modelling and optimization of call distribution systems. In *Proceedings of the 15th International Teletraffic Conference*, 133–144, 1997. 5.2, 7.1

[31] 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

[32] R.G. Brown. *Smoothing, Forecasting and Prediction of Discrete Time Series*. Prentice-Hall, 1963. 3.3

[33] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: a queueing science perspective. Working Paper, The Wharton School, 2002. (document), 1.1, 3.2, 6, 4.7.1, 1, 1, 6.3.1, 6.3.3, 20, 6.3.3, 6.3.3, 21, 6.3.4, 6.3.4, 6.4, 7.2

[34] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Empricial analysis of a network of retail-banking call centers. Work in Progress, The Wharton School, 2002. 3.2, 8, 18

[35] M.J. Brusco and L.W. Jacobs. Optimal models for meal-break and start-time flexibility in continuous tour scheduling. *Management Science*, 46:1630–1641. 4.5

[36] E.S. Buffa, M.J. Cosgrove, and B.J. Luce. An integrated work shift scheduling system. *Decision Sciences*, 7:620–630, 1976. 1.1, 3.1, 5, 3.2

[37] Z. Carmon and D. Kahneman. The experienced utility of queuing: experience profiles and retrospective evaluations of simulated queues. Working Paper, INSEAD, 2002. 2.5

[38] A. Charnes, W.W. Cooper, and R.J. Niehaus, eds. *Management Science Approaches to Manpower Planning and Organizational Design*. North-Holland Publishing, 1978. 4.6

[39] E. Chlebus. Empirical validation of call holding time distribution in cellular communications systems. In *Proceedings of the 15th International Teletraffic Conference*, 1997. 6.3.2

[40] B.P.K. Chen and S.G. Henderson. Two issues in setting call centre staffing levels. *Annals of Operations Research*, 108):175–192, 2001. 4.4

[41] H. Chen and D.D. Yao. *Fundamentals of Queueing Networks*. Springer-Verlag, 2001. 4.1

[42] B. Cleveland and J. Mayben. *Call Center Management on Fast Forward*. Call Center Press, 1997. 1.1, 2.5, 7.1

[43] Datamonitor. As reported on `resources.talisma.com/ver call statistics.asp`. 1

[44] 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

[45] S.G. Eick, W.A. Massey, and W. Whitt. The physics of the $M_t/G/\infty$ queue. *Operations Research*, 41:731–742, 1993. 4.3

[46] S.G. Eick, W.A. Massey, and W. Whitt. The $M_t/G/\infty$ queue with sinusoidal arrival rates. *Management Science*, 39:241–252, 1993. 4.3

[47] A.K. Erlang. On the rational determination of the number of circuits. In *The life and works of A.K. Erlang*. E. Brockmeyer, H.L. Halstrom and A. Jensen, eds. Copenhagen: The Copenhagen Telephone Company, 1948. 4.1.1, 4.2.1

[48] 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

[49] G.I. Falin and J.G.C. Templeton. *Retrial Queues*. Chapman & Hall, 1997. 6.3.3

[50] A. Federgruen, and H. Groenevelt. M/G/c systems with multiple customer classes: characterization and achievable performance under nonpreemptive priority rules. *Management Science*, 34:1121–1138, 1988. 5.1.2

[51] M.A. Feinberg. Performance characteristics of automated call distribution systems. In *GLOBECOM '90*, IEEE, 415–419, 1990. 4.2.1

[52] P.J. Fleming, A. Stolyar, and B. Simon. Heavy traffic limits for a mobile system model. *Second International Congress on Telecommunication Systems, Modeling, and Analysis*, 158–176, 1994. 4.7.1

[53] 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. 4.5

[54] N. Gans. Customer loyalty and supplier quality competition. *Management Science*, 48:207-221, 2002. 7.4

[55] N. Gans and G. van Ryzin. Optimal control of a multiclass, flexible queueing system. *Operations Research*, 45:677–693, 1997. 5.1.1, 5.1.3

[56] N. Gans and Y.-P. Zhou. Managing learning and turnover in employee staffing. To appear in *Operations Research*. 4.6, 4.7.2, 7.2

[57] N. Gans and Y.-P. Zhou. A call-routing problem with service-level constraints. To appear in *Operations Research*. 5.1.2, 5.2

[58] O. Garnett and A. Mandelbaum. An introduction to skills-based routing and its operational complexities. Teaching note, Technion, 2001. Full version available upon request, from AM. Downloadable from ie.technion.ac.il/serveng/Homeworks/HW9.pdf. 16, 5.1.2

[59] O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. Working Paper, Technion, 2000. 1.1, 4.2.2, 4.7.1

[60] J.J. Gordon and M.S. Fowler. Accurate force and answer consistency algorithms for operator services. In *Proceedings of the 14th International Teletraffic Conference*, 339–348, 1994. 6.3.1

[61] W.K. Grassmann. Transient solutions in Markovian queueing systems. *Computers & Operations Research*, 4:47–53, 1977. 4.3, 4.3

[62] W.K. Grassmann. Finding the right number of servers in real-world systems. *Interfaces*, 18:94–104, 1988. 4.4

[63] L. Green and P. Kolesar. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science*, 37:84–97, 1991. 4.3

[64] L. Green and P. Kolesar. The lagged PSA for estimating peak congestion in multiserver markovian queues with perioding arrival rates. *Management Science*, 43:80–87, 1997. 4.3

[65] L. Green, P. Kolesar, and J. Soares. Improving the SIPP approach for staffing service systems that have cyclic demands. *Operations Research*, 49:549–564, 2001. 4.3

[66] R.C. Grinold and K.T. Marshall. *Manpower Planning Models*. ORSA Publications in Operations Research, 1977. 4.6

[67] 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. 1.1

[68] H.W. Gustafson. Force-loss cost analysis. In W.H. Mobley, *Employee Turnover: Causes, Consequences, and Control*. Addison-Wesley, Reading, Massachusetts, 1982. 7.2

[69] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29:567–587, 1981. 4.1.1, 4.2.2, 4.7.1

[70] J. Hall and E. Porteus. Customer service competition in capacitated systems. *Manufacturing & Service Operations Management*, 2:144–165, 2000. 7.4

[71] C.M. Harris, K.L. Hoffman, and P.B. Saunders. Modeling the IRS telephone taxpayer information system. *Operations Research*, 35:504–523, 1987. 6.3.2

[72] J.M. Harrison and M.J. Lopez. Heavy traffic resource pooling in parallel-server systems. *Queueing Systems*, 33:339–368, 1999. 5.1.3, 5.1.3

[73] J.M. Harrison and A. Zeevi. Dynamic scheduling of a multi-class queue in the Halfin-Whitt heavy traffic regime. Working Paper, Columbia University, 2002. 5.1.4

[74] S.G. Henderson. Estimation for nonhomogeneous poisson processes from aggregated data. Working Paper, Cornell University, 2002. 6.3.1

[75] W.B. Henderson and W.L. Berry. Heuristic methods for telephone operator shift scheduling: an experimental analysis. *Management Science*, 22:1372–1380, 1976. 4.5

[76] K.L. Hoffman and C.M. Harris Estimation of a caller retrial rate for a telephone information system. *European Journal of Operational Research*, 27:207–214, 1985. 6.3.3

[77] C.C. Holt, F. Modigliani, J. Muth, and H.A. Simon. *Planning Production, Inventory and Work Force*. Prentice-Hall, Englewood Cliffs, N.J., 1960. 4.6

[78] A. Ingolfsson, E. Akhmetshina, S. Budge, Y. Li, and X. Wu. A survey and experimental comparison of service level approximation methods for non-stationary M/M/s queueing systems. Working Paper, School of Business, University of Alberta, 2002. 4.3

[79] A. Ingolfsson and E. Cabral. Combining integer programming and the randomization method to schedule employees. Working Paper, School of Business, Unibersity of Alberta, 2002. 4.3, 4.5

[80] A. Ingolfsson, M.A. Haque, and A. Umnikov. Accounting for time-varying queueing effects in workforce scheduling. *European Journal of Operational Research*, 139:585–597, 2002. 4.3, 4.5

[81] D.L. Jagerman. Some properties of the Erlang loss function. *Bell Systems Technical Journal*, 53:525–551, 1974. 4.2.1, 4.7.1

[82] P. Jelenkovic, A. Mandelbaum, and P. Momcilovic. The GI/D/N queue in the QED regime. In Preparation, 2002. 4.7.1

[83] O.B. Jennings, A. Mandelbaum, W.A. Massey, and W. Whitt. Server staffing to meet time-varying demand. *Management Science*, 42:1383–1394, 1996. 4.3, 4.7.1, 4.7.2

[84] G. Jongbloed and G.M. Koole. Managing uncertainty in call centers using Poisson mixtures. *Applied Stochastic Models in Business and Industry*, 17:307- 318, 2001. 4.4, 6.3.1, 6.4

[85] J.F.C. Kingman. On queues in heavy traffic. *Journal of the Royal Statistical Society*, Series B, 24:383–392, 1962. 4.1, 5.1.2

[86] L. Kleinrock. A delay-dependent queue discipline. *Naval Research Logistics Quarterly*, 11:59–73, 1964. 5.1.2, 5.1.3

[87] L. Kleinrock and R.P. Finkelstein. Time dependent priority queues. *Operations Research*, 15:104–116, 1967. 5.1.2, 5.1.3

[88] 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. 5.3

[89] G.M. Koole and A. Mandelbaum. Queueing models of call centers: an introduction. *Annals of Operations Research*, to appear, 2002. 1.1, 3, 4, 10, 11, 12

[90] 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. 3.2, 4.5

[91] G.M. Koole and J. Talim. Exponential approximation of multi-skill call centers architecture. In *Proceedings of QNETs 2000*, 23/1–10, 2000. 5.1.2

[92] B.W. Kort. Models and methods for evaluating customer acceptance of telephone connections. *GLOBECOM '83*, IEEE, 706–714, 1983. 6.3.2, 6.3.3, 6.3.3, 6.4

[93] J.C. Larréché, C. Lovelock, and D. Permenter. First Direct: Branchless banking. INSEAD Case 01/97-4660, 1997. 1

[94] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. ASA-SIAM Series on Statistics and Applied Proability, 1999. 6.3.2

[95] A.M. Lee and P.A. Longton. Queueing processes associated with airline passenger check-in. *Operational Research Quarterly*, 10:56–71, 1959. 4.1.1

[96] S. Makridakis, S.C. Wheelwright, and R.J. Hyndman. *Forecasting: Methods and Applications*, 3rd ed. John Wiley & Sons, 1998. 3.3

[97] A. Mandelbaum. Call centers (centres): Research bibliography with abstracts. Version 3, 137 pages, 2002. Downloadable from ie.technion.ac.il/serveng/References/ccbib.pdf. 1, 1.1, 6, 7.3

[98] A. Mandelbaum and W.A. Massey. Strong approximations for time-dependent queues. *Mathematics of Operations Research*, 20:33–64, 1995. 4.3

[99] A. Mandelbaum, W.A. Massey, and M.I. Reiman. Strong approximations for Markovian service networks. *Queueing Systems*, 30:149–201, 1998. 5.2

[100] A. Mandelbaum, W.A. Massey, and B. Rider. The $M_t/M/N_t$ in the QED Regime. Work in Progress, 2002. 4.3, 4.7.1

[101] 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. 4.3, 13, 4.7.1

[102] A. Mandelbaum, W.A. Massey, M. Reiman, and A. Stolyar. Waiting time asymptotics for time varying multiserver queues with abandonment and retrials. *Proceeding of the 37th Allerton Conference*, 1999. 4.3, 4.7.1

[103] 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, Technion, 2000. 4.3, 4.7.1, 7.3

[104] A. Mandelbaum and M. Reiman. Dimensioning finite-buffer queues with abandonment. Work carried out at Bell Labs, 2000. 4.2.1

[105] A. Mandelbaum and A. Ruszczynski. Staff scheduling problems with large, random staffing requirements. Work in Progress, 2002. 3.2, 4.5

[106] A. Mandelbaum, A. Sakov, and S. Zeltyn. Empirical analysis of a call center. Technical Report, Technion, 2001. Downloadable from ie.technion.ac.il/serveng/References/references.html. (document), 1.1, 5, 1, 6.3.1, 6.3.2, 19, 6.3.3, 20, 6.3.3, 6.3.3, 6.3.4, 6.3.4, 6.4, 7.1

[107] A. Mandelbaum and R. Schwartz. Simulation experiments with M/G/100 queues in the Halfin-Whitt (Q.E.D.) regime. Technical Report, Technion, 2002. Downloadable from ie.technion.ac.il/serveng/References/references.html. 4.7.1, 14, 4.7.1

[108] A. Mandelbaum and N. Shimkin. A model for rational abandonments from invisible queues. *Queueing Systems*, 36:141–173, 2000. 6.3.4, 7.3

[109] A. Mandelbaum and A.L. Stolyar. Scheduling flexible servers with convex delay costs: heavy-traffic optimality of the generalized $c\mu$-rule. Working Paper, Technion, 2002. 5.1.3, 5.1.3, 5.1.4

[110] A. Mandelbaum and S. Zeltyn. Exploring queueing systems with impatient customers: Empiricallly-based analysis of call centers. Work in Progress, Technion, 2002. 6.3.3, 6.3.4, 6.4

[111] W.A. Massey and R.B. Wallace. An optimal design of the M/M/C/K queue for call centers. Working Paper, Princeton University, 2002. 4.2.1, 5.2

[112] W.A. Massey, G.A. Parker, and W. Whitt. Estimating the parameters of a nonhomogeneous Poisson process with a linear rate. *Telecommunications Systems*, 5:361–388, 1996. 6.3.1, 6.4

[113] W.A. Massey and W. Whitt. Peak congestion in multi-server service systems with slowly varying arrival rates. *Queueing Systems*, 25:157–172, 1997. 4.3, 4.7.2

[114] V. Mehrotra. Ringing up big business. *OR/MS Today*, 18–24, August 1997. 1.1

[115] G.L. Nemhauser and L.A. Wolsey. *Integer and Combinatorial Optimization*. Wiley Interscience, 1988. 4.5

[116] C. Palm. Intensitätsschwankungen im Fernsprechverkehr. *Ericsson Technics*, 44:1–189, 1943. 4.2.2, 6.3.3, 6.4

[117] C. Palm. Methods of judging the annoyance caused by congestion. *Tele*, 4:189–208, 1953. 6.3.3

[118] M. Perry and A. Nilsson. Performance modeling of automatic call distributors: assignable grade of service staffing. In *XIV International Switching Symposium*, 294–298, 1992. 5.1.2, 5.1.3

[119] 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

[120] E. Pinker and R. Shumsky. The efficiency-quality tradeoff of crosstrained workers. *Manufacturing & Service Operations Management*, 2:32–48, 2000. 7.4

[121] 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. 4.7.1

[122] P. Quinn, B. Andrews, and H. Parsons. Allocating telecommunications resources at L.L. Bean, Inc. *Interfaces*, 21:75–91, 1991. 1.1

[123] J. Riordan. *Stochastic Service Systems*. Wiley, 1961. 4.2.2

[124] J.W. Roberts. Recent observations of subscriber behavior. In *Proceedings of the 9th International Teletraffic Conference*, 1979. 6.3.3, 6.3.3, 6.3.4, 6.4

[125] A.M. Ross. Queueing systems with daily cycles and stochastic demand with uncertain parameters. Ph.D. Dissertation, University of California, Berkeley, 2001. 4.4

[126] D.A. Samuelson. Predictive dialing for outbound telephone call centers. *Interfaces*, 29:66–81, 1999. 2.1

[127] B. Schneider, S.S. White, and M.C. Paul. Linking service climate and customer perceptions of service quality: test of a causal model. *Journal of Applied Psychology*, 83:150–163, 1998. 7.3

[128] M. Segal. The operator-scheduling problem: A network-flow approach. *Operations Research*, 24:808–823, 1974. 4.5

[129] 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. 5.3

[130] N. Shimkin and A. Madelbaum. Rational abandonment from tele-queues: nonlinear waiting costs with heterogenous preferences. Working Paper, Technion, 2002. 6.3.4, 7.3

[131] R.A. Shumsky. Approximation and analysis of a queueing system with flexible and specialized servers. Working Paper, University of Rochester, 2000. 5.1.2

[132] R. Shumsky and E. Pinker. Gatekeepers and referrals in services. Working Paper, University of Rochester, 2002. 7.3

[133] R. Srinivasan and J. Talim. Performance analysis of a call center with interacting voice response units. Technical Report, University of Saskatchewan, 2001. 1.1, 7.1

[134] D.A. Stanford and W.K. Grassmann. Bilingual server call centres. In *Analysis of Communication Networks: Call Centres, Traffic and Performance*, D.R. McDonald and S.R.E. Turner (eds.). Fields Institute Communications 28:31–48, 2000. 5.1.2

[135] A.L. Stolyar. MaxWeight scheduling of a generalized switch: state space collapse and workload minimization in heavy traffic. Working Paper, Bell Laboratories, Lucent Technologies, 2002. 5.1.3

[136] A.L. Stolyar. Private communication. 5.1.2

[137] D.Y. Sze. A queueing model for telephone operator staffing. *Operations Research*, 32:229–249, 1984. 4.1.1, 4.1.2, 4.7.1, 7.2

[138] G.M. Thompson. Improved implicit optimal modeling of the labor shift scheduling problem. *Management Science*, 41:595–607, 1995. 4.5

[139] G.M. Thompson. Assigning telephone operators to shifts at New Brunswick Telephone Company. *Interfaces*, 27(4):1–11, 1997. 4.5

[140] U.S. Bureau of Labor Statistics. Table B-1: Employees on Nonfarm Payrolls by Major Industry, 1950 to Date. As reported on www.bls.gov. 1

[141] J.A. van Mieghem. Dynamic scheduling with convex delay costs: the generalized $c\mu$ rule. *Annals of Applied Probability*, 5:809–833, 1995. 5.1.3, 5.1.4

[142] W. Whitt. Understanding the efficiency of multi-server service systems. *Management Science*, 38:708–723, 1992. 4.1.1

[143] W. Whitt. Approximations for the GI/G/m queue. *Productions and Operations Management*, 2:114–161, 1993. 4.1

[144] W. Whitt. Improving service by informing customers about anticipated delays. *Management Science*, 45:192–207, 1999. 6.3.3

[145] W. Whitt. How multi-server queues scale with growing congestion-dependent demand. Working Paper, AT&T Research, 2001. 4.1.2, 7.3

[146] W. Whitt. *Stochastic-Process Limits*. Springer-Verlag, 2002. 4.1, 4.7.1

[147] W. Whitt. Heavy-traffic limits for the $G/H_2^*/n/m$ queue. Working Paper, Columbia University, 2002. 4.2.1

[148] W. Whitt. A diffusion approximation for the $G/GI/n/m$ queue. Working Paper, Columbia University, 2002. 4.2.1, 4.7.1

[149] R.J. Williams. On dynamic scheduling of a parallel server system with complete resource pooling. In *Analysis of Communication Networks: Call Centres, Traffic and Performance*, D.R. McDonald and S.R.E. Turner (eds.). Fields Institute Communications 28:49–71, 2000. 5.1.3, 5.1.3

[150] R.W. Wolff. Poisson arrivals see time averages. *Operations Research* 30:223–231, 1982. 3.2

[151] J, Yoo. Queueing models for staffing service operations. Ph.D. Dissertation, University of Maryland, 1996. 4.3, 4.5

[152] E. Zohar, A. Mandelbaum, and N. Shimkin. Adaptive behavior of impatient customers in tele-queues: Theory and emperical support. *Management Science*, 48:566-583, 2002. 6.3.3, 6.3.3, 6.3.4, 7.2, 7.3