STOCHASTIC MODELS FOR THE DESIGN AND MANAGEMENT OF CUSTOMER CONTACT CENTERS: SOME RESEARCH DIRECTIONS

by

Ward Whitt

Department of Industrial Engineering and Operations Research
Columbia University, New York, NY 10027

March 5, 2002

Abstract

A (customer) *contact center* is a collection of resources providing an interface between a service provider and its customers. The classical contact center is a *call center*, containing a collection of service representatives (reps) who talk to customers over the telephone. In a call center, the service reps are supported by quite elaborate information-and-communication-technology (ICT) equipment, such as a private branch exchange (PBX), an interactive voice response (IVR) unit, an automatic call distributor (ACD), a personal computer (PC) and assorted databases. With the rapid growth of *e-commerce*, contact between the service provider and its customers if often made via e-mail or the Internet instead of by telephone. Thus the general interface between a service provider and its customers is now often called a contact center.

The design and management of contact centers is important, and worthy of research, because contact centers comprise a large, growing part of the economy and because they are quite complicated. Classic call centers are complicated because there are often many types of calls, requiring different service skills, such as knowledge of different languages and knowledge of different special promotions. The ACD is able to respond through skill-based routing, but there remains an opportunity to determine better routing algorithms and more effective methods for staffing. The difficulties are largely due to uncertainty about future arrivals of service requests of various kinds and the time and resources that will be required to respond to those requests.

Given that contact with customers no longer need be by telephone, it is necessary to consider the other media, e.g., e-mail, fax, Internet, etc. When service reps perform several different kinds of work, there is a need to allocate effort to the different tasks, which presents opportunities for improved efficiency if some of the work can be postponed. Multimedia also lead to more complicated contact experiences. For example, customers might alternate between periods of interaction with service reps and periods of browsing on the Internet. Thus, the full contact experience might take place over several disjoint subintervals of time. Consequently, service reps might effectively be serving several different customers at once.

This paper describes some research directions related to stochastic models of contact centers, aiming to respond to these challenges. The goal is to uncover fundamental principles and develop performance analysis tools that will make it possible to improve the design and management of contact centers. We should seek conditions for *resource pooling* in contact centers, where the efficiency of a large single-skill center is achieved in a multi-skilled center with minimal flexibility. Staffing should be studied in each of three important time scales: real-time, short-term and long-term. Effort should be made to understand the contribution of different sources of uncertainty: model uncertainty, parameter uncertainty and process uncertainty. Simulation should be used as a tool and should be the subject of study: Effort should be made to determine the required simulation runlengths for call-center simulations. Effort should be made to investigate operational real-time simulations by conducting simulations with nested simulations. Finally, an effort should be made to integrate research and education by making contact centers a context focus in courses on stochastic models and simulation.

1. Introduction

1.1. Customer Contact Centers as Interfaces for Teleservices

Many services - from emergency to retail - are largely *teleservices*, in that customers and service providers are remote from each other, at least when service is initiated. For example, emergency medical service is often provided by emergency-medical-service (EMS) personnel who arrive in an ambulance, but the medical service is usually requested by making a 911 telephone call. With retail, the customer is likely to remain remote from the service provider even when the service is performed. For example, a customer may order a coat from a catalog by making a telephone call or from the Internet by viewing an online catalog and entering appropriate credit-card information. Then the coat may be delivered to the customer by a shipper such as United Parcel Service (UPS).

With a teleservice, the delivery of service is provided or enabled by a (customer) contact center. A *contact center* is a collection of resources providing an interface between the service provider and its remote customers. For the emergency-medical-service example above, the contact center may be a group of EMS dispatchers. For the retail example above, the contact center can be either a telephone call center, receiving telephone orders, or a more general contact center, with equipment and service representatives responding to Internet orders.

1.2. Classical Telephone Call Centers

The classical contact center is the (telephone) *call center*, containing a collection of customer service representatives (service reps, i.e., people) who talk to customers over the telephone. In a call center, the service reps are supported by quite elaborate information-and-communication-technology (ICT) equipment, such as a private branch exchange (PBX), an automatic call distributor (ACD), a personal computer (PC) and assorted databases.

The environment of a typical call center is a large room filled with cubicles, in which service reps wearing telephone headsets sit before computer screens, which provide supporting information. If we ignore the ICT support, a call-center service rep is similar to an old-time telephone operator. Today, we have relatively little experience with telephone operators, but forty years ago telephone operators were as prevalent as call-center service reps are today. Back then, it was projected that every living person would soon be a telephone operator, if the number of people employed as telephone operators kept growing at the prevailing rate. The spirit of the telephone-operator culture is well captured by Lily Tomlin's unforgettable character, Ernestine. For a replay of the video "Ernestine Meets Mr. Veedle" from the television program *Laugh-In*, see http://www.lilytomlin.com .

There are different kinds of call centers: Call centers may have only inbound traffic (customer-generated calls), only outbound traffic (rep-generated calls, e.g., telemarketing) or a combination of these. Inbound call centers are usually supported by interactive voice response (IVR) units, which serve as elaborate answering machines. Through a selection of menus, IVR's attempt to respond to the customer's needs and, if necessary, help route the call to an appropriate service rep.

Outbound call centers may have predictive dialers, which dial before a rep is available, attempting to anticipate when one will be available, thus reducing rep idle time, but unfortunately leading to some calls without a calling party, thus annoying the called parties [58]. (It would be nice to reduce that annoyance, and thus improve system performance from the customer's perspective.)

Call centers are pervasive and growing rapidly. Hence, just as with operators forty years ago, we could project that every living person would soon become a call-center service rep, if the number of people employed as service reps kept growing at current rates. From the perspective of job satisfaction, that would not be so bad, because the service rep's job can be much more exciting than the job of an old-time telephone operator (even allowing that Ernestine seemed to be having a pretty good time). With all the supporting ICT equipment at their fingertips, the service reps have a greater opportunity to "bust out with mad skills." In the more challenging settings, the job of a service rep can approach that of an airplane pilot or an air-traffic controller.

1.3. The E-Commerce Revolution

However, it is evident that, just as with telephone operators before, dramatic technological changes are currently underway, which will reduce the need for so many service reps. The telephone is no longer the only means of communication. The telephone is rapidly being replaced by electronic mail and the Internet. Indeed, we are in the midst of a communication revolution, with service being provided more and more via electronic mail and the Internet. We are experiencing a rapid growth of (electronic) *e-commerce*. As an illustration, Weinberg [70] describes the flurry of online spending for Valentine's Day as an "E-Commerce Love Letter:"

> Even in a shaky economy, Americans still had time and money to send virtual love on Valentine's Day. Valentine's Day 2002 results have online retailers whistling love songs to the tune of $3 billion. Americans spent nearly 50% more online than last year in the 20 days before Valentine's Day, according to e-commerce tracker Bizrate.com. More and more customers are seeing the Internet as a primary shopping resource, not just as a place to compare prices.

Since the interface between service providers and their customers is less likely to involve telephone calls, the interface is now often referred to as a customer contact center or just a contact center. We want to better understand traditional telephone call centers and respond to new challenges posed by the emerging, more general, contact centers.

1.4. Economic Importance of Contact Centers

The introduction to the Program Announcment for Exploratory Research on Engineering the Service Sector, NSF 02-029, vividly describes the importance of the service sector in the economy and the need for productivity growth in that sector. Since contact centers are used thoughout the service sector, those general observations apply to contact centers as well.

As indicated above, the contact-center industry itself is vast and rapidly expanding. It is estimated that about four million people in the United States - 3% of the workforce - work in contact centers, with the number growing by about 20% per year. Many statistics have been collected demonstrating the economic importance of contact centers; for a sample, see http://www.callcenternews.com/resources/shtml .

1.5. History of Research on Telephone Call Centers

1.5.1. The Long Tradition at AT&T

There is a long tradition of research on telephone call centers, much conducted at AT&T, where the author worked for the last twenty-five years. Just as we propose to do, most previous researchers have exploited stochastic (or probabilistic) models to help improve the design and management of call centers. For a sample of the published research on call centers from AT&T, see Brigandi et al. [14], Durinovic-Johri and Levy [19], Kogan et al. [36], Levy et al. [40], Milito et al. [52], Segal [59] and Sze [64].

However, the published research is only the tip of the iceberg; there was much more unpublished research on call centers at AT&T. A strong point of internal research at AT&T over the years has been the extensive use of call-detail data.

1.5.2. Empirical Studies Exploiting Call-Detail Data

At AT&T it has long been recognized that it is necessary to exploit the full call-detail data in order to properly understand call-center performance. Fortunately, AT&T has had access to all the necessary call-detail data. the call-detail data can come from several sources: the customer billing system, the network switches, the PBX, and the computer data network supporting the PC's of the service reps.

The call-detail data were used to help manage the AT&T customer-care centers and to advise customers with call centers connected by toll-free (1-800 service) telephone lines about system performance. That advice would often justify selling the customer more lines, but more importantly the advice constituted a significant component of the 1-800 service provided by AT&T to its customers, enabling the AT&T customers in turn to provide better service to their customers.

Unfortunately, there is relatively little in the open literature on empirical studies of call centers. Usually, the data were regarded as proprietary information. Indeed, it is difficult to obtain call-detail data from operating call centers, even though much is available from the ACD.

Of course, call centers do perform data analysis and make performance reports. However, the widely available reports produced for call center managers are sample averages over intervals, usually ranging from five to thirty minutes. It evidently has not been deemed worthwhile to collect and maintain full call-detail data.

It is thus very helpful that significant empirical studies of call centers have recently been conducted and reported by Mandelbaum et al. [47] and Zohar et al. [86]. The analysis in [47] is based on the full call-detail record of 444,000 telephone calls to a bank over a twelve-month period. The empirical studies are clearly critical. There is a need for more empirical work.

1.5.3. The Trade Literature

There is also a vast professional trade literature on call centers and the more general contact centers, which can be accessed easily from the Internet, e.g., from http://www.pipkins.com/resources.asp and http://www.callcenternews.com . Good overviews of current approaches to contact-center management can be obtained from Anton [4], Barber et al. [7] and Cleveland and Mayben [17].

1.5.4. Stochastic Models

As part of the long research tradition, it has been standard to use stochastic models, especially queueing models. The workhorse queueing models have been the Erlang B (loss) and C (delay) models, known as $M/M/s$ in the standard Kendall queueing notation. The most common extensions considered attempt to account for customer abandonment, customer retrials, non-exponential call-holding-time distributions and time-varying arrival rates, but even these familiar phenomena pose serious analysis challenges.

Largely stimulated by the rapid growth of contact centers, the more academic published research literature on contact centers is now growing rapidly, as can be seen from the recent survey on queueing models of call centers by Koole and Mandelbaum [37], the longer tutorial paper by Gans, Koole and Mandelbaum [22] and research bibliography by Mandelbaum [41]. Nevertheless, much remains to be done.

1.6. Research Directions

We describe research directions related to stochastic models of contact centers, aiming to uncover fundamental principles and develop performance-analysis tools that will make it possible to improve the design and management of contact centers. The goal is to improve the quality and the efficiency of service via contact centers. Indeed, we would like to simultaneously achieve both objectives: We would like to make it possible for service providers to provide higher quality service more efficently. However, we recognize that it is often necessary to balance these competing objectives. We would like to be able to consult and advise contact-center managers today, but our main aim is to uncover and address fundamental problems that meet long-term needs. We should seek additional exposure to operating contact centers, but primarily to gain a better understanding of the important fundamental issues. We want to significantly advance the state of the art of stochastic modelling in the area of customer contact centers.

2. Research Challenges

In this section we describe some of the exciting research challenges we see in developing and analyzing insightful stochastic models of contact centers.

2.1. Skill-Based Routing

2.1.1. Multiple Types

In modern call centers it is common to have multiple types of calls and multiple types of service reps. One way to classify calls is by language: With the globalization of many businesses, call centers often receive calls in several different languages. The callers and the service reps each may speak one or more of several possible languages, but not necessarily all of them. And, of course, it is not practical for the service reps to learn all the languages. (For research on methods to analyze multilingual telephone call centers, see Stanford and Grassmann [63].)

Another classification involves special promotions. The callers may be calling a special 800 number designated for the special promotion. Correspondingly, the service reps typically are trained to respond to inquiries about some of the available promotions, but not all of them. Learning about special promotions is certainly less difficult than learning entire languages, but it tends to be prohibitive to teach all service reps about all special promotions. Similarly, in technical (e.g., computer) help desks, service reps may only be able to help customers with some of their technical problems.

Thus, frequently, the calls have different requirements and the service reps have different skills. Fortunately, modern automatic call distributors (ACD's) have the capability of routing calls to service reps with the appropriate skills. Thus the call-center ICT equipment can allow for the generalization to multiple types. That capability of ACD's is called *skill-based routing* (SBR).

In fact, there may be additional flexibility. First, there may be multiple skill levels: Some skills are primary skills, while other skills are secondary skills. The idea is that calls requiring a particular skill should first be routed to a service rep with that skill as a primary skill. Only if all service reps with the skill as a primary skill are busy should service reps with that skill as a secondary skill be considered. Thus, the skill level provides a service priority for assigning calls to service reps.

Second, several different call centers, each with their own ACD, might be linked together to form a network of call centers. There is thus the possibility of rerouting traffic from one of these call centers to another in order to better balance the loads.

2.1.2. The Challenge: To Do Skill-Based Routing Well

However, it is difficult to manage the multi-type system effectively. First, for any given collection of service reps, it is difficult to do the routing of calls to service reps in an effective manner, let alone an optimal manner. With the ICT equipment, it is possible to obtain relevant data for making the routing decision, but so far only rudimentary algorithms are used. While there are skill-based-routing (SBR) algorithms in place, there remains a great opportunity for devising better routing algorithms.

Second, it remains to determine how many service reps with special skill sets are needed; i.e., there is a more complicated staffing problem in the multi-type setting. The problem of forecasting is much more difficult. With multiple types, we need to predict, not only the total call arrival rate, but also the call arrival rate for each call type. Moreover, with the promotions, the demand may be transient, so that there may be little history to rely upon.

Even after estimating the call arrival rates by type, there remain difficulties in determining the appropriate staffing levels in different categories. Of course, the long-term staffing problem is connected to the short-term routing problem, so that a full solution requires solving both problems, not just one.

Garnett and Mandelbaum [24] conducted a simulation study to show some of the operational complexities of skill-based routing. Research on the difficult problem of skill-based routing has been done by Perry and Nilsson [56], Koole and Talim [38] and Borst and Seri [13]. There is clearly room for much more research.

2.2. Resource Pooling

2.2.1. The Goal: Efficiency with Multiple Types

An extreme case with different skills is to have separate groups of service reps dedicated to each different skill set. Then, instead of one big call center, we would effectively have a large number of much smaller independent call centers. Upon each call arrival, the call can be routed to the appropriate group of service reps. With a Poisson arrival process to the full call center, and with independent Bernoulli routing to the subgroups, the separate arrival processes to the subgroups become mutually independent Poisson processes. So, indeed, the smaller groups do act as separate independent call centers.

The difficulty with this partitioning of the center into smaller subcenters is that it tends to be much less efficient. In operation, some of the smaller service groups are likely to be overloaded, while others are underutilized. Indeed, it is well known that larger service groups are more efficient. For some supporting theory establishing the efficiency of larger service groups (under regularity conditions), see Smith and Whitt [60] and Whitt [74]. The paper [60] exploits stochastic-order relations to show that performance improves with scale, e.g., when two systems with common customer holding-time distributions are combined. When the customer holding-time distributions are very different, partitions can be advantageous, as with supermarket express checkout lanes. For a study investigating how to determine good partitions, when partitions of the servers are appropriate, see Whitt [80].

By having more flexible service reps with multiple skills, it is possible to meet the special needs of customers and yet still have some of the efficiency of a single large call center. In fact, a small amount of flexibility may go a long way. It may be possible to achieve almost the full efficiency of a single service group by providing only a modest amount of overlap. That general efficiency phenomenon is called *resource pooling.*

2.2.2. Insights from Stochastic-Process Limits

Significant early theoretical work exposing the resource pooling phenomenon was done by Azar et al. [6], Vvedenskaya et al. [68], Turner [65] and Mitzenmacher [53]. An abstract problem considered in [68] involves assigning arriving customers, immediately upon arrival, to one of several identical queues. In the model there is a single stream of customers arriving in a Poisson process. Upon arrival, each customer must join one of a large number of independent, identical single-server queues with exponential service times and unlimited waiting space. The standard approach is to examine the state of all the queues and join one of the queues with the fewest customers. Indeed, the join-the-shortest-line rule is optimal [85]. (However, optimality can cease if some of the assumptions are changed, e.g., if the service-time distribution need not be exponential [72].) A difficulty with the join-the-shortest-queue rule, however, is that it may require obtaining a large amount of state information. In particular, we need to know the number of customers in each of the queues, which may be difficult if there are many widely-distributed queues.

So it is natural to consider alternative approaches that require less information. A simple alternative is to assign each arrival randomly to one of the queues, with each server being selected having equal chance. However, as noted above, a random assignment makes the individual servers act as separate $M/M/1$ queues. If we assign successive arrivals randomly to one of the servers, then we lose all the efficiency of the multiserver system.

The key idea in [6] [68], [65] [53] is to allow just a little choice. In particular, it suffices to compare just two queues: An alternative "lightweight" approach is to pick two of the queues at random and join the shorter of those two. The remarkable fact is that this lightweight "low-information" decision rule does almost as well as the join-the shortest-queue rule, requiring full information, which in turn does nearly as well as the large combined group of servers with a single queue, serving in first-come first-served order. That conclusion was formalized by considering the asymptotic behavior as the number of servers gets large (with the arrival rate kept proportional).

Thus, mathematical analysis reveals the possibililty for effective resource pooling with relatively little flexibility. Further related research on resource pooling has been done by Mandelbaum and Reiman [46], Mitzenmacher [54], Mitzenmacher and Vöcking [55], Turner [66] [67] and Williams [84].

It remains to establish comparable theoretical results for skill-based routing in call centers. In the context of skill-based routing in call centers, it remains to determine how much efficiency can be achieved by various amounts of flexibility (service skill overlap). Both mathematical analyses and simulation studies would be interesting.

2.3. More Insights from Stochastic-Process Limits

As indicated above, the mathematical results supporting resource pooling in multiserver service systems primarily have been established by considering the asymptotic regime in which the number of servers is allowed to approach infinity. More generally, we believe that important insights about queueing systems can be gained from such *stochastic-process limits* [83]. For understanding the performance of call centers, it is natural to focus on the behavior of multiserver service systems in which the number of servers is allowed to grow.

For the standard Erlang C ($M/M/s$ delay) model and its $GI/M/s$ generalization having a general renewal arrival process, such asymptotic results were established by Halfin and Whitt [30]. They obtained a useful limiting regime by letting the number of servers, s, approach infinity, while letting the probabilty of delay approach a limit strictly between 0 and 1. They showed that those conditions are equivalent to letting the traffic intensity, ρ, approach 1 while $s \to \infty$, so that $(1-\rho)\sqrt{s} \to \beta$ for some positive constant β. That limit in turn quantifies the economies of scale in an important way. In that limiting regime, a properly scaled version of the queue-length process converges to a tractable diffusion process, and the associated sequence of properly scaled steady-state distributions converges to a tractable nondegenerate limiting distribution, providing the basis for useful approximations; see Chapter 10 of Whitt [83] for a review.

More recently, others have obtained additional limiting results in the "Halfin-Whitt" limiting regime. The generalization to the $GI/PH/s$ model, having a more general phase-type service-time distribution, was established by Puhalskii and Reiman [57]. Generalizations accounting for costs and customer abandonment were established by

Borst et al. [12] and Garnett et al. [25].

Corresponding results for Erlang loss systems were obtained by Jagerman [33], Borovkov [10] [11] and Srikant and Whitt [61], as discussed in [71] and [61]. Related limits for queueing networks, state-dependent queues and time-dependent queues were established by Armony and Maglaras [5], Borst et al. [12], Mandelbaum et al. [42] [43] and Mandelbaum and Pats [45].

Even though, by now, there is a substantial literature on many-server asymptotics, it remains a fruitful direction for research. There may even be interesting new limiting regimes!

2.4. Workforce Management: Staffing

Even in the relatively simple case in which each service rep can handle all calls, there are challenges in staffing call centers. To understand the staffing challenges, it is helpful to identify *three different time scales for staffing*: real-time, short-term and long-term.

2.4.1. Real-Time Staffing

By *real-time staffing*, we mean dynamic staffing in the time scale of the length of a single call, done in response to observed system state, including information about the history of the call center on that day and information about calls currently in process. The idea is to have sufficient flexibility to be able to add service reps when they are needed and to pull them off to do alternative work when they are not needed. Of course, call-center managers routinely do some form of real-time staffing, but systematic real-time staffing based on substantial data and analysis so far is only a dream.

Whitt [78] discusses how real-time staffing might be done. Given that there is indeed the necessary flexibility in staffing, the idea is to exploit available information. First, the ICT equipment makes it possible to obtain the information. For each call, we can know the calling and called parties, we know the service rep handling the call, and we can know much about the purpose of the call. Thus we can often classify the call.

By carefully classifying the call in process and by using the known length of time that call has already been in service, it may be possible to predict how long the call will remain in service before service is complete. Or, more precisely, it may be possible to predict the conditional probability distribution of the remaining call holding time, given the information. And that conditional distribution may be much more predictive than the unconditional call holding-time distribution.

By combining such information over many calls and service reps, it may be possible to accurately predict the service demand in the near future, e.g., 3-30 minutes in the future, thus providing a basis for real-time staffing. The paper [78] shows how stochastic models can be exploited to facilitate the process. The idea deserves to be explored more fully.

Real-time staffing places great challenges upon queueing theory, because it requires that we consider the time-dependent behavior of the queueing system. Usually, performance analysis is confined to a description of the steady-state behavior. However, recent research has begun seeking algorithms to describe the time-dependent behavior of queueing systems. Particularly promising in this regard are applications of numerical transform inversion; e.g., see Abate, Choudhury and Whitt [1], Abate and Whitt [2] [3], Choudhury et al. [16] and Whitt [76]. Further exploration of that approach is warranted.

2.4.2. Short-Term Staffing

By *short-term staffing*, we mean the daily staffing done in reponse to forecasted demand and knowledge of the available agents. By *long-term staffing*, we mean the staffing done in the time scale of the length of time required to hire and train service reps. The time required to hire a new service rep is filled with uncertainty and might need to be analyzed probabilistically. Under favorable circumstances, the time to hire may be relatively short. However, even if the time to hire is short, the time to train often is a significant bottleneck, preventing the call center from responding quickly (in the time scale of days or weeks) to a perceived increase in demand.

A significant challenge in short-term staffing is that the call arrival rate varies significantly over the day. In some cases, the call holding times are sufficiently short that the time-dependence can be safely ignored. Then, it is appropriate to use a dynamic steady state, using the parameters that are appropriate at any time instant (a short-term average), rather than the long-run average parameters over an entire day. However, we actually do not understand well enough when each kind of analysis is appropriate. We would project that research in this area is now only in its infancy.

In some instances of short-term staffing, it is clearly important to directly analyze the system with a time-varying arrival rate. And, indeed, a significant effort has been made to do so; e.g., see Abate and Whitt [2], Green et al. [27] [28], Grier et al. [29], Jennings et al. [34], Mandelbaum et al. [42] [44] [45], Massey and Whitt [50] [51] and Whitt [79]. Again, further research in that direction is warranted.

A significant component of short-term staffing is scheduling work shifts for the service reps, including breaks for coffee and lunch. Mathematical programming has been successfully applied to shift scheduling for some time; e.g., see Segal [59].

2.4.3. Long-Term Staffing

There are different challenges in long-term staffing when it takes a relatively long time to train new service reps. Then training of new service reps becomes an important decision variable, which might be exploited in a dynamic programming approach. With a large variety of skills, there are correspondingly a large number of training targets.

Over the longer time scale, it is also important to address service-rep attrition and service-rep career paths. The ideal is to have both satisfied service reps and satisfied customers. Gans and Zhou [23] have recently developed a Markov decision process model for call-center staffing in which they address learning and turnover. They also cite earlier related literature for the long-term staffing problem.

2.5. Different Sources of Uncertainty

In any setting where stochastic models are being applied, it is important to consider what are the most important sources of uncertainty. That seems very important for contact centers.

2.5.1. Sources of Uncertainty for the Arrival Process

For example, it is natural to model the arrival of initial service requests as a Poisson process, possibly with a time-varying arrival rate. What are the sources of uncertainty? First, there is *model uncertainty*, because the arrival process might actually not be well modeled by a Poisson process; e.g., see Jongbloed and Koole [35]. It is appropriate to investigate how well the model fits.

Given that a Poisson process model is deemed appropriate, there remains *parameter uncertainty*. In practice we do not know the arrival-rate function and must estimate it. The parameter uncertainty is addressed by doing forecasting. With a time-varying arrival rate, the estimation problem is more complicated, because we must estimate the entire arrival-rate function, not just one or two parameters. One can return to the parametric setting by assuming a parametric form for the arrival-rate function, such as linear or quadratic. Massey et al. [49] investigate ways to estimate the coefficients of linear arrival-rate functions from nonhomogeneous-Poisson-process data. Further study is warranted.

In the call-center setting, there are a variety of data sources to use for the forecasting, so the forecasting is more complicated. It is natural to investigate the quality of the forecasting. We should aim to quantify the error resulting from forecasting. For settings in which the error is significant, it may be appropriate to include that error in the analysis. Jongbloed and Koole [35] show how congestion can be described when there is uncertainty about the arrival rate in the Erlang C model.

Given that a Poisson process model is deemed appropriate and the arrival-rate function has been estimated by forecasting (and is thus presumed known), there remains the *process uncertainty*, i.e., the inherent randomness associated with the fully specified Poisson-process model. For example, the number of arrivals in any given time interval is not known in advance, but instead is random, having a Poisson distribution (with a variance always equal to its mean, which can serve as the single parameter of the Poisson distribution). The process uncertainty gives us the random variable instead of its realization, which is unknown when we are performing the analysis.

An interesting research problem is to compare these three forms of uncertainty - model uncertainty, parameter uncertainty and process uncertainty - in actual application settings. First, we want to know if one form of uncertainty is dominant; then it pays to pay most attention to it. More generally, we want to quantify the contribution to the total uncertainty due to each component. That requires a view of the entire process, rather than just one component. That may require developing some new methods as well.

2.5.2. Sources of Uncertainty for the Service Process

In addition to uncertainty about arrivals, there is uncertainty about the service process. The basic model for customer telephone call holding (service) times is a sequence of independent and identically distributed (IID) random variables, each with an exponential distribution, but as with the arrival process, there is model uncertainty. For example, the actual service-time distribution might not be exponential. In various studies, Bolotin [9] made a case for a lognormal distribution and a mixture of lognormal distributions. Moreover, the holding-time distribution may depend upon the call type, the service rep and the time of day.

Just as with the arrival process, there is parameter uncertainty about the call-holding-time distribution. Given that an exponential distribution is deemed appropriate (possibly after conditioning upon the call type, the service rep and the time of day), it remains to estimate the single parameter of the exponential distribution, which can be done by analyzing historical data. In cases of special promotions, however, there may be very little historical data. Thus, parameter uncertainty may be significant for call holding times too.

Finally, there is process uncertainty for the holding times as well. Given that an exponential distribution is deemed appropriate for the call holding times and the single exponential parameter is known, there remains the inherent randomness of the exponential distribution: The service times are realizations from the exponential distribution.

2.5.3. Sources of Uncertainty for the Whole Model

When we put the model primitives together to obtain a model for the performance of the entire call center, we again have the three forms of uncertainty described above, namely, model uncertainty, parameter uncertainty and process uncertainty. Unfortunately, the relative importance of these different sources of uncertainty is not nearly as well understood as it should be. It is natural to ask which form of uncertainty is most critical in any application.

And there are still other forms of uncertainty that should be considered. For example, poor performance might occur because of system failures. Equipment might break down; service reps might get sick. For example, on the Internet it is not yet clear whether the primary cause of congestion is the occurrence of occasional bursts of customer demand or the occurrence of occasional system failures.

In the context of contact centers, the goal is to better understand these different sources of uncertainty and to determine the importance of each.

2.6. Simulation

2.6.1. An Important Performance Analysis Tool

Since contact centers are complicated, especially with skill-based routing, it is natural to rely on simulation to analyze the resulting stochastic models. And, indeed, simulation has often been applied to analyze call-center models. A simulation tool called the Call Processing Simulator (CAPS) was created and extensively applied at AT&T [14].

Simulation remains an attractive approach to predict and control contact center performance, but there are challenges to make simulation more effective. There is a need for new efficency-improvement techniques, so that the complex systems can be analyzed effectively. There is a need for more sophisticated simulation output analysis in order to understand the statistical precision of estimates. And there is a need for analysis tools to help simulators plan their simulation experiments.

2.6.2. Planning Contact-Center Simulations

Since contact-center simulations may need to be done quite frequently, it is natural to try to determine how best to carry out these simulations. For both natural methods and clever indirect methods, it is natural to try to determine how long the simulation runs need to be in order to estimate the performance measures of interest with any desired statistical precision. Moreover, it is natural to seek variance-reduction or efficiency-improvement techniques to reduce the required run lengths for any given desired statistical precision.

Whitt [73] applied diffusion approximations to obtain simple heuristic formulas to estimate the simulation run length required to obtain any desired statistical precision when simulating a single-server queue. That analysis also can be applied to networks of single-server queues and multiserver queues with relatively few servers, but it does not apply to queues with a large number of servers, and thus not to most contact centers.

Srikant and Whitt [61] carried out a similar program for estimating the long-run average blocking probability in a multiserver loss model. First they identified two different methods that are more efficient in different regions. The direct method, which is just the ratio of the observed losses to the observed arrivals, is much more efficient under light load, while an indirect method based on estimating the mean occupancy, and applying Little's conservation law, $L = \lambda W$, is much more efficient under heavy loads. For both methods, relatively simple formulas were determined for the asymptotic variance, which in turn leads to simple formulas for the required simulation run length. Interestingly, the formulas for the multiserver loss models differ dramatically from the corresponding formulas for the single-server models.

Srikant and Whitt [62] also continued to look for even better methods that produced greater variance reduction. In [62] they showed that a combined estimator, constructed by taking a weighted combination of the direct estimator and the indirect estimator, provided additional variance reduction.

It is natural to investigate the same questions for multiserver delay models and their more complicated contact-center relatives. Intuitively, it would appear that the required simulation run lengths for multiserver delay models would be similar to the required run lengths for the corresponding multiserver loss models, but the analysis is yet to be done. Moreover, it is natural to conjecture that the required run lengths for more complicated contact-center models would be similar to the required run lengths for the multiserver delay models, but again the analysis is yet to be done. Those are promising directions of research.

2.6.3. Operational Real-Time Simulation

Simulation is almost always applied off line to make system studies. However, with the steady increase of computer power, it is becoming possible to perform simulations dynamically in real time to predict and control congestion. For contact centers, the idea is to exploit the ICT equipment to obtain the current system state, and then perform multiple replications of simulations over a short-term horizon in order to predict the transient behavior for various scenarios of interest.

To evaluate the performance of real-time simulation, we can perform a system simulation accompanied by additional replications of transient simulations performed through simulation time in the main run. We hope to investigate these "simulations with nested simulations."

2.7. Multimedia

Given that contact with customers no longer need be by telephone, it is necessary to consider the other media - email, fax, Internet, etc. The new media perform in different ways, requiring new stochastic models.

2.7.1. Work that can be Postponed

When the new media are introduced, it may happen that the service reps that were only answering the telephone are now required to repond to service requests arriving by the other media as well. With e-mail and Internet, the required response might be regarded as instantaneous, so that the new job can be assigned to an available service rep, just as if it were a telephone call, with the service rep required to do only one thing at a time. However, in some cases, e.g., with fax, the response may not be required to be instantaneous. Indeed, for some of the media there may be much more flexibility in the required response time.

In fact, service reps in traditional call centers have already been faced with several kinds of alternative work, some of which is quite non-time-critical. First, in many call centers the service reps are required to perform "after-call work," such as completing the entry of customer orders. Second, different forms of work exist in call centers that have both inbound and outbound calls. The inbound calls require immediate response, but the timing of the outbound calls is much more flexible. The question of how to allocate service rep time between inbound and outbound calls is known as the "call-blending problem;" e.g., see Bhulai and Koole [8].

When there are several forms of work arriving, some of which can be postponed, there is an opportunity to smooth the workload by postponing work that can be postponed when the workload is relatively high. Whitt [81] showed how relatively simple models can be used to estimate the reduced overall required capacity that is needed to meet demand, when some of the less-time-critical work can be postponed. Armony and Maglaras [5] consider the possibility of a call-back option. With the new multimedia contact centers, the possibility of postponing some work is worth more careful study.

2.7.2. More Complicated Contact Experiences

Until recently, the customer would make their service requests over a single medium. The customer might call on the telephone or make an Internet order, but the customer would use ony a single medium at any one time to carry out the transaction. In a multimedia contact center, the service reps might be working with several different media, but they would be using each only one at a time, because customers were using only a single medium.

However, that is rapidly changing. Multimedia applications where audio, video, graphics and text are synchronized are becoming very important. The most familiar example is Macromedia Flash Player, evidently present in 97% of all browsers. Work is underway to replace HTML and XML with the programming infrastructure to support multimedia applications. Prominent approaches are the Synchronized Multimedia Integration Language (SMIL) and Speech Application Language Tags (SALT); see http://www.w3.org/AudioVideo and http://www.saltforum.org .

Thus, soon, the customer and the service rep will be able to simultaneously be using multiple media to carry out the transaction. A customer might be on the Internet looking at an online catalog and then, while connected to the Internet, place a telephone call, and so be in contact with the service rep simultaneously by telephone and the Internet. Alternatively, the session might start on the Internet and then a telephone connection might be added.

When both the customer and the service rep are simultaneously in multimedia contact, the customer contact experience can get much more complicated. It now becomes possible for service reps to time-share over several different customers that they are in contact with. The customer might alternate between interactions with the service rep, either on the telephone or via one of the other media, and periods of browsing on the Internet, just as customers do in large department stores.

In large department stores, the salesperson now typically needs to be in contact at the final step in order to complete the sale at the cash register, but that final step can be avoided with a remote contact center, because the sale can be completed by standard procedure over the Internet. The function of the service rep is escalated to a higher level, with the routine low-value-added tasks being eliminated (done by computers). In the emerging multimedia contact center, service reps will be only providing customers important information, helping them achieve their objectives.

It should be evident that it is much more challenging to construct suitable stochastic models to help design and manage contact centers when customers have such more elaborate contact experiences. For one thing, the customer contact time associated with one contact experience may not be a simple interval. It will be interesting to model even a single contact experience!

2.8. Non-Temporal, Non-Congestion Measures of Quality

Call-center performance is routinely characterized by the numbers of calls handled, blocked and abandoned, and by associated averages - the percentage of calls answered, average speed to answer (ASA) and average after call handling time. The attention is focused on throughput and congestion. A common delay target is the 80/20 rule - to have 80% of the calls answered within 20 seconds.

However, for effective management, it is important to also consider non-temporal, non-congestion measures of service quality. The obvious one-dimensional alternative is cost (actually revenue minus cost). Significant steps in that direction were made by Borst et al. [12] and Garnett et al. [25]; they included costs in many-server asymptotics.

From the customer's perspective, there is more: There is the quality of service received. How do we rate the quality of the service experience received [69]? It seems appropriate to go beyond congestion and cost. One candidate, but elusive, measure is brand loyalty, which might increase with good service experiences and decrease with bad ones. Brand loyalty should change in a slower time scale and significantly affect the customer service-request rates. Focusing on brand loyalty could capture effects that are important over longer time scales.

2.9. The Human Element

Perhaps the most attractive feature of contact centers, making them so interesting to study, is the central role played by people. The customers are people and the service reps are people. We can easily relate to contact centers because we ourselves often are customers of contact centers.

Given the important human element, it is natural to try to better understand human behavior in contact centers. It is interesting to study the psychology of queueing [39] [21]. It is interesting to consider why service reps might slow down in certain circumstances, as Sze observed [64]. It is interesting to investigate why customers abandon and retry. Such studies have been made at AT&T, but little if any are published. Recently, Avishai Mandelbaum and his colleagues have made a serious study of customer abandonment [48] [86] [47].

3. Integration of Research and Education

It is desirable to have a research program that helps provide valuable and exciting educational experiences for students, both graduate and undergraduate. This paper describes many possible research directions, which should be suitable for students with varying backgrounds.

Contact centers can be made a context focus in courses on stochastic models and simulation, to provide additional motivation for learning the technical material. Contact-center simulations can make good simulation course projects. By having students read the trade literature, visit operational contact centers and work with contact-center data, we can better prepare students for future work and motivate the value of the underlying theory that we focus on.

References

[1] Abate, J., G. L. Choudhury and W. Whitt (1999) An introduction to numerical transform inversion and its application to probability models. *Computational Probability*, W. Grassman (ed.), Kluwer, Boston, 257-323.

[2] Abate, J. and W. Whitt (1998) Calculating transient characteristics of the Erlang loss model by numerical transform inversion. *Stochastic Models* 14, 663-680.

[3] Abate, J. and W. Whitt (1999) Computing Laplace transforms for numerical transform inversion via continued fractions. *INFORMS Journal on Computing* 11, 394-405.

[4] Anton, J. (2000) The past, present and future of customer access centers. *International Journal of Service Industry Management* 11, 120-130.

[5] Armony, M. and C. Maglaras. 2001. On customer contact centers with a call-back option: customer decisions, sequencing rules, and system design. New York University.

[6] Azar, Y., A. Broder, A. Karlin and E. Upfal (1994) Balanced allocations. *Proceedings of the 26th ACM Symposium on the Theory of Computing (STOC)* 593-602.

[7] Barber, G., B. Cleveland, H. Dortmans, G. Levin, G. MacPherson and A. Smith (2000) *Call Center Forecasting and Scheduling: The Best of Call Center Management Review*, Call Center Press, Annapolis, MD.

[8] Bhulai, S. and G. M. Koole (2000) A queueing model for call blending in call centers. *Proceedings of the 39th IEEE CDC* 1421-1426.

[9] Bolotin, V. A. (1994) Telephone circuit holding time distributions. *Proceedings of the 14th International teletraffic Congress, ITC-14*, 125-134.

[10] Borovkov, A. A. (1976) *Stochastic Processes in Queueing Theory*, Springer, New York.

[11] Borovkov, A. A. (1984) *Asymptotic Methods in Queueing Theory*, Wiley, New York.

[12] Borst, S. C., A. Mandelbaum and M. I. Reiman (2000) Dimensioning large call centers. Bell Laboratories, Lucent Technologies, Murray Hill, NJ.

[13] Borst, S. C. and P. Seri (2000) Robust algorithms for sharing agents with multiple skills. Bell Laboratories, Murray Hill, NJ.

[14] Brigandi, A. J., D. R. Dargon, M. J. Sheehan and T. Spencer (1994) AT&T's call processing simulatior (CAPS): Operational design of inbound call centers. *Interfaces* 24, 6-28.

[15] Champy, J. A. and M. Hammer (2001) *Reingineering the Corporation: A Manifesto for Business revolution*, second edition, Harper Business Books, New York.

[16] Choudhury, G. L., D. M. Lucantoni and W. Whitt (1997) Numerical solution of $M_t/G_t/1$ queues. *Operations Research* 45, 451-463.

[17] Cleveland, B. and J. Mayben (1997) *Call Center Management on Fast Forward*, Call Center Press, Annapolis, MD.

[18] Duffield, N. G. and W. Whitt (1997) Control and recovery from rare congestion events in a large multi-server system. *Queueing Systems* 26, 69-104.

[19] Durinovic-Johri, S. and Y. Levy (1997) Advanced routing solutions for toll-free customers: algorithm design and performance. *Proceedings of the International Teletraffic Congress, ITC-15*, Elsevier, Amsterdam, 157-167.

[20] Falin, G. and J. G. Templeton (1997) *Retrial Queues*, Chapman and Hall.

[21] Gail, T. and L. Scott (1997) A field study investigating the effect of waiting time on customer satisfaction. *Journal of Psychology*, November 1997, 655-660.

[22] Gans, N., G. Koole and A. Mandelbaum (2002) Telephone call centers: A tutorial and literature review. The Wharton School, University of Pennsylvania. To appear

[23] Gans, N. and Y. Zhou (1999) Managing learning and turnover in employee staffing. The Wharton School, University of Pennsylvania.

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

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

[26] Green, L. and P. Kolesar (1989) Testing the validity of a queueing model of police patrol. *Management Science* 35, 127-148.

[27] Green, L. V., P. J. Kolesar and J. Soares (2001) Improving the SIPP approach for staffing service systems that have cyclic demands. *Operations Research* 49, 549-564.

[28] Green, L. V., P. J. Kolesar and J. Soares (2001) An improved heuristic for staffing telephone call centers with limited operating hours. Graduate School of Business, Columbia University.

[29] Grier, N., W. A. Massey, T. McKoy and W. Whitt (1997) The time-dependent Erlang loss model with retrials. *Telecommunication Systems* 7, 253-265.

[30] Halfin, S. and W. Whitt (1981) Heavy-traffic limits for queues with many exponential servers. *Operations Res.* 29, 567-588.

[31] Harris, C. M., C. L. Hoffman and P. B. Saunders (1987) Modeling the IRS telephone texpayer information system. *Operations Research* 35, 504-523.

[32] Hassin, R. and M. Haviv. 1995. Equilibrium threshold strategies: the case of queues with priorities. *Operations Research* 45, 966-973.

[33] Jagerman, D. L. (1974) Some properties of the Erlang loss function. *Bell System Technical Journal* 53, 525-551.

[34] Jennings, O. B., A. Mandelbaum, W. A. Massey and W. Whitt (1996) Server staffing to meet time-varying demand. *Management Science* 42, 1383-1394.

[35] Jongbloed, G. and G. M. Koole (2000) Managing uncertainty in call centers using Poisson mixtures. Free University, Amsterdam.

[36] Kogan, Y., Y. Levy and R. A. Milito (1997) Call routing to distributed queues: Is FIFO really better than MED? *Telecommunication Systems* 7, 299-312.

[37] Koole, G. and A. Mandelbaum (2001) Queueing models of call centers: an introduction. Free University, Amsterdam.

[38] Koole, G. M. and J. Talim (2000) Exponential approximation of multi-skill call centers architecture. *Proceedings of QNETs 2000*, Ilkley, UK, 23, 1-10.

[39] Larson, R. C. (1987) Perspectives on queues: social justice and the psychology of queueing. *Operations research* 895-905.

[40] Levy, Y., S. Durinovic-Johri and R. A. Milito (1994) Dynamic network call distribution with periodic updates. *Proceedings of the Fourteenth International Teletraffic Congress, ITC-14*, Elsevier, Amsterdam, 85-94.

[41] Mandelbaum, A. (2001) Call Centers (Centres): Research bibliography with abstracts. Faculty of Industrial Engineering and Management, The Technion, Haifa.

[42] Mandelbaum, A., W. A. Massey and M. I. Reiman (1998) Strong approximations for Markovian service networks. *Queueing Systems* 30, 149-201.

[43] Mandelbaum, A., W. A. Massey, M. I. Reiman, B. Rider and A. Stolyar (2000) Queue length and waiting times for multiserver queues with abandonments and retrials. *Proceedings of the 5th INFORMS Telecommunications Conference*, Boca Raton, FL.

[44] Mandelbaum, A., W. A. Massey, M. I. Reiman and B. Rider (1999) Time varying multiserver queues with abandonments and retrials. *Proceedings of the 16th International Teletraffic Congress*, P. Key and D. Smith (eds.).

[45] Mandelbaum, A. and G. Pats (1995) State-dependent queues: approximations and applications. *Stochastic Networks*, IMA Volumes in Mathematics and its Applications, F. P. Kelly and R. J. Williams, eds., Springer, New York, 239-282.

[46] Mandelbaum, A. and M. I. Reiman (1998) On pooling in queueing networks. *Management Science* 44, 971-981.

[47] Mandelbaum, A., Sakov, A. and S. Zeltyn (2001) Empirical analysis of a call center. The Technion, Haifa, Israel.

[48] Mandelbaum, A. and N. Shimkin (2000) A model for rational abandonments from invisible queues. *Queueing Systems* 36, 141-173.

[49] Massey, W. A., G. A. Parker and W. Whitt (1996) Estimating the parameters of a nonhomogeneous Poisson process with a linear rate. *Telecommunication Systems* 5, 361-388.

[50] Massey, W. A. and W. Whitt (1993) Networks of infinite-server queues with non-stationary Poisson input. *Queueing Systems* 13, 183-250.

[51] Massey, W. A. and W. Whitt (1997) Peak congestion in multi-server service systems with slowly varying arrival rates. *Queueing Systems* 25, 157-172.

[52] Milito, R. A., Y. Levy and Y. Arian (1991) Dynamic algorithms for distributed queues with abandonments. *Proceedings of the Thirteenth International Teletraffic Congress*, North-Holland, Amsterdam, 329-334.

[53] Mitzenmacher, M. (1996) *The Power of Two Choices in Random Load Balancing*, Ph. D. dissertation, University of California at Berkeley.

[54] Mitzenmacher, M. (1997) How useful is old information? *Proceedings of the 16th ACM Symposium on Principles of Distributed Computing*, 83-91.

[55] Mitzenmacher, M. and B. Vöcking (1999) The asymptotics of selecting the shortest of two, improved. *Proceedings of the 1999 Allerton Conference on Communication, Control and Computing*, University of Illinois.

[56] Perry, M. and A. Nilsson (1992) Performance modeling of automatic call distributors: assignable grade of service staffing. *Proceedings of the 14th International Switching Symposium*, 294-298.

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

[58] Samuelson, D. A. (1999) Predictive dialing for outbound telephone call centers. *Interfaces* 29, 66-81.

[59] Segal, M. (1974) The operator-scheduling problem: a network-flow approach. *Operations Research* 22, 808-823.

[60] Smith, D. R. and W. Whitt (1981) Resource sharing for efficiency in traffic systems. *Bell System Technical Journal* 60, 39-55.

[61] Srikant, R. and W. Whitt (1996) Simulation run lengths to estimate blocking probabilities. *ACM Transactions on Modeling and Computer Simulation* 6, 7-52.

[62] Srikant, R. and W. Whitt (1999) Variance reduction in simulations of loss models. *Operations Research* 47, 509-523.

[63] Stanford, D. A. and W. K. Grassmann (2000) Bilingual server call centres. *Analysis of Communication Networks: Call Centres, Traffic and Performance*, D. R. McDonald and S. R. E. Turner (eds.), Fields Institute Communications 28, The American Math. Society, Providence, RI, 31-48.

[64] Sze, D. Y. (1984) A queueing model for telephone operator staffing. *Operations Research* 32, 229-249.

[65] Turner, S. R. E. (1996) *Resource Pooling in Stochastic Networks*, Ph.D. dissertation, University of Cambridge.

[66] Turner, S. R. E. (1998) The effect of increased routing choice on resource pooling. *Probability in the Engineering and Informational Sciences* 12, 109-124.

[67] Turner, S. R. E. (2000) A join the shortest queue model in heavy traffic. *Journal of Applied Probability* 37, 212-223.

[68] Vvedenskaya, N. D., R. L. Dobrushin and F. I. Karpelovich (1996) Queueing systems with selection of the shortest of two queues: an asymptotic approach. *Problems in Information Transmission* 32, 15-27.

[69] Waite, A. J. (1996) *Customers Arriving with a History and Leaving with an Experience*, Telecom Books, New York.

[70] Weinberg, A. (2002) E-commerce love letter. http://www.Forbes.com/technology/ecommerce/2002/02/20/0220valentine.html

[71] Whitt, W. (1984) Heavy-traffic approximations for service systems with blocking. *AT&T Bell Laboratories Technical Journal* 63, 689-708.

[72] Whitt, W. (1986) Deciding which queue to join: some counterexamples. *Operations Research* 34, 55-62.

[73] Whitt, W. (1989) Planning queueing simulations. *Management Science* 35, 1341-1366.

[74] Whitt, W. (1992) Understanding the efficiency of multi-server service systems. *Management Science* 38, 708-723.

[75] Whitt, W. (1993) Approximations for the $GI/G/m$ queue. *Production and Operations Management* 2, 114-161.

[76] Whitt, W. (1999) Improving service by informing customers about anticipated delays. *Management Science* 45, 192-207.

[77] Whitt, W. (1999) Predicting queueing delays. *Management Science* 45, 870-888.

[78] Whitt, W. (1999) Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters* 24, 205-212.

[79] Whitt, W. (1999) Decomposition approximations for time-dependent Markovian queueing networks. *Operations Research Letters* 24, 97-103.

[80] Whitt, W. (1999) Partitioning customers into service groups. *Management Science* 45, 1579-1592.

[81] Whitt, W. (1999) Using different response-time requirements to smooth time-varying demand for service. *Operations Research Letters* 24, 1-10.

[82] Whitt, W. (2001) How multiserver queues scale with growing congestion-dependent demand. *Operations Research*, to appear. (http://www.research.att.com/~wow/demand3no.pdf)

[83] Whitt, W. (2002) *Stochastic-Process Limits*, Springer, New York.

[84] Williams, R. J. (2000) On dynamic scheduling of a parallel server system with complete resource pooling. *Analysis of Communication Networks: Call Centres, Traffic and Performance*, D. R. McDonald and S. R. E. Turner (eds.), Fields Institute Communications 28, The American Math. Society, Providence, RI, 49-72.

[85] Winston, W. (1977) Optimality of the shortest line discipline. *Journal of Applied Probability* 14, 181-189.

[86] Zohar, E., A. Mandelbaum and N. Shimkin (2000) Adaptive behavior of impatient customers in tele-queues: theory and empirical support. The Technion, Haifa, Israel.