Service Engineering (Science, Management)

Mini-Course / Workshop

] **Avi(shai) Mandelbaum**

Faculty of Industrial Engineering and Management
Technion - Israel Institute of Technology

e.mail: avim@tx.technion.ac.il
Office phone: (972) 4-829-4504
Course site: http://ie.technion.ac.il/serveng (References/Mini-Courses menu)

Contents

1 Introduction

This is a description of a graduate mini-course, that I believe fits the present research agenda of Business and Engineering Schools. Prerequisites consist of merely a course in Stochastic Processes, at any level, with the hope that it covered Markov Chains and Poisson Processes. Acquaintance with basic Queueing Theory is helpful but not essential. Following the course description, I add some notes that summarize my thinking about Services, Service Engineering (Science and Management), Queueing Theory (and Science), and the likes.

The subject of the course are *Service Networks*: these include public service centers (municipal, justice, government), telephone services (business and marketing, emergency, assistance), banks and insurance (front and back office), hospitals (emergency rooms, outpatient clinics, operating rooms, internal wards), airports, supermarkets, maintenance and field-service operations, some transportation systems, and even more. In many such systems, the network-view, as opposed to that of a one-stop service-station, is useful. For example, in a hospital a patient could start at the emergency room, then visit the x-ray department, move on to be hospitalized in an internal ward and then, possibly, be operated on. In a telephone call-center, a customer often starts with self-identification at the Interactive Voice Response (IVR) unit and then moves on to queue for an agent's service.

Significant motivators for the theory that will be described are tele-services, in which customers and servers are remote from each other. Communication in tele-services is through snail-mail, fax, electronic-mail, IVR, telephone and increasingly the Internet. However, existing tele-services are predominantly telephone-based, hence our heavy emphasis on *Telephone Call Centers*.

Successful design, analysis and management of services must often be multi-disciplinary, fusing ingredients from Operations Research, Statistics, Industrial Engineering, Sociology, Psychology, Game Theory, Economics, Management Information Systems, and more. The significant relevance of most of these disciplines will be clear throughout the course. However, my background and interests render my research, and hence my methodological lectures, biased towards service *operations*, viewing the latter through the mathematical lenses of a *queueing scientist*.

2 Course Description

The course will start with a self-contained seminar, which will also serve as a course-introduction. The seminar will be followed by two sessions, each divided into two lectures.

Under the time-constraint of a mini-course, I shall have time to only expose you to "what can be done" as opposed to actually "how to do it". In particular, mathematical theory will be always well motivated but then not often rigorously proved. (I trust the interested audience to complete the details in case of need.)

Within the seminar and lectures, I hope to be able to cover most of the following:

2.1 Seminar (Location and Time TBA)
"Data-Based Science for Service Engineering & Management" or: "Empirical Adventures in Call-Centers and Hospitals"

This seminar is a self-contained presentation, starting with a bird's-eye view of Service Engineering, Science and Management. Then some empirical findings of service systems will be presented, which motivate or are motivated by (or both) interesting research questions. These findings give rise to model-features that are essential to incorporate into useful service models. Examples include customers' (im)patience, time-varying service demand (predictable variability), heterogeneity of customers and servers (skills-based routing), over-dispersion in Poisson arrivals, generally-distributed (as opposed to exponential) service- and patience duration, and more. Empirical analysis also enables validation of prevalent models and protocols, either supporting or refuting their relevance and robustness.

My main data-source is a unique data repository, which is maintained at the Technion's SEE Laboratory (SEE = Service Enterprise Engineering:http://ie.technion.ac.il/Labs/Serveng/). It is unique in that it is transaction-based: it details the individual operational history of *all* service transactions (e.g. calls in a call center or patients in an emergency department). For example, one source of data is a network of 4 call centers of a U.S. east-coast bank, spanning close to 3 years and covering over 800 agents; there are 218,047,488 telephone calls overall, out of which 41,646,142 where served by agents, while the rest were handled by an answering machine (IVR = Interactive Voice Response).

Data-bases within most call centers are inadequate for operations research. Hence a universal data-structure (schema) had to be designed and implemented, under the heading **DataMOCCA** = Data Models for Call Centers Analysis. (See
http://iew3.technion.ac.il/serveng/References/DataMOCCA.pdf)
A friendly flexible user-interface is accompanying DataMOCCA, which we call **SEEStat**. This interface will be demonstrated throughout the seminar, in particular its ability to support *online* Exploratory Data Analysis (EDA) analysis, at resolutions that span the whole range from a single second through minutes, hours, weeks and up to months.

The SEE repositories and the scope of DataMOCCA and SEEStat have now been extended to accommodate *healthcare*, and are planned to expand to *internet* services. The underlying theme is again the archiving of transaction-based operational history of patients (eg. within the ED = Emergency Department) or surfers (within websites, via clickstream data), all interfacing with SEEStat for online analysis, similarly to call centers.

The above-mentioned U.S. Bank data, as well as SEEStat (the DataMOCCA inter-

face) and some relevant tutorials and documentation, are publicly available after registration: go to the SEELab server: http://seeserver.iem.technion.ac.il/see-terminal , then register, get a password, login, go over SEEStat's introductory tutorial, and you are now ready to experience your own empirical adventures.
Note: A full version of the raw (cleaned) U.S. Bank data takes close to 40GB. SEEStat, therefore, will be accessing, in real-time, "only" an abridged version of about 7GB. This is still a significant challenge, and I hope that you will appreciate the way SEEStat overcomes it. (Please let me know if you are interested in the complete 40GB data version.)

2.2 Lectures 1+2 (Location and Time TBA)

2.2.1 *Introduction to Services and Service Engineering (Science, Management).*

The ultimate goal of Service Engineering, as I perceive it, is to develop scientifically-based design principles and tools (often culminating in software), that support and balance service quality, efficiency and profitability, from the likely conflicting perspectives of customers, servers, managers, and often also society. I find that queueing-network models constitute a natural convenient nurturing ground for the development of such principles and tools. However, the existing supporting (Queueing) science of such models has been somewhat lacking. Hence, advances in Service-Engineering and Service-Science must go hand in hand. (Quoting from the lecture notes subsequent to this course description)

In the first lecture, I introduce Service Engineering" (and Science) through representative examples of service operations. I shall start with a detailed description of Service Engineering of a Telephone Call Center, immediately comparing it to Service Engineering of an Emergency Department. Then, the structure of the lecture will follow that of a course under the same title, taught at the Technion during recent years (http://ie.technion.ac.il/serveng). Biased by my background, I emphasize services in which the *queueing phenomena is a modeling-must*. Hence, descriptions of my service examples take the view-point of queueing-network models, but here they will be mostly informal and empirically-based. We are thus setting up the stage for the lectures to come (and for future readings on the subject, for those seeking deeper understanding).

Parts of this introductory lecture is devoted to some fundamental laws of congestion. In particular, I shall (tentatively) explain queue-drivers (scarce resources, synchronization gaps) in terms of fork-join networks (dynamic-stochastic project (PERT/CPM) networks), then possibly continue with empirical "proofs" of some classical congestion laws (Little, Khintchin-Pollatcheck, Kingman, Palm).

2.2.2 *Operational Regimes: QD, ED, QED.*

A tradeoff that achieves the "right" balance between service quality and efficiency is a fundamental operational challenge. (Recently, quality and efficiency have been interwoven also with profitability, but we shall leave this aside for now.) Typically, high levels of efficiency, as measured say through

high servers utilization, come at the cost of low service quality, for example long delays prior to service. But we shall discover and understand when alternative scenarios are feasible, given appropriately accommodating circumstances.

Queueing models are ideally suited to quantitatively capture the quality-efficiency tradeoff. I shall demonstrate this by introducing three operational regimes for a queueing system: Quality-Driven (QD) regime, where the organization focus is service-quality at the cost of efficiency, Efficiency-Driven (ED) where the focus is reversed, and an intermediate **QED** regime, where quality and efficiency are delicately balanced. Server *staffing* (determining service capacity) is the "knob" in terms of which an organization calibrates its operational preferences - a high staffing level relative to service demand is QD and low level is ED. Economies of scale enable an organization to be QED, with well-run medium-to-large call centers being prevalent convincing QED examples. Other examples include, perhaps surprisingly, subsystems of medium-to-large hospital which consist of, say, an Emergency Room plus the Internal Wards it is feeding; or parking systems in the downtown of large metropolitan areas. (The theory of Operational Regimes in Queueing Systems is by now rather advanced. For example, at least *ten* regimes have been developed by now for *stationary* queues, which could potentially double to *twenty* in *time-varying* environments.)

2.2.3 *A (Pre-)Basic Model for a Service Station: Erlang-C.*

The mathematical framework for QD/ED/QED analysis is *asymptotic* queueing theory, where limits are taken as the number of servers increases indefinitely, in a way that is carefully balanced against offered loads. We shall initially demonstrate these regimes within our pre-basic model for a service station - the $M/M/n$ queue, or Erlang-C in telecommunication terminology. ("Pre-basic" because it assumes out operationally significant options that customers enjoy - first, the ability to abandon a queue if, while waiting, service turns out to be unworthy of its wait; and second, the option of returning for additional service, either as a must of by choice - more on these options momentarily.)

Limits of $M/M/n$, as $n \uparrow \infty$, can be taken either in steady-state or process-wise. *Steady state* limits are obtained via regeneration analysis of busy- and idle-excursions: in the former all servers are busy, while in the latter at least one of the servers is idle. Consider, for example, the probability that a customer is delayed prior to being served (the "Erlang-C" delay formula): it is obtained through dividing the expected duration of a busy-excursion by the expected duration of the busy+idle cycle. In the QD and ED regime, the probability of delay is asymptotically 0 or 1 respectively. On the other hand, busy- and idle-excursions in the QED regime are *both* of order $n^{-1/2}$ (both fast, but "equally" fast), hence this probability of delay is asymptotically non-degenerate.

The 0,1 and non-degenerate limits of the delay probability can also serve as characterizations of the QD, ED and QED regimes, respectively. A non-degenerate delay-probability turns out equivalent to the well-known "square-root" staffing rule: with R denoting the offered-load, $n \approx R + \beta\sqrt{R}$, for some constant β (positive to ensure stability). Square-root staffing was discovered already by Erlang (around 1910), but its mathematical substantiation had to await the seminal paper by Halfin and Whitt (1981), which constitutes our theoretical starting point.

Process limits provide first-order *fluid approximations* through Strong Laws of Large Numbers,

and second-order *diffusion refinements* through Central Limit Theorems. Some of this theory will hopefully also be touched upon.

2.3 Lectures 3+4 (Location and Time TBA)

2.3.1 *Two Basic Models for a Service Center: Erlang-A, or a Call Center with Abandoning Customers; and Erlang-R, or an Emergency Department with Recurrent Patients.*

In this lecture, I emphasize the interface between *operational* and *human* aspects. One operationally significant aspect is customer *(im)patience* (as already acknowledged by Palm in the 40's). The other is *returns* to service, which turns out to be operationally significant especially in a time-varying environments.

Abandonments: I shall start with an empirical and statistical analysis of (im)patience. Then I describe operational models that acknowledge (im)patience: in steady state (Erlang-A and relatives), in the ED and QED regime, and possibly also in Nash equilibrium (due to adaptive customers). While the discussion is motivated by abandonments in call centers, the (im)patience phenomenon is prevalent and significant beyond call centers; for example in IVR services, electronic commerce and even in hospital Emergency Departments (EDs), where the term "LWBS = Left Without Being Seen" has been coined.

Returns: Patients in EDs, after first being seen by a physician, typically undergo a series of tests and reexaminations, before being either released or hospitalized. The situation is similar yet different for Oncology patients, who are scheduled to return for hospitalization and treatment, say on a monthly basis. Yet another case is elderly patients who, during flu-season, recycle between the ED and their nursing home. Recycling customers exist also in telephone call centers, where some returns are positive (eg. repeat purchase) and others are negative (eg. failing to achieve "first-call resolution"). In all of these cases, the service process consists of "needy" and "content" cycles, and a natural question arises as to how to acknowledge such cycles in a queueing model, if at all.

2.3.2 *Predictable Variability in Time-Varying Services: Fluid Models and Offered Loads; Staffing Time-Varying Queues to Achieve Time-Stable Performance.*

Time-varying demand and capacity are common-place in service operations. Sometimes, *predictable* variability (eg. peak demand of about 1250 calls on Mondays between 10:00- 10:30, on a regular basis) dominates stochastic variability (i.e. random fluctuations around the 1250 demand level). In such cases, it is useful to model the service system as a deterministic *fluid model*, which transportation engineers standardly practice. I shall describe such fluid models, focusing on their accuracy (under appropriate circumstances) and relating to their theoretical justifications (Functional Strong Laws of Large Numbers).

One way to cope with time-varying demand is to time-vary staffing levels. This is common

practice in many service operations, notably call centers and hospitals. Such practice raises a multitude of research challenges. I shall address mainly one of them: how to achieve, via appropriate staffing, *time-stable* performance in the face of *time-varying* demand. It turns out that the square-root rule, properly applied, provides a remarkably simple and robust solution to this seemingly difficult staffing problem. Here "properly applied" means applying, as the backbone of square-root staffing, the *offered-load* function. This function is derived from a corresponding time-varying ample-server queueing system, and it is related to the fluid approximation mentioned above, as well as to a time-varying version of Little's Law.

Anecdotely, square-root staffing, that acknowledges abandoning customers, helped me understand a phenomenon that had frustrated me over some years, which I summarize as "The Right Answer for the Wrong Reasons": how come so many call centers enjoy a rather acceptable and often good performance (i.e. apply proper staffing), despite the fact that the managers of these call centers are "stochastically-ignorant".

2.3.3 *Addendum*

This last lecture will be used for

- Buffer-time for topics that were not covered adequately in previous lectures, or
- Special requests that will arise during the mini-course, or
- Deeper analysis via SEEStat, or
- Some more empirical adventures, for example throughput degradation in the ED, fairness in the routing of patients from the ED to Internal Wards, asymptotic regimes in practice, data-based SBR = Skills-Based Routing, practical manifestation of congestion laws (Little's Law, PASTA, state-space collapse,...), or
- Additional theoretical topics.

In the text that follows, I start with a macro-view of Services in our society. Next, I gradually narrow the focus down to Service Engineering, Service Networks (Stochastic and Deterministic-Fluid), and Queueing Theory and Science. The discussion will be then specialized to Tele-Networks and further to Call/Contact Centers. I conclude with some details on relevant Service Engineering research, emphasizing the contributions of my students and colleagues that have made this course, and the research supporting it, both possible and fun.

3 Some Background on Services

The phenomena and statistics, here and later, are mainly from Israel and the U.S.A. Some data is somewhat old. Nevertheless, I have sound reasons to believe that the facts are representative of today's reality and they apply to Europe as well.

- *Scope - Services are Central in our Life*: Service functions include financial services (eg., banking, insurance, real-estate), distributive services (transportation, information), utility, social (medical, education, government), hospitality and entertainment, wholesale and retail trade, professional (legal, engineering), and more. Service interfaces include face-to-face, telephone, internet, chat, fax, snail-mail, and more.

- *Economics - Services are Vital for Economic Viability*:

  U.S.A. (Excerpts from the Economist, September 29th, 2005) "For the first time since the industrial revolution, fewer than 10% of American workers are now employed in manufacturing. And since perhaps half of the workers in a typical manufacturing firm are involved in service-type jobs, such as design, distribution and financial planning, the true share of workers *making things you can drop on your toe* may be only 5%. Our figure of 10% comes from dividing the number of manufacturing jobs - just over 14m, say the latest figures - by an estimated total workforce (including the self-employed, part-timers and the armed forces) of 147m. Indeed, most people today work in services: in America, as many as 80%. But this trend is hardly new. As early as 1900, America and Britain already had more jobs in services than in industry."

  Israel: In 1995, the total number of employed civilians in Israel amounted to about 2 million people. Out of these, 68.2% (about 1.4 million) were employed in Services, 28.9% in Industry and 2.9% in Agriculture. Furthermore, between 1995 and 1996, the sectors with the largest increase in the number employed were Communication and Transportation (about 10%) and Business Activities and Banking (8%). Health and Welfare services also enjoyed an increase of about 4%, while Industry was stable and Agriculture declined by about 11%. This profile is consistent across other economic measures (for example GDP).

- *Productivity - Services are Lagging Behind Agriculture and Manufacturing*:

  U.S.A. During 1980-1990, annual productivity growth rate averaged 3.3% in manufacturing (recovering from 1.4% during 1970-1980) but it was only 0.8% in services (stagnating from 0.7% over the previous 10 years).

  Israel: Between 1986 and 1996, Israels productivity growth averaged annually 8% in agriculture, about 1% in Industry, 1.5% in Services and Commerce and about 3% in Communication and Transportation.

- *Trends - Convergence of Services and Manufacturing around the Customer*: Given the compression of product life-cycles (due to time-based competition), explosion of product variety (due to required customization), and heightened expectations for after-sale support, the manufacturing supply-chain has been moving closer to the service-model in which the (production) *process* and the *product* essentially coincide. In other words, products are increasingly service-intensive in that customers' interaction with the manufacturer or its service representative (contact-time) prevails throughout the products' life-cycle. (See also the discussion below on Outsourcing). This amplifies customers' *contact-time* as a fundamental product attribute, just as in services.

  In parallel, insatiable customer demand for services has led to scales and scope that necessitate frequent redesign of existing services and creation of new ones, all enabled through information and automation technologies. These technologies are capital-intensive enough to deserve sound management, engineering and scientific principles which, traditionally, "only manufacturing was acknowledged as being worthy of."

- *Trends - Outsourcing*: Rather than buying and maintaining a car-fleet, why not let a leasing company do it for you? Rather than setting-up and running a help desk for technical support, with its costly fast-to-obsolete hardware, growing-sophisticated software, highly skilled peopleware and ever-expanding infoware, why not let an outsourcing company do it all for you? Indeed, "everything is becoming a service" in that, more and more, customers are buying the *services* that products render, rather than buying the *products* themselves.

- *Trends - Manufacturing or Services? A redundant distinction?* (More excerpts from the Economist, September 29th, 2005) "Any analysis of labour-market trends soon gets bogged down in a statistical swamp. For instance, a small part of the fall in manufacturing jobs is a statistical illusion caused by manufacturers contracting out services. If a carmaker stops employing its own office cleaners and instead buys cleaning services from a specialist company, then output and employment in the service sector appear to grow overnight, and those in manufacturing to shrink, even though nothing has changed.

  More generally, the line between manufacturing and services is blurred. McDonalds counts as a service company, but a visit to any of its restaurants puts one in mind of an industrial assembly line, turning out cooked meat products. Similarly, an increasing slice of value-added in manufacturing consists of service activities, such as design, marketing, finance and after-sales support. Last but not least, Britains number-crunchers stick The Economist, along with the whole publishing and printing industry, in manufacturing, even though almost all our staff are engaged in service-like activities. The division between manufacturing and services has become redundant. A more sensible split now is between *low-skilled* and *high-skilled* jobs. Neither manufacturing nor services is inherently better than the other; they are interdependent. Computers are worthless without software writers; a television has no value without programmes. The issue is not whether people work in factories or not, but whether they are creating wealth. Manufacturing once delivered the highest value-added; high-tech industries, such as drugs and aerospace, still do. But in developed economies today, telecoms, software, banking and so on can create more wealth than making jeans or trainers. Writing a computer program creates more value than producing a computer disc. Before long no one will much care whether firms are classified under manufacturing or services. Future prosperity will depend not on how economic activity is labelled, but on economies' ability to innovate and their capacity to adjust."

Relevance to Engineering, Science, Management and more specifically to *Operations Research, Industrial Engineering, Statistics, Marketing, Computer Science, Information Systems, Psychology,...*: Consider the centrality of Services in our life and economy, the yet superior efficiency of manufacturing and agriculture, the trends described above, and the fact that so many university graduates are employed in the service sector. All this, plus intrinsic interest in service-topics, highly suggest that the Science, Engineering and Management of Service Networks, as has been and will now be described, should occupy a central role in our teaching and research agenda.

4 Service Networks: Models of Congestion-Prone Service Operations

The title of the present section, as well as the title of this note, reflects my (biased) angle on service operations - I often view them as stochastic (random) or deterministic (fluid) systems, within the Operations Research paradigm of Queueing Networks. To support this view, let me first present my conception of the role of Queues in services, from the perspectives of customers, servers and managers. I shall then describe Service Networks, continuing with relevant queueing-theory history and concluding with the fluid-view of service networks.

4.1 On Queues in Service

Queues in services are often the arena where customers, service-providers (servers) and managers interact (establish contact), in order to jointly create the service experience. Process-wise, queues play in services much the same role as inventories in manufacturing (see JIT = Justin- Time, TBC = Time-based-Competition, etc.) But, in addition, "human queues" express preferences, complain, abandon and even spread around negative impressions. Thus:

- *Customers* treat the queueing-experience as a window to the service-providing party, through which their judgement of it is shaped for better or worse.
- *Servers* can use the queue as a clearly visible proxy for the state of the system based on which, among other things, service protocols can be exercised (eg. customers priorities).
- *Managers* can use queues as indicators (queues are the means, not the goals) for control and improvement opportunities. Indeed, queues provide unbiased quantifiable measures (these are not abundant in services), in terms of which performance is relatively easy to monitor and goals (mainly tactical and operational, but sometimes also strategic) are naturally formulated.

My point of view is thus clear: the design, analysis and management of queues in service operations could and should constitute a central driver and enabler in the continuous pursuit of service quality, efficiency and profitability.

4.2 On Service Networks and their Analysis

Service Networks here refer to *dynamic (process)* models (mostly analytical, sometimes empirical, and rarely simulation) of a service operation as a *queueing network*. The dynamics is that of serving *human customers*, either directly face-to-face or through phone-calls, email, internet etc. Informally, a *queueing network* can be thought of as consisting of interconnected service stations. Each station is occupied by servers who are dedicated to serve customers queued at the station. In the simplest version, the evolution over time is stationary as statistically-identical customers arrive to the station either exogenously or from other stations. Upon arrival, customers join a queue and

get served first-come-first-served. Upon service completion, customers either leave the network or move on to another station in anticipation of additional service. Extensions to this simplest version cover, for example, models with non-stationary arrivals (peak-loads), multi-type customers that adhere to alternative service and routing protocols, customers abandonment while waiting, finite waiting capacities that give rise to blocking, splitting and matching of customers and more.

In analyzing a Service Network, I find it useful to be guided by the following four steps (though, unfortunately, most often only the first three are applied/applicable):

- *Can we do it?* Deterministic capacity analysis, via process-flow diagrams (spreadsheets, linear programming), which identifies resource-bottlenecks (or at least candidates for such) and yields utilization profiles.
- *How long will it take?* Typically stochastic response-time analysis, via analytical q-net models (exact, approximations) or simulations, which yields congestion curves. Note: When predictable variability prevails and dominates then the Fluid View is appropriate; the analysis is then deterministic, via inventory buildup diagrams. (e.g., The trucks of National Cranberries.)
- *Can we do better?* Sensitivity and Parametric ("what-if") analysis, of Measures of Performance (MOP's) or Scenarios, which yields directions and magnitudes for improvements.
- *How much better can we do?* or put simply: What is optimal to do? via Optimal Control (exact, asymptotic) that is typically difficult but becoming more and more feasible.

I usually demonstrate these four steps in class via models of Dynamic-Stochastic (DS) PERT/CMP networks (sometimes referred to as fork-join or split-match networks). These are also convenient means to expose the two major types of *operational queues*: there are either *resource* queues, where the wait is for a resource to become available, or *synchronization* queues, where the wait is for a precedence constraint to be fulfilled.

4.3 Some Relevant History of Queueing-Theory

The father of Queueing Theory is the Danish Telecommunication Engineer Agner Krarup *Erlang* who, around 1910-20, introduced and analyzed the first mathematical queueing models. Erlang's models are standardly taught in elementary/introductory academic courses (for example $M/M/n, M/M/n/n$), as they are still corner-stones of today's telecommunication models (where $M/M/n/n$ is known as Erlang-B, "B" apparently for Blocking - the central feature of this model, and $M/M/n$ is referred to as Erlang-C, "C" conceivably because it is a subsequent to "B"). Moreover, and more relevant to our present discussion, $M/M/n$ is still the work-horse that supports workforce decisions in telephone call centers.

Another seminal contributor to Queueing Theory, Scandinavian (Swedish) as well, is Conny Palm, who in 1940-50 added to Erlang's $M/M/n$ queue the option of customers abandonment. I shall refer to Palm's model as Palm/Erlang-A, or just Erlang-A for short (unfortunately and perhaps unjustly to Palm, but Erlang "was there first".) The "A" stands for Abandonment, and

for the fact that Erlang-A is a mathematical interpolation between Erlang-B and Erlang-C. Palm, however, has been mostly known for his analysis of time-varying systems, also of great relevance to service operations.

A next seminal step (one might say a "discontinuity" in the evolution of Queueing Research) is due to James R. Jackson, who was responsible for the mathematical extension of Erlang's model of a *single* queueing-station to *networked* queueing stations, or Queueing Networks, around 1955-1965. Jackson was motivated by manufacturing systems and actually analyzed open and semi-open networks. Closed networks, relevant to healthcare as it turns out, were analyzed in the mid 60's by William J. Gordon and Gordon F. Newell. Interestingly, Newell, who passed away only recently, was a Transportation Engineer at Berkeley (I am not sure if and why closed networks are natural for transportation) that was the earliest influential advocator of incorporating Fluid Models as a standard part of Queueing Theory - see his text book (Applications of Queueing Theory, 1982). A student of Newell, Randolph W. Hall, who is currently a Professor at USC, wrote an excellent Queueing book (Queueing Methods for Services and Manufacturing, 1991) that has greatly influenced my teaching of Service Engineering; Hall is currently working on healthcare systems, adopting the fluid-view (described below) to model the flows of patients in hospitals.

Jackson networks are the simplest theoretically tractable models of queueing networks. (Their simplicity stems from the fact that, in steady state, each station in the network behaves like a naturally-corresponding birth-death model, independently of the other stations.) The next step beyond Jackson networks are BCMP/Whittle/Kelly networks, where the heterogeneity of customers is acknowledged by segregating them into classes. But service operations often exhibit features not captured by Jackson and BCMP/Whittle/Kelly networks. Further generalizations are therefore needed, which include precedence constraints (fork-join, or split-match networks), models with one-to-many correspondence between customer types and resources (skills-based routing, agile workforce), and models that exhibit transient behavior.

The key tradeoff in running a service operations is that between service efficiency and quality, which queueing models are ideal to accommodate. This tradeoff is most delicate in large systems (many servers), but here exact analysis of queueing models turns out limited in its insight. This was already recognized by Erlang who thus resorted to approximations. However, the first to put Erlang's insight on a sound mathematical footing were Shlomo Halfin and Ward Whitt, at the early 80's, in the context of mainly Erlang-C. They introduced what we shall call QED Queues, which stands for queues that are *both* Efficiency- and Quality-Driven, hence their name. QED Q's emerge within an asymptotic framework that theoretically and insightfully supports the analysis of the efficiency-quality tradeoff of *many-server* queueing systems.

Prime examples of QED Q's are well-run telephone call centers; but to properly model these, one must generalize the Halfin-Whitt framework to allow for customers' impatience. This was done in a Technion M.Sc. thesis by Ofer Garnett, in the late 90's, and later published with Marty Reiman. At that same time, a generalization of the Halfin-Whitt framework to time-varying Jackson-like networks was carried out with Bill Massey and Marty Reiman (under the name Markovian Service Networks). In analogy to QED Q's, there are also ED (Efficiency-Driven) and QD (Quality-Driven) queues, all arising from asymptotic analysis as well: Erlang-C, in these three operational regimes, was treated with Sem Borst and Marty Reiman; generalizations to Erlang-A and relatives is the

subject of a Technion Ph.D. thesis by Sergey Zetlyn. The research-part of the website of Ward Whitt is recommended for further references on QED/ED/QD Q's.

Thus, since the 60's, queueing networks have been successfully used to model systems of manufacturing, transportation, computers and telecommunication. For us they are models of service systems, in which customers are human and queues, broadly interpreted, capture prevalent delays in the service process. The service interface could be phone-to-phone (naturally measured in units of seconds), or face-to-face (in minutes), fax-to-fax (hours) letter-to-letter (days), face-to-machine (e.g., ATM, perhaps also Internet), etc. The finer the time-scale, the greater is the challenge of design and management. Accordingly, the greater is the need for supporting rigorous models, a need that further increases with scale, scope and complexity.

4.4 The Fluid View - Flow Models of Service Networks

Most queueing-network models are *stochastic* (random), in that they acknowledge *uncertainty* as being a central characteristic. In recent years, it has turned out that viewing a q-net through a "deterministic eye", animating it as a *fluid network*, is often appropriate and useful. For example, the Fluid View often suffices for bottleneck analysis (the "Can we do it?" step, mentioned above), motivating congestion laws (eg. Little's Law) and crude staffing.

Some illuminating "Fluid" quotes:

- "Reducing letter delays in post-offices": "Variation in mail flow are not so much due to random fluctuations about a known mean as they are time-variations in the mean itself ... Major contributor to letter delay within a postoffice is the shape of the input flow rate: about 70% of all letter mail enters a post office within 4-hour period". (From Oliver and Samuel, a classical 1962 OR paper).
- " ... a busy freeway toll plaza may have 8000 arrivals per hour, which would provide a coefficient of variation of just 0.011 for 1 hour. This means that a non-stationary Poisson arrivals pattern can be accurately approximated with a deterministic model". (Hall's textbook, pages 187-8). Note: the statement is based on a Poisson model, in which mean = variance.

There is a rich body of literature on Fluid Models. It originates in many sources, it takes many forms, and it is powerful when used properly. For example, the classical EOQ model takes a fluid view of an inventory system, and physicists have been analyzing macroscopic models for decades. Not surprisingly, however, the first explicit and influential advocate of the Fluid View to queueing systems is a Transportation Engineer (Gordon Newell, mentioned previously). To understand why this view was natural to Newell, just envision yourself sitting in an airplane that is landing in an airport of a large city, during a nightly rush-hour - the view of the network of highways that surrounds the airport, as seen from the airplane, is precisely this fluid-view. (The influence of Newel1 is clear in Hall's text book.)

Some main advantages of fluid-models, as I perceive them, are:

- They are simple (intuitive) to formulate, fit (empirically) and analyze (elementary).

- They cover a broad spectrum of features, relatively effortlessly.
- Often, they are all that is needed, for example in analyzing capacity, bottlenecks or utilization profiles.
- They provide useful approximations that support both performance analysis and control. (The approximations are formalized as first-order deterministic fluid limits, via Functional (Strong) Laws of Large Numbers.)

Fluid models are intimately related to Empirical Models, which are created *directly* from measurements. As such, they constitute a natural first step in modeling a service network. Indeed, refining a fluid model with the outcomes of Work (Time and Motion) Studies (classical Industrial Engineering), captured in terms of say histograms, gives rise to a (stochastic) service network model, as described previously.

5 Service Engineering (Science and Management)

I have been advocating the terminology *"Service Engineering"* to describe my research, teaching and consulting on (tele-)services. (Service Engineering is to be compared against the traditional *Industrial Engineering*. It is to provide an essential support and supplement to *Service Management*, while drawing it scientific foundation from *Service Science*.)

5.1 Challenges and Goals

Research, teaching and practice of Service Engineering, as I perceive it, should take a *designer's view*. Design challenges pertain, for example, to

- Service strategy: determinants of service-quality levels, full- vs. self-service, customization vs. standardization, warranty (after-sales support depth), ...
- Service Interface (channel): by phone and/or by email, fax, letter, ..., or perhaps face-to-face, ...
- Service Process: front- vs. back-office or possibly both, sequential or parallel tasks, ...
- Control: who to admit, priority scheduling, skills-based-routing, exploiting idleness, ...
- Resources: staffing - how many servers, off- or on-line, shifts structure, ...
- Environment: waiting experience, busy-signal vs. music, information (eg. predicting delay durations), ...
- Marketing: customer segmentation, cross- or up-selling, marketing-operations interfaces, ...

- Information Systems: data-base design of call-by-call operational and business data, off-line and on-line queries, ...
- Human factors: career paths, incentives, hiring policies, FTE's vs. actual workforce levels, ...

The ultimate goal of Service Engineering is to **develop scientifically-based design principles and tools (often culminating in software), that support and balance service quality, efficiency and profitability, from the likely conflicting perspectives of customers, servers, managers, and often also society.** I find that queueing-network models constitute a natural convenient nurturing ground for the development of such principles and tools. However, the existing supporting (Queueing) theory has been somewhat lacking, as will now be explained.

5.2 Scientific Perspective

The bulk of what is called Queueing Theory consists of research papers that formulate and analyze queueing models with realistic flavor. Most papers are knowledge-driven, where "solutions in search of a problem" are developed. Other papers are problem-driven, but most do not go far enough to a practical solution. Only some articles develop theory that is either rooted in or actually settles a real-world problem, and scarcely few carry the work as far as validating the model or the solution. In concert with this state of affairs, not much is available of what could be called *"Queueing Science"*, or perhaps the Science of Congestion, which should supplement traditional Queueing Theory with data-based models, observations and experiments. In service networks, such "Science" is lagging behind that in telecommunications, transportation, computers and manufacturing. Key reasons seem to be the difficulty to measure services (any scientific endeavor ought to start with measurements), combined with the need to incorporate human factors (which are notoriously difficult to quantify). Since reliable measurements ought to constitute a prerequisite for proper management (see TQM = Total-Quality-Management, for example), the subject of measurements and proper statistical inference is important in our context.

5.3 Re-Engineering Perspective

Service networks provide a platform for advancing, what could be described as, Queueing Science and Management Engineering of Sociotechnical Systems. Management Engineering links Management Science with Management Practice, by "solving problems with existing tools in novel ways". Quoting the late Robert Herman, acknowledged as the "father of Transportation Science", Sociotechnical systems are to be distinguished from, say, "physical and engineering systems, as they can exhibit incredible model complexity due to human beings expressing their microgoals". (Significantly, Herman's models of complexity were nevertheless "tractable through remarkable collective effects"; in other words "laws of large numbers" which, for services as well, turn out to play a central explanatory role.) The approach and terminology that I have been using, namely *Service Engineering*, is highly consistent with the once influential BPR (=Business-Process-Reengineering) evolution, as well as with ERP (=Enterprise-Resource-Planning) and CRM (= Customer-Relations-Management), placing heavy emphasis on the process-view and relying heavily on the accessibility of information technology.

5.4 Phenomenology, or Why Approximate

Service systems often operate over finite-time horizons (the notion of steady-state then requires re-interpretation). They employ heterogeneous servers, whose service capacities are time and state-dependent. Their customers are "intelligent", who typically (but not always) prefer short queues; they jockey, renege and, in general, react to state-changes and learn with experience. Finally, service systems suffer from high variability – both predictable and unpredictable, and diseconomies of scale – when being decentralized and inefficient (e.g., often FCFS/FIFO is the only option). Such features render the modeling of service networks a challenge and their exact analysis a rarity. This leads to research on *approximations*, typically short but also long-run fluid and diffusion approximations. Approximations also enhance exact analysis by simplifying calculations and exposing operational regimes that arise asymptotically. (Recall the QD/ED/QED operational regimes.)

The "ultimate products" of approximations are scientifically-based practically-useful rules-of-thumb. Here are three (of what I believe to be) convincing examples.

5.4.1 Square-Root (QED) Staffing Rules for Moderate-to-Large Telephone Call Centers, and beyond to Healthcare

In the context of call centers, the square-root Staffing rule asserts that in a call center which experiences an offered load of R Erlangs (this will be explained momentarily), an appropriate staffing level is about $R + c\sqrt{R}$, for some constant c, positive or negative: c is a calculable *quality-of-service* parameter that reflects the balance between service-level and operational-efficiency (the larger the c the more weight is placed on service quality vs. efficiency).

An implementation of the rule could run as follows: Suppose that 1,000 telephone calls "arrive" to a call center every hour, on average, and that essentially all calls remain online until being served (as opposed to some abandoning due to impatience); suppose that average call duration is 3 minutes, and that its standard deviation is of that same order; finally, assume that an agent's hourly salary is comparable to the cost of one n-th of a customer's hour waiting (the latter cost being at least its 1-800 waiting cost). Then, with an offered load on the system of $R = 1000 \times \frac{3}{60} = 50$ hours-of-service per hour (or 50 Erlangs), the average operating costs are minimized by following the square-root staffing rule with $c = \sqrt{2n/\pi}$, namely with about $50 + 10\sqrt{\frac{n}{\pi}}$ agents (or rather FTE's, namely Full-Time-Equivalent service positions).

The square-root staffing rule helps explain a phenomenon that puzzled me for some time: how do call centers that are run by "stochastic-ignorant" managers perform rather acceptably and even better? taking $c = 0$ in the "recipe" $R + c\sqrt{R}$ provides an answer, since then the recommended staffing level is simply R, which is what one would get from merely a naive reasoning (stochastic-ignorant, as previously referred to). One can describe this as obtaining "the right answer for the wrong reasons". To be concrete, in the above example, with offered load of 50 units of service-time per unit of time, assume that callers are moderately patient. Then, staffing with 50 agents will give rise to close to 50% of the callers answered without delay, around 5.5% callers abandoning due to impatience, and those who stay on line served within less than 10 seconds on average; "increasing the operation 8-fold", with an offered load of 400 that is attended to by 400 agents, only 2% would

abandon and the rest answered within 3.5 seconds on average (economies of scale).

The square-root staffing rule aims at balancing service quality and efficiency, hence it leads to what one might call a Quality and Efficiency Driven (QED) operational regime. Theory tell us that, in that regime, one could expect abandonment rates of about 2-3%, for example. Alternatively, one can derive staffing rules that are Efficiency Driven (ED, leading to 10-20% abandonment) or Quality Driven (QD, with essentially no delays hence no abandonment). The mathematical framework for supporting these staffing rules is asymptotic analysis of *many-server* queues, which turns out to provide remarkably accurate insights (for moderate to small systems as well).

Practice and recent theory have revealed that square-root staffing enjoys a remarkable level of robustness. Indeed, it has been proved applicable over a wide spectrum of scenarios, from small single-queue systems (with single-digit number of servers) to complex queueing networks. This observation is important beyond its theoretical significance. Indeed, its confirms the relevance of the QED regime to healthcare systems, which are typically smaller and more complex than call centers.

5.4.2 Routing Rules for Efficiency-Driven Email Operations, or Patients in Emergency-Departments

Suppose that a service operation caters to several types of emails. Specifically, there are several pools of servers working in parallel, and each pool serves its own constituency of email types, with possibly overlaps of constituencies. Control of such a system amounts to, first, routing of emails to pools (either upon arrivals or taken from types-designated queues) and, second, assignments of servers to emails upon service completion.

More formally, let i denote email types (i-emails) and j stand for server pools (j-servers). Let μ_{ij} be the average service rate of i-emails by j-servers. (μ_{ij} is the reciprocal of an average service duration; $\mu_{ij} = 0$ indicates that j-servers cannot serve i-emails.) Consider an i-email whose sojourn time (waiting + service time) in the system is W; then, upon service completion, such an i-email incurs a waiting cost of $C_i(W)$, where the cost function $C_i(\cdot)$ is increasing and convex. (Convexity is natural - the longer the sojourn time W the higher is the marginal delay cost $C_i'(W)$ – the derivative of C_i at time W; a good example to to have in mind are costs that are quadratic in the delay.)

Assume that such an email system is "well-balanced" and "efficiency-driven" (both concepts can be made mathematically rigorous). Then, the following remarkably-simple strategy turns out essentially cost-optimal (again, in a mathematically precise way): when becoming free at time t, a j-server chooses for service an i-email from its constituency for which $C_i'(W_i(t))\mu_{ij}$ is maximal; here $W_i(t)$ is the waiting time at time t of the longest-waiting (head-of-the-line) i-email. (In the special case of costs that are quadratic in the sojourn time, this translates to serving the email with the maximal $W_i(t)\mu_{ij}$ - in words, the email chosen for service is the one with longest weighted waiting time.) It also turns out that the efficiency-driven regime renders irrelevant the decisions about emails that encounter idle servers upon arrival.

Remarkably, the routing rule just described can be simply generalized to accommodate feedback.

Specifically, and changing terminology from emails to generic customers, assume that an i-customer, after being served, has a probability P_{ik} to return (immediately) to the system as a k-customer (and with probability $1 - \sum_k P_{ik}$ leave the system). Then the above routing is still cost-optimal, if one replaces μ_{ij} by the reciprocal of the average *total* service time rendered by a j-server to an i-customer, covering all services from entry, as an i-customer, to exit. (For now, this statement has been proved only for a single pool of servers.) Such a generalization is useful for designing control of patient-flow through physicians in Emergency Departments. Here one must also adhere to deadline (triage) constraints on the time that it takes an arriving patient till first seen by a physician. These triage deadlines depend on clinical severity; for example, the Canadian scale (used in Israel) acknowledges 5 levels of severity: 1 and 2 are most severe, and 3-5 correspond to walking patients. The ED problem, which turns out to enjoy a simple solution, is thus to minimize congestion of in-process patients, subject to adhering to triage-constraints. (Note that cost convexity is natural or, strongly put, essential for healthcare applications.)

5.4.3 (Im)Patience While Waiting

The third example somewhat differs from the previous two. Its relation to Approximations is that it shows how to take a *complicated reality* and reduce it to (approximate it by) a *tractable model*, then design a *measurable* rule-of-thumb that captures its essence.

I am sure that many of you, at some point, summarized their waiting-for-service experience in terms close to the following: "I expected to wait 10 minutes, I felt like I waited 20 minutes, but after the fact I realized that I actually waited 15 minutes." Since the waiting experience is an important part of the service experience, both psychologically and operationally, this is worth elaborating on. To this end, the waiting experience of a customer can be broken down into the following five components:

- Time that a customer *expects* to wait;
- Time that a customer is *willing* to wait (patience, need) - denote it by τ;
- Time that a customer is *required* to wait (offered wait) - denote it V;
- Time that a customer *actually* waits - denote it W;
- Time that a customer *perceives* waiting.

Each of these five "measures" of waiting-time is relevant and significant in its own right. But accounting for all of them in a model is difficult. Here is an attempt at a simplifying approximation.

For a customer that is *experienced*, it is plausible that the expected-wait, based on previous perceived-waitings, is close to the offered-wait V. To a *rational* customer, for lack of a better terminology, it is plausible that the perceived-wait equals the actual-wait. Finally, the actual-wait W is clearly the minimum between τ and V: indeed, if $\tau < V$ the customer abandons and if $\tau \geq V$ the customer reaches service. One is thus left with the pair (τ, V), which determines all

else. And here our operational queueing models, for example Erlang-A, come to the rescue: they accept the time-willing-to-wait τ as a model-input and they produce the offered-wait V, and hence also $W = \min\{\tau, V\}$, as a model-output. Having W, or more precisely the distribution of W, one can plan staffing levels to satisfy service-quality constraints, for example: find the least number of servers so that at least 80% of the *served* customers will be waiting 20 seconds or less (formally, $P\{V \leq 20; \tau > V\} \geq 0.8$), jointly with no more than 3% abandoning $P\{\tau \leq V\} \leq 0.03$).

The above also helps answer the following question: Can one refer to a customer that is willing to wait 10 minutes as being "patient"? Well, if that customer expected to wait 1 hour, than willing to wait only 10 minutes is a manifestation of **im**patience; but if the expected wait was 2 minutes, that 10 minutes manifests patience. This suggests to measure "patience" in relative terms, for example through a *Patience Index* defined as follows:

$$\begin{aligned} \text{Patience Index} &:= \frac{\text{time willing to wait}}{\text{time required to wait}} \\ &= \frac{\text{average patience}}{\text{average offered wait}} = \frac{E[\tau]}{E[V]}\,. \end{aligned}$$

(In the above, we implicitly assumed an experienced customer, for which the time *expected* to wait equals that *required* to wait.) The larger the Patience Index the higher the tolerance with waiting. But, with such a definition, a natural question arises: how would one go about measuring or estimating patience indices? For that, we have developed statistical survival-analysis-based techniques, which require data at the call-by-call (transactional) level. But these are not easy to implement and, perhaps more importantly, transactional data is unfortunately unavailable in most circumstances. Therefore, we wish to introduce an *Empirical* Patience Index, our *rule-of-thumb*, which will serve as an auxiliary measure for the above theoretical patience index. The following has been found very useful *and* accurate:

$$\text{Empirical Patience Index} := \frac{\%\text{ served}}{\%\text{ abandoned}}\,.$$

For example, with an abandonment rate of 20%, our rule-of-thumb suggests that customers are willing to wait 4 times their expected-wait. As a final comment, under certain circumstances we have been able to explain the closeness of the theoretical and empirical indices. However, these explanations are incomplete and hence open up a very interesting research direction.

6 Telephone-Based Services: Scope, Significance and Relevance

6.1 On Call/Contact Centers

Call Centers are telephone-based service centers. Contact Centers are their extensions with additional multimedia communication channels, for example emails, internet, chats, etc. Either are viewed by some as the **business-frontier** and by others as the **sweat-shops** of the 21-st century. Indeed:

- **Scope:** The Call Center Magazine is a U.S. monthly magazine (there are several others, for example Call Center Europe) that is dedicated to telephone services. Its readers are typically professionals in the call center industry. They are asked by the magazine to classify themselves according to the following business categories, which amply demonstrate the scope of telephone-services: advertising, banking, catalog retailer, computing, electronics or software, consulting, credit collection, direct mail marketer, dealer or distributor, entertainment, finance, securities or mutual funds, fund-raising, government, health-care, hospitality, information services, insurance, list or database supplier, manufacturer, market research, professional services, publishing or broadcasting, retailing, telecommunications, telemarketing, transportation, travel or recreation, utility, wholesaler or others.

- **Scale:** In the U.S., annual telephone sales far exceed 50% of the total business volume. The universal accessibility, time sensibility and cost efficiency in conducting business over the phone has given rise to a huge growth industry (**20% growth rate**) - the (telephone) **call center industry**. There are anywhere between 70,000 up to 200,000 call centers, which employ anywhere between 3 to 6.5 million people (more than the entire agriculture sector). Annual expenditures on call centers are estimated between US$100 to US$300 billion, with 60-75% labor cost.

- **The Agents:** Call Centers' agents often suffer from, what could be called, **tele-stress**, which leads to unsatisfied employees who tend to perform at below acceptable standards. Several reasons have been acknowledged as leading to this state of affairs: frequent and intense **interpersonal contacts** with customers; **efficiency demands** excessive work load (e.g., intensive call volume, strict response and waiting time metrics); **highly-monitored environments** with "unseen audience"; advanced information and computer technology, leading to **tightly defined dialogues** in the form of "screen pops" that contain standard communication scripts, and little or no control over work methods and procedures; finally, often **repetitive** work that is mostly carried out independently, allowing for little socialization.

- **The Challenge of Human Resource Management:** A large-scale national survey of management practices and outcomes in the U.S. call center industry, published in 2004, reported that total **annual turnover** (including quits, layoffs, dismissals, and retirements) **average** 33%. Outsourced call centers have the highest turnover rates (51%) followed by retail call centers (47%). For grasping these findings, take retail call centers as an example: with a national average of 47% yearly turnover, it is likely that there are call centers with close to 100% turnover rates, meaning that the whole workforce turns over within 1 year. Add to that the fact that, in retail call centers, only 8% have discretion over their pace of work, the average call handling time per customer is 4.7 minutes, close to 10% of the workforce is absent on a typical day and, finally, it takes an average of 3 month to become proficient on the job. The challenge in managing a call center hence clearly manifests itself, which explains the terminology *" sweatshops of the 21st century"*, "assembly-lines in the head" or *"modern form of Taylorism"*, "female ghettos" (73% female workforce in retail call centers), etc.

- **The Customers:** Numerous customer surveys are conducted on the quality of telephone services. For example, a 2005 survey of over 2,000 UK mobile users found that a quarter of the young people switched mobile service provider as a result of bad call center customer service, with the situation almost as severe across the broader population. (With UK mobile saturation at over 90%, customer defection in the lucrative 18-29 age bracket is of particular concern to operators

who look to this demographic to make up a significant part of their revenues. This age group is also the heaviest users of call centers). Indeed, overall dissatisfaction was high, with over 40% of users polled saying they were unhappy with the customer service they receive from their mobile operator's call center. The five major customer pain-points were, as identified by respondents: Having to repeat a query to more than one agent (41%); Being kept on hold too long (32% said they were kept on the line for more than ten minutes while an agent dealt with their enquiry.); Being asked for the same details again and again (29%); The agent lacked the necessary knowledge to deal with my query (27%); And finally, it took a long time to deal with my query (26%)

- **National Success:** Telecom Ireland, Ireland's premier telecommunications provider, and the Industrial Development Agency of Ireland (IDA Ireland), a government agency that provides assistance for overseas companies setting up in Ireland, jointly created a partnership to ensure that Ireland is Europe's #1 international call center location. And indeed, numerous companies, ranging from Fortune 500 firms to start-ups, have established centralized multilingual call centers that serve Europe, the Middle East, Africa and now even the U.S. markets. In fact, call centers are expanding in most of western Europe. For example, in 2001, an estimated number of 265,000 agents worked in about 2,900 call centers in Germany, at an annual growth rate of about 20%.

- **Scale and Scope again, but also Quality:** A U.S. **sales-company** has a call center that attends to 15,000 calls daily (on average); the average duration of a call is about 4 minutes, customers essentially never get a busy-signal and the average wait on the line is below 1 second. A U.S. **health-insurance** company has more than 40 call centers spread over the country; the largest dozen are networked to allow for centralized load-balancing, thus yielding an average abandonment rate that varies from the negligible to 3% at the most. Finally, Customer Service and Support is an integral part of a large **U.S. bank**, employing about 10,000 highly skilled associates in contact centers located in twenty cities across the United States. These associates provide service and financial solutions to more than 130 million phone-calls and 1.74 million e-mail customers each year; the Interactive Voice Response (IVR) units of that bank handled over 500 millions calls.

- **Technical-Support Crisis:** In October 1996, the **Help Desk Institute** had 5,339 members in the U.S. and Canada. The Institute publishes an annual report, which provides a comprehensive look at current practices in the help desk and customer support industry. A typical help desk provides a "single point of contact and responsibility for rapid closure of technology problems," catering to both internal and external customers. The preferred mode of receiving technical services are by telephone, fax or mail. Advice is sought on bug fixes, configuration utilities, product usage tips, software upgrades and product training. According to the 1996 report, help desks are prevalent in manufacturing, computer software, banking, insurance, government, healthcare and more. It is estimated that over 80% of the help desks are experiencing increase in call volume, so much so that observers claim "customer support is at present in crisis." (The three predominant reasons for the increase are "newer, more complex technology", "more customers" and "changes: upgrades, conversions, installations".) Crisis in the provision of customer support could prove a bottleneck in the evolution and adaption of new technologies.

- **What does is take to become a Call Center Manager?** A leading Israeli provider of Internet services has a technical support center (Help-Desk) that employs about 400 people, many of whom are Technion students. They work part-time and cover 3 shifts, 7 days a week,

occupying at any given time over 50 agent-positions that provide on-line technical assistance. The founding manager of this help-desk had a bachelor degree in Political Science. He started working as a manager with just a few student-agents, continuing until the call center grew to the above-mentioned scale. According to him, he got the job because the company liked his philosophy of customer service, having worked previously in marketing. He has **no technical background** and he learns hands-on.

6.2 Tele-Nets: Models of Telephone-Based Service Operations

A *call center* is in fact a service network, as discussed perviously, in which agents provide *tele-services* (here to be interpreted mainly as *telephone-based services* or, sometimes more generally, *online-services with customers and servers being remote from each other*). As mentioned, call Centers that accommodate telephone, internet, chat, e.mail and fax services are often referred to as *(Customer) Contact Centers* - this terminology will not be used here.

Call centers are thus modeled by (queueing) networks of tele-services, which can be referred to as *tele-nets*. In tele-nets, the customers are callers, servers (resources) are telephone agents (operators) or communication equipment, and *tele-queues* consist of callers that await service by a system resource. The network-view is often essential to capture transfers of customer among service resources, for example a caller that is referred to a specialist or is transferred to an IVR (Interactive Voice Response) unit and then switches back (often frustrated) to a human operator, or a customer who opts to abandon due to limited patience (and disturbing music or commercials) and then calls back later. Tele-queues differ from, say, queues in a bank-branch in that they are mostly *invisible* (phantom queues) and hence amenable to management control without visibly violating fairness principles.

Call Centers typify an emerging business environment in which Information Technologies enable the simultaneous attainment of superb service quality with extreme operational efficiency. Call centers vary greatly in functionality (support, sales, information), size (up to many thousands of agents per center), technology, customer profiles and agents skills. The future call center, as I perceive it, will cater to a vast customer-base. It will be connected externally to the Telephone and Internet networks and internally, through CTI (Computer Telephony Integration), to an enterprise-wide computer database. Customers will receive multi-media information via the phone (upon request or call-backs), a Web site, IVR, e.mail or fax. Future ACD's (Automatic Call Distributors) will increasingly route requests to electronic agents — yet, I believe, *the human-service* is here with us to stay.

Sound scientific principles are prerequisites for sustaining the complex socio-technical enterprise of the call center. In my research, I seek to contribute to the theory that supports these principles and to the creation of new ones.

7 A Sample of Coauthored Service-Engineering Research, in the Context of Call Centers

I conclude with a brief description of some theoretical and empirical research projects, jointly with students and colleagues. The relevant papers (or technical reports) all appear in http://ie.technion.ac.il/serveng/References/references.html.

From here on we focus on *Telephone Call Centers*. (Future versions of this document will extend to cover Healthcare Operations as well.) As an introduction to call-centers research, or at least its early stages, we recommend the review
http://ie.technion.ac.il/serveng/References/Gans-Koole-Mandelbaum-CCReview.pdf
which surveys a significant part of the research that will be now described.

7.1 Design of Call Centers

A central goal of Service Engineering is to develop practically useful rules-of-thumb, but these must be based on rigorous models and analysis. The starting point is the classical M/M/n queue, which must be extended to accommodate non-negligible phenomena within call centers. Relevant research-lines are, for example:

- "Rules for Designing Call Centers with *Impatient* Customers", starting with Ofer Garnett's M.Sc thesis and then published jointly with Marty Reiman. This research builds on research of Palm (who first introduced Garnett's queueing model, in the 40-50's), Riordan, Halfin and Whitt and Fleming, Stolyar and Simon. It accounts for a fundamental feature of service operations - waiting customers can typically abandon and seek alternatives. Palm's model, which has been denoted M/M/s+**M** and referred to as Erlang-A, assumes exponentially-distributed patience (the +M). Practically, however, patience has been demonstrated to be non-exponential. This motivated the Ph.D. thesis of Sergey Zeltyn, who extended Garnett's insights to accommodate generally distributed patience. Zeltyn's underlying queueing model, denoted M/M/s+**G** (G for General (Im)Patience) was first developed by Baccelli and Hebuterne (1981) and later re-analyzed and extended by A. Brandt A. and M. Brandt.

- "*Predicting Delays* under Prioritized Skills-based Routing", that was Efrat Nakibli's M.Sc thesis, jointly supervised with Isac Meilijson. This research enables online prediction of delay durations - a feature often sought-after by waiting customers that are trapped in listening to music, commercials or, at best, miscellaneous trivia. This subject of delay prediction has attracted a great deal of attention in recent years; see, for example, the PhD thesis of Rouba Ibrahim, under the supervision of Ward Whitt; and three recent projects by Alon et al., Aksin et al. and Armony et al.

- *"Dimensioning of Moderate-to-Large Call Centers"*, first jointly with Sem Borst and Marty Reiman, then continued with Sergey Zeltyn. Here one seeks to characterize, via asymptotic analysis, operating regimes for large call centers, specifically, Efficiency-Driven (ED), Quality-Driven

(QD) and rationalized regimes that thrive for *both* quality and efficiency (QED). The latter gives rise to the *square-root staffing rule*, putting it on a concrete mathematical footing and exhibiting for it robustness and accuracy at such an incredible level that invited further research to explain it. And indeed, in the Erlang-C context, recent research (by Janssen, van Leeuwaarden and Zwart) did provide the explanation, while developing a refinement of the square-root staffing rule (through a corrected diffusion approximation for the Erlang-C formula). Another recent refinement, now for Erlang-A, is a regime labeled ED+QED, which calls for staffing at the level of $R \cdot (1-\gamma) + c \cdot \sqrt{R}$, for some constants $\gamma \in (0,1)$ and c. This regime arose in research with Zeltyn, as a means for controlling the most popular operational performance measure, namely the fraction of customers that are provided with timely service (formally, $P\{Wait > T\} \leq \epsilon$, with T being in the order of a service-duration.)

- *"Designing a Call Centers with an Interactive Voice Response (IVR) units"*, which is Polyna Khudiakov's M.Sc. thesis. (Polyna then continued for a Ph.D at the Technion, under the supervision of Malka Gorfine and Paul Feigin, doing statistical analysis of customers' redials in call centers; graduating in 2010, she is now a postdoc at Harvard U.) In her M.Sc., Polyna modeled a call center with an IVR, analyzed it in steady-state and then approximated the model in the QED regime for ample additional insights.

  Interestingly, Polyna's model was adapted by Galit Yom-Tov (another past Ph.D. student at the Technion, now a postdoc at Columbia U.) in order to capture the operational reality of internal wards in hospitals. Such models, as first observed by de Vericourt and Jennings, provide theoretical support for square-root staffing of physicians and nurses - a problem that is, of course, of no less significance than that of call center agents. Galit's model is referred to as Erlang-R, where the "R" stands for ReEntrant or Returning or Recurrent customers/services (as opposed to the service being continuously provided over a single customer-server encounter).

- *"Skills-Based Routing"*, which is the difficult problem of matching customers and agents, taking into account agents capabilities (cross-trained, specialized) and customers profiles (eg. VIP, regulars). This research has been carried out with Sasha Stolyar (for efficiency-driven services) and Mor Armony, Rami Atar, Itay Gurvich (a past M.Sc. Technion student, who got his Ph.D. at Columbia under Ward Whitt, and then joined Kellogg), Marty Reiman, and Gennady Shaikhet (a past Technion Ph.D. student, continuing for a postdoc at CMU and now at Carleton U.). For example, in his M.Sc thesis, written jointly with Mor Armony, Gurvich considered a single pool of many iid servers that attend to several customer classes; he solved, to asymptotic optimality, *jointly* the two problems of staffing (how many servers?) *and* scheduling (how to differentiate class service-levels?), in terms of a simple square-root staffing rule, accompanied with a simple service discipline that is threshold-based. Similar research, with Armony, solved the same problem for a model that has a single-customer class with several pools of agents. (Gurvich's Columbia Ph.D. thesis covered the two latter models and much more, and his research since has also contributed significantly to our understanding of SBR.)

- *"Operational Models in Healthcare"*, it turns out, can gain from the experience gathered in call-center modeling. As a hint to these gains, several examples were already mentioned: square-root staffing of physicians and nurses, and control of patient flow through physicians in EDs. Another example is the routing of patients from emergency departments to hospital wards, a problem notorious for its complexity (operational, human, political), and (too) frequently acknowledged

as a hospital bottleneck. A Technion M.Sc. thesis by Yulia Tseytlin, jointly supervised with Petar Momcilovic, captured some of this complexity by balancing operational issues (delays) with fairness (being careful not to punish the wards which are most efficient). Yulia is now a member of IBM Haifa Research Lab.

7.2 Behavioral Operational Models

Data-Based (Empirical) research in Operations Management is gaining importance, and rightly so. This goes hand in hand with acknowledging psychological aspects that are significant operationally, for example customer-patience, mechanisms that trigger abandonment, preferences as to what information customers seek and when, and design interface of IVR (Interactive-Voice-Response) to minimize OOR (Opt-Out-Rate), namely the fraction of customers that opt to human servers.

A fundamental issue here, for which I believe no definite answer is yet available, is the understanding (quantification) of the "Cost of Delay". This is especially significant in (phantom) tele-queues such as waiting at the phone, "conversing" with an IVR or a computer terminal. A related question is the following: given the individual cost of waiting and abandonment triggers, predict the ensuing system (Nash) equilibrium, in particular accounting for learning due to accumulated experience. The above was joint research with Nahum Shimkin and the M.Sc student Ety Zohar. The mathematical framework is the Erlang-A queue with *general* abandonment (M/M/n + G), as analyzed by Baccelli and Hebuterne. In parallel to these "queueing-theory" efforts, Anat Rafaeli (Technion Psychology professor) and her graduate students continued, in both theoretical and laboratory research, with psychological research that aimed at explaining customers reaction to information, provided to them while waiting.

7.3 Predictable Variability

Many service operations operate over a finite horizon, during which operating characteristics vary predictably with time. In order to account properly for this predictable variability, models must sometimes be transient, which renders impossible their exact analysis. One thus resorts to alternative models, based on the fluid view and its diffusion refinement. This work started with Bill Massey, and continued with the Bell-Labs group of Marty Reiman and Sasha Stolyar, greatly assisted by then the post-doc Brian Rider (now a Mathematics Professor at U. of Colorado, Boulder).

One way to cope with predicable variability is to predictably vary staffing levels. It is then possible, via a surprisingly simple adaptation of the square-root staffing rule, to achieve time-stable performance in the face of time-varying demand. This was first done with the then-student Otis Jennings (now a Business School Professor at Duke), advised by Bill Massey and Ward Whitt; recently, the research has been significantly expanded by the same team, except that the student "changed" to Zohar Feldman, a past Techion M.Sc. student (now a member of IBM Haifa Research).

A central ingredient of time-varying staffing is the *offered-load* function. This function summarizes the time-varying average work (measured in time-units of service), which arrives to the system per unit of time. To this end, the offered-load appropriately combines arrival counts with the ser-

vice requirements of these arrivals. Time-varying staffing levels are planned to be above or below the offered-load, depending on the relative importance of *customer waiting* to *agent waiting* (idleness). Interestingly, and not surprisingly after comprehending the relevant facts, the time-varying offered-load turns out to be nothing but a time-varying version of the classical Little's Law.

7.4 Statistical Inference

Service data is often vast, yet incomplete and inaccurate. We are thus looking for tools that statistically summarize the available as well as infer significant missing components.

One example started with consulting engagements that utilized a unique measurement system for face-to-face services. It was based on a network of bar-code readers that recorded individual service transactions. This system triggered the M.Sc thesis by Sergey Zeltyn, who has since been a major contributor to Service Engineering research and teaching at the Technion. Sergey developed a Queueing Inference Engine (QIE) for queueing networks, as originally developed by R. Larson for isolated stations.

Another example is the inference of customers' (im)patience, modelling the latter as the distribution of the time to abandon. (Here we need techniques from Survival Analysis, since the data is censored: the time-to-abandon for customers that get served is censored by their waiting time.) This has been ongoing research that started with Yaakov Ritov, then continued with Anat Sakov and Sergey Zeltyn, where we carried out a descriptive analysis of a data-base with about 450,000 telephone calls (all calls to a small Israeli call center during 1999). The analysis has advanced understanding of the operational characteristics of the center, the behavioral characteristics of its customers and the interaction of the two. The research then continued in two directions, both with the Wharton team of Larry Brown, Noah Gans, Haipeng Chen and Linda Zhao: first, statistical analysis (estimation and prediction) of the small-bank data-base mentioned above; and second, collecting and analyzing telephone-calls to a much larger U.S. bank (a network of four call centers) that caters to about 400,000 calls per week. These efforts culminated in the Technion's SEE Laboratory/Center, which will be described below.

7.5 Hierarchical Modeling of Stochastic Networks

Stochastic networks model environments in which uncertainty is a dominant factor. Such models are typically set up in terms of microscopic primitives, and hence are difficult to analyze. For many purposed, however, cruder descriptions suffice. These are provided through long-run and short-run fluid approximations (deterministic models at a macroscopic level) and corresponding diffusion approximations (Brownian-like models at a mesoscopic level). This can be all integrated into a five-level hierarchy of models for stochastic networks. Research that supported the hierarchy started in the Ph.D. theses of Hong Chen and Gennady Pats. It has continued in joint research with Bill Massey, Brian Rider, Marty Reiman, Kavita Ramanan and Sasha Stolyar.

7.6 Call Center Data: The First Steps

Call centers have been accumulating vast quantities of data (operational, marketing, survey), but these have been *inaccessible* to academic research, at least at the level of the individual transaction (call-by-call data). As already mentioned, with Anat Sakov and Sergey Zeltyn, we were fortunate to obtain and analyze operational transaction-data of a small banking call center in Israel (covering the full year of 1999). Our findings were then applied towards supporting "Queueing Science", and extended to further statistical/operational insight and larger call/contact centers. This was a joint effort with a Wharton group that consisted of Larry Brown, Noah Gans, Linda Zhao, Haipeng Shen (Larry's Ph.D. then, now a professor at UNC) and Yotam Shlomay (a Technion student then visiting Wharton). This initial success paved the way to the first major data adventure, again with the Wharton group: obtaining call-by-call data from a large (800 agents working in parallel) east-coast U.S. bank, and covering a period of over 2.5 years.

8 The Technion SEE Center/Laboratory (SEE = Service Enterprise Engineering)

The data-collection effort with Wharton has been greatly extended at the Technion, with the partnership of my colleague Paul Feigin, and the technical leadership of Valery Trofimov. This project goes under the heading DataMOCCA (Data MOdels for Call Centers Analysis), which I shall now elaborate on.

DataMOCCA is a universal schema for accommodating transaction-level data from service operations. A central part of DataMOCCA is its graphical user interface, SEEStat, which enables online EDA (Exploratory Data Analysis) that spans seconds-to-months resolutions. Specifically, DataMOCCA accommodates several large call centers: the U.S. bank mentioned above, an Israeli Cellular company (about 700 agents at peak times), and two Israeli banks; all of these cover at least 2 years worth of data. As an example, the U.S. bank data has close to 220 million calls, out of which about 40 million were served by agents, and the rest by the Interactive Voice Response (IVR) system. And one of the Israeli banks (about 400 agents), after its regular nightly data-archiving, is depositing the previous day's data at a SEEStat data-safe. This *daily* transaction-level data is then fused automatically to the SEELab repositories.

All of the above data-centered activities are conducted at the Technion's SEE Laboratory. This research lab was created in 2007, through a generous donation by Hal and Inge Marcus, and it has since regularly employed 3-4 full-time researchers, postdocs and students (graduate and undergraduate). The SEElab serves as the focal point for Service Research and Teaching at the Technion. Its main goal has been designing, maintaining and analyzing an accessible repository of resources and data from service systems, preparing the data to support research and teaching.

The scope of DataMOCCA has now extended to cover also Hospitals and Web Services. For example, its healthcare component contains patient-level data from Emergency Wards of six hospitals in Israel, each covering periods from 12 to 52 months; there is, all in all, data of about one million patient arrivals, individually for each patient from entry to departure - similarly to the

call center data. Recently, a Technion-affiliated hospital (over 1000 beds) contributed 4 years of transaction-level data, covering its ED, most hospital wards, operating rooms, and more. This hospital is also contemplating a patient-tracking system, based on Radio-Frequency IDs (RFID). Such a tracking system will continuously provide the SEELab with patient flow data; the hope is for this system to eventually cover, to some extend, also nurses and doctors.

The original goal of the SEELab was the creation of a data-repository, drawn from call centers of various functionality, and accessible for analysis to researchers world-wide. This goal is now partially achieved - at least the part of universal accessibility: see
http://seeserver.iem.technion.ac.il/see-terminal
The latter is a link to the server of the SEELab. It serves as a gate, through SEEStat, to a very detailed summary (7GB) of the transaction-level data (40GB), from the U.S. Bank mentioned above.

The link to the website of the SEELab is http://ie.technion.ac.il/Labs/Serveng

9 Teaching Service Engineering

A course on "Service Engineering" has been taught at the Technion since the early 90's, starting as a seminar for graduate students and culminating in the present compulsory course, documented in *http://ie.technion.ac.il/serveng*.
This web-site includes teaching notes, homework assignments and a full-year operational data of a small Israeli call center. In addition, at the References menu of the site, there are downloadable research papers, theses, technical reports and overheads of lectures.

The Service Engineering course was condensed into a mini-course that has been taught first at INSEAD, twice at Columbia University, three times at Wharton, twice at Stanford, and it is now planned to be taught at HKUST (September 2011).

10 Some Downloadable References

I shall now list several papers that provide helpful background (in the context of call centers):

First, two surveys (one general and one specific):

- Early Survey of Call Centers (Research and Practice):
  Gans, N., Koole, G., M. "Telephone Call Centers: Tutorial, Review and Research Prospects." Invited review paper by Manufacturing and Service Operations Management (M&SOM), 5, 79141, 2003.
  http://iew3.technion.ac.il/serveng/References/Gans-Koole-Mandelbaum-CCReview.pdf
- A "Light" Teaching Note on Erlang-A:
  M. and Zeltyn S. "The Palm/Erlang-A Queue, with Applications to Call Centers."
  http://iew3.technion.ac.il/serveng/References/Erlang_A.pdf

Next, starting points to four research directions:

- Statistical Analysis of Transaction-Level Call Centers Data:
  Brown, L., Gans, N., M., Sakov, A., Zeltyn, S., Zhao, L. and Haipeng, S. "Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective." Journal of the American Statistical Association (JASA), 100: 36-50, 2005.
  http://iew3.technion.ac.il/serveng/References/JASA_callcenter.pdf
- Model for The Basic Call Center: Erlang-A
  Garnett O., M. and Reiman M. "Designing a Call Center with Impatient Customers." Manufacturing and Service Operations Management, 4, 208-227, 2002.
  http://iew3.technion.ac.il/serveng/References/abandon.pdf
- Staffing of a Stationary Erlang-C (preliminary to Erlang-A):
  Borst S., M. and Reiman M. "Dimensioning Large Call Centers." Operations Research, 52, 17-34, 2004.
  http://iew3.technion.ac.il/serveng/References/Dimensioning.pdf
- Staffing in the Face of Time-Variable Demand:
  Feldman Z., M., Massey W.A. and Whitt W. "Staffing of Time-Varying Queues to Achieve Time-Stable Performance." Management Science. Management Science, 54, 324-338, 2008.
  http://iew3.technion.ac.il/serveng/References/Feldman2008pub.pdf

And, finally, a source for references to the research literature:

- Bibliographical Support, till 2006:
  M. "Call Centers. Research Bibliography with Abstracts." Version 7, May, 2006.
  http://iew3.technion.ac.il/serveng/References/US7_CC_avi.pdf

11 Preparatory Readings

Active learning is superior to passive learning. Thus, as a preparation for my lectures, I recommend that you read (at least "diagonally") the following material:

1. *Background and Introduction:*

   - Read the present note, at the depth-level that suits your personal interests.
   - "Telephone Call Centers: Tutorial, Review and Research Prospects," 2003, by Gans et al. Start reading Sections 1,2,3. Then go over the Table of Contents, to see what is in there, skimming through what you find most interesting. Section 4 is closely related to the call-center component of my lectures.
     http://iew3.technion.ac.il/serveng/References/Gans-Koole-Mandelbaum-CCReview.pdf
   - "Using Operations Research to Reduce Delays for Healthcare," 2008, INFORMS Tutorials, by Linda V. Green (Columbia U.). This is interesting reading throughout, but make sure you read Section 1 (Sources of Healthcare Delays) and Section 4 (Obstacles to using OR in Healthcare), especially &4.1 (Lack of data).
     http://www1.gsb.columbia.edu/mygsb/faculty/research/pubfiles/3874/OR

2. *A Data-View*:

   - "Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective," 2005, by Brown et al. Skim over the paper, focusing more on the last Section 7. (You will recognize this article as the source for some of the figures in my lectures.)
     http://iew3.technion.ac.il/serveng/References/JASA_callcenter.pdf
   - "Patient Flow in Hospitals: A Data-Based Queueing-Science Perspective," 2011, by Armony et al. Skim over the paper, reading at least Section 1 (Introduction) and Section 2 (Summary of Results).
     http://ie.technion.ac.il/serveng/References/Patient%20flow%20main.pdf
   - "The Workload Process: Modelling, Inference and Applications," 2011, M. Reich (advised jointly with Y. Ritov). The Offered-Load is a concept that is central for understanding the operational characteristics of a service system. For our purposes, the most relevant parts are the Introduction, Section 3 and briefly Section 5.
     http://ie.technion.ac.il/serveng/References/Michael_Reich_Thesis_withlinks.pdf
   - "Data Stories about (Im)Patient Customers in Tele-Queues," 2012, with S. Zeltyn. This is an empirical view of customers' impatience, while waiting for a phone-service. (The notion of impatience plays an important role beyond tele-services: see, for example, LWBS in hospitals.) This is easy non-technical reading, at your leisure.
     http://ie.technion.ac.il/serveng/References/impatience_data_stories_final.pdf

3. *Two Basic Models*:

   - "The Palm/Erlang-A Queue, with Applications to Call Centers," a teaching note with S. Zeltyn. Mathematically, Erlang-A, with "A" standing for Abandonments, is **only** a special

case of an ergodic Birth & Death model. But practically, it is (or should be) the central model in support of workforce management of call centers.
http://iew3.technion.ac.il/serveng/References/Erlang_A.pdf

- "Erlang-R: A Time-Varying Queue with ReEntrant Customers, in Support of Healthcare Staffing," 2011, with G. Yom-Tov and based on her PhD thesis. Erlang-R, with "R" standing for Returning or Recurrent customers, is a simple 2-station open queueing network. It has significant modeling powers, which are most pronounced in time-varying environments. If fact, it has been conceived to help stabilize performance, via staffing (e.g. physicians) in the face of time-varying demand (e.g. patients arriving to an Emergency Department): See Section 6, after reading Sections 1 and 3.
http://iew3.technion.ac.il/serveng/References/Erlang_R.pdf

4. *Beyond Basics: Staffing & Control of Some Simple Systems*: The role of the papers, from here to the end, is mainly to have you become aware of their existence.

   - "Service Level Differentiation in Call Centers with Fully Flexible Servers," 2008, with Gurvich and Armony: Skills-Based Routing (SBR) of a single-pool system.
   http://iew3.technion.ac.il/serveng/References/Vdesign_pub.pdf
   - "On Fair Routing From Emergency Departments to Hospital Wards: QED Queues with Heterogeneous Servers," 2011, with Momcilovic and Tseytlin, 2010: An example of Fair Routing from the Emergency Department to Internal Wards.
   http://iew3.technion.ac.il/serveng/References/fr6.pdf
   - "Routing and Staffing in Large-Scale Service Systems: The Case of Homogeneous Impatient Customers and Heterogeneous Servers," 2010, with Armony: Routing and Staffing *jointly*.
   http://www.stern.nyu.edu/om/faculty/armony/research/IV_abandon.pdf

5. *(More) Complex Models*: As mentioned, just skim in order to become aware of existence.

   - "Control of Many-Servers Queueing Systems in Heavy Traffic," 2007, G. Shaikhet (advised jointly with Atar): Especially Part 3 on "Simplifying controls of the full SBR topology".
   http://iew3.technion.ac.il/serveng/References/PhD_gennady.pdf
   - "Fork-Join Networks in Heavy Traffc: Diffusion Approximations and Control," 2011, A. Zviran (advised jointly with R. Atar): We are capturing here a prevalent feature in heathcare (Fork-Join), within a control-model that thrives to maximize throughput.
   http://iew3.technion.ac.il/serveng/References/ZviranMScThesis.pdf
   - "Excursion-Based Universal Approximations for the Erlang-A Queue in Steady-State," 2012, with I. Gurvich and J. Huang. A major reason for the success of Erlang-A and Erlang-R is that they are amenable for practical approximations (fluid and diffusion). In the case of Erlang-A, this has given rise to a plethora of asymptotic regimes, and Universal Approximations provide a way to circumvent the challenge of choosing which regime is most appropriated for a particular application. To understand the idea, it suffices to go over Sections 1 and 2. The results and their applications appear in Section 3.
   http://ie.technion.ac.il/serveng/References/Universal_040712_Final.pdf