Service Engineering of Stochastic Networks

Background, with a focus on Tele-Services

Avishai 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
http://ie.technion.ac.il/serveng2006W

The subject of this note, and a central theme of my present research, are *Service Networks*: these include public service centers (municipal, government), telephone services (business and marketing, emergency, assistance), banks and insurance (front and back office), hospitals (emergency rooms, outpatient clinics, operating rooms), 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 essential.) Significant motivators for my research efforts have been *tele-services*, in which customers and servers are remote from and invisible to each other. Communication in tele-services is through snail-mail, fax, electronic-mail, interactive-voice-response, telephone and increasingly the Internet. However, existing tele-services are predominantly telephone-based hence my heavy emphasis on telephony.

When lecturing on Stochastic Service Networks, I typically divide my presentation into three parts:

- *Introduction to Service Engineering and Management.* Ample examples are described, based on my experience in project-supervision and consulting, with an emphasis on the practical significance of basic theoretical research. The examples are mainly of congestion-prone service networks, in particular their measurements, time-varying behavior, and the controllable drivers of delay-queues (synchronization gaps, scarce resources).

- *Service-driven Theory.* Successful service analysis and management must often be multi-disciplinary, fusing ingredients from Operations Research, Statistics, Industrial Engineering, Sociology, Psychology, Game Theory, Economics, Management Information Systems, and even more. In this part, theoretical examples are surveyed that support design/engineering (for example pooling of service components), control (for example skills-based-routing to match demand with supply) and management (for example staffing scenarios.)

- *Tele-services and Call/Contact Centers.* This part deals with telephone call centers, or more broadly contact centers. Of importance is on the interface between human and operational aspects, most notably customers' impatience. For example, I have been seeking to characterize and measure human patience while waiting in phantom (invisible) queues. With the lessons learned, one could then incorporate patience into operational models and derive comprehensive performance measures.

The text in the sequel provides some background on Services and Stochastic Networks, followed by (my conception of) Service Engineering, more background on Call Centers and Tele-Nets, and finally a sample of some research projects.

Background on Services

The phenomena and statistics below are mainly from the U.S.A. and Israel. I have sound reasons to believe that they are representative of Europe as well.

- *Scope — Services are Central in our Life*: Services 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.

- *Economics — Services are Vital for Economic Viability*: 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), as well as across all industrialized nations (for example, 73.1% of the U.S. labor force in 1995 was employed in service jobs).

- *Productivity — Services are Lagging Behind Agriculture and Manufacturing*: Between 1986 and 1996, Israel's productivity growth averaged annually 8% in agriculture, about 1% in Industry, 1.5% in Services and Commerce and about 3% in Communication and Transportation. In the U.S., during 1980–1990, annual 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).

- *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 coincided. 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.

  Conversely, 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, until recently, "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.

Relevance to Engineering and Management: Consider the centrality of Services in our life and economy, the 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 highly suggests that the Science, Engineering and Management of Service Networks, as will now be described, should occupy a central role in our teaching and research agenda.

Service Networks: Models of Congestion-Prone Service Operations

The above title reflects my (biased) angle on service operations - I often view them as stochastic 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.

Queues in services are often the arena where customers, service-providers (servers) and managers establish contact, in order to jointly create the service experience. Process-wise, queues play in services much the same role as inventories in manufacturing (see JIT = Just-in-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 priorities can be exercised, among other things. *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 also strategic) are naturally formulated. In summary, the design and management of queues in service operations could and should constitute a central driver and enabler, in the continuous pursuit of service quality and efficiency.

Service Networks here refer to process models (mostly analytic, sometimes empirical, and rarely simulation) of a service operation as a queueing network. A *queueing network* consists 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, non-stationary arrivals (peak-loads) multi-type customers that accommodate alternative service and routing disciplines, splitting and matching of customers, customers' abandonment while waiting, finite waiting capacities that give rise to blocking, etc.

Since the 50's, queueing networks have been successfully used to model systems of manufacturing, transportation, computers and telecommunication. Here they are used as 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.

Service Engineering (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 and Service Marketing.) Research, teaching and practice of Service Engineering, as I perceive it, should take a designer's view. Design challenges pertain to service strategy (e.g.,

determinants of service-level, full- vs. self-service, customization vs. standardization), service interface (by phone and/or by fax, Internet, letter, ..., or perhaps face-to-face), process (front vs. back office - or both, sequential or parallel tasks,...), control (who to admit, skills-based-routing, priorities, exploiting idleness, ...), environment (waiting, information to and from customers - for example busy-signal vs. music, ...), resources (staffing - how many agents, adaptive or off-line, shifts, ...), human factors such as career paths and incentives, marketing factors such as pricing, after-sales relations, and much more.

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.

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 empirically-based models, observations and experiments. In service networks, such "Science" is lagging behind that in telecommunications, computers, transportation 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.

(Re-)Engineering perspective: Service networks provide a platform for advancing, what might 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, the acknowledged "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". The approach and terminology that I have been using, namely *Service Engineering*, is highly consistent with the 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.

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 rates 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—being decentralized and inefficient (e.g., often FCFS/FIFO is the only option). Such features render the modeling of service net-

works 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.

The ultimate "products" of approximations are scientifically-based practically-useful rules-of-thumb. An example is the *"square-root rule of safety-staffing"* for medium-to-large telephone call centers, which asserts that in a call center that experiences an offered load of R Erlangs, an appropriate staffing level is about $R + c\sqrt{R}$, for some constant c, positive or negative. One variation of this is the following: Suppose that 1,000 telephone calls arrive to a call center every hour, on average; 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 fraction of a customer's hour waiting (the latter being at least its 1-800 waiting cost). Then the offered load on the system is 50 hours-of-service per hour, and average operating costs are minimized with approximately $50 + 10\sqrt{\frac{n}{\pi}}$ agents.

Telephone-Based Services: Scope, Significance and Relevance

Call Centers are telephone-based service centers. They are viewed by some as the business-frontier and by others as the sweat-shops of the 21-st century. Indeed:

- 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.

- In the U.S. alone, annual telephone sales are estimated by some at about US$500 billion worth of goods and services, growing at a rate of about 8% and anticipated to reach 50% of the total business volume after the year 2000. 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 from 20,000 to 350,000 call centers, which employ anywhere between 4 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 50-75% labor cost.

- 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, have 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.

- 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 2 second.

- 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 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".

- A leading Israeli provider of Internet services has a technical support center (Help-Desk) that employs about 250 people, the vast majority of which are Technion students. They work part-time and cover 3 shifts, 7 days a week, occupying at any given moment between 50 to 60 agent-positions that provide on-line technical assistance. The manager of the help-desk has a bachelor degree in Political Science. He started working, as a manager, a few years ago, when there were about 20-plus students/students total. 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.

Tele-Nets: Models of Telephone-Based Service Operations

A *call center* is a service network 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*. (Call Centers that cater to telephone, internet, e.mail and fax services are often referred to as *(Customer) Contact Centers* - this terminology will not be used here.)

Call centers are modelled by *Tele-nets*. Theses are (queueing) networks of tele-services, in which the customers are callers, servers (resources) are telephone agents (operators) or communication equipment, and queues consist of callers that await service by a system resource. The network-view is often essential to capture transfer 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. Telephone queues differ from, say, queues in a bank in that mostly they are 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 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 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.

A Sample of Coauthored Research on Modelling, Inference, Analysis

Here is a brief description of some theoretical and empirical research projects, jointly with students and colleagues. The relevant papers all appear in
http://ie.technion.ac.il/serveng/References/references.html.
Of special interest to the call-centers research is the review
http://ie.technion.ac.il/serveng/References/references.html.

- *Incorporation of psychological aspects into operational models.* Example are characteristics of customer-patience, mechanisms that trigger abandonment, preferences as to what information customers seek and when, 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 answer is available, is the understanding (quantification) of the "Cost of Delay". This is especially significant in remote (phantom) queues such as waiting at the phone, "conversing" with an IVR or a computer terminal. A related question, that has been addressed, 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. This is joint research with Nahum Shimkin and Ety Izhar, with help from Ziv Carmon, Dan Zakay, Sergey Zeltyn and Ilan Guedj. The mathematical framework is the M/M/S queue with general abandonment, as analyzed by Baccelli and Hebuterne.

- *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/S 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", with Ofer Garnett, Marty Reiman and Sergey Zeltyn, and "*Estimating* Waiting Times in Telephone Call Centers", with Efrat Nakibli and Isac Meilijson. The former research, building on research of Palm, Riordan, Halfin and Whitt and Fleming, Stolyar and Simon, accounts for a fundamental feature of service operations - waiting customers can typically (but not always) abandon and seek alternatives. The latter 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. A third current research is "Dimensioning of Large Call Centers", jointly with Sem Borst and Marty Reiman. Here one seeks to characterize, via asymptotic analysis, operating regimes for large call centers, specifically, efficiency-driven, quality-driven and rationalized regimes. And lastly, a challenging line of research is on matching customers and agents (skills-based routing), taking into account agents' capabilities (cross-trained, specialized) and customers' profiles (VIP, regulars). This research is with Sasha Stolyar (for efficiency-driven services) and Mor Armony, Rami Atar, Marty Reiman and Ward Whitt (if also quality-driven, namely in the Halfin-Whitt regime).

- 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 is the inference of customers' impatience, say via 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 is

ongoing research that started with Yaakov Ritov, Anat Sakov and Sergey Zeltyn, where we carried out a descriptive analysis of a data-base with about 450,000 telephone calls (all calls during 1999, to a small Israeli call center). 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 continues 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; second, collecting and analyzing telephone-calls to a large U.S. bank (a network of four call centers) that caters to about 400,000 calls per week. Another example, with Sergey Zeltyn, is on approximations and inference that are inspired by the Queueing Inference Engine (QIE). QIE was originally developed by R. Larson in order to infer congestion (queues) from transactional data. Larson developed his algorithms for isolated stations, and we extended his QIE to a network setting.