Applied Probability in Operations Research: A Retrospective[*]

Shaler Stidham, Jr.
Department of Operations Research
CB #3180, Smith Building
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-3180
sandy@email.unc.edu

November 16, 2001
(revised, March 5, 2002)

[*] This paper is an earlier, unabridged version of the paper "Analysis, Design, and Control of Queueing Systems", which appears in the 50th Anniversary Issue of *Operations Research*.

1 Introduction

As the reader will quickly discover, this article is a joint effort, unlike most of the articles in this special issue. It covers a very broad topic – applied probability – and I could not have attempted to do it justice without the help of many of my colleagues. It might have been more appropriate, in fact, to list the author as the Applied Probability Society of *INFORMS*. I solicited the help of the members of the Society, as a group and in some cases individually, and I am grateful for the help I received. In some cases, this help took the form of reminiscences – about the origins of applied probability as a recognized discipline, about the founding of several journals devoted to applied probability, and about the origins of what is now the Applied Probability Society. Others have offered more personal accounts of their experiences in the field. Collectively (and in some cases individually) these span the last fifty years – the fifty years since the first issue of *Operations Research*.

These reminiscences are at the end of the article, in alphabetical order. I have taken advantage of my status as author of record to include some of my own memories from nearly forty years in the profession. I hope they will be of interest to the reader, not because my career has been remarkably long nor, indeed, remarkable in any way, but because my experiences have perhaps been representative of those who were attracted to *OR* in the 60's – the "glory days", as they seemed then (and perhaps were).

The first part – indeed, the larger part – of the article is a retrospective account of some of the significant achievements of research in applied probability over the past fifty years. A frustration for me was deciding how much to say about each specialty within applied probability. This frustration was in part due to the breadth of the area. (A cursory review of the *OR* courses at a typical university in North America will find that about half of them cover topics in applied probability.) Obviously, it is impossible to be encyclopedic in an article of this nature. Instead, I have chosen to cover in some depth what I believe to be the most important accomplishments in applied probability. My *AP* colleagues have been very helpful in guiding me through areas with which I am not intimately familiar, but the choice of areas has been for the most part my own and it is no doubt controversial.

That this article is appearing in *Operations Research* has helped make my task easier: it has given me an excuse to limit my coverage to topics that have consistently attracted the attention of researchers within the *OR* community. The topics covered are, by and large, those that have received the most attention in the pages of *Operations Research*, *Management Science*, and *OR*-oriented journals such as the *Journal of Applied Probability*, *Advances in Applied Probability*, *Queueing Systems: Theory and Applications*, *Stochastic Models*, and *Probability in the Engineering and Information Sciences*, as well as topics that have generated sessions at *INFORMS* National Meetings and the conferences sponsored by the Applied Probability Society. I have also been helped, indirectly, by Peter Whittle, whose article in this issue on "Applied Probability in Great Britain" has relieved me of the task of discussing the impressive achievements of British researchers. To paraphrase Winston Churchill: "Never have so many important results come from so few."

One cannot spend much time looking at the history of Applied Probability without becoming impressed with the overwhelming contribution of one man: Joe Gani. Not only was Joe responsible for much of the early work on dams and reservoirs and later made important contributions to biological modelling, not only did he establish and nurture several outstanding academic departments and collaborate with and train some of the best minds in the field, but he gave the field professional legitimacy when he founded the *Journal of Applied Probability* in 1964 (and *Advances in Applied Probability* in 1969). As important as all these contributions has been Joe's generous spirit and warm personal concern for his colleagues.

An excellent source for additional "musings" is the book that Joe put together in 1986, *The Craft of Probabilistic Modelling* [115], which contains the reminiscences of twenty eminent applied probabilists. See also Joe's "autobiography" in *Adventures in Applied Probability: A Celebration of Applied Probability*, published in 1988, the twenty-fifth anniversary of the founding of the *Journal of Applied Probability*.

2 Descriptive Models

2.1 Classical Queueing Theory

The postwar period saw a maturation of "classical" queueing theory – a discipline that had already reached adolescence, if not adulthood, before it was adopted by the infant field of operations research in the early 50's. Here I use the term "classical" queueing theory to refer to descriptive models of queueing systems, usually based on Markovian assumptions, in which the goal is to derive an explicit expression for the queue-length or waiting-time distribution (or its transform), usually in steady state.

As I write this, sitting here at the beginning of a new century, I am struck by how narrow this definition now seems. And yet this is how queueing theory was usually characterized when I entered the profession in the mid '60's. Some exceptionally gifted operations researchers, applied probabilists, and mathematicians had achieved impressive results within this framework: explicit steady-state solutions were available for essentially all single-server queueing models with independent inputs (interarrival times and service times), at least if one counts a transform solution as explicit. Results for multi-server systems, mainly with exponential service-time distributions, were available. Priority queues were pretty well understood, at least for Poisson arrivals and fixed priorities. Those who promulgated, or at least accepted, this narrow definition could rightly be proud of these accomplishments. It was only a short step from this pride to the assertion – often heard in those days – that "queueing theory is dead". (This thesis was seriously debated in the applied probability community as late as 1973, as Ralph Disney has observed [83].)

In any case, the contributions to queueing theory in the first two decades of operations research were both broad and deep, and in the context of the narrow definition given above, they came close to saying the last word. Landmarks (by no means an exhaustive list) included the work in Britain of Kendall [185], [186], Lindley [217], Smith [285], Cox [64], and Loynes [222], [223], discussed by Peter Whittle [341] in his article in this issue. A unifying theme was finding (or constructing) a Markov process in (or from) a non-Markov process. Kendall's approach to the *M/GI/1* queue involved finding an "imbedded" Markov chain (see, e.g., [185]), whereas Cox [64] saw that many queueing processes could be made Markovian by adding supplementary variables. The book by Cox and Smith [65] provided a remarkable, concise summary of the state of the art in 1961. The analysis of the *GI/GI/1* queue by Lindley [217] exploited the property that the waiting-time process is a random walk with a barrier at zero. This initiated a fertile line of research which included refinements, such as the Wiener-Hopf factorization by Smith [285], Spitzer's identity [289], and generalizations by Prabhu [250] and others. Kiefer and Wolfowitz [187] extended the Lindley approach to a vector-valued Markov process in their analysis of the *GI/GI/c* queue. Loynes [222] provided the generalization to the *G/G/1* queue with stationary input.

In a series of papers, Takács characterized the transient and steady-state behavior of a variety of queueing systems, using transforms and generating functions. His 1962 book [307] is a good source for most of this work. The late 1950's and 1960's saw the publication of several important texts on queueing theory, including Morse [231], Syski [287], Saaty [269], Beneš [22], Prabhu [250], Cohen [62], and Gnedenko and Kovalenko [127]. The landmark two-volume work by Feller [103], [104] provided a wealth of theoretical insights and applications to queueing systems, as well as other areas in applied probability. Jaiswal [173] wrote the first book completely devoted to priority queues.

2.2 Networks of Queues

Most real-life queueing systems have more than one service facility. The output of one facility may proceed to another facility for further processing, or return to a facility already visited for re-work or additional work of a different type. Examples abound: assembly lines, flow shops, and job shops in manufacturing, traffic flow in a network of highways, processing in order-fulfillment systems, client-server computer systems, telecommunications networks – both *POTS* ("plain old telephone service) and the emerging high-speed packet-switched networks designed to carry a variety of services, including data transfers, conferencing, web browsing, and streaming audio/video.

Early in the postwar period, several researchers in queueing theory turned their attention to networks of queues, recognizing the importance of the applications. Early work by Koenigsberg (e.g., [201]) and others considered special cases, such as cycles or series of exponential queues. An emerging theme was that the steady-state joint distribution of the number of jobs at the various facilities was found to be of *product form*, with each individual factor corresponding to the queue-length distribution in an isolated exponential queue. In the case of a series of queues, this phenomenon could be explained by noting that the output from a *M/M/1* queue is a Poisson process with the same rate as the arrival process and that departures before a given time point are independent of the number in the system at that time point. These surprising properties follow from the reversibility of the queue-length process in a *M/M/1* queue and were proved in the early 1960's by Reich and Burke (see, e.g., [260], [49]).

In 1957 Jackson [170] had considered an arbitrary open network of queues, each with an exponential server and an exogenous Poisson process, and with Markovian routing of jobs from one node to another. Jackson showed that in this case too the steady-state distribution has a product form. The marginal distribution of the number of jobs at the i-th node is the same as it would be in an isolated *M/M/1* queueing system with the service rate at node i and an arrival rate equal to the equilibrium total flow through node i, that is, the sum of the exogenous arrival rate and the flow rates into node i from all the nodes of the network. The vector of total flows at the nodes is the unique solution to a linear system of equations, usually referred to as the *traffic equations*.

The product form for the stationary probabilities in a Jackson network tells us that in steady state the network behaves as if the nodes were isolated, independent facilities, each fed by an external Poisson process with arrival rate equal to the flow rate found from solving the traffic equation. It is a remarkable result, because it is demonstrable that the arrival processes at the nodes are *not* in general Poisson processes. (They are Poisson in a feed-forward network – that is, a network without cycles – which includes as a special case a series of queues. In this case the Reich-Burke results apply.)

Later [171] Jackson extended this result to an open network with multi-server nodes and arrival processes that can depend, in a restricted way, on the state of the system. This extension includes closed networks of exponential servers, in which there is a fixed number of jobs circulating among the nodes. In this case the stationary distribution has a truncated product form, without (of course) the independence between nodes that holds for open networks. (Gordon and Newell [128] re-discovered this property in 1967.) The expression for the stationary distribution is once again explicit in terms of the problem parameters, but efficient computation of the normalization constant is a non-trivial task, to which much effort was subsequently devoted. Recursive schemes, in which the constant for a network with population N is expressed in terms of that for a network of size $N-1$, gained favor. A particularly effective technique is the convolution algorithm of Buzen [51].

By the mid-70's researchers interested in modelling the performance of computer systems had discovered the Jackson network and its variants and come to appreciate the versatility and applicability of such models. In those days, the emphasis was on systems consisting of a mainframe computer with disk storage and satellite terminals – a precursor to what we would now call a local area network *LAN*. The idea of a communication network tying together widely separated computers was just a gleam in the eye of Leonard Kleinrock and a few other visionaries (Al Gore?). It was Kleinrock who, more than anyone else, was responsible for spreading the word among computer scientists about Jackson networks in particular and queueing theory in general. (Others included Reiser, Kobayashi, Sevcik, Lavenberg, Zahorjan, and Chandy. See the article by Kleinrock [199] in this issue for a personal account.) It is significant that one of the major texts in queueing theory published around this time was Kleinrock's two-volume work [198], the second volume of which was devoted to computer applications. (Others included Cooper [67], Borovkov [39], and Gross and Harris [132].)

2.2.1 The Insensitivity Phenomenon; Point Processes and Sample-Path Analysis

Kleinrock observed that the Jackson network model, in spite of its unrealistic exponential assumptions, provided a surprisingly good fit for observed data from computer systems, regardless of the actual distributions of processing times. Only the mean service rates seemed to be relevant. He conjectured that this property, which became known as the "Kleinrock assumption", was due to the superposition of traffic from many sources within the network. The ground-breaking paper by Baskett, Chandy, Muntz, and Pallacios [16] (henceforth referred to as *BCMP*) shed further light on possible reasons for this property. They showed that the product-form stationary distribution characteristic of a Jackson network in fact holds for arbitrary service-time distributions with the same means, provided the service discipline belongs to a certain class, which includes last-come, first-served and processor-sharing (a continuous limit of the round-robin discipline that was in wide use in time-sharing systems in the 70's). This property has come to be referred to as the *insensitivity phenomenon*. The *BCMP* paper used a phase-type approximation (see Section 2.5) of the service-time distribution at each node to model the network as a continuous-time Markov chain and then showed that the conjectured product-form solution satisfies certain local (or partial) balance equations, which are sufficient for the global balance equations that uniquely characterize the stationary probabilities.

The concept of partial balance had in fact been suggested previously by Whittle [337]. It was exploited, along with reversibility and quasi-reversibility, by Kelly [180], who simultaneously and independently proved the insensitivity properties demonstrated by *BCMP*. (See Kelly [181] and Peter Whittle's article in this issue [341] for more on this line of research.)

The insensitivity phenomenon had in fact been well known for a long time in the context of some specific single-facility queueing models. The Erlang and Engset formulas are examples. The Erlang formula gives the steady-state probability of the number of busy servers in a $M/GI/c/c$ system, a model of interest in telephony, where a call that arrives when all c channels are busy is lost. Its insensitivity was discovered gradually for various special cases of the service-time distribution (see Franken et al [110], p. 187, for a chronology). The Engset formula is the counterpart of the Erlang formula in the case of a finite number of sources. See Cohen [61] for an early proof of its insensitivity. In the early 60's, researchers in East Germany began a fertile line of research on insensitivity in the abstract setting of random marked point processes (*RMPP*) and processes with imbedded point processes (*PMP*). The 1967 monograph by König, Matthes, and Nawrotzki [204] demonstrated the potential for this approach. Subsequent papers revealed the utility of the *RMPP*/*PMP* framework for establishing other useful results in queueing theory in a general (stationary) setting. These included relations between customer and time averages, such as $L = \lambda W$ and its generalizations ($H = \lambda G$), using Campbell's theorem, as well as relations between time-stationary and customer-stationary (Palm) probability distributions. The book by Franken, König, Arndt, and Schmidt [110] provides a good summary of the research by this group through the 70's.

Much of the East German work on insensitivity initially became known in the West through Rolf Schassberger's papers [273], [274], [275], [276], in which he developed many of the same ideas in the less abstract setting of a generalized semi-Markov process (*GSMP*), where more familiar, Markovian techniques can be used.

It turned out that many of the results that were proved using the *RMPP*/*PMP* or *GSMP* approach have sample-path counterparts. This was certainly the case with $L = \lambda W$ and $H = \lambda G$, as shown by Stidham [294], [296]. Later work by Stidham, El-Taha, and others showed that the same was true with relations between time-stationary and Palm distributions and (to a lesser extent) insensitivity. The book by El-Taha and Stidham [98] offers a compendium of results like these (as well as other results) achievable by sample-path analysis. See also Glynn and Whitt [125], [126], Whitt [332], [333], Sigman [284], Serfozo [280], [281]. For a personal perspective on sample-path analysis, see Section 4.6.

2.2.2 Networks with Finite Queues and Blocking

At the other extreme from Jackson networks and insensitivity are a class of networks that are nearly intractable from the perspective of analytical characterization of stationary, let alone transient, distributions. These are networks in which some of the nodes have finite-capacity queues and hence jobs that are routed to these nodes may be lost or *blocked*. In the latter case, these jobs continue to occupy the station at which they just finished being processed (or in some cases cannot even begin processing at that station) until the destination station has a free space in its buffer. Such situations can occur in manufacturing and communication networks. There are a limited number of situations in which networks with these characteristics still admit product-form solutions. Good sources for these types of results are the books by Serfozo [281] and Chao et al [54]. Otherwise, the only alternative is to resort to approximations, numerical solutions, or simulation. The book by Perros [247] is a good source. See, e.g., Buzacott and Shanthikumar [52] and Altiok [3] for applications of these models and other stochastic models to performance analysis and control of manufacturing systems.

2.3 Heavy Traffic Theory

A queue is most annoying (and most expensive) when it is long. Consistently long queues are characteristic of a queueing system that is operating close to saturation, or in *heavy traffic*, that is, a system in which jobs are arriving nearly as fast as the service mechanism can process them. The ability to model the heavy traffic regime accurately is therefore crucial for the designer or operator of a queueing system. For complicated queueing systems, often the only modelling alternatives are numerical methods or simulation, both of which are subject to inaccuracies and/or slow convergence in heavy traffic. An alternative is to look for approximations – preferably approximations that become more, rather than less, accurate as the system approaches saturation.

Heavy traffic analysis of queueing systems (and other stochastic input-output systems) had its origin in the work of J.F.C. Kingman in the 1960's [188], [189], [190], [191]. From the perspective of applications as well as pedagogy, perhaps the most important of Kingman's contributions is the simplest – his formula [188] for the heavy-traffic approximation to the expected steady-steady waiting time in the queue in a *GI/GI/1* queueing system – a single-server queue with i.i.d. interarrival times and i.i.d. service times:

$$\mathrm{E}[W_q] \approx \tau \left(\frac{\rho}{1-\rho} \right) \left(\frac{c_a^2 + c_s^2}{2} \right) .$$

Here τ is the mean service time, $\rho = \lambda\tau$ is the traffic intensity (with λ the arrival rate), and c_a and c_s are the coefficients of variation of the interarrival times and service times, respectively. This simple formula, which is exact for the *M/GI/1* queue, tells the two most important stories of queueing theory. First it reveals the dependence of the *mean* waiting time (and hence the mean queue length, via $\mathrm{E}[L_q] = \lambda E[W_q]$) on the *variation* in both the arrival and service processes. Second, it demonstrates the explosive nonlinear growth in waiting time as $\rho \uparrow 1$, that is, as the system approaches saturation (100% utilization). In layman's terms: (1) variation is bad, and (2) high utilization comes at a (very) high cost.

Kingman [188], [189] extended this approximation to multi-server systems and showed that the distribution of the steady-state waiting time in the queue, apart from the mass $1-\rho$ at zero, is approximately exponential in heavy traffic.

The next major steps were taken by Prohorov [255], Borovkov [36], [37], [37] and by Iglehart and Whitt [166], [167], [168], who extended Kingman's results from steady-state random variables to processes. (An excellent account of this early work, as well as Kingman's, is contained in Ward Whitt's survey [331] in the proceedings of the conference on Mathematical Methods in Queueing Theory at Western Michigan University (Kalamazoo) in 1973. See Lemoine [215] for a slightly later survey. The recent book by Kushner [214] provides comprehensive coverage of the current state of the art.)

Kingman's approximations were based on the central limit theorem. They exploited the intuition that the waiting-time process in heavy traffic spends most of its time away from the barrier at zero and hence should behave probabilistically like the random walk from which it is constructed (cf. Lindley [217]). The central limit theorem implies that this random walk, properly normalized, has a Normal limiting distribution. Iglehart and Whitt used the theory of weak convergence (cf. Billingsley [31]) to extend this result to the entire time-dependent waiting-time process.

The theory of weak convergence tells us that the random walk, considered as a process, converges weakly (that is, in distribution) to Brownian motion when time is scaled by n^{-1} and space by $\sqrt{n}^{-1}$. Since the waiting-time process is defined in terms of a continuous mapping of the random walk process, the continuous-mapping theorem of weak convergence theory (cf. [31]) implies that the waiting-time process converges weakly to a process that is the same continuous mapping of Brownian motion. This process is *reflected Brownian motion* (*RBM*), that is, Brownian motion restricted to lie in the positive orthant. Since this process is an example of a diffusion process, the resulting heavy-traffic approximation is an example of a *diffusion approximation*. The weak-convergence theory which forms the basis for establishing that the waiting-time process converges to reflected Brownian motion is often referred to as a *functional central-limit theorem* (*FCLT*), since it is a function-space generalization of the classical *CLT*.

2.3.1 Single-Class Queueing Networks

Generalization of these results to an open networks of queues began with the PhD dissertation of Marty Reiman [262] in 1978 (see also Harrison and Reiman [140], Reiman [263], Harrison and Williams [143]). The type of network considered has been called a *generalized Jackson network*. Like a Jackson network, it has a single class of jobs, all following the same Markovian routing between nodes, but job interarrival and service times are no longer required to be exponentially distributed. (They are required to be i.i.d. at each node and independent between nodes.)

The research program established in Reiman's dissertation has since become standard operating procedure in papers on diffusion approximations to queueing systems. There are two distinct steps: (1) proving limit theorems; and (2) defining the limiting process. Reiman was able to carry out both steps in the case of a generalized Jackson network. (In some more complicated models studied since then, it has been possible only to carry out the second step.

That is, one can conjecture the form of the limiting diffusion process and then characterize its probabilistic structure – in terms of its Ito equation (cf. Harrison [142], [146]) or the partial differential equations satisfied by its probability measure – but not prove that the original process weakly converges to the diffusion process in question.) The general form for the limiting diffusion process in a generalized Jackson network is a *semi-martingale reflected Brownian motion* (*SRBM*) – so called because the stochastic differential equation (*SDE*) or equivalently the stochastic integral equation defining the process has a Brownian term plus a term that is of bounded variation and adapted to the filtration (loosely speaking, the history) of the process itself. This term is sometimes called the compensator. The Brownian motion process is the limit of the *netput* process, which represents (in vector form) the difference between the cumulative input at each node and the cumulative output that would occur if none of the queue lengths were constrained to be non-negative. The compensator is a vector process each of whose components increases only when the corresponding component of the queue-length process equals zero. It is best understood by the physical interpretation that it "pushes" up on the contents of a node just enough to keep it from becoming negative.

In some simple cases, the diffusion-process limit can be characterized analytically, with closed-form expressions at least for the steady-state probabilities. Where this is not possible, one can resort to numerical calculations. Considerable effort has gone into developing efficient numerical methods. Most prominent are the *QNET* procedure (see Harrison and Nguyen [148], Dai and Harrison [72], [73], [74]), which solves the differential equations for the stationary density function and expresses the solution as a series of polynomial functions, and the methods of Kushner and Dupuis [213], [214] based on approximation of the process by a discrete-state Markov process.

2.3.2 Multi-Class Queueing Networks

In a multi-class queueing network (*MCQN*) there are different types of jobs, whose service requirements and routes through the network may depend on the job type. The by-now-standard model is one proposed by Mike Harrison in a talk delivered at the *IMA* Workshop on Stochastic Differential Systems, Stochastic Control Theory and Applications at the University of Minnesota in 1986 and subsequently published as a paper in the workshop proceedings [144]. This paper set the agenda for future research in heavy traffic analysis of a *MCQN*, both with and without control. Indeed, the suggested modelling framework has since been adopted by most researchers concerned with multi-class stochastic networks, whether in their original versions or in the form of a Brownian or fluid approximation.

In this modelling framework, the *class* of a job indicates both its type and the node at which it is currently being processed. Thus each node corresponds to a set of job classes, while each job class corresponds to exactly one node. Type-dependent routing and service requirements are handled by allowing the routing probabilities and service requirements to depend upon the class, not just the node. In all other respects, the *MCQN* behaves like a generalized Jackson network. The *MCQN* model is quite versatile, allowing for modelling of complex telecommunication systems and manufacturing systems, including job shops and flexible manufacturing systems, in which different product types follow different paths through the system, and *re-entrant lines*, in which all jobs follow the same path, but re-visit certain machines for different operations.

An example of the latter is semiconductor wafer fabrication.

The generalized Jackson network (homogeneous jobs: Reiman [263]) is the special case of a *MCQN* in which there is a one-one correspondence between stations and classes. Another special case is the model of Kelly [181], in which jobs belong to different types, each type has a Poisson arrival process and a fixed route through the network, and all jobs served at a particular station have a common exponential service-time distribution. The Markovian routing assumption is essentially without loss of generality, inasmuch as the class designation can incorporate information about past processing history and there can be arbitrarily many classes.

Peterson [248] developed the heavy-traffic theory for a multi-class feedforward network and conjectured that his results would extend to an arbitrary multi-class network under similar conditions, including the "obvious" condition that the total work input to each node should be less than the processing capacity of the node. (The conjecture was later shown to be false. Indeed, the "obvious" condition referred to above is not even sufficient for stability. See Section 2.4.) Harrison and Williams [149] focused on the steady-state distribution for a feedforward network in heavy traffic and gave conditions under which it has a product form. Under these conditions (which include the condition that the coefficients of variation of the service times be equal at all but the last node) the nodes of the network are quasi-reversible. Taylor and Williams [315] considered a general multi-class Brownian network (an *SRBM* in the orthant) and gave necessary and sufficient conditions for the existence and uniqueness of the *SRBM* representation of the process. The conditions include the requirement that the matrix of processing coefficients that characterizes the class and network structure be of a particular form (*completely S*). Bramson [46] and Williams [342] provide a survey of the state of the art as of 1998. The recent books by Kushner [214], Chen and Yao [58], and Whitt [336] also provide comprehensive coverage of heavy traffic theory.

We shall revisit *MCQN*'s in Section 3.2 when we discuss models for control of queueing systems.

2.3.3 The Halfin-Whitt Regime

In all of the heavy traffic research described thus far, the number of servers per service station is assumed to be small, and a general principle for that "conventional" parameter regime is that the heavy-traffic behavior of a multi-server system is indistinguishable from that of an equivalent and readily identified single-server system. (Early work by Kingman along these lines, for *GI/GI/s* systems, was cited earlier.) Approximations based on this principle are generally not very accurate, although they may be provably "correct in the limit," and for systems with many servers (not just two or three), they are so crude as to be useless.

In response to this state of affairs, Halfin and Whitt [134] proposed a "many-server" heavy-traffic regime that is most easily explained in terms of $M/M/c$ systems. Let us suppose that the arrival rate, λ, and the number of servers, c, both approach infinity while the service rate, μ, remains constant, in such a way that the traffic intensity $\rho := \lambda/c\mu$ approaches one and more specifically, $\sqrt{c}(\rho - 1) \to \theta$ (a finite constant). This is the Halfin-Whitt heavy-traffic parameter regime, and denoting by $Q(t)$ the number of customers in the system at time t, those authors studied the centered and scaled process,

$$Z(t) = s^{1/2}[Q(t) - s] \ , \ t \geq 0 \ .$$

In the limiting regime identified immediately above, Halfin and Whitt showed that Z is well approximated by a diffusion process $Z*$ which behaves like Brownian motion with drift θ to the right of the origin, but behaves like an Ornstein-Uhlenbeck process to the left of the origin. In the stable case where $\theta < 0$ (that is, $\rho < 1$), the stationary distribution of $Z*$ can be written out in an explicit formula, and the corresponding Halfin-Whitt approximations for steady-state performance measures are generally very accurate. This work on heavy-traffic approximations for many-server systems did not attract a great deal of attention at the time of its appearance, but interest in the Halfin-Whitt regime is now surging because of its relevance in telephone call-center applications.

2.4 Stability and Fluid Models

One of the articles of faith of queueing theory is that any system in which each service facility has sufficient capacity to process all work that passes through it (on the average) should be stable. The exact meaning of stability depends on the problem context and the model under consideration. At the very least, "stable" means that the contents (e.g., queue length, work) at each facility should be $o(t)$ as $t \to \infty$. Equivalently, the output rate should equal the input rate at each facility. This is called *rate stability* and, because it deals in long-run averages, can be analyzed on a pathwise basis (i.e., deterministically) – see El-Taha and Stidham [98]. For systems with sufficient probabilistic structure, "stable" can be strengthened to mean that there exists a proper limiting or at least stationary distribution for the system state (e.g., the vector of contents at each facility).

In classical Markov models proving stability is often a by-product of finding the (unique) stationary distribution. For example, an irreducible *DTMC* or *CTMC* is stable (positive recurrent) if and only if there exists a proper probability distribution that satisfies the stationary equations, in which case this is the unique stationary distribution (and the unique limiting distribution if the system is aperiodic). The necessary and sufficient condition for stability (e.g., $\rho < 1$ in an *M/M/1* queue) often "falls out" of the solution of the stationary equations. An alternative approach to proving stability is to look for a Lyapunov function (e.g., Foster's criterion [109] – roughly speaking, a potential (reward) function of the system state that monotonically increases over time and hence approaches a limit if and only if the system is stable. The situation in Markov processes with general state space is more delicate, but the basic approach is similar. Meyn and Tweedie [229] is an excellent reference for this subject.

Establishing stability in non-Markovian systems is trickier. In the context of queues and related fields (e.g., Petri nets) the approach of Loynes [222] can be useful. Loynes analyzed a *G/G/1* queue with strictly stationary interarrival and service times. Using the pathwise recursion established by Lindley [217] for the waiting time in the queue, Loynes showed that the waiting-time process (starting from an empty system) monotonically increases over time and (provided $\rho < 1$) approaches a strictly stationary process. Extending this approach to more complicated (e.g., multi-dimensional) systems depends on finding similar monotonicity properties in the recursive operator relating the system states at successive points in time.

(Kingman provided a unifying treatment of this approach in his “algebra of queues” [192], [193], [194], [196].) For generalizations see Baccelli and Liu [12], Foss [108], Baccelli and Brémaud [13].

An alternative (in principle) is to enlarge the state space of the process in order to render it Markovian and then apply the well-developed theory for stability in Markov processes with a general state space (cf. Meyn and Tweedie [229]). This technique, sometimes called the *method of supplementary variables*, has a long history in queueing theory (cf. Cox [64] for the case of discrete supplementary variables). For example, in a multi-class queueing network (*MCQN*: see Section 2.3) one can add the elapsed times since the last arrival and last service completion at each job class to the vector of number of jobs in each class to obtain a Markovian state descriptor. While often not a fruitful approach for deriving explicit expressions for time-dependent or steady-state probabilities, this approach can lead to useful results regarding stability, when combined with other techniques (see below).

A significant new addition to the arsenal of techniques for proving stability resulted in the early 90's from a negative, and very surprising result. Kumar and Seidmann [208] provided an example (albeit in the framework of a deterministic model) of a *MCQN* in which the traffic intensity at each node is less than one but the system is not stable. Indeed, the queue lengths fluctuate dramatically, with buffers alternately emptying and growing and the peak buffer levels approaching infinity. (Here the traffic intensity at a node is the average total rate at which work of all classes arrives at the node, from both inside and outside the network. Service times of different classes of jobs at a node may have different distributions. Work is measured in units of service time and the server at each node is assumed to process work at unit rate.) At about the same time Dai and Wang [75], in the course of trying to develop heavy-traffic approximations (Brownian networks) for multi-class queueing networks, found examples in which there is no Brownian limit when one examines the usual sequence of processes indexed by n in which time is scaled by n^{-1} and space by $\sqrt{n}^{-1}$. Rybko and Stolyar [268] developed a stochastic model for a multi-class queueing network which exhibits the same anomalous behavior as the deterministic example of Kumar and Seidmann [208], namely, the system is not stable although the traffic intensity at each node is less than one. They used a fluid model to analyze the stability of their model. Bramson [45] was a particularly significant contribution, in that it presented the first example of this phenomenon in the context of a *MCQN* with *FIFO* discipline at each node.

2.4.1 Deterministic Fluid Models

Dai [76], [77], [78], [79], [80] developed this idea in the context of a general multi-class queueing network, producing a useful and widely applied technique for analyzing stability. He showed that a *MCQN* is stable (in the sense of having a proper limiting distribution for the state of the system) if and only if the “equivalent fluid system” is stable (in the sense that the contents of the buffers at all nodes eventually reach zero from any starting state). The equivalent fluid system is a system with the same network and class topology and routing probabilities, in which discrete jobs are replaced by continuous fluid and inputs and outputs are at deterministic rates equal to the reciprocals of the corresponding mean interarrival and service times in the original stochastic system.

One can think of the equivalent fluid system as playing a role similar to the Brownian (heavy-traffic) approximation (see Section 2.3), except that one uses the space scaling n^{-1} associated with the law of large numbers, rather than the scaling $\sqrt{n}^{-1}$ associated with the central limit theorem. As a result the approximating system is deterministic rather than stochastic. There is now a well-developed literature on deterministic fluid systems and their behavior. The monograph by Chen and Mandelbaum [55] established notation and some basic theory. (Companion papers [56], [57] discuss the Brownian approximation and heavy traffic.)

2.4.2 Stochastic Fluid Models

A somewhat different line of research has focussed on stochastic fluid models. In these models the input and/or output rates of fluid are allowed to depend on an exogenous, random environment (usually at *CTMC*). Such models have been proposed for flexible manufacturing systems and telecommunication networks, in which individual units (products, packets) are processed so rapidly that they can reasonably be modelled as fluid instead of discrete units. The state of the environment can represent, for example, the number of operating machines (in a manufacturing setting) or the number of active sources (in a telecommunication setting). The papers of Kosten [206] and Anick, Mitra, and Sondhi [5] were influential in setting the research agenda in this area. The emphasis in these models has been on deriving explicit expressions for (transforms of) steady-state distributions. For a survey of the state of the art as of 1996, see Kulkarni [207].

2.5 Computational Probability

In the early days of applied probability researchers paid little attention to numerical computations. This was particularly true of queueing theory. It was standard practice to end the analysis of a model with a formula for the generating function or Laplace transform for a random variable of interest, such as the steady-state waiting time or queue length. A comment to the effect that the transform can be inverted by "standard techniques" might have followed the formula, but in many cases there was no such comment. One simply took it for granted that the transform was in effect the solution to the problem. Without taking away from the mathematical accomplishments of the early researchers in applied probability, it is fair to say (and it was frequently said in the late 60's and early 70's) that the "Laplacian curtain" kept a lot of good theory from being understood, let alone applied.

One of the first researchers to do more than complain about this situation was Marcel Neuts – one of most prolific developers (by his own admission) of the classical, transform-based theory. At the conference on Mathematical Methods in Queueing Theory at Western Michigan University (Kalamazoo) in 1993, Marcel began to set forth his philosophy about the need for "numerical probability", or perhaps more accurately, "computational probability". Perhaps Marcel's most important contribution to applied probability was to make computational probability respectable, at least in queueing theory, both as a topic for research and as a tool for experimentation with models that do not have analytical solutions. Marcel has provided some personal reminiscences, which appear later in this article, as well as on several previous occasions (see [235], [237], [239], [240], [242]), so I will be brief in my review of his work.

Within the general framework of computational probability, Marcel proposed two basic models, which are the subject of his two influential books on the subject [234], [238]. The first of these focusses on queueing systems that have a transition structure that generalizes that of *GI/M/1* queue. The second book considers systems with a transition structure that generalizes that of the *M/GI/1* queue. The unifying theme in both books – as well as in most of Marcel's papers on computational probability – is the phase-type approximation, in which non-exponential distributions are approximated by distributions that are built up from exponentially distributed phases. The idea – which goes back (at least) to Cox [64] – can best be understood by giving it a physical interpretation in the particular setting of a single-server queue.

Imagine that a job beginning service enters a network with a finite number of nodes, choosing the first node to visit according to a given distribution. The job spends an exponentially distributed length of time at a node (with a parameter that may be node-dependent) and then either exits the network (which corresponds to completion of service) or goes to another node for more processing. The node next visited is chosen according to a Markov transition probability matrix. The structure of this network is exactly that of a Jackson network (see Section 2.2 and Jackson [170], [171]) with single-server nodes, with the important distinction that in the present setting only one job may occupy this network at a time (since only one job can be in service at a time). The time spent at each node is called a *phase* and the distribution of the total time spent by a job in the network is called a *phase-type* distribution. The phases might correspond to different tasks that must be performed as part of the service of a job, with branches in the network corresponding to different job types or (in the case of feedback) the need for re-work when a task is not properly completed the first time. But it is important to note that this physical interpretation need not correspond to a physical system in order to be useful.

Phase-type distributions include the extreme cases of a k-Erlang distribution, which corresponds to the sum of i.i.d. exponential-$k\mu$ phases (a series network), and the hyperexponential distribution, which can be modelled as a network in which an entering job chooses randomly among parallel exponential nodes and leaves the network upon completion of a single phase.

The distribution of any non-negative random variable can be approximated arbitrarily closely by a phase-type distribution. (In fact this is also possible with a subset of phase-type distributions – the *mixed generalized Erlang* (*MGE*) distributions. A distribution is *MGE* if it can be represented by a network consisting of a series of exponentially distributed phases, with the property that a job may leave the network upon completion of a particular phase with a specified probability. Equivalently, it is Coxian in the sense of Cox [64], but with real rates.) For numerical calculations, it is important to find an approximation with as small a number of phases as possible, since the size of the resulting system model (see below) is proportional to the number of phases. Considerable work has been directed toward this issue.

The approximation of non-exponential distributions by phase-type distributions allows one to model a non-Markovian system (approximately) as a continuous-time Markov chain (*CTMC*), in which the state variable has been augmented to include the phase currently occupied by the job in service (or in the "pre-arrival" process, if it is the inter-arrival-time distribution that is being approximated by a phase-type distribution). The resulting *CTMC* models have special structure which makes it possible to solve them numerically in an efficient way, even (indeed especially) in the case of an infinite state space. For example, when one applies a phase-type approximation to a *GI/M/1* queue (or a more complicated variant) one finds that the resulting steady-state distribution is of *matrix-geometric* form – a matrix generalization of the geometric distribution in the *GI/M/1* queue. Similarly, when the process resembles an *M/GI/1* queue, the steady-state solution is a matrix generalization of the distribution in an *M/GI/1* queue.

As an example of a system that is amenable to phase-type methods, consider a $M/M/c$ queue operating in a random environment. In such a system, the arrival and/or service rate are allowed to depend on the state of the environment, which itself evolves as an exogenous finite-state *CTMC*. Such a model is useful in the analysis of telecommunication systems, in which the number of active sources (each generating a Poisson process of messages) varies over time, or in manufacturing systems, in which the number of active servers may similarly vary because of breakdowns and repairs. In this case the phases have a direct physical interpretation, with each phase corresponding to a particular state of the environment.

Marcel Neuts's tireless proselytizing on behalf of computational probability has benefited not only those using the phase-type approach, but also those committed to research on other techniques for doing efficient numerical computations in applied probability. First note that the phase-type approach allows one to do numerical computations for problems with an infinite number of states without truncating the state space, provided there is sufficient structure. If one restricts attention to irreducible Markov models (in discrete or continuous time) with a finite number of states, however, then other approaches are possible. In particular, the problem of calculating the steady-state distribution becomes a problem in numerical linear algebra, since the steady-state probabilities are the unique solution to the stationary equations – a linear system. Thus all the machinery of numerical linear algebra can be brought to bear on this problem. One can use direct methods, such as Gauss elimination, L-U decomposition or indirect (successive-approximation) methods, such as the power method (which corresponds to iterative computation of the time-dependent probabilities in the case of a discrete-time Markov chain (*DTMC*)), Gauss-Seidel, Jacobi, successive over-relaxations. Many researchers from numerical analysis have been drawn to Markov-chain models as a rich source of applications, and the result has been an abundance of experimentation with different algorithms.

An excellent reference for numerical analysis of Markov chains is the book by Stewart [291]. See also Stewart [290], [292], which are volumes of proceedings of a series of conferences on numerical methods in Markov models, organized by Stewart and colleagues.

Within the OR community there have also been a number of innovative approaches to computational applied probability, in addition to the work of Neuts. Grassmann, Taksar, and Heyman [130] have proposed iterative numerical schemes that exploit the structure of a Markov chain, using an imbedding technique in which one successively solves for the steady-state distributions for the system imbedded at points of transition within larger and larger subsets of the state space. This method turns out to be a variant of Gauss elimination, but the derivation from a probabilistic motivation results in an algorithm with more stable numerical properties than ordinary Gauss elimination.

Recent work on efficient numerical techniques for inversion of transforms has breathed new life into the use of generating functions and transforms in applied probability. Much of this work was carried out by Joseph Abate, Gagan L. Choudhury, Kin K. Leung, David M. Lucantoni and Ward Whitt in a sequence of fifteen papers on numerical techniques for transform inversion and application of those techniques to various stochastic models, which received Honorable Mention for the *INFORMS* Lanchester Prize in 1997. For an overview of these techniques see the survey by Abate, Choudhury, and Whitt [1] in the recent book edited by Grassmann [129], which provides an indication of the state of the art in computational applied probability.

2.6 Priority Queues, Polling Systems, and Queues with Vacations

Priority queues have been studied since the early days of queueing theory. The monograph on queues by Cox and Smith [65] gives a concise summary of the early work. The book by Jaiswal [173] provides a compendium of known results as of the late 60's. Most of the research until then concerned single-server queues with fixed priorities, operating under preemptive or non-preemptive disciplines. In a fixed priority scheme, the server always selects a job from the class with the highest priority among those in the queue. Under a pre-emptive discipline, the job currently in service is removed and replaced by a job with higher priority if such a job arrives during its service. Under a non-preemptive discipline, a service cannot be interrupted once started. Priority queues are most amenable to analysis when the arrivals come from Poisson processes, so most models have been of *M/GI/1* type.

A polling system is a single-server, multi-class queueing system in which the server "polls" the various classes, serving jobs in each class for a certain length of time, then switching to another class, and repeating this process until all classes have been polled, at which point the polling process begins again. The most common models assume *cyclic polling*, that is, the server visits the m classes in a fixed order, which can be labelled $1, 2, \ldots, m$, without loss of generality. There are several possible rules for determining how long the server spends at each class, the most common being the *exhaustive* and *gated* disciplines. The exhaustive discipline continues serving a class until no jobs are present, and then switches to the next class. The gated discipline serves only those jobs that are present at the instant the server begins serving that class and then switches. Variants include the *globally gated* discipline, in which during a particular cycle the server serves only the jobs that were present in each class at the beginning of the cycle, and *limited* (exhaustive or gated) disciplines, in which at most a certain fixed number of jobs are served in each class in each cycle. Models also differ according to whether or not there is a switchover time (sometimes called the "setup" or "walking" time) associated with the server moving from one class to another.

Applications of polling systems occur in many areas. The term "polling system" apparently originated in the telephone industry, where there are several applications. For example, a switch may poll each of the input channels to see if there is incoming traffic, or a telephone handset may poll the line to determine whether a number is being dialled. The earliest application of queueing theory to a polling system (although the term "polling" was not used in the paper) is the paper by Eisenberg [97], which was motivated by telephony applications.

In traffic-flow theory, a vehicle-actuated traffic signal at the intersection of two one-way streets may be modelled as a polling system with two classes. The exhaustive discipline can be implemented if the intersection has a detector in each street just before the intersection. The traffic light changes when the detector fails to detect a vehicle in the street that is currently in the green phase. The yellow phase (plus a start-up delay) is the switchover time in this case. Traffic-flow models of this type were developed beginning in the 60's. Almost all models in both the telephone-system and traffic-flow literatures assume Poisson arrivals and are thus *M/GI/1* type models.

In the papers in the traffic-flow literature, the exhaustive discipline is often called the *zero-switch* rule or the *alternating priority* queue discipline. Indeed, an *M/GI/1* queue operating under the exhaustive discipline resembles a queue with a fixed, non-preemptive priority rule, and the solution techniques are also similar. In both cases, when the server switches to the class with highest priority, it serves all k (say) jobs present, plus future arrivals, until there are no jobs present, and then switches to another class. The time spent serving that class is thus distributed as a k-busy period – a busy period initiated by k jobs – and it follows from an argument due to Takács [307] that the length of such a busy period is distributed as the sum of k ordinary busy periods. The k jobs present at the beginning of this period are just those that arrived over the time interval since the server's last visit to this class. Since the arrivals are from Poisson processes, the distribution of k can be calculated from the distribution of this time interval, which in turn is made up of one or more busy periods for other classes. The only difference between the alternating-priority and fixed-priority disciplines is that in the former each class in its turn has the highest priority.

These considerations make it possible to set up a system of equations satisfied by the mean workloads or the transforms of the workload distributions, using *PASTA* (*P*oisson *A*rrivals *S*ee *T*ime *A*verages: cf. [346]). These equations can be solved for closed-form expressions. For mean values, this was done by Wolff [345] for fixed priorities and Stidham [295] for the alternating-priority discipline. The techniques – which are simple and intuitive and have been frequently re-discovered – have since become known under the rubric *mean-value analysis* (cf. Reiser and Lavenberg [265] and applied in a variety of contexts, including networks of queues. For transform solutions, see Jaiswal [173] for fixed priorities and Avi-Itzhak, Maxwell and Miller [10], Neuts and Yadin [233] for the alternating-priority discipline.

The earliest papers in which the term "polling" is used appear to be Hashima [154], Gaver [117], and Konheim and Berndt [203]. Beginning in the early 1980's, there was an explosive growth of research on polling systems, motivated apparently by the increasing number of applications to computer and communications systems. The gated discipline and other alternatives to the exhaustive (alternating-priority) discipline were soon introduced, as well as non-cyclic polling and the effect of switching times. The book [308] and survey paper [309] by Takagi provide an overview of the models, techniques, and results as of the mid-to-late 1980's. Takagi [310] and [312] are subsequent updates. See also Levy and Sidi [216]. Research continues to the present – a testimony to the versatility of the polling system model – motivated by applications to modern communication networks with differentiated services. Traditional descriptive modelling, in which steady-state distributions or transforms are derived, is still the focus of a significant number of papers. Meanwhile, other lines of research have opened up, including decompositions (see below), optimization of polling schemes (discussed below in Section 3), stability analysis and bounds (discussed above in Section 2.4), and heavy-traffic approximations to polling systems, both without control (see Section 2.3) and with control (see Section 3.2).

Research on vacation systems has paralleled and frequently overlapped research on polling systems. A vacation system is a queueing system in which the server intermittently spends time away from the queue, perhaps because of a breakdown and repair or because it is serving other jobs. Thus a polling system, seen from the vantage point of a particular job class, is an example of a vacation system, in which the vacation corresponds to the time spent serving other job classes (plus switchover times). The methods of analysis are similar to those used for polling systems, but, because of the focus on one class, more general results have been obtained. One of the most significant results of the research on vacation systems has been the discovery that the waiting time in the queue in a *M/GI/1* queue with vacations is distributed as the sum of two independent components, one distributed as the waiting time in the queue in the corresponding *M/GI/1* queue without vacations and the other as the equilibrium residual time in a vacation. Fuhrmann [111] (see also Fuhrmann and Cooper [112], Fuhrmann [113]) was apparently the first to identify and prove this *decomposition* property, which holds under very general conditions. (For example, the length of the vacation can depend on the previous evolution of the queue and the amount of work to be served when the server returns from vacation can depend on the length of the vacation.) Subsequent generalizations were by Doshi [88], [89], [90], who gave a very general version of the decomposition property.

The robustness of the property in its mean-value form can best be understood by recognizing that it is a *sample-path property*, that is, it holds on every sample path (realization) of the processes involved for which the relevant limiting averages (e.g., the limiting average waiting time in the queue) are well defined and finite. For a simple proof that exploits this fact, see Bertsekas and Gallager [23], pp. 147–149. (For a compendium of sample-path results in queueing theory, see El-Taha and Stidham [98].)

Boxma and Groenendijk [43] established pseudo-conservation laws for polling and vacation systems with switching times (see also Boxma [44]). Recent papers on decomposition in polling and vacation models include Borst and Boxma [42] and Bertsimas and Mourtzinou [30] and Takine [313], who use the distributional form of Little's Law (cf. Haji and Newell [133], Keilson and Servi [178], [179], Bertsimas and Nakazato [28]).

2.7 Effective Bandwidth, Large Deviations, and Heavy Tails

Often in queueing models, means and variances do not tell the whole story. Indeed, sometimes one is most interested in tail behavior, in particular the probability that a random variable or stochastic process in question exceeds a specified extreme value.

2.7.1 Exponential Tails

When tail probabilities are exponentially bounded, one is assured that extreme values are very unlikely. A classical route to exponential tails is via moment generating functions. Let X be a non-negative random variable and suppose

$$M(s) := \mathrm{E}[e^{sX}] < \infty \ , \ 0 \leq s < \delta \ . \tag{1}$$

Then the tail probabilities of X are exponentially bounded. Specifically, applying Chebyshev's inequality to the random variable e^{sX} yields

$$\mathrm{P}\{X > a\} \leq e^{-sa} M(s) \ , \ a > 0 \ . \tag{2}$$

This result has been applied in many different settings, some of which are illustrated below.

2.7.2 Effective Bandwidth

In an integrated-services, high-speed communication network, some classes of traffic are very sensitive to delays or losses, and a network designer would like to be able to guarantee that the probability of a large delay or a buffer overflow is smaller than some pre-specified limit. As an example, consider a *M/GI/1*-type model for a buffer in a communication network (cf. Kelly [183]). The cumulative input from a source (or sources) to the buffer (measured, e.g., in bits or packets) comes from a compound Poisson process, in which batches (bursts) of size X arrive at rate λ. The buffer is drained by an output channel at constant rate c. The buffer-content process behaves like the workload process in an *M/GI/1* queue.

The well-known duality between the *M/GI/1* workload process and the assets process in the classical insurance-risk model (cf., e.g., Chapter 1 of Embrechts, Klüppelberg, Mikosch [99]) makes it possible to apply the Cramér-Lundberg approximation to the tail of the buffer-contents distribution in steady state. First suppose that $\lambda \mathrm{E}[X] < c$. This is just the natural stability condition for steady state: the expected input rate to the buffer must be less than the processing rate of the channel. Let ν be the (unique) solution to the equation

$$\lambda \int_0^\infty e^{\nu x} \mathrm{P}\{X > x\} dx = c \, . \tag{3}$$

Then

$$\mathrm{P}\{Q > b\} \sim C e^{-\nu b}$$

as $b \to \infty$ (more formally, $\lim_{b\to\infty} e^{\nu b}\mathrm{P}\{Q > b\} = C$), where Q is the steady-state buffer contents and C is a finite positive constant. A solution to (3) does not exist unless (1) holds, and we have seen that condition (1) holds only if the service-time distribution has an exponentially bounded tail.

The equation (3) satisfied by ν can be written in equivalent form as

$$\alpha(\nu) = c \, , \tag{4}$$

where $\alpha(s) := (\lambda/s)(M(s) - 1)$. Following Kelly [183] (and for reasons that will become clear presently) we shall call $\alpha(s)$ the *equivalent bandwidth* (*EBW*) of the (compound Poisson) input process. Equation (4) reveals that the critical exponent ν that characterizes the exponential tail behavior of the buffer-contents distribution is found by equating the *EBW* of the input process to the (constant) bandwidth c of the output channel. Another way of interpreting this result (which helps motivate the definition of effective bandwidth) is to observe that, in order to guarantee that the tail of the buffer-contents distribution is asymptotically exponential with exponent no smaller than ν, the bandwidth c of the output channel must be at least equal to the effective bandwidth $\alpha(\nu)$ of the input process. (This follows from the fact that $\alpha(s)$ is non-decreasing in s.)

Now, if $\{A(t), t \geq 0\}$ is a compound Poisson process with arrival rate λ and batch-size m.g.f. $M(s)$, the moment-generating function of $A(t)$ is

$$\mathrm{E}[e^{sA(t)}] = \exp\{\lambda t(M(s) - 1)\} = \exp\{ts\alpha(s)\} \, .$$

More generally, this form for the moment-generating function holds for any process $\{A(t), t \geq 0\}$ with stationary independent increments (Lévy process), for an appropriately defined function

$\alpha(s)$. Hence the superposition, $A(t) = \sum_j A_j(t)$, of independent Lévy processes is a Lévy process with *EBW* equal to the sum of the *EBW*'s of the component processes:

$$\alpha(s) = \sum_j \alpha_j(s) \; .$$

This property makes it possible to analyze models for *multiplexing* – serving several input sources by a single channel with a common buffer – provided the sources are independent and generate input according to Lévy processes. For example, if one wishes to guarantee that the tail of the buffer-contents distribution is exponential with exponent no smaller than ν, then it follows from the above observations that new sources can be added as long as the sum of their *EBW*'s (with $s = \nu$) does not exceed c.

To what extent do results like these depend on having Lévy input processes? This question motivates the following extension of the definition of effective bandwidth (cf. Kelly [183]). Consider a source with a cumulative input process $\{A(t), t \geq 0\}$ with stationary (but not necessarily independent) increments. Define the *effective bandwidth* of the source by

$$\alpha(s,t) := \frac{1}{st} \log \mathrm{E}[e^{sA(t)}] \; . \tag{5}$$

(Note that, for a source with a Lévy input process, $\alpha(s,t)$ does not depend on t, in which case the definition coincides with the previous definition, that is, $\alpha(s,t) = \alpha(s)$.) The extended definition (5) preserves the additive property of effective bandwidths of independent multiplexed sources; that is, if $A(t) = \sum_j A_j(t)$, where the $\{X_j(t)\}$ are independent, then

$$\alpha(s,t) = \sum_j \alpha_j(s,t) \; .$$

To exploit these properties in the setting of the above single-buffer model, suppose the input process $\{A(t), t \geq 0\}$ has stationary and ergodic increments, that $\lim_{t\to\infty} \alpha(s,t) = \alpha(s)$, there exists a finite constant ν such that $\alpha(\nu) = c$, and $\alpha'(\nu)$ is finite. Then

$$\log \mathrm{P}\{Q > b\} \sim -\nu b \; , \tag{6}$$

as $b \to \infty$ (cf. Chang [53]). Hence the tail probabilities for the buffer contents are once again asymptotically exponential, with parameter ν found by equating the effective bandwidth, $\alpha(\nu)$, of the total input process to the bandwidth, c, of the output channel.

The derivation of (6) uses the theory of large deviations (see, e.g., Chang [53], Theorem 3.9). A good source for this theory and its applications to queueing systems is the book by Shwartz and Weiss [283]. The next section summarizes some of the key results.

2.7.3 Large Deviations

As we have seen, if X is a non-negative random variable satisfying condition (1), then the tail probabilities are exponentially bounded: $\mathrm{P}\{X > a\} \leq e^{-sa}M(s)$, $a > 0$. Indeed, this is a family of upper bounds, one for each s. For a particular fixed value of a, the tightest upper bound is given by

$$\mathrm{P}\{X > a\} \leq \inf_s e^{-sa} M(s) \; , \; a > 0 \; .$$

Define $l(a) := -\log(\inf_s e^{-sa} M(s))$, $a > 0$. The function $l(a)$ is called the *rate function* associated with X. It follows that

$$\mathrm{P}\{X > a\} \leq e^{-l(a)} \; , \; a > 0 \; . \tag{7}$$

An equivalent representation for the rate function is: $l(a) = \sup_s\{s(a - \alpha(s))\}$, where in this setting

$$\alpha(s) := \frac{1}{s} \log M(s) = \frac{1}{s} \log \mathrm{E}[e^{sX}] \; .$$

Note that this definition is consistent with (5) when X is the cumulative input from a source over the time interval $[0, 1]$ ($X = A[0, 1]$), so we may legitimately interpret $\alpha(s)$ as the *EBW* associated with X.

Now suppose we wish to bound the tail probability, $\mathrm{P}\{X > a\}$ by $e^{-\gamma}$ for some positive γ. From (7) we see that it suffices to have $l(a) \geq \gamma$, or equivalently,

$$\inf_s\{s(\alpha(s) - a)\} \leq -\gamma \; .$$

Kelly [183] has shown how to apply this inequality to the case of multiplexed sources. Suppose

$$X = \sum_{j=1}^{J} \sum_{i=1}^{n_j} X_{ji} \; ,$$

where the X_{ji} are independent, and for each j the X_{ji} have the same distribution, with *EBW*'s $\alpha_j(s) = \frac{1}{s} \log [e^{sX_{ji}}]$. So there are J types of sources, with n_j of type j. It follows that $\alpha(s) = \sum_{j=1}^{J} n_j \alpha_j(s)$. In this case the tail-probability constraint will be satisfied if the vector $n = (n_1, \ldots, n_J)$ belongs to the set

$$A := \left\{ n : \inf_s\{s(\sum_{j=1}^{J} n_j \alpha_j(s) - a)\} \leq -\gamma \right\} \; .$$

Since A has a convex complement, any supporting hyperplane to this complement can be used to generate a conservative global bound. That is, if the linear inequality

$$\sum_{j=1}^{J} n_j \alpha_j(s^*) \leq a - \frac{\gamma}{s^*}$$

holds, then $\mathrm{P}\{X > a\} \leq e^{-\gamma}$. Here s^* is the value of s that attains the infimum at the point where the hyperplane touches the boundary between A and its complement.

Large-deviation theory is typically used to obtain asymptotic expressions for probabilities of the form

$$\mathrm{P}\{X_1 + \ldots + X_N > Na\} \; ,$$

where the X_i are i.i.d. A repetition of the argument that led to (7) yields the following upper bound:

$$\mathrm{P}\{X_1 + \ldots + X_N > Na\} \leq e^{-Nl(a)} \; .$$

A more refined analysis (cf. Shwartz and Weiss [283], Chapter 1) shows that in fact

$$\mathrm{P}\{X_1 + \ldots + X_N > Na\} = e^{-Nl(a)+o(N)} .$$

An equivalent representation of this result (the *large deviation principle*) is the following:

$$\lim_{N\to\infty} \frac{1}{N} \log \mathrm{P}\{X_1 + \ldots + X_N > Na\} = -l(a) .$$

We can use this result to get an exact asymptotic expression for the probability that the total input from a source exceeds a certain level, as both the number of sources of each type and the level are scaled and the scale factor goes to infinity:

$$\lim_{N\to\infty} \frac{1}{N} \log \mathrm{P}\{\sum_{j=1}^{J} \sum_{i=1}^{Nn_j} X_{ji} > Na\} = -l(a) .$$

It follows that the approximation that led to the region A becomes exact in this asymptotic regime.

2.7.4 Heavy Tails

Recently empirical evidence has suggested that many of the distributions encountered in queueing applications do not have exponential tails. Examples are file sizes in telecommunication transmissions (see, e.g., Willinger [344], Crovella and Taqqu [70]), which are often more accurately modeled by distributions such as the lognormal, Pareto, and Weibull. (The same is true of claim sizes in insurance-risk applications. See Embrechts et al[99].) Such distributions are often called *heavy-tailed* (although this term has been used somewhat loosely in the literature). A useful sub-class of heavy-tailed distributions is the class of *sub-exponential* distributions. For the *M/GI/1* model discussed above, when the batch-size distribution is sub-exponential, there are alternatives to the Cramér-Lundberg approach which yield estimates of the tail probability of the buffer contents process (again, see Embrechts et al[99].)

Roughly speaking, a distribution is sub-exponential if, for all n, the tail of the n-fold convolution behaves asymptotically like n times the tail of the distribution itself. An equivalent characterization is that the tail of the maximum determines the tail of the sum of n i.i.d. random variables with the given distribution. A distribution F with power-law tail behavior ($\bar{F}(x) = x^{-\alpha}L(x)$ where $\alpha \geq 0$ and L is slowly varying) is an example of a sub-exponential distribution. (For a formal definition of sub-exponentiality, as well as a taxonomy of the various sub-classes of distributions with heavy tails, see Embrechts et al [99].)

For the *M/GI/s* queueing system, the effect of a heavy-tailed distribution for the service time (equivalently, the amount of work brought by each arrival) on the distribution of waiting time or queue size has been investigated by Whitt [335].

While it is often true that heavy-tailed distributions have an infinite variance, this is not necessary. For example, a distribution with $\bar{F}(x) = K \cdot x^{-\alpha}$ for sufficiently large x (a special case of power-law tail behavior) has a finite variance if (and only if) $\alpha \geq 2$. An important class of heavy-tailed distributions in which the variance may be infinite is the class of stable laws. There are various equivalent definitions of a stable distribution. A useful characterization is the following (see, e.g., Embrechts et l [99], Theorem 2.2.2): the class of stable distributions coincides with the class of all limit laws for properly normalized and centered sums of i.i.d. random variables. That is, a distribution is stable if and only if it is the limiting distribution of $b_n^{-1}(S_n - a_n)$ for a sequence of constants a_n and $b_n > 0$, where $S_n = X_1 + \ldots + X_n$ and the X_i are i.i.d. The most famous example, of course, is the normal distribution, in which $a_n = n\mu$ and $b_n = \sigma\sqrt{n}$, where μ and σ are, respectively, the mean and standard deviation of X_1. In this case the central limit theorem is the basis for the convergence. The normal is the only example of a stable law with a finite variance.

In Section 2.3 we saw that the central limit theorem can be extended to a functional central limit theorem (*FCLT*) to establish the weak convergence of certain stochastic processes to a Gaussian process (Brownian motion). Similarly, when a different normalization is used, weak convergence to stable processes can be proved. See, e.g., Samorodnitsky and Taqqu [270], Embrechts et al [99], and Whitt [334].

3 Control Models

3.1 Markov Decision Processes (Stochastic Dynamic Programming)

Richard Bellman usually gets credit for inventing dynamic programming. While one can debate issues of precedence (earlier work by Wald [327] and Massé (see Gessford and Karlin [120]) deserves mention), there is no question that Bellman coined the terms “dynamic programming” and “principle of optimality” and constructed a coherent framework for studying certain classes of sequential decision processes, which he set forth in a series of papers in the 1950’s (see, e.g., [19]) and an influential book, *Dynamic Programming* [20]. A sequential decision process is a system that evolves over time, with opportunities to influence its future evolution by taking actions at various (discrete or continuous) points in time. There may be costs and/or rewards, incurred continuously and/or at discrete points in time, that depend on the actions taken and the way in which the process evolves. The objective of the decision maker may be to maximize the total (expected) reward or minimize the total (expected) cost over a certain time horizon. If the horizon is infinite, then one may need to use discounting or long-run averaging to come up with a finite-valued objective. (An exception is the class of problems, including optimal stopping problems, in which a costless absorbing state is entered in a finite length of time with probability one.)

Dynamic programming is an approach to sequential decision processes that exploits a simple truism: nothing is ever lost by postponing a decision until the last possible moment. In doing so, one may gain information that will make a more intelligent decision possible. This information may, for example, be about the evolution of the process up to and including the point at which the decision must be made. Armed with this information, we may be able to make a more accurate prediction about the future evolution of the process and hence about future values (costs and/or benefits), given each possible choice for the action to be taken. Dynamic programming is most effective when the information needed to predict the future is contained in a simple sufficient statistic, called the *state* of the system. Of course, this is just a version of the Markov property, extended in this case to require that the future should depend only on the current action as well as the current state, but not on past actions or states. A sequential decision process with this property is called a *Markov decision process* (*MDP*).

Rather than a specific sequence of actions to be taken, the appropriate concept of a *solution* to an *MDP* is a *policy*, that is, a set of rules specifying the action to be taken at each of the possible observation points for each possible state (or more generally for each possible history). In a stochastic system, this is essential; since the future evolution of the system is unknown, we cannot specify future actions unconditionally. All we can do is answer "what-if" questions, e.g., what action should we take at time t if the state is x? Even in a deterministic system, formulating solutions in terms of policies turns out to be a useful approach.

Since the state contains all the information required to predict the future, it follows that the optimal action at a particular point in time depends *only* on the state at that time point, and not on the path taken by the process to reach this state. (A rigorous proof of this property in a general setting takes some effort, however.) Indeed, the same applies to the optimal policy: the optimal policy for the remainder of the problem horizon and its associated (expected) value depend only on the current state and (possibly) the current time point.

These observations lead to Bellman's Principle of Optimality. For processes which are observed at discrete time points, the Principle of Optimality says that, however we may have reached a particular state at a particular time point, an optimal policy will take an action now that optimizes the sum of the immediate value and the (expected) value associated with following an optimal policy from whatever state is reached at the next observation point. The mathematical expression of the Principle of Optimality is the optimality equation of dynamic programming, often called the Bellman equation. For example, in the case of an infinite-horizon discrete-time, discrete-state *MDP* with discounting and stationary rewards and transition probabilities, the optimality equation takes the form,

$$v(i) = \max_{a \in A} \{ r(i,a) + \beta \sum_{j \in S} p_{ij}(a) v(j) \} \ , \ i \in S \ .$$

Here S is the state space, A is the action space, $\beta < 1$ is the one-period discount factor, and $r(i,a)$ and $p_{ij}(a)$ are, respectively, the reward received at time t and the probability that the state at time $t+1$ is j, given that the state at time t is i and the action is a. The optimal value function v gives the maximal expected discounted reward over the infinite horizon, from each starting state i. In the case of a finite-horizon problem, the optimality equation is a recursive counterpart to the above equation, with v replaced by v_n (v_{n-1}) on the l.h.s. (r.h.s.), where the subscript indicates the number of stages remaining in the horizon, and v_0 is a given terminal reward function..

For discrete-time *MDP*'s, Howard [161] provided an accessible introduction to the finite-state case, both with and without discounting. Bellman had proposed two algorithms for this case, which he called *approximation in function space* and *approximation in policy space*. The former is often called simply *successive approximations* or *value iteration*. The latter is often referred to as *policy iteration*, or (thanks to Howard's proselytizing efforts) Howard's algorithm. We shall discuss both these algorithms presently. Another alternative is to use linear programming. The idea is to consider randomized policies and define *LP* decision variables that are (in the average-return case) the joint probability of being in a particular state and taking a particular action. (In the discounted case the decision variables can be interpreted as discounted frequencies with which states are visited and actions taken over the infinite horizon.) The optimality equation is relaxed to a set of linear inequalities, one for each state and action, and an objective function introduced that forces at least one inequality to be satisfied with equality for each state. At the optimal solution to the *LP* this inequality indicates the optimal action for that state. Manne [225] was the first to suggest the *LP* approach for solving *MDP*'s. Hordijk and Kallenberg [159], [160], among others, have extended this approach. The *LP* approach is especially well suited to constrained *MDP*'s, in which the optimal policy must satisfy side constraints. For example, one might wish to optimize with respect to a particular criterion (e.g., minimize delay in a queueing system) subject to a constraint that a second criterion (e.g., throughput) be no less than a specified value. See Altman [4] for an overview of research in this area.

Blackwell and Strauch, among others, put Bellman's theory on a firm mathematical footing, first for the case of a finite state space [33], and then for a general (measurable) state space, with discounting and bounded rewards [34], and with no discounting and positive rewards [35] or negative rewards [305]. Denardo [82] revisited the discounted, bounded-reward model of Blackwell [34] and elaborated on the fact that the operator implicit in the right-hand side of the optimality equation is a contraction mapping on the Banach space of bounded functions and hence has a unique fixed point, which is necessarily the optimal value function.

Hinderer [157] and Schäl [272] considered a general model which includes all three cases considered by Blackwell and Strauch – dynamic programming with discounting and bounded rewards, positive dynamic programming, and negative dynamic programming – as well as the *essentially negative* case, in which there is strict discounting and one-stage rewards are bounded above. This model provides a framework for establishing the "standard" results of dynamic programming: that the optimal value functions for the finite and infinite horizon are well defined and satisfy the appropriate optimality equations, that a Markov (stationary) policy is optimal for the finite-horizon (infinite-horizon) problem if and only if it achieves the maximum in the associated optimality equation, and that the finite-horizon optimal value functions converge to the infinite-horizon optimal value function when the terminal value function is identically zero. The case of a non-zero terminal reward function was studied by Hordijk [158], van Hee, Hordijk, and van der Wal [318], Whittle [338], [339], [340], Stidham [300], and van der Wal [317], all of whom established conditions under which the finite-horizon optimal value functions converge to the infinite-horizon optimal value function. An equivalent characterization is to view this convergence as that of a successive-approximation algorithm (value iteration) for calculating the infinite-horizon optimal value function, v, in which v_0 is an initial approximation of v, which is refined by successive applications of the optimality operator, yielding the sequence $\{v_n\}$ converging to v.

The references just cited established pointwise convergence of successive approximations. The model setting is often a generalization of the *essentially negative* model of Hinderer, a model combining some of the elements of the discounted case studied by Blackwell [34] and the negative case studied by Strauch [305]. (An example is a model with discounting and rewards that are uniformly bounded above, but not necessarily below. Many problems in the control of queues have this structure.) An alternative approach seeks conditions under which the convergence is uniform with respect to some norm. In the model of Blackwell [34], with discounting and bounded rewards, the convergence is uniform with respect to the sup norm. The assumption of bounded rewards is too restrictive, however, in many applications, including control of queues. Lippman [219] introduced a model with *weighted supremum norms*, in which the reward function is can be unbounded as long as it is $O(u)$, where u is a weighting function. The contraction-mapping theory can still be applied in this case, provided that the one-step transition operator satisfies a related boundedness condition relative to the weighting function u. This concept was further refined and extended by Jaap Wessels and his students and colleagues at the Technical University of Eindhoven in the 1970's. (See Wessels [330].) This approach can also be combined with the use of a *shift* function, in which all value functions are measured relative to a given function (see van Nunen [319], [320]). Indeed, in some cases the use of a shift function alone makes it possible to apply the classical contraction-mapping theory with the ordinary sup norm. This approach is discussed in Stidham and van Nunen [301] and Stidham [302], with a focus on examples from control of queues. In this case a natural candidate for the shift function is often the value function for an extremal policy, such as the policy that admits no jobs in an arrival-control problem or the policy that always serves at the maximal rate (the *full-service* policy) in a service-control problem. The use of such shift functions is intimately connected to the so-called equivalent charging scheme proposed by Lippman and Stidham [220] (see Section 4.6).

Jaap Wessels's group at Eindhoven was a focus for prolific research on *MDP*'s in the late 1970's and early 1980's. A primary emphasis was on efficient numerical algorithms for calculating the optimal value function. This work paralleled the established research on numerical methods for the simpler but related problem of solving a system of linear equations. Variants of Gauss-Seidel, Jacobi, and successive over-relation schemes an the use of bounds, extrapolations, and elimination of non-optimal actions were studied by van Nunen [320], van Nunen and Wessels [321], and van der Wal [317]. A hybrid method combining value iteration and policy iteration was simultaneously proposed by van Nunen [320] (who called it the *value-oriented* algorithm and by Puterman and Shin [256], [257] (who called it *modified policy iteration*).

3.1.1 Average-Return Criterion

Markov decision processes with the long-run average-return criterion require a different analysis. The most common approach has been to find conditions under which the average-return problem can be solved by letting the discount rate approach one in the problem with discounting, using Tauberian theorems. The situation is complicated by the fact that the results obtained depend on the structure of the state space – in particular, the number of closed communicating classes. For the case of a finite state space Blackwell [33] used analytical properties of the optimal value function for the problem with discounting to establish the existence of a deterministic, stationary policy that is discount-optimal for all discount factors in a neighborhood of 1. Such a policy – subsequently called *Blackwell optimal* – is necessarily also average-return optimal, as can be readily verified by a Tauberian argument. Veinott [322] extended Blackwell's result, using a Laurent expansion of the discount-optimal value function in terms of the discount rate to show that a Blackwell-optimal policy is optimal for the undiscounted problem in a *lexicographical* sense. Specifically, it is average-return optimal, and among all such policies, it maximizes the *bias*, roughly speaking, the rate at which the system earns return in excess of the average. And among all such policies, it is optimal with respect to another, even more refined criterion, and so forth. Such a policy can be found, in principle, by solving a sequence of functional equations, starting with the familiar average-return optimality equation. In the so-called *unichain* case, in which all policies give rise to just one closed communicating class of states, this algorithm reduces to policy iteration (Howard's algorithm).

For problems with a non-finite state space, the average-return problem is more difficult to solve. For countable-state problems, Ross [267] provided a set of conditions under which the average-return optimality equation has a solution and the solution can be used to find a stationary policy that is optimal. Ross's approach generalized that of Taylor [314] for a replacement problem, which used Tauberian arguments. Lippman [219] relaxed Ross's conditions, which required bounded reward functions and a uniform (with respect to both state and discount factor) bound on the difference between the discount-optimal value function and a reference function. Weber and Stidham [323] relaxed these conditions further and produced a set of conditions that are satisfied by most queueing control problems. Simultaneously and subsequently, Linn Sennott established even weaker conditions in a series of papers, extending the results to semi-Markov decision processes and processes with uncountable state spaces. Her book [277] is a good source for this theory and for applications to control; of queues.

The book by Marty Puterman [258] provides an excellent survey (as of 1994) of the theory, algorithms, and applications for discrete-time, countable-state *MDP*'s, with both the discounted and average-return criteria. For a brief survey of numerical techniques see Stidham [303] (Chapter 9 in the recent book on computational probability edited by Grassmann [129]).

3.1.2 Models with Continuous Time and General State Space; Applications to Finance

For problems in continuous time with a general state space, the methods of stochastic control theory provided an alternative, Most of the research in this area has been done outside the *OR* community. (An exception is the work by Doshi in the late 1970's [84], [85], [86], [87].) A good (although somewhat dated) reference is Fleming and Rishel [105]. When the state trajectories are both Markov and continuous in time, one has a controlled diffusion process, governed by a stochastic differential equation. Such processes arise, for example, in queueing theory as heavy traffic limits. They are discussed in Section 3.2 below.

Modelling of financial asset prices as a diffusion process (usually geometric Brownian motion) has generated a voluminous literature, which has a limited but significant and growing intersection with the *OR* community. Space does not permit me to enter the fascinating realm of arbitrage theory for pricing options and other derivatives, initiated by the work of Black and Scholes [32] and Merton [228] in the early 1970's, except to say that the problem can be formulated as an optimal stopping problem and hence falls within the purview of Markov decision processes, albeit with continuous time and state if one accepts the received wisdom in the finance community. Versions of the option pricing problem in discrete time and space have gained momentum, however, since the introduction of the binomial-tree model by Cox, Ross, and Rubenstein [66] and Rendleman and Bartter [266]. (The book by Pliska [249] is a good source for discrete-time, discrete-state finance models.) In this case, the theory and solution techniques discussed in this section are applicable.

In both the continuous and the discrete models of derivative pricing the insight of arbitrage theory is that the appropriate *MDP* model is one in which the "true" transition probabilities must be replaced by the *risk-neutral* probabilities – the probabilities that would have to hold in a world with only risk-neutral investors in order for there to be no opportunities for arbitrage (riskless profit). The generality of this principle, and its connection to martingale theory (the risk-neutral probabilities are in fact the equivalent martingale measure), was elucidated by Harrison and Kreps [138] (see also Harrison and Pliska [139]).

3.2 Controlled Brownian Networks

The extension of heavy traffic analysis to control problems was initiated by Mike Harrison in a talk delivered at the *IMA* Workshop on Stochastic Differential Systems, Stochastic Control Theory and Applications at the University of Minnesota in 1986 and subsequently published as a paper in the workshop proceedings [144]. Harrison set the agenda for future research in heavy traffic analysis of multi-class queueing networks (*MCQN*'s), both with and without control. Indeed, his suggested modelling framework has since been adopted by most researchers concerned with multi-class stochastic networks, whether in their original versions or in the form of a Brownian or fluid approximation.

The model (for an open network) has a finite number I of single-server stations (nodes) and a finite number K of job classes. Associated with each job class j is a particular station i where that class is served (denoted $j \in i$). A station may serve several different job classes, however. Interarrival times and service times for jobs in a particular class are both i.i.d. sequences. Upon completion of service, jobs change class according to a transient Markovian routing matrix, $P = (P_{jk})$, independent of the current state and history of the system. The assumption that the routing matrix is transient insures that all customers eventually leave the system. At least one job class j has a positive arrival rate: $\lambda_j > 0$. These two assumptions together characterize the system as an open network.

The model is quite versatile. The generalized Jackson network (that is, the case of homogeneous jobs: Reiman [263]) is the special case in which there is a one-one correspondence between stations and classes. Another special case is the model of Kelly [181], in which jobs belong to different types, each type has a Poisson arrival process and a fixed route through the network, and all jobs served at a particular station have a common exponential service-time distribution. The Markovian routing assumption is essentially without loss of generality, inasmuch as the class designation can incorporate information about past processing history and there can be arbitrarily many classes.

Harrison suggests thinking of a *MCQN* as a special case of a stochastic processing network. The numbers of jobs in the different classes play the role of *stocks*, the stations may be thought of as *resources* and the servicing of a particular job class as an *activity*. Activity j consumes resource i at rate $A_{ij} = 1$ if job class j is served by station i; otherwise, $A_{ij} = 0$. The average rate of depletion of stock k by activity j is

$$R_{kj} = \mu_j(\delta_{jk} - P_{jk}) \ ,$$

where μ_j^{-1} is the mean service time for class-j jobs and $\delta_{jk} = 1$ if $j = k$ and $\delta_{jk} = 0$ otherwise. Since P is transient, there is a unique vector $\beta = (\beta_k)$ satisfying the *traffic equations*

$$\lambda = R\beta \ ,$$

where $\lambda = (\lambda_j)$ and $R = (R_{jk})$. Thus β_k represents the average fraction of time that the server for job class k must allocate to that class in order to maintain material balance. Define

$$\rho_i = \sum_{k:k\in i} \beta_k \ ,$$

so that ρ_i is the traffic intensity at station i – the fraction of time that the server at station i must be working, on the average, in order to maintain material balance.

The corresponding Brownian network (heavy traffic) model for this system is given by

$$\begin{aligned} Z(t) &= X(t) + RY(t) \in S \ , \quad \text{for all } t \geq 0 \\ U(t) &= AY(t) \text{ is a non-decreasing process with } U(0) = 0 \ . \end{aligned}$$

Here $Z(t)$ is the vector of stock levels, with state space S (e.g., the nonnegative orthant), $X(t)$ is a K-dimensional Brownian motion, and $Y(t)$ is the control process. The control $Y(t)$ must be non-anticipating with respect to X. The vector $U(t)$ measures the cumulative amount of *unused* capacity at each resource (station). (Note that control is exercised, in effect, by deciding how much capacity *not* to devote to a particular job class.)

Heavy-traffic analysis of *MCQN*'s involves the same basic steps as in the special case of a generalized Jackson network, extended to allow for control (cf. Williams [342]):

1. formulate the model for the *MCQN*;
2. construct the corresponding Brownian network;
3. deduce the structure of an optimal policy for the Brownian network;
4. use this policy to construct a "good" policy for the original *MCQN*;
5. show that a sequence of *MCQN*'s, operating under this "good" policy converges weakly in heavy traffic to the corresponding Brownian network, operating under its optimal policy.

A key step in Harrison's development involves replacing this Brownian model by an equivalent model for the vector *workload* process,

$$W(t) = MZ(t) \ ,$$

where $M := AR^{-1}$. Here $W_i(t)$ can be interpreted as the expected total amount of work for server i that will be generated by any of the jobs in the network at time t (in scaled units appropriate for the Brownian approximation of the original network). The equivalent model takes the form

$$\begin{aligned} W(t) &= B(t) + U(t) \ , \ t \geq 0 \\ W(t) &= MZ(t) \ , \ t \geq 0 \\ Z(t) &\geq 0 \ , \ t \geq 0 \end{aligned}$$

where $B(t)$ is I-dimensional Brownian motion, U is non-decreasing with $U(0) = 0$ and non-anticipating with respect to B. The relationship between the Brownian motions B and X

is given by $B(t) = MX(t)$. Note that this control problem has dimension I (the number of stations) rather than J (the number of classes). Typically I is significantly smaller than J. In terms of the original queueing system, the problem in this reduced form can be interpreted as follows. First, the decision maker chooses a cumulative idleness process U, which reflects when servers will be working or not. This choice must be such that $Z(t) \geq 0$, which means that the server can work only when work is available. Second, the decision maker can allocate busy time at each station among the classes so as to realize any queue length process that is consistent with the workload process defined by the above equations.

The research program implicit in the framework suggested by Harrison [144] has been carried out by several researchers. Examples include a series of papers authored or co-authored by Larry Wein [145], [147], [324], [244], [245], [325].

So far the approach has been successful primarily in the solution of problem in which the optimal control of the Brownian network is pathwise optimal; that is, there exists a (greedy) policy that simultaneously minimizes the total cost up to all time points t for all realization of the processes involved. Although this class of problems is limited, it is still significantly larger than the class for which a pathwise optimal policy exists for the original queueing network. The process of going from the optimal policy for the Brownian network to a "good" policy for the original network can be tricky. It is easy to keep servers from being idle in the Brownian model, whereas it may be difficult in the original network. Ad hoc adjustments may be necessary, such as deviating from the greedy solution in order to replenish a nearly empty buffer to keep a particular server from becoming starved (see, e.g., Harrison [152].

For analyzing the performance of a Brownian network under a specific policy, one can resort to the numerical algorithms mentioned earlier, specifically, the *QNET* algorithm [148], [72], [73], [74], and the discretization algorithms of Kushner and Dupuis [213], [214]. For solving the control problem, various numerical methods are under development. This development is rendered more challenging by the fact that many control problems involve free boundaries. In principle, one can discretize time and space and approximate the control problem for a Brownian network by a Markov decision process. But this approach is limited by the familiar "curse of dimensionality" (even though the resulting problem may have significantly simpler structure than the control problem for the original *MCQN*).

A significant feature of models for control of Brownian networks is the phenomenon of "state-space collapse". We have already seen that the transformation to the workload formulation reduces the dimension of the problem from J (the number of job classes), I, the number of stations. In some cases, the special structure of the problem results in an additional reduction in dimension. Probably the first to point out this phenomenon were Foschini [106] and Foschini and Salz [107], who analyzed a queueing system with a single input stream and parallel servers, each with its own queue, in heavy traffic. They showed that, if the system operates under the "join-the-shortest-queue" policy, the heavy traffic approximation collapses to a one-dimensional problem, in which the queue lengths are all equal. The intuitive reason for this is that arrivals occur so rapidly in heavy traffic that any deviation from equality is immediately removed by the "join-the-shortest-queue" policy. Kelly and Laws [182] analyzed a four-node network in which a more subtle form of state-space collapse (from four to two dimensions) results from the special structure of the network. Recent research has revealed an extreme form of state-space collapse (to one dimension) which occurs in scheduling problems with "complete resource pooling" (see, e.g., Harrison and Lopez [153], Williams [343], Kushner [214], Chapter 12). In these models, the servers are "agile", to use a term from research on manufacturing models, and can be moved rapidly from one class to another.

Scheduling *MCQN*'s in which there is a setup (or switching) cost or time involved in moving a server from one job class to another are notoriously difficult. Control of polling systems (cf. Section 2.6 is an example of a problem in which this issue must be confronted. A breakthrough – the discovery of the "averaging principle" by Coffman, Puhalskii, and Reiman [60] has made heavy-traffic analysis of such systems feasible. See also Reiman and Wein [264], Markowitz and Wein [226]. In some cases the heavy-traffic approximation involves an Ornstein-Uhlenbeck process rather than Brownian motion.

3.3 Conservation Laws and Achievable Region Approach

Stochastic scheduling problems arise when one must decide how to allocate a resource (such as a server in a queueing system) dynamically to competing demands (such as jobs of different classes). Associated with each scheduling rule is a *performance vector*, whose i^{th} component is the performance measure for class i (for example, the steady-state expected waiting time for class-i jobs). The *achievable region*, or *performance region*, of a stochastic scheduling problem is the set of performance vectors of all admissible policies. The definition of "admissible" depends on the specific problem.

The idea for characterizing the achievable region of a stochastic scheduling problem had its origin in the work of Coffman and Mitrani [59] (see also Gelenbe and Mitrani [118]), who studied a multi-class *M/GI/1* queueing system in which the performance measure for each class is the steady-state expected waiting time for jobs in that class. Each class has an associated arrival rate and service-time distribution. Admissible policies are scheduling rules that are non-anticipative (they use no information about future arrivals nor their service times), non-idling (the server always works when any jobs are present), and regenerative (they only use information from the current busy cycle). Coffman and Mitrani [59] showed that the achievable region is a polyhedron. They accomplished this by identifying a set of linear constraints (called *conservation laws*), which are satisfied if and only if the vector is the performance vector of an admissible policy. Moreover, the extreme points of the polyhedron are the performance vectors of strict priority policies: policies that give absolute preference to jobs in a particular class. Since a linear objective achieves its minimum (or maximum) over a polyhedron at an extreme point, it follows that a strict priority policy is optimal among all admissible scheduling rules, if the objective is a linear combination of the performance measures of the various classes. In particular, this result provides a proof of the optimality of the $c\mu$ rule for the problem of minimizing the expected steady-state total number of jobs in an *M/GI/1* queueing system. (The $c\mu$ rule schedules jobs in strict priority according to the ordering, $c_1\mu_1 \geq c_2\mu_2 \geq \ldots \geq c_m\mu_m$, where c_i is the waiting cost per unit time per class-i job in the system and μ^{-1} is the mean service time for a class-i job.)

Federgruen and Groenevelt [100, 101] showed that, for certain queueing models including those studied in [59] and [118], the polyhedron characterizing the performance region is of special type, called the *base of a polymatroid*. Shanthikumar and Yao [288] extended these results by introducing the concept of *strong conservation laws* and proving a powerful result about the achievable region when strong conservation laws hold. When the performance vectors satisfy strong conservation laws, the following properties concerning the achievable region hold:

(i) the achievable region is completely characterized by the strong conservation laws;

(ii) the achievable region is the base of a polymatroid; and

(iii) the set of vertices of the achievable region is equivalent to the set of performance vectors obtained by all strict-priority rules.

Polymatroids have special properties that can be exploited in the context of queueing systems. First, the vertices of the base of a polymatroid can be easily identified from its half-space representation. Next, optimization of a linear objective over a polymatroid is very efficient; it can be accomplished by a greedy algorithm. Since the solution of a linear objective optimized over a polyhedron is found at a vertex of the polyhedron, the optimal policy for a multi-class queueing system that satisfies strong conservation laws will always be a strict-priority rule. In addition, because of the properties of polymatroids, the particular strict-priority rule that is optimal is easily identified. Strong conservation laws are particularly applicable in scheduling problems in multi-class, single-station queueing systems, but other applications have also been discovered.

The achievable-region approach has progressed significantly beyond the development of strong conservation laws. Bertsimas [25] and Bertsimas and Niño-Mora [26] generalized these results to more complex stochastic and dynamic scheduling problems called *indexable systems*. An indexable system is a dynamic scheduling system in which the following policy is optimal. To each job class j attach an index γ_j. At each decision epoch select a job with largest current index. Bertsimas and Niño-Mora showed that, if the performance vector satisfies certain *generalized conservation laws*, the performance region is a special polyhedron they call an *extended polymatroid*. When the performance space is an extended polymatroid, a linear objective can be optimized over the performance space by an adaptive greedy algorithm. Moreover, the optimal policy is a strict-priority rule (which is an index policy).

The optimality of index rules for a variety of scheduling and related problems (such as bandit processes) was discovered by Gittins [122], [123], who used quite different techniques. Peter Whittle's paper [341] in this issue reviews Gittins's work and that of many authors (himself included) who were inspired by Gittins's ground-breaking efforts.

Several authors have made attempts to generalize the achievable-region method to other queueing systems. Bertsimas [25] applied the approach to several different types of stochastic control problems, including multiarmed bandit problems, restless bandit problems and polling systems. Bertsimas [25] and Bertsimas, Paschalidis and Tsitsiklis [27] generalized the approach to queueing networks (both open and closed), and develop two methods for bounding the region of achievable performance. In this context the inequality constraints do not provide an exact characterization of the performance space. Instead, they characterize a larger region that contains the performance space, so optimizing over this larger set results in a lower bound on the minimal cost. If this lower bound is fairly tight, it can be very useful for measuring the degree of suboptimality for heuristic policies in these queueing networks.

Most of the applications of the achievable-region approach have focussed on performance measures that are expectations of steady-state random variables, such as the waiting time in the system or in the queue. By contrast, Green and Stidham [131] and Bäuerle and Stidham [17] have derived strong conservation laws for sample paths. These conservation laws hold at every finite time point t and for every realization of the input workload process in systems that are generalizations of the *GI/GI/1* queue as well as in fluid models. Green and Stidham [131] analyzed a single-server system without feedback and Bäuerle and Stidham [17] generalized their results to a single-server system with feedback (the Klimov problem), in which a job (or unit of fluid) that completes service make change class and return to the queue for more service.

Glazebrook and Garbe [124] investigated systems that "almost" satisfy generalized conservation laws. For such systems, they analyze the index policies produced by the adaptive greedy algorithm and provide bounds on their degrees of suboptimality. In addition, Bertsimas, Paschalidis and Tsitsiklis [27] addressed problems that satisfy conservation laws, but also have additional side constraints. Dacre, Glazebrook and Niño-Mora [71] have provided an excellent review of the achievable-region approach and extensions to scheduling in multi-class systems with parallel servers, distribution of workload across a network of interconnected stations, and a class of intensity control problems. Work continues in this fertile area.

4 Personal Reminiscences

4.1 U. Narayan Bhat

I chaired the founding group that carried on most of the negotiations with *ORSA* and then with *TIMS*, which ultimately led to the establishment of a Technical Section (*ORSA*) and College (*TIMS*) on Applied Probability – the precursor to the current Applied Probability Society of *INFORMS*. When the Technical Section was proposed in December, 1971, the Technical Sections Committee of *ORSA* did not consider it favorably. At that time they argued that technical sections should be used only for applications areas. Even then we pushed for it. In 1974, Ralph Disney asked me whether we would consider being a College under *TIMS*. We gladly accepted the offer and by May of 1975 the College on Applied Probability was established. Spurred by this development *ORSA* approved the establishment of a *SIG* in November of that year. It was not until 1981 that the group was recognized as a technical section. On the whole we were treated much better by *TIMS* than by *ORSA*. The irony is that ORSA was then considered as dominated by academicians. Fortunately *TIMS* didn't have any inhibitions in welcoming us. If *ORSA* hadn't changed its mind we would have simply remained as a College in *TIMS*.

4.2 Marcel Neuts

I am ambivalent toward writing about history, especially at the arbitrary marks on the time axis that call for special anniversary issues. Historical writing tends to freeze a complex, dynamic reality into a confined mental construct. A historical account becomes a substitute for the real experience, a substitute that can never live up to the real thing.

I also doubt that reminiscences about the past are of much use to the young, even to those that read them with interest. Innovations of twenty, thirty years ago are considered well known, even those that are continually rediscovered. The highest distinction of a theorem or a mathematical method is to be included in textbooks. When it does, it matters little by whom or when it was first conceived.

What I gladly share are memories: those sessions around the faded photo album with comments and anecdotes from the elders that are, always, selective and subjective. In sharing memories, I like to impart the feeling that "it was great being part of it all" and "look how exciting it was! With luck, 't may get better yet."

I learned probability models from Samuel Karlin and Joseph Gani at Stanford during my graduate studies, but it was only in 1963 that applied probability became my primary research interest. A research seminar at Purdue by the late James McFadden and by Douglas Lampard had shown me how exciting and methodologically versatile that field could be.

Looking back at 1963, not quite fifty years into the past, yet long enough for a deeper time perspective, the field of applied probability has changed profoundly. Our methodological arsenal is now much better stocked and with impressive mathematical weaponry. Applied probability contributes, quite often indirectly, to many areas of engineering, biology, and technology. In 1963, articles on stochastic models had to find places in general mathematical or statistical journals; now, our research contributions are published in some twenty periodicals of high quality that are fully dedicated to our field.

Once in a while, in the odyssey of life, it is good to pause and to look back at the path already travelled. For applied probability, that journey was very good indeed. My colleagues and I can give ourselves a collective, well-deserved "pat on the back" for a job well done; the field is much richer because of our joint endeavors. It is a personal privilege to have a share in that success.

Sandy Stidham specifically asked me to "muse" about computational probability, a sub discipline whose paternity – barring appropriate DNA tests – is often ascribed to me. Sandy also mentioned that, already on many occasions, I have written or held forth about the origins of computational probability and that, indeed, is true. The persistent reader will find more in Neuts [235, 237, 239, 240, 242].

Before I launch myself in yet another account, let me mention – as I have done before – that I fervently hope the designation "computational" will soon disappear. I do not wish that the field will disappear – how could any "father" want that? – but that instead, algorithmic concerns will become pervasive. I hope that all serious papers on applied stochastic models will become algorithmic! If, as we assert, our field is applied and deals with probability models, our primary objective ought to be the enhanced physical understanding of the process that we are modelling. If our model's complexity exceeds that of a textbook example, then simple analytic solutions are as rare as white ravens and asymptotic results are only mildly informative. How else than by substantial, serious computation can we gain deeper and especially detailed physical understanding? The tools of computation are now omnipresent and they are magnificent. Past excuses of computational drudgery no longer have merit. That I am still, on many occasions, asked to expound my "computational philosophy" some thirty years after computers became universal is beyond my comprehension. It all became obvious long ago - or did it?

In the long forgotten times "Before the Computer", algorithmic work was done, competently (it had to be) and slavishly (there was no other way) provided that the need was great enough. Astronomy, statistics, and actuarial science have a long tradition of numerical computation. In astronomy and statistics, substantial results were mostly unattainable without numerical work.

Actuarial science had the advantage that money was involved and, as always, it spoke. My only formal course on computational methods was part of an actuarial program in Belgium in the 1950s. With a handful of other students, I spent Friday afternoons figuring out the cost of annuities and the yield of government bonds. Our only tools were mechanical calculators (electrical ones were too costly for student use!) and the Pereire tables that no actuary ever went without.

People who did actual computation never looked down on that task. They knew how much care and mathematical knowledge was involved. That was brought home to me when, in the 1980s, I looked up the nearly inaccessible original papers of C. D. Crommelin [68, 69] on the $M/D/c$ queue. To obtain the additional equations for the first c terms of the steady-state distribution, Crommelin used the now familiar analyticity argument based on Rouché's theorem. In discussing matrix-analytic solutions, I had pointed out that when the Rouché roots coincide or are close together, the method of roots could be numerically inaccurate. Many colleagues dismissed my concern as a mathematician's obsession and a few asserted omnisciently that such difficulties never arose "in practice". The concern is real – see e.g., Dukhovny [94].

When I finally got copies of Crommelin's papers, I was elated to read that, for the same reasons as I, he was concerned about the clustering of roots. In 1932, Crommelin knew; in 1980, many of my colleagues did not. I believe that the difference lies in the fact that Crommelin, a distinguished actuary of his day, did actual computation. A biologist friend once told me: "Anyone who thinks that empirical work is easy and routine has never done it." Substitute "numerical computation" for "empirical work" and you get another truism.

By now, attitudes in the applied probability community have greatly changed. Where once algorithmic papers were rare – I was asked a few times to "remove all that computational stuff" from some of my early papers – there is now vigorous research of excellent quality on algorithmic methods for various probability structures. Particularly since the early 1980s, algorithmic know-how has increased very fast.

On one hand, that should please me greatly, yet, on the other, my satisfaction with the state of the discipline is incomplete. I am quite willing to blame that on my impatient nature, yet there is more to it than that. Let me start my explanation with an anecdote. At the end of 2000, I was in China where I met many enthusiastic younger colleagues. One day, I was describing a recent research project to one of them. Having my portable computer at hand, I showed him the code and ran a few illustrative examples. I noticed the young man staring at me with wide-open eyes. Suddenly, he exclaimed: "But, Professor, do you actually write computer programs??" The answer is: "Of course, I do" – but to me, the real question is why that should so greatly astonish a young colleague. Is there still the perception that the priveleged ought not soil their hands with actual work?

Once, at the end of a lecture, I was asked to give one example of some probabilistic knowledge that had been obtained by computer experimentation. On the spur of the moment, I could not give a good example. I would be almost as hard pressed if asked to give an example of knowledge on stochastic models gained from numerical computation. There are, of course, examples of both, but clearly they are not widely known. Polling systems, for example, have been the subject of extensive research since the 1960s. Many techniques such as classical analytic methods, vacation models, and extensive simulation have been brought to bear on polling systems. Yet, I find that, even among experts, there is very little concrete knowledge about the operation and design of polling systems, comparable, say, to what one would find about laminar flow and airfoil design among aerodynamicists. Often, after a few days at a conference on applied probability, I ask myself what more I have learned about the behavior or the design, say, of service systems after hearing many learned, mathematically sound communications in queueing theory. I shall keep my honest, but pithy answer to myself.

Let me briefly come back to the question about knowledge gained from simulation or algorithmic exploration. If one means knowledge of theorem quality, then neither can directly supply that. A good algorithm serves the applied probabilist in the same manner that a fine instrument aids the experimental physicist. It is a tool of exploration that can help us understand the model over significant parameter ranges. As with all knowledge, the results of such an exploration are limited. I do not see why that acceptance of reality should diminish the merits of that methodology below those of articles with pages of abstruse, inapplicable theory.

In looking at algorithmic methods in a constructive spirit, I have reflected on the difference between the methodological maturity of our field and the perception of its relevance to physical understanding and concrete application. For a detailed discussion, see [242]. However, briefly stated, my thesis is as follows:

In all research, contributions fall into two broad, overlapping categories. The first strengthen the methodology of the field through better instruments, broader theorems, or more efficient algorithms. The second category encompasses the knowledge gained about the objects of study. In physics, chemistry, and biology, the second category is most prominent. The tools and techniques of measurement necessarily evolve with the need for increasingly refined observation, but they are primarily means to the end of understanding physical reality.

Until the computer era, and particularly with the historical distancing of mathematics from its immediate applications, applied mathematical research primarily focused on the first category. I am not talking here about the common, often artificial distinction between theory and application, but about emphasis.

Already in courses, we teach students to solve, say, differential equations, but we rarely ask for interpretations of their solutions; that is to be learned in a physics course, if ever. Students have come to dread “word problems” where they need to formulate equations and learn something from their solutions. Many a mathematics course deals with rote formalism only, not with the formulation of equations and the interpretation of results! How misdirected that is! Surely, no one learns a natural language merely to speak correctly and to have a good vocabulary. We learn a language to communicate substantive information. Should the same not hold for mathematics, the most pristine of languages?

Much of the literature on stochastic models places major, if not exclusive emphasis on methodology needed to study the model. Physical insights gained from applying that methodology get only passing attention, if even that. We write about algorithms or about simulation methods for queues, not – in the first place – about queues themselves. This is an observation, not a criticism. Rigorously validated methodology is essential, but in deepened physical understanding we reap the most pleasing rewards of our endeavors and we strengthen the real applicability of our field.

To me, that was the fundamental reason for doing algorithmic probability. In the early 1970s, I became interested in numerical computation out of frustration with the limited understanding that I gained from formal solutions.

Clearly, establishing the mathematical validity of algorithms, finding better and faster algorithms, clarifying the structural properties of models that facilitate their algorithmic solutions, all these are essential steps in a sound methodology. By far, most of my articles also deal with such matters. A few, such as [236], offer a larger share of physical insight. I wish I could have written more such papers.

Greater emphasis on the qualitative behavior of systems would enhance the appeal of our published work. That will require discussing systematic explorations of models, something that is far more detailed than a few illustrative numerical examples. However, such an emphasis will be important far beyond its appeal to the readers of our work. The difficulty of fully capturing the complexity of behavior, even of simple stochastic systems, will soon be evident. We will recognize the need for new informative quantitative descriptors of random behavior. The familiar ones, such as steady-state distributions, moments, and asymptotic parameters, tell only a small part of the story.

The past fifty years have witnessed immense progress in stochastic modelling. If a lesson remains to be learned, it is that my generation has not availed itself as fully of the computer's capabilities as it might have. I hope that those who are now shaping the discipline will do so. For the sake of applied probability, I hope that they will. Let them look beyond the vision of those who hold that computation is unimportant – and beyond the elitism of those who still say that it is lackey's work. Let them not accept the defeatism of some who say that nobody takes applied probability seriously anyway. For, if that were true, why do it in the first place? Let us instead go forth, with luck 't may get better yet.

4.3 Uma Prabhu

In the late 1960s there were two main problems facing applied probabilists in the United States: the lack of a strong professional identity for their research area and the lack of adequate communication among themselves and with other research groups. Now, more than thirty years later, I believe these problems have been successfully resolved. In 1971 (the late) Julian Keilson organized the first in the series of conferences on Stochastic Processes and their Applications (*SPA*); in 1973 he co-founded with me the journal with the same title; in 1971 Narayan Bhat and his associates laid the foundations of the *INFORMS* Applied Probability Society; in 1986 I founded the journal *Queueing Systems: Theory and Applications (QUESTA)*; in 1985 Marcel Neuts founded the journal *Stochastic Models*; and in 1991 the Institute of Mathematical Statistics (*IMS*) started its journal, *Annals of Applied Probability.*

In published literature the term applied probability seems to have been first used by the American Mathematical Society (*AMS*) as the title of its 1955 symposium on applied mathematics [224]. The emphasis at this symposium was on problems in classical and modern physics. The same was the case with a later *AMS* symposium [21]. A somewhat different trend had already been in evidence at the 1949 Royal Statistical Society symposium on statistical physics, population growth models and non-stationary stochastic processes [3]. There was a more comprehensive coverage of problems in applied probability at the Third Berkeley Symposium on Mathematical Statistics and Probability held in 1954. As a further indication of the growing research activity in applied probability, a series of monographs [7, 8] began to appear in 1962, reporting the investigations of K. J. Arrow, S. Karlin, H. Scarf and their co-workers.

Applied probability received its baptism when Joe Gani started his *Journal of Applied Probability* (*JAP*) in 1964, with the aim of publishing research and review papers on applications of probability theory to the biological, physical, social and technological sciences. The basic motivation was Joe's feeling that it "may have a constructive influence in bridging the gap between those probabilists who regard abstract theory as more valuable, and those who view applications as more worthwhile" (see [114]). *JAP* was followed five years later by *Advances in Applied Probability* (*AAP*). Joe has given an account of his involvement in applied probability in [116].

Julian's objectives in starting the *SPA* conferences were to encourage communication between abstract and applied probabilists, and to provide greater visibility to younger research workers in probability. Both these objectives were successfully met at the early conferences in this series. In 1975 the Committee that organized these conferences was formally affiliated with the Bernoulli Society as a subject area committee. (I was chairman of this Committee during 1975-1979). By then it was generally recognized that applied probability had come of age, but the affiliation with Bernoulli Society gave the *SPA* conferences an added stature, viewed as an activity of the International Statistical Institute. To put this in less political terms, the Bernoulli Society hijacked both the *SPA* journal and *SPA* conferences, activities that were started as personal enterprises and driven by professional idealism and hard work.

I have given brief histories of *SPA* conferences and journal in [7] and [8]. For Julian's version see [9]. *SPA* now seems to be following scientific directions that are less consistent with Julian's objectives. Ah well, such is life! In any case applied probabilists are now better served by the series of international conferences organized by the INFORMS Applied Probability Society. But that part of the story should be told by Narayan Bhat and his associates. In summary, the problem of communication has been successfully solved for applied probabilists.

The motivation for the starting of *QUESTA* was the following. While it does not appear to be particularly difficult to publish queueing theory papers in existing journals, there would certainly be advantages to publishing in a specialized journal, rather than an OR or applied probability journal. *QUESTA* is now the definitive journal on queueing systems; it has given queueing researchers greater professional visibility and established a firm identity for queueing systems as a distinct area of applied probability.

When the IMS started its Annals of Applied Probability the expectation was (I was on the IMS Committee that recommended its publication) that it would actively encourage new (at that time) areas of research such as finance models, probabilistic algorithms and inferences from stochastic processes. This has not happened, and in the meantime journals have been started on finance models and on inference from stochastic processes. These newly established journals, along with JAP, AAP, and *QUESTA* provide a unique forum for applied probabilists working in a rather wide spectrum of research areas. Let us celebrate!

4.4 Dick Serfozo

My story is about some frustrations in queueing before the discovery of Palm probabilities.

In 1973, Don Gross and Carl Harris were preparing their book *Fundamentals of Queueing Theory* [132], which is a major text to this day. One of their aims was to find a rigorous proof of the result that, for a stationary $M/M/1$ queue, the sojourn time of a typical customer is exponentially distributed with rate $\mu - \lambda$, where λ and μ are the arrival and service rates. This was an old result at that time, but its proof involved hand waving.

Carl asked me to help him with this task. We worked on it for several days, wrestling with the conditional probability of the queue length conditioned on the event that a customer arrives at a certain instant; we knew that this event has a zero probability, but tried to deal with it anyway. We failed to find a rigorous proof and were humbled by the experience. The reason for our failure is that the result is not true under the underlying probability measure of the process, which we were trying to use. The result is true, however, under the Palm probability associated with the arrival process, which is not the same as a conditional probability with respect to the underlying probability measure.

A second frustration was with Little's law, $L = \lambda W$, that the average queue length L equals the arrival rate λ times the average waiting time W for a customer. This law was quite well understood when the quantities are "long-run averages". However, the law also holds for stationary systems when the quantities are interpreted as "expected values". As a junior professor in the 1970's, I attempted to give a rigorous proof of this for my students but failed. I later found out that the law is not exactly true for expected values as many queueing analysts thought. The caveat is that L and λ are expectations under the underlying probability measure for the process, but W is the expected value under the Palm probability measure associated with the arrival process.

Similar frustrations occurred in the development of Jackson queueing networks in the 1970's and 1980's. The culprits were the following results for a stationary open Jackson network:
(1) When a customer (or unit) moves in the network, the distribution of the rest of the network is the same as the limiting distribution for the network (i.e., moving units see time averages). This *MUSTA* property is related to the *PASTA* property (Poisson arrivals see time averages) for single queues.
(2) When a unit traverses a non-overtaking route in the network, its sojourn time at the nodes on the route are independent, exponentially distributed random variables.
There were several articles giving different and rather lengthy proofs in terms of limits of probabilities that mimic conditional probabilities.

The theory of point processes and Palm probabilities had been developed by the time these results were proved (see, e.g., [48], [110], [81], [176], [47]) but only later was this theory used to simplify their proofs. Buried in a lengthy paper on networks in 1993 [279], I showed that the *MUSTA* property is a one-line statement of a Palm probability, and that many other statements of this sort follow similarly. Another paper with Kwang Ho Kook [205] on travel times in networks shows that the sojourn times in result (2) are actually under the Palm probability that a unit begins traversing the route at time 0 and the unit does indeed complete the route. This Palm probability involves "conditioning on the future" that the unit completes the route.

The novelty of conditioning on the future as well as the past was developed further in my book *Introduction to Stochastic Networks* [281] in 1999. It contains an extended Levy formula for expected values of functionals of Markov processes, which in turn leads to closed-form expressions of Palm probabilities for stationary Markov processes. This new sub-theory of Palm probabilities for Markov processes is considerably simpler than that for non-Markovian systems where Palm probabilities have abstract characterizations that are difficult to apply. Now many results as above for Markov systems can be obtained by simply plugging in the appropriate system data in the closed-form expressions for the Palm probabilities.

4.5 Les Servi

In the late 1980's, Julian Keilson and I were analyzing a complex queueing system based on the local telephone switch developed by GTE, the GTD-5. After we had a solution, Julian conjectured that its waiting time and queueing distribution were related by the same distributional form as a single server queue, *M/G/1.* The algebra required to test this conjecture was tedious and for a while impenetrable. In the middle of all this I had a dentist appointment. Upon arriving I was told that the dentist would be quite late so I passed the time with this algebra. Perhaps due to the change in scenery I succeeded. As a result, we suspended working on the queuing paper and started a paper related to this distributional form of Little's Law. This included an explicit formula for each waiting time moment in terms of queue length moments. When we subsequently gave university talks about this work we included a triangle, reminiscent of the Pascal triangle, containing the coefficients characterizing this relationship. People invariably asked whether there was a recursive relationship between the elements in this triangle and our consistent answer was "one would think so but none is known". One day, Ushio Sumita, then a professor at the University of Rochester, visited us at GTE and immediately guessed the pattern after being shown the triangle. Armed with a good conjecture, the three of us quickly derived a rigorous proof. Unfortunately, (or perhaps fortunately), days before sending this note out we discovered that the three of us had inadvertently rediscovered Stirling numbers and our recursive formula proof was scooped by James Stirling over three hundred years ago.

4.6 Sandy Stidham

My career in applied probability began almost accidentally. It was the summer of 1966. I had finished my course work and passed the PhD exams in the Stanford OR program. Now I faced my next hurdle: finding a dissertation topic and advisor. I had enjoyed all the OR subjects about equally well and could find no compelling reason to choose one specialty over the others. Finally, almost in desperation, I remembered a paper I had written in a course on queueing theory at another university. The class had read a paper by Ernie Koenigsberg [201] on a cycle of queues with exponential servers, in which he shows that the stationary probability distribution has a product form. In my paper I showed that this result extends to an arbitrary closed network of queues with exponential servers. The instructor had been quite pleased and had urged me to submit it for publication. So I dug up a copy of the paper, took it to Fred Hillier, and asked him if he thought it could be a starting point for a dissertation. He was very diplomatic: he said yes, except that the result was already known. Specifically, it followed from a more general result proved by Jackson [171] in his famous 1963 paper on open networks of queues. I knew that Jackson had shown that an open network has a product-form solution, but it had escaped my attention (and that of my previous instructor) that the corresponding result for a closed network followed as a special case of Jackson's generalization to open networks with certain kinds of state-dependent arrival rates. (As it turned out, I was in good company. Gordon and Newell [128] published their paper on closed queueing networks without being aware that the product-form property had already been demonstrated by Jackson. They later graciously acknowledged their oversight, but this hasn't keep subsequent authors from referring to such networks as "Gordon-Newell networks".)

Disappointed, I asked Fred if he had any ideas for a dissertation topic in queueing theory. He mentioned several promising areas in which little work had been done, but the one that caught my attention was *optimization of queueing systems*. Fred had recently written two papers on optimal design of queueing systems. (The main results in these papers are nicely summarized in Chapter 18 of Hillier and Lieberman [156].) One may legitimately argue that the field of queueing optimization began with the publication of these papers. Using simple *M/M/1* and *M/M/s* models, Fred effectively raised the consciousness of the OR community to the idea that explicit optimization models had a legitimate place in the analysis of queueing systems, no less than in other OR applications. For example, in deciding how fast a machine to buy, one should balance the purchase and operating costs (which are increasing in the service rate) with the average cost of having jobs waiting in the system (which is decreasing in the service rate, since waiting times are lower with a fast server than a slow one). Of course, any self-respecting industrial engineer knew this already. What Fred had demonstrated was that such trade-offs could be effectively captured in simple economic models, often with explicit solutions for the optimal values of the decision variables. For example, in the case of an *M/M/1* model with linear service and waiting costs, the optimal service rate is given by a square-root formula reminiscent of the venerable *EOQ* formula of inventory theory. As with the *EOQ* formula (still widely used, often in applications that do not satisfy the assumptions on which it is based), the significance of such formulas lies not in their in their numerical accuracy, but in their structural properties and their qualitative implications for optimal decision making.

The problem Fred suggested to me was to prove a folk theorem of queueing theory: the optimality of the single-server system. In the simplest version of this problem one has to choose both the number of servers and the rate at which each serves, assuming that any given service capacity costs no more when it is concentrated in one server than when spread among several servers. The folk theorem says that the single fast server is always better. Here "better" means a lower total service cost plus waiting cost. Under the given assumption about service costs, the result is true if the average waiting cost is also lower when using a single fast server. For the case in which the waiting cost per unit time is proportional to the number of s in the system, the result had long been known to be true for *M/M/s* systems (see, for example, Morse [231]) and was conjectured to be true with more general arrival processes and service-time distributions. The intuition was simple and persuasive: whenever at least one customer is present, the system with a single fast server always uses its total service capacity, whereas the system with several slow servers can have idle capacity (if the number of customers present is smaller than the number of servers). If both systems are fed the same input, then the single-server system, having a (stochastically) more efficient output process, will have (stochastically) fewer customers present at any point in time. As a consequence the long-run average (or expected steady-state) number present will also be smaller. It was Pete Veinott (a member of my doctoral committee) who drew my attention to the possibility of using the theory of stochastic ordering to prove this result, which may have been the first application of this powerful theory to queueing systems. (Nowadays one would use a coupling argument. See Shaked and Shanthikumar [282].)

Thus began my career-long interest in optimization of queueing systems. In 1968 I went to

Cornell University as an assistant professor in the Department of Operations Research. Shortly after my arrival at Cornell, Paul Naor, the Israeli economist, gave a seminar in which he presented his now-famous results concerning control of arrivals to a queueing system. Naor was the first to point out that customers concerned only with their own welfare will join a queueing system more often than is socially optimal, that is, optimal for the collective of all customers. The reason for this phenomenon is that an individually optimizing customer takes into account her own waiting time (what economists call the *internal effect*), but not the increase in other customers' waiting times that results from her joining the system.

Welfare economists were familiar with this type of phenomenon – an example of the "Tragedy of the Commons" – but this concept had not yet become a meme for the general public (as it would in the 70's with the raising of environmental consciousness). Among queueing theorists it amounted to a paradigm shift, the significance of which was twofold. First, it started us thinking about *feedback* and *adaptation*: rather than coming from an exogenous process, the arrivals to a queueing system may be affected by the congestion within the system. We recognized that we should be modelling this phenomenon. Second, if in fact customers are behaving adaptively, then there is no particular reason why this behavior should be optimal, either from the point of view of the service provider or society as a whole. This realization leads directly to the idea of controlling arrivals by *pricing* – an idea that has recently begun to gain favor in the telecommunications community. Indeed, the *TCP/IP* protocol that we all use when accessing the Internet is an excellent example of a feedback mechanism for restricting arrivals (in this case, packet transmissions) to a queueing system when congestion is detected (in this case, by means of signals of packet losses). Recent research (see, e.g., Kelly [184]) has exploited the recognition that *TCP/IP* is in essence an implicit pricing mechanism and that explicit modelling of Internet flow control as a problem in queueing optimization can be fruitful.

But I am getting ahead of my story. As a junior faculty member at Cornell in the early 70's, I was taken under the wing of Uma Prabhu, who shared my interest in optimization of queues. Uma organized a weekly seminar on queueing optimization. Other participants included Ham Emmons and a young economist visiting from Denmark, Niels Christian Knudsen. Inspired by Naor's seminar, we set about extending his results on social vs. individual optimization. Naor's model was as simple as it could be. Customers arrive at an *M/M/1* queue and decide whether or not to join. A joining customer receives a reward (utility) r and incurs a holding (waiting) cost h per unit of time in the system (in queue plus in service). The performance measure (objective function) for social optimality is the expected steady-state net benefit (reward minus cost) earned per unit time, whereas each individually optimizing customer is concerned only with maximizing her own net benefit (reward minus expected holding cost). A basic question was: how robust were Naor's results and how much did they depend on the special structure, probabilistic and economic, of his model?

Several of us wrote papers that grew out of our seminar. Uma and I wrote a survey paper on optimal control of queues, which ultimately appeared in the proceedings [297] of the conference on Queueing Theory, held at Western Michigan University in Kalamazoo in 1973. Niels's paper [200] extended Naor's results to multi-server systems, while simultaneously generalizing the benefit-cost structure. Meanwhile, other researchers were similarly inspired by Naor. Yechiali [348], [349] extended Naor's result to queues with general (that is, not necessarily exponentially distributed) interarrival times. Adler and Naor [2] studied the *M/D/1* queue. It was becoming clear that the phenomena discovered by Naor were not confined to single-server queues, nor were they an accidental property of the exponential distribution.

The analysis in all these papers, however, was based on classical steady-state models of finite-capacity queueing systems, in which the capacity parameter is induced by the customers' behavior. There was an implicit assumption that optimal admission policies – whether optimal from the point of view of society, the individual, or the service provider – would be of *threshold* form. That is to say, customers join the system if and only if the number already in the system is below a certain threshold, which is the implied capacity parameter. While intuitive, this property is not completely obvious. It seemed to many of us that a more satisfying modelling framework would be one in which the threshold property was a consequence of the model structure and the economic assumptions, rather than an assumption.

A natural approach for establishing the threshold property, as well as other monotonicity properties of optimal policies, is to model the system as a Markov decision process (*MDP*: cf. Section 3.1) and use backward induction on the number of stages (observation points) remaining in the horizon to prove that an optimal policy has the desired form. The monotonicity properties established by this approach could then be extended to infinite-horizon problems with discounting and thence to infinite-horizon problems with the average-return criterion by standard limiting arguments. Arrow, Karlin, and Scarf had provided the precedent and template for such an approach in their pioneering papers on inventory theory, many of them reprinted in their influential book [6]. The general idea is to discover properties of the optimal value (sometimes called "cost-to-go") function that are (minimally) sufficient for the policy to have the desired form, then show by backward induction (using the dynamic-programming optimality equation – sometimes called the "Bellman" equation) that the optimal value function has these properties. In processes with additive or nearly additive transitions (such as occur in inventory and queueing systems) the crucial property of the optimal value function is typically concavity (or convexity in the case of cost minimization).

Work began in earnest in the late 60's on modelling queueing control problems as *MDP*'s. My own contributions to this effort started with the above-mentioned survey paper, followed by a joint paper with Steve Lippman [220], in which we showed that a socially optimal policy for controlling Poisson arrivals to an exponential service system has a monotonic structure and that Naor's result once again holds: an individually optimal policy admits more customers than is socially optimal. Novel features of our model were: (i) generalization to a random reward, non-linear (convex) holding costs, and a more general service mechanism with a state-dependent (concave) service rate; and (ii) the use of an "equivalent charging scheme" to facilitate the comparison between socially and individually optimal policies in the case of a finite horizon and/or discounting. The equivalent charging scheme is based on exploitation of the principle that fixed costs do not affect optimal decisions, combined with recognition that the expected (discounted) cost associated with holding and serving all customers present when an arrival occurs is indeed fixed, that is, independent of current and future decisions. As a result, this cost can be removed from the optimal value function. The resulting value function turns out to be equivalent to that for an *MDP* model in which the expected (discounted) cost of holding and serving a customer is charged as a lump sum if and when that customer joins the system. In this model, an individually optimizing customer joins if and only if its reward exceeds this lump-sum cost, whereas the optimality equation reveals that for social optimality one must add to this cost an additional cost, namely, the difference between the value function evaluated at the current state and with one more customer present. The latter cost is just the external effect of the arriving customer's decision to join the system – the expected loss in aggregate net benefits to all future arrivals. It is straightforward to show (again by an inductive argument) that this expected loss is always non-negative, from which Naor's result immediately follows: an individually optimal policy admits more customers than is socially optimal. In economic terms, therefore, the equivalent charging scheme leads to the "right" model – one in which the proper balance of internal and external effects is seen clearly as an instance of the classic trade-off between an immediate benefit and a future cost, made explicit (as is always the case in an *MDP* model) by the optimality equation.

A key to the success of the inductive analysis in [220] was the use of a clever idea that Steve had recently developed and applied to a number of queueing control problems. Initially referred to as "Lippman's new device" (from the title of his paper [218]), the idea is best understood in its original context, a queue with a single exponential server. Steve observed that the optimality of a policy is not affected by pretending that the server continues to work even when no customers are present, completing fictitious services (so-called "null events") that do not change the state of the system. This may introduce additional time points at which the system is observed and/or actions are taken, but the memoryless property of the exponential distribution insures that the underlying sample path, and thus the optimal policy, is not affected.

It was later recognized (cf. Serfozo [278]) that this approach is equivalent to extending the idea of *uniformization* from continuous-time Markov processes to continuous-time Markov decision processes. In the context of a backward induction to prove that the optimal value function has certain properties (e.g., concavity), the advantage of uniformization is that it makes the expected time until the next state transition (and hence the denominator in the right-hand side of the optimality equation) independent of the state and action, which makes the inductive step much easier. An equivalent way of understanding the advantage of uniformization is to note that the resulting optimality equation is equivalent to that of a discrete-time *MDP*. It is not an exaggeration to say that Lippman's "new device" opened the gates for the application of *MDP* theory, specifically the backward induction approach, to queueing control problems. Virtually every paper using the backward induction approach since Lippman [218] has used uniformization.

Niels Knudsen and I became close friends during his visit to Cornell and our friendship continues to this day, despite Niels's defection from academia a few years later to embark on a second career – first as a spokesman for the Danish savings banks and ultimately as president of the largest savings bank on the island of Fyn. Niels arranged for me to spend a sabbatical year (1971-72) at his university – the University of Aarhus – in the newly formed OR group in the Matematisk Institut. I gave a weekly seminar on queueing optimization and began a fruitful collaboration with Søren Glud Johansen, which led (several years later) to a joint paper on control of arrivals to a stochastic input-output system [175]. Again the motivation was to see how far one could extend Naor's results.

In the course of preparing the seminar on queueing optimization, I decided to revisit a problem that I had first encountered in my doctoral dissertation: when can we assert that the average cost per unit time in a queueing system equals the arrival rate times the average cost per customer? This intuitive relation (now usually denoted $H = \lambda G$) is a generalization of Little's Law – $L = \lambda W$ – and reduces to Little's Law when the cost per customer is linear in the customer's time in the system.

My first step was to think hard about Little's Law itself. One traditionally encounters $L = \lambda W$ in a steady-state context in a Markov setting, in which each of the three quantities is an expected value of a random variable. The great contribution of Little's proof [221], in my opinion, was that it recast the problem in terms of limiting averages along sample paths, rather than means of limiting or stationary distributions. In this interpretation, the relation has an intuitive appeal. Indeed, if one adopts the cost interpretation, as above, it sounds almost too "obvious" to need a proof. Moreover, it should be true regardless of what is going on inside the "black box", that is, the queueing system itself. It should apply to any system in which customers (discrete units) arrive, spend some time, and then leave. The traditional interpretations in terms of expected values can then be recovered by invoking a law of large numbers or ergodic theorem – depending on the stochastic assumptions in one's model – to assert that the limiting average equals the expectation of the quantity in question, with probability one. Little's proof, however, did not follow this approach completely, but instead mixed sample-path arguments with stochastic arguments, the latter exploiting the strict stationarity of the processes involved. (In addition, there was a subtle error in his application of strict stationarity. In order for $L = \lambda W$ to hold as a relation between expected values in a stationary setting, one must interpret W as the mean waiting time taken with respect to the Palm (customer-stationary) distribution, rather than the continuous-time stationary distribution. See Section 2.2 and Dick Serfozo's comments in Section 4.4.)

It seemed to me that it should be possible to construct a "pure" sample-path proof of $L = \lambda W$. Such a proof would, properly speaking, fall outside the discipline of Applied Probability, inasmuch as it would not use any probabilistic reasoning. A sample path is, by its nature, a deterministic entity – a given sequence of numbers and/or a given function of time. In talking about limiting averages along a sample path, we are adopting a God-like perspective: we are standing at the end of time looking back over the history of the universe. Everything that was ever going to happen has already happened; since there is no future, there is no uncertainty. (And hence there is no longer a market for Applied Probability!) So (I thought) a proof of $L = \lambda W$, as a relation between limiting averages along a fixed sample path, should adopt exactly this "end-of-time" perspective. It should not go beyond the sample path in question, because at the end of time there is only the one, fixed sample path. In particular, it should not use the terms "almost surely" or "with probability one", because such terms seduce us back into the world of uncertainty – the world of many sample paths.

Armed with this determination to be deterministic, I first thought of using the cost interpretation and relating the limiting average costs to discounted costs via Tauberian theorems. The starting point was a clever argument by Colin Bell [18] for re-expressing the discounted cost in a queueing system (an integral over time) as a summation of discounted costs associated with each customer, each of these discounted from the customer's arrival point back to zero. This relation, which follows from simple interchanges of integration and summation, can be viewed as a discounted version of $L = \lambda W$. It should be straightforward (I thought) to let the discount rate go to zero, apply a Tauberian argument, and get $L = \lambda W$ as a relation between limiting averages. My first attempt at this made an (unconscious) assumption that the departures occurred in the same order as arrivals. This was kindly pointed out to me by my colleagues at Aarhus. So I tried again. After several false starts, it finally dawned on me that the crucial requirement was that, whatever the queue discipline, the waiting times of the customers must grow more slowly than their arrival times ($W_n = o(A_n)$). If this was the case then the limits (as the discount rate goes to zero) should be the same, whether one thinks of a customer's costs being incurred (i) at the instant of arrival, (ii) the instant of departure, or (iii) anywhere in between. Assumption (i) yields λW and assumption (iii) yields L, in the limit. The requirement that $W_n = o(A_n)$ turned out to be relatively innocuous. It holds, for example, if both λ and W are simply well defined, finite limiting averages. The results of this analysis appeared in [294].

Others (in particular, Maxwell [227], Eilon [96], and Newell [243]) had preceded me in the insight that $L = \lambda W$ should be provable on sample paths, but each of their proofs required assumptions beyond just the existence and finiteness of the averages. (Eilon's proof actually required re-defining W in such a way that $L = \lambda W$ becomes a tautology, and leaving out the hard part: showing that his new definition coincides with the standard one.) I recall meeting Gordon Newell and Ron Wolff in a bar at the Atlantic City ORSA/TIMS meeting in 1972. I told them about my pure sample-path proof. When I told them it used a discounted analogue and Tauberian theorems, they both winced. I could see that I had not accomplished my objective: providing a proof that was both rigorous and intuitive. On the way back to Ithaca, I figured out how to recast the proof without the detour through discounting and the result was my paper, "A last word on $L = \lambda W$" [296]. Of course, it was not a last word, but just the beginning of a twenty-five year research program concerned with applying sample-path analysis to a wide variety of problems in queueing theory and related fields, including relations between continuous-time and imbedded discrete-time distributions, insensitivity, and conservation principles. Much of this work was done with my colleague, Muhammad El-Taha, who was my PhD student at N.C. State University. Our collaboration culminated in our 1999 book, "Sample-Path Analysis of Queueing Systems" [98].

It occurs to me that over the course of my career I have been very astute (or fortunate) in my choice of collaborators. In addition to my students, these have included Steve Lippman, Søren Glud Johansen, Dick Serfozo, Dan Heyman, Jo van Nunen, Jacob Wijngaard, Tomasz Rolski, Colin Bell, Richard Weber, Eitan Altman, and Nicole Bäuerle. The opportunity to work with fine people like these, and to travel and live in many different parts of the world, has meant that I have never once regretted my choice of *OR* as a career.

REFERENCES

[1] Abate J., Choudhury G. L., Whitt W. (2000) An introduction to numerical tranform inversion and its application to probability models. In *Computational Probability.* W. K. Grassmann, Editor, Kluwer Academic Publishers, Boston, 257–324.

[2] Adler I., Naor P. (1969) Social optimization vs. self-optimization in waiting lines. Technical Report No. 4, Department of Operations Research, Stanford University.

[3] Altiok T. (1997) *Performance Analysis of Manufacturing Systems.* Springer-Verlag, New York.

[4] Altman E. (1999) *Constrained Markov Decision Processes.* Chapman & Hall/CRC, London/Boca Raton.

[5] Anick D., Mitra D., Sondhi M. M. (1982) Stochastic theory of a data-handling system with multiple sources. *Bell Sys. Tech. J.* **61** 1871–1894.

[6] Arrow K. J., Karlin S., Scarf H. (1958) *Studies in the Mathematical Theory of Inventory and Production.* Stanford Mathematical Studies in the Social Sciences, Stanford University Press, Stanford, Calif.

[7] Arrow K.J., Karlin S., Scarf H. (eds.) (1962) *Studies in Applied Probability and Management Science, Vol. 1.* Stanford University Press, Stanford, CA.

[8] Arrow K.J., Karlin S., Scarf H. (eds.) (1963) *Studies in Applied Probability and Management Science, Vol. 2.* Stanford University Press, Stanford, CA.

[9] Arrow K. J. (2002) The genesis of "optimal inventory theory". *Operations Research* **000** 000–000.

[10] Avi-Itzhak B., Maxwell W. L., Miller L. W. (1965) Queuing with alternating priorities. *Operations Research* **65** 306–318.

[11] Baccelli F., Brémaud P. (1987) *Palm Probabilities and Stationary Queues.* Lecture Notes in Statistics 41, Springer-Verlag, Berlin/New York.

[12] Baccelli F., Liu Z. (1992) On a class of stochastic recursive sequences arising in queueing theory. *Ann. Probab.* **20** 350–374.

[13] Baccelli F., Brémaud, P. (1994) *Elements of Queueing Theory. Palm-Martingale Calculus and Stochastic Recurrences.* Applications of Mathematics **26** Springer-Verlag, Berlin.

[14] Balsamo S., de Nitto Personé V., Onvural R. (2001) *Analysis of Queueing Networks with Blocking.* Kluwer Academic Publishers, Boston.

[15] Barford P., Crovella M. (1998) Generating representative Web workloads for network and server performance evaluation. In *Proc. 1998 ACM Sigmetrics*, 151–160.

[16] Baskett F., Chandy K., Muntz R., Palacios F. (1975) Open, closed, and mixed networks of queues with different classes of customers. *J. Assoc. Comput. Mach.* **22** 248–260.

[17] Bäuerle N., Stidham S. (2001) Conservation laws for single-server fluid networks. *Queueing Systems: Theory and Applications* **38** 185–194.

[18] Bell C. (1971) Characterization and computation of optimal policies for operating an *M/G/1* queuing system with removable server. *Operations Research* **19** 208–218.

[19] Bellman R. E. (1954) The theory of dynamic programming. *Bull. Amer. Math. Soc.* **60** 503–516.

[20] Bellman R. E. (1957) *Dynamic Programming.* Princeton University Press, Princeton, NJ.

[21] Bellman R. (ed.) (1964) *Stochastic Processes in Mathematical Physics and Engineering: Proc. Symp. Appl. Math. XVI*, Am. Math. Soc., Providence, R.I.

[22] Beneš, V. E. (1963) *General Stochastic Processes in the Theory of Queues.* Addison-Wesley, Reading, MA.

[23] Bertsekas D., Gallager R. (1987) *Data Networks.* Prentice-Hall, Englewood Cliffs, NJ.

[24] Bertsekas D. (1987) *Dynamic Programming: Deterministic and Stochastic Models.* Prentice Hall, Englewood Cliffs, NJ.

[25] Bertsimas D. (1995) The achievable region method in the optimal control of queueing systems; formulations, bounds and policies. *Queueing Systems: Theory and Applications* **21** 337–389.

[26] Bertsimas D., Niño-Mora, J. (1996) Conservation laws, extended polymatroids and multiarmed bandit problems; a polyhedral approach to indexable systems. *Operations Research* **21** 257–306.

[27] Bertsimas D., Paschalidis I., Tsitsiklis J. (1994) Optimization of multiclass queueing networks: polyhedral and nonlinear characterizations of achievable performance. *Ann. Appl. Prob.* **4** 43–75.

[28] Bertsimas D., Nakazato D. (1995) The distributional Little's law and its applications. *Operations Research* **43** 298–310.

[29] Bertsimas D., Mourtzinou G. (1997) Multiclass queueing systems in havy traffic: an asymptotic approach based on distributional and onservation laws. *Oper. Res.* **45** 470–487.

[30] Bertsimas D., Mourtzinou G. (1999) Decomposition results for general polling systems and their applications. *Queueing Systems Theory Appl.* **31** 295–316.

[31] Billingsley P. (1968) *Convergence of Probability Measures.* Wiley, New York.

[32] Black F., Scholes M. (1973) The pricing of options and corporate liabilities. *J. Polit. Econ.* **81** 637–654.

[33] Blackwell D. (1962) Discrete dynamic programming. *Ann. Math. Statist.* **33** 719–726.

[34] Blackwell D. (1965) Discounted dynamic programming. *Ann. Math. Statist.* **36** 226–235.

[35] Blackwell D. (1965) Positive dynamic programming. In *Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability,* Vol. 1, University of Caifornia Press, Berkeley, CA, 415–418.

[36] Borovkov A. A. (1964) Some limit theorems in the theory of mass service I. *Theory of Probability and its Applications.* **9** 550–565.

[37] Borovkov A. A. (1965) Some limit theorems in the theory of mass service II. *Theory of Probability and its Applications.* **10** 375–400.

[38] Borovkov A. A. (1967) On the convergence to diffusion processes. *Theory of Probability and its Applications.* **12** 405–431.

[39] Borovkov A. A. (1972) *Stochastic Processes in the Theory of Mass Service.* Nauk, Moscow.

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

[41] Borovkov A. A. (1986) Limit theorems for queueing networks. *Theory of Probability and its Applications.* **31** 413–427.

[42] Borst S. C., Boxma O. J. (1997) Polling models with and without switchover times. *Oper. Res.* **45** 536–543.

[43] Boxma O. J., Groenendijk W. P. (1987) Pseudoconservation laws in cyclic-service systems. *J. Appl. Probab.* **24** 949–964.

[44] Boxma O. J. (1989) Workloads and waiting times in single-server systems with multiple customer classes. *Queueing Systems Theory Appl.* **5** 185–214.

[45] Bramson M. (1994) Instability of *FIFO* queueing networks. *Ann. Appl. Probab.* **4** 414–431.

[46] Bramson M. (1998) Stability of two families of queueing networks and a discussion of fluid limits. *Queueing Systems Theory Appl.* **28** 7–31.

[47] Brandt A., Last G. (1995) *Marked Point Processes on the Real Line: the Dynamic Approach.* Springer-Verlag, New York.

[48] Brémaud, P. (1981) *Point Processes and Queues. Martingale Dynamics.* Springer Series in Statistics, Springer-Verlag, New York-Berlin.

[49] Burke P. J. (1964) The dependence of delays in tandem queues. *Ann. Math. Statist.* **35** 874–875.

[50] Burke P. J. (1968) The output process of a stationary *M/M/s* queueing system. *Ann. Math. Statist.* **39** 1144–1152.

[51] Buzen J. P. (1973) Computational algorithms for closed queueing networks with exponential servers. *Comm. ACM* **16** 527–531.

[52] Buzacott J. A., Shanthikumar J. G. (1993) *Stochastic Models of Manufacturing Systems.* Prentice-Hall, Englewood Cliffs, NJ.

[53] Chang C.-S. (1994) Stability, Queue Length, and Delay of Deterministic and Stochastic Queueing Networks. *IEEE Trans. Auto. Control* **39** 913–931.

[54] Chao X., Miyazawa M., Pinedo M. (1999) *Queueing Networks: Customers, Signals, and Product Forms.* Wiley, New York.

[55] Chen H., Mandelbaum A. (1991) Discrete flow networks: bottleneck analysis and fluid approximations. *Math. Oper. Res.* **16** 408–446.

[56] Chen H., Mandelbaum A. (1991) Leontief systems, RBVs and RBMs. In *Applied stochastic analysis (London, 1989).* Stochastics Monogr. **5** Gordon and Breach, New York, 1–43.

[57] Chen H., Mandelbaum A. (1991) Stochastic discrete flow networks: diffusion approximations and bottlenecks. *Ann. Probab.* **19** 1463–1519.

[58] Chen H., Yao. D.D. (2001) *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization.* Springer, New York.

[59] Coffman E., Mitrani I. (1980) A characterization of waiting-time performance realizable by single-server queues. *Operations Research* **28** 810–821.

[60] Coffman E. G., Puhalskii A. A., Reiman M. I. (1995) Polling systems with zero switchover times: a heavy-traffic averaging principle. *Ann. Appl. Probab.* **5** 681–719.

[61] Cohen J. W. (1957) The generalized Engset formulae. *Philips Telecomm. Rev.* **18** 158–170.

[62] Cohen J. W. (1969) *The Single Server Queue.* North Holland, Amsterdam.

[63] Cooper R. B., Murray G. (1969) Queues served in cyclic order. *Bell Syst. Tech. J.* **48** 675–689.

[64] Cox D. R. (1955) The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables. *Proc. Camb. Phil. Soc.* **51** 433–441.

[65] Cox D. R., Smith W. L. (1961) *Queues.* Methuen, London.

[66] Cox J., Ross S., Rubenstein M. (1979) Option pricing: simplified approach. *J. Finan. Econ.* **7** 229–264.

[67] Cooper R. B. (1972) *Introduction to Queueing Theory.* Macmillan, New York.

[68] Crommelin C. D. (1932) Delay probability formulae when the holding times are constant. *Post Office Electrical Engineer's Journal* **25** 41–50.

[69] Crommelin C. D. (1934) Delay probability formulae. *Post Office Electrical Engineer's Journal* **26** 266–74.

[70] Crovella M. E., Taqqu, M. S. Estimating the heavy tail index from scaling properties. *Methodol. Comput. Appl. Probab.* **1** 55–79.

[71] Dacre M., Glazebrook K. D., Niño-Mora J. (1999) The achievable region approach to the optimal control of stochastic systems. *J. Royal Stat. Soc.* **B 61** 747–791.

[72] Dai J. G., Harrison J. M. (1991) Steady-state analysis of *RBM* in a rectangle: numerical methods and a queueing application. *Ann. Appl. Probab.* **1** 16–35.

[73] Dai J. G., Harrison J. M. (1992) Reflected Brownian motion in an orthant: numerical methods for steady-state analysis. *Ann. Appl. Probab.* **2** 65–86.

[74] Dai J. G., Harrison J. M. (1993) The *QNET* method for two-moment analysis of closed manufacturing systems. *Ann. Appl. Probab.* **3** 968–1012.

[75] Dai J. G., Wang Y. (1993) Nonexistence of Brownian models for certain multiclass queueing networks. *Queueing Systems Theory Appl.* **13** 41–46.

[76] Dai J. G. (1995) On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Ann. Appl. Probab.* **5** 49–77.

[77] Dai J. G. (1995) Stability of open multiclass queueing networks via fluid models. In *Stochastic networks*, IMA Vol. Math. Appl. **71** Springer, New York, 71–90.

[78] Dai J. G., Meyn S. P. (1995) Stability and convergence of moments for multiclass queueing networks via fluid limit models. *IEEE Trans. Automat. Control* **40** 1889–1904.

[79] Dai J. G. (1996) A fluid limit model criterion for instability of multiclass queueing networks. *Ann. Appl. Probab.* **6** 751–757.

[80] Dai J. G., Weiss G. (1996) Stability and instability of fluid models for reentrant lines. *Math. Oper. Res.* **21** 115–134.

[81] Daley D. J., Vere-Jones D. (1988) *An Introduction to the Theory of Point Processes.* Springer Series in Statistics, Springer-Verlag, New York.

[82] Denardo E. V. (1967) Contraction mappings in the theory underlying dynamic programming. *SIAM Review* **9** 165–177.

[83] Disney R. L. (1986) The making of a queueing theorist. In *The Craft of Probability Modelling.* J. Gani, Editor, Springer-Verlag, New York, 196–212.

[84] Doshi B. T. (1976) Continuous time control of Markov processes on an arbitrary state space: discounted rewards. *Ann. Statist.* **4** 1219–1235.

[85] Doshi B. T. (1976) Continuous time control of Markov processes on an arbitrary state space: average return criterion. *Stochastic Processes Appl.* **4** 55–77.

[86] Doshi B. T. (1977) Continuous time control of the arrival process in an *M/G/1* queue. *Stochastic Processes Appl.* **5** 265–284.

[87] Doshi B. T. (1978) Controlled one dimensional diffusions with switching costs—average cost criterion. *Stochastic Process. Appl.* **8** 211–227.

[88] Doshi B. T. (1985) A note on stochastic decomposition in a *GI/G/1* queue with vacations or set-up times. *J. Appl. Probab.* **22** 419–428.

[89] Doshi B. T. (1986) Queueing systems with vacations—a survey. *Queueing Systems Theory Appl.* **1** 29–66.

[90] Doshi B. T. (1990) Generalizations of the stochastic decomposition results for single server queues with vacations. *Comm. Statist. Stochastic Models* **6** 307–333.

[91] Down D., Meyn S. P. (1995) Stability of acyclic multiclass queueing networks. *IEEE Trans. Automat. Control* **40** 916–919.

[92] Down D., Meyn S. P. (1997) Piecewise linear test functions for stability and instability of queueing networks. *Queueing Systems Theory Appl.* **27** 205–226.

[93] Dreyfus S. E. (2002) Richard Bellman on the genesis of dynamic programming. *Operations Research* **000** 000–000.

[94] Dukhovny A. (1998) Multiple roots in queueing equations: how realistic must examples get? *Stochastic Models* **14** 763–765.

[95] Edmonds J. (1970) Submodular functions, matroids and certain polyhedra. In *Combinatorial Structures and Their Applications: Proceedings of the Calgary International Conference on Combinatorial Structures and Their Applications.* Gordon and Breach, Science Publishers, 69–87.

[96] Eilon S. (1969) A simpler proof of $L = \lambda W$. *Operations Research* **17** 915–917.

[97] Eisenberg M. (1972) Queues with periodic service and change-over times. *Oper. Res.* **20** 440–451.

[98] El-Taha M., Stidham, S. (1999) *Sample-Path Analysis of Queueing Systems.* Kluwer Academic Publishers, Boston.

[99] Embrechts P., Klüppelberg C., Mikosch T. (1997) *Modelling Extremal Events for Insurance and Finance.* Springer-Verlag, New York.

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

[101] Federgruen A., Groenevelt H. (1988) Characterization and optimization of achievable performance in general queueing systems. *Operations Research* **36** 733–741.

[102] Federgruen A., Groenevelt H. (1987) The impact of the composition of the customer base in general queueing models. *J. Appl. Prob.* **24** 709–724.

[103] Feller W. (1957) *An Introduction to Probability Theory and Its Applications. Vol. I* (3rd ed., 1968) Wiley, New York.

[104] Feller W. (1966) *An Introduction to Probability Theory and Its Applications. Vol. II* Wiley, New York.

[105] Fleming W. H., Rishel R. W. *Deterministic and Stochastic Optimal Control.* Springer-Verlag, New York.

[106] Foschini G. (1977) On heavy traffic diffusion analysis and dynamic routing in packet-switched network. In *Computer Performance,* K. M. Chandy and M. Reiser, Editors, North-Holland, Amsterdam, 499–513.

[107] Foschini G., Salz J. (1978) A basic dynamic routing problem and diffusion. *IEEE Trans. Commun.* **26** 320–327.

[108] Foss S. G. (1992) On the ergodicity conditions for stochastically recursive sequences. *Queueing Systems Theory Appl.* **12** 287–296.

[109] Foster F. G. (1953) On stochastic matrices associated with certain queueing processes. *Ann. Math. Statist.* **24** 355–360.

[110] Franken P., König D., Arndt U., Schmidt V. (1981) *Queues and Point Proceses.* Akademie-Verlag, Berlin.

[111] Fuhrmann S. W. (1985) Symmetric queues served in cyclic order. *Oper. Res. Letters* **4** 139–144.

[112] Fuhrmann S. W., Cooper R. B. (1985) Stochastic decompoitions in a *M/G/1* queue with gneralized vacation. *Oper. Res.* **33** 1117–1129.

[113] Fuhrmann S. W. (1992) A decomposition result for a class of polling models. *Queueing Systems Theory Appl.* **11** 109–120.

[114] Gani J., Spier A. (1965) The birth of the Journal of Applied Probability. *Am. Statist.* **19** 18–22.

[115] Gani J. (ed.) (1986) *The Craft of Probabilistic Modelling.* Springer-Verlag, New York.

[116] Gani J. (ed.) (1988) *Adventures in Applied Probability: A Celebration of Applied Probability. J. Appl. Probab.* **25A** 3–23.

[117] Gaver D. P. (1971) Analysis of remote terminal backlogs under heavy demand conditions. *J. Assoc. Comput. Mach.* **18** 405–415.

[118] Gelenbe E., Mitrani I. (1980) *Analysis and Synthesis of Computer Systems.* Academic Press, London.

[119] Gershwin S. B. (1994) *Manufacturing Systems Engineering.* Prentice-Hall, Englwood Cliffs, NJ.

[120] Gessford J., Karlin S. (1958) Optimal policies for hydroelectric operations. In *Studies in the Mathematical Theory of Inventory and Producton.* K. J. Arrow, S. Karlin, and H. Scarf, Editors, Stanford University Press, Stanford, CA, 179–200.

[121] Ghoneim H., Stidham S. (1985) Control of arrivals to two queues in series. *European J. Operational Research* **21** 399–409.

[122] Gittins J. C., Jones D. M. (1974) A dynamic allocation index for the sequential design of experiments. *Progress in Statistics,* J. Gani, Ed., North-Holland, Amsterdam, 241–266.

[123] Gittins J. C. (1979) Bandit processes and dynamic allocation indices. *J. Roy. Statist. Soc. B* **41** 148–177.

[124] Glazebrook K. D., Garbe R. (1997) Almost optimal policies for stochastic systems which almost satisfy conservation laws. *Annals of Operations Research*

[125] Glynn P., Whitt W. (1986) A central limit theorem version of $L = \lambda W$. *Queueing Systems: Theory and Applications.* **2** 191–215.

[126] Glynn P., Whitt W. (1989) Extensions of the queueing relations $L = \lambda W$ and $H = \lambda G$. *Operations Research.* **37** 634–644.

[127] Gnedenko B. V., Kovalenko I. N. (1989) *Introduction to Queueing Theory.* Second Edition. Birkhäuser, Boston.

[128] Gordon W. J., Newell G. F. (1967) Cyclic queueing systems with restricted queue lengths. *Operat. Res.* **15** 266–278.

[129] Grassmann W. K. (ed.) (2000) *Computational Probability.* Kluwer Academic Publishers, Boston.

[130] Grassmann W. K., Taksar M. I., Heyman D. P. (1985) Regenerative analysis and steady state distributions for Markov chains.

[131] Green T. C., Stidham S. (2000) Sample-path conservation laws, with applications to scheduling queues and fluid systems. *Queueing Systems: Theory and Applications* **36** 175–199.

[132] Gross D., Harris C. M. (1974) *Fundamentals of Queueing Theory.* Wiley, New York.

[133] Haji R., Newell G. (1971) A relation between stationary queue and waiting time distributions. *J. Appl. Probab.* **8** 617–620.

[134] Halfin S., Whitt W. (1981) Heavy traffic limits for queues with many exponential servers. *Operations Research* **29** 567-587.

[135] Harrison J. M. (1973) The heavy traffic approximation for single server queues in series. *J. Appl. Probability* **10** 613–629.

[136] Harrison J. M. (1973) A limit theorem for priority queues in heavy traffic. *J. Appl. Probability* **10** 907–912.

[137] Harrison J. M. (1978) The diffusion approximation for tandem queues in heavy traffic. *Adv. in Appl. Probab.* **10** 886–905.

[138] Harrison J. M., Kreps D. M. (1979) Martingales and arbitrage in multiperiod securities markets. *J. Econom. Theory* **20** 381–408.

[139] Harrison J. M., Pliska S. R. (1981) Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Process. Appl.* **11** 215–260.

[140] Harrison J. M., Reiman M. I. (1981) Reflected Brownian motion on an orthant. *Ann. Probab.* **9** 302–308.

[141] Harrison J. M., Reiman M. I. (1981) On the distribution of multidimensional reflected Brownian motion. *SIAM J. Appl. Math.* **41** 345–361.

[142] Harrison J. M. (1985) *Brownian motion and stochastic flow systems.* Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, Inc., New York.

[143] Harrison J. M., Williams, R. J. (1987) Brownian models of open queueing networks with homogeneous customer populations. *Stochastics* **22** 77–115.

[144] Harrison J. M. (1988) Brownian models of queueing networks with heterogeneous customer populations. In *Stochastic Differential Systems, Stochastic Control Theory and Applications.* W. Fleming, Editor, IMA Vol. Math. Appl. **10** Springer, New York, 147–186.

[145] Harrison J. M., Wein, L. M. (1989) Scheduling networks of queues: heavy traffic analysis of a simple open network. *Queueing Systems Theory Appl.* **5** 265–279.

[146] Harrison J. M. (1990) *Brownian motion and stochastic flow systems.* (Reprint of the 1985 original.) Robert E. Krieger Publishing Co., Inc., Malabar, FL.

[147] Harrison J. M., Wein, L. M. (1990) Scheduling networks of queues: heavy traffic analysis of a two-station closed network. Oper. Res. **38** 1052–1064.

[148] Harrison J. M., Nguyen, V. (1990) The *QNET* method for two-moment analysis of open queueing networks. *Queueing Systems Theory Appl.* **6** 1–32.

[149] Harrison J. M., Williams R.J. (1992) Brownian models of feedforward queueing networks: quasireversibility and product form solutions. *Ann. Appl. Probab.* **2** 263–293.

[150] Harrison, J. M., Nguyen V. (1993) Brownian models of multiclass queueing networks: current status and open problems. *Queueing Systems Theory Appl.* **13** 5–40.

[151] Harrison J. M. (1995) Balanced fluid models of multiclass queueing networks: a heavy traffic conjecture. In *Stochastic Networks*, IMA Vol. Math. Appl. **71**, Springer, New York, 1–20.

[152] Harrison J. M. (1996) The *BIGSTEP* approach to flow management in stochastic processing networks. In *Stochastic Networks: Theory and Applications,* F. P. Kelly, S. Zachary, and I. Ziedins, Editors, Oxford University Press, Oxford.

[153] Harrison J. M., Lopez, M. J. (1999) Heavy traffic resource pooling in parallel-server systems. *Queueing Systems Theory Appl.* **33** 339–368.

[154] Hashida O. (1970) Gating multiqueues served in cyclic order. *Systems-Computers-Controls* **1** 1–8.

[155] Heyman D. P., Stidham S. (1980) The relation between customer and time averages in queues. *Operations Research* **28** 983–994.

[156] Hillier F. S., Lieberman G. J. (2001) *Introduction to Operations Research*, McGraw-Hill, New York.

[157] Hinderer K. (1970) *Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameter.* Springer-Verlag, New York.

[158] Hordijk A. (1974) *Dynamic Programming and Markov Potential Theory.* Mathematical Centre Tract **51** Amsterdam.

[159] Hordijk A., Kallenberg L. C. M. (1979) Linear programming and Markov decision chains. *Management Sci.* **25** 352–362.

[160] Hordijk A., Kallenberg L. C. M. (1981) On solving Markov decision problems by linear programming. In *Recent Developments in Markov ecision Processes.* R. Hartley, L. C. Thomas, and D. J. White, Editors, Academic Press, New York.

[161] Howard R. (1960) *Dynamic Programming and Markov Processes.* MIT Press, Cambridge, MA.

[162] Howard R. A. (2002) Comments on the origin and applications of Markov decision processes. *Operations Research* **000** 000–000.

[163] Iglehart D. L. (1965) Limit theorems for queues with traffic intensity one. *Ann. Math. Statist.* **36** 1437–1449.

[164] Iglehart D. L. (1965) Limiting diffusion approximations for the many server queue and the repairman problem. *J. Appl. Probability* **2** 429–441.

[165] Iglehart D. L. (1968) Diffusion approximations in applied probability. In *Mathematics of the Decision Sciences,* Part 2, American Mathematical Society, Providence, R.I., 235–254

[166] Iglehart D. L., Whitt, W. (1970) Multiple channel queues in heavy traffic. I. *Advances in Appl. Probability* **2** 150–177.

[167] Iglehart D. L., Whitt, W. (1970) Multiple channel queues in heavy traffic. II. Sequences, networks, and batches. *Advances in Appl. Probability* **2** 355–369.

[168] Iglehart D. L., Whitt, W. (1971) Multiple channel queues in heavy traffic. IV. Law of the iterated logarithm. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **17** 168–180.

[169] Iglehart D. L., Whitt, W. (1971) The equivalence of functional central limit theorems for counting processes and associated partial sums. *Ann. Math. Statist.* **42** 1372–1378.

[170] Jackson J. R. (1957) Networks of waiting lines. *Operations Res.* **5** 518–521.

[171] Jackson J. R. (1963) Jobshop-like queueing systems. *Management Science* **10** 131–142.

[172] Jackson J. R. (2002) How networks of queues came about. *Operations Res.* **000** 000–000.

[173] Jaiswal, N. K. (1968) *Priority Queues.* Academic Press, New York.

[174] Jewell W. (1967) A simple proof of $L = \lambda W$. Operations Research **15** 1109–1116.

[175] Johansen S. G., Stidham S. (1980) Control of arrivals to a stochastic input-output system. *Advances in Applied Probability* **12** 972–999.

[176] Karr A. (1991) *Point Processes and Their Statistical Inference.* Second Edition. Marcel Dekker, New York.

[177] Keilson J. (1986) Return of the wanderer: a physicist becomes a probabilist. In *The Craft of Probabilistic Modelling.* J. Gani, Editor, Springer Verlag, New York.

[178] Keilson J., Servi L. (1988) A distributional form of Little's law. *Oper. Res. Lett.* **7** 223–227.

[179] Keilson J., Servi L. (1990) The distributional form of Little's law and the Fuhrmann-Cooper decomposition. *Oper. Res. Lett.* **9** 239–247

[180] Kelly F. P. (1975) Networks of queues with customers of different types. *J. Appl. Prob.* **12** 542–554.

[181] Kelly F. P. (1979) *Reversibility and Stochastic Networks.* Wiley, New York.

[182] Kelly F. P., Laws C. N. (1993) Dynamic routing in open qeueing networks: Brownian models, cut constraints, and resource pooling. *Queueing Systems Theory Appl.* **13** 47–86.

[183] Kelly F. P. (1996) Notes on effective bandwidths. In *Stochastic Networks: Theory and Applications*, F.P. Kelly, S. Zachary, and I.B. Ziedins, Editors, *Royal Statistical Society Lecture Notes Series* **4**, Oxford University Press, 141–168.

[184] Kelly F. (2000) Mathematical modelling of the Internet. *Proc. 4th International Congress on Industrial and Applied Mathematics.*

[185] Kendall D. G. (1951) Some problems in the theory of queues. *J. R. Statist. Soc.* **B 13** 151–173.

[186] Kendall D. G. (1953) Stochastic processes occurring in the theory of queues, etc. *Ann. Math. Statist.* **24** 338–354.

[187] Kiefer J., Wolfowitz J. (1955) On the theory of queues with many servers. *Tran. Amer. Math. Soc.* **78** 1–18.

[188] Kingman J. F. C. (1961) The single server queue in heavy traffic. *Proc. Cambridge Philos. Soc.* **57** 902–904.

[189] Kingman J. F. C. (1962) On queues in heavy traffic. *J. Roy. Statist. Soc. Ser. B* **24** 383–392.

[190] Kingman J. F. C. (1965) The heavy traffic approximation in the theory of queues (with discussion). In *Proc. Symposium on Congestion Theory.* W. Smith and W. Wilkinson, Editors, Univ. North Carolina Press, Chapel Hill, N.C., 137–169.

[191] Kingman J. F. C. (1965) Approximations for queues in heavy traffic. In *Queueing Theory: Recent Developments and Applications.* R. Cruon, Editor, Elsevier, New York.

[192] Kingman J. F. C. (1966) On the algebra of queues. *J. Appl. Probability* **3** 285–326.

[193] Kingman J. F. C. (1966) On the algebra of queues. *Methuen's Supplementary Review Series in Applied Probability*, Vol. 6, Methuen and Co., Ltd., London.

[194] Kingman J. F. C. (1968) The ergodic theory of subadditive stochastic processes. *J. Roy. Statist. Soc. Ser. B* **30** 499–510.

[195] Kingman J. F. C. (1970) Inequalities in the theory of queues. *J. Roy. Statist. Soc. Ser. B* **32** 102–110.

[196] Kingman J. F. C. (1976) Subadditive processes. *Lecture Notes in Math.,* Vol. 539, Springer, Berlin.

[197] Kingman J. F. C. (1982) Queue disciplines in heavy traffic. *Math. Oper. Res.* **7** 262–271.

[198] Kleinrock L. (1975) *Queueing Systems, Vols. I, II.* Wiley Intersciences, New York.

[199] Kleinrock L. (2002) Creating a mathematical theory of computer networks. *Operations Research* **000** 000–000.

[200] Knudsen N. C. (1972) Individual and social optimization in a multiserver queue with a general cost-benefit structure. *Econometrica* **40** 515–528.

[201] Koenigsberg E. (1958) Cyclic queues. *Operat. Res. Quart.* **9** 22–35.

[202] Koenigsberg E. (1960) Finite queues and cyclic queues. *Operations Res.* **8** 246–253.

[203] Konheim A. G., Meister, B. (1974) Waiting lines and times in a system with polling. *J. Assoc. Comput. Mach.* **21** 470–490.

[204] König D., Matthes K., Nawrotzki K. (1967) *Verallgemeinerungen der Erlangschen und Engsetschen Formeln.* Akademie-Verlag, Berlin.

[205] Kook K., Serfozo R. F. Travel and sojourn times in stochastic networks. *Ann. Appl. Probab.* **3** 228–252.

[206] Kosten L. (1974) Stochastic theory of a multi-entry buffer (1). *Delft Progress Report* **1** 10–18.

[207] Kulkarni V. G. (1997) Fluid models for single buffer systems. In *Frontiers in Queueing,* J. Dshalalow, Ed., CRC, Boca Raton, FL, 321–338.

[208] Kumar P. R., Seidman T. I. (1990) Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems. *IEEE Trans. Automat. Control* **35** 289–298.

[209] Kumar P. R. (1993) Re-entrant lines. *Queueing Systems Theory Appl.* **13** 87–110.

[210] Kumar P. R. (1995) Scheduling queueing networks: stability, performance analysis and design. In *Stochastic networks,* IMA Vol. Math. Appl. **71** Springer, New York, 21–70.

[211] Kumar P. R., Meyn S. P. (1995) Stability of queueing networks and scheduling policies. *IEEE Trans. Automat. Control* **40** 251–260.

[212] Kumar P. R., Meyn S. P. (1996) Duality and linear programs for stability and performance analysis of queuing networks and scheduling policies. *IEEE Trans. Automat. Control* **41** 4–17.

[213] Kushner H., Dupuis P. (1992) *Numerical Methods for Stochastic Control Problems in Continuous Time.* Springer-Verlag, New York. Second Edition, 2001.

[214] Kushner H. (2001) *Heavy Traffic Analysis of Controlled Queueing and Communication Networks.* Springer-Verlag, New York.

[215] Lemoine, A. (1978) Networks of queues: a survey of weak convergence results. *Management Sci.* **24** 1175–1193.

[216] Levy H., Sidi M. (1990) Polling systems: applications, modeling and optimization. *IEEE Trans. Commun.* **38** 1750–1760.

[217] Lindley D. V. (1952) The theory of queues with a single server. *Proc. Camb. Phil. Soc.* **48** 277–289.

[218] Lippman S. A. (1975) Applying a new device in the optimization of exponential queuing systems. *Operations Res.* **23** 687–710.

[219] Lippman S. A. (1975) On dynamic programming with unbounded rewards. *Management Sci.* **21** 1225–1233.

[220] Lippman S. A., Stidham S. (1977) Individual versus social optimization in exponential congestion systems. *Operations Research* **25** 233–247.

[221] Little J. D. C. (1961) A proof of the queuing formula: $L = \lambda W$. *Operations Research* **9** 383–387.

[222] Loynes, R. M. (1962) The stability of a queue with non-independent inter-arrival and service times. *Proc. Camb. Phil. Soc.* **58** 497–520.

[223] Loynes, R. M. (1962) Stationary waiting-time distributions for single-server queues. *Ann. Math. Statist.* **33** 1323–1339.

[224] MacColl L.A. (ed.) (1957) *Applied Probability: Proc. Symp. Appl. Math. VII*, Am. Math. Soc., Providence, R.I.

[225] Manne A. (1960) Linear programming and sequential decisions. *Management Sci.* **6** 259–267.

[226] Markowitz D. M., Wein L. M. (2001) Heavy traffic analysis of dynamic cyclic policies: a unified treatment of the single machine scheduling problem. *Oper. Res.* **49** 246–270.

[227] Maxwell W. (1970) On the generality of the equation $L = \lambda W$. *Operations Research* **18** 172–174.

[228] Merton R. C. (1973) Theory of rational option pricing. *Bell J. Econ. Management Sci.* **4** 141–183.

[229] Meyn S. P., Tweedie R. L. (1993) *Markov chains and stochastic stability.* Communications and Control Engineering Series, Springer-Verlag, London.

[230] Meyn S. P., Down D. (1994) Stability of generalized Jackson networks. Ann. Appl. Probab. **4** 124–148.

[231] Morse P. M. (1958) *Queues, Inventories, and Maintenance.* Wiley, New York.

[232] Naor P. (1969) On the regulation of queue size by levying tolls. *Econometrica* **37**.

[233] Neuts M. F., Yadin, M. (1968) The transient behavior of the queue with alternating priorities, with special reference to the waiting times. *Bull. Soc. Math. Belg.* **20** 343–376.

[234] Neuts M. F. (1981) *Matrix-Geometric Solutions in Stochastic Models. An Algorithmic Approach.* Johns Hopkins Series in the Mathematical Sciences **2** Johns Hopkins University Press, Baltimore, MD.

[235] Neuts M. F. (1986) An algorithmic probabilist's apology. In *The Craft of Probabilistic Modelling: A Collection of Personal Accounts.* J. Gani, Editor, Springer-Verlag, New York, 213–221.

[236] Neuts M. F. (1986) The caudal characteristic curve of queues. *Adv. Appl. Prob.* **18** 221–54.

[237] Neuts M. F. (1988) Computer experimentation in applied probability. In *A Celebration of Applied Probability.* Special volume (25A) of *J. Appl. Prob.*, J. Gani, Editor, 31–43.

[238] Neuts M. F. (1989) *Structured Stochastic Matrices of $M/G/1$ Type and Their Applications.* Probability: Pure and Applied **5** Marcel Dekker, New York.

[239] Neuts M. F. (1990) Probabilistic modelling requires a certain imagination. In *Proceedings of the Third International Conference on Teaching Statistics* (ICOTS 3), August 19–24, 1990, Dunedin, New Zealand, D. Vere-Jones, Editor, Vol. 2, 122–131.

[240] Neuts M. F. (1992) Algorithmic probability: a survey and a forecast. In *Proceedings of the Second Conference of the Association of Asian-Pacific Operational Research Societies*, Beijing, Cang-Pu Wu, Editor, Peking University Press, 13–25.

[241] Neuts M. F. (1995) Matrix-analytic methods in queueing theory. In *Advances in queueing,* J. Dshalalow, Editor, Probab. Stochastics Ser., CRC Press, Boca Raton, FL, 265–292.

[242] Neuts M. F. (1998) Some promising directions in algorithmic probability. In *Advances in Matrix Analytic Methods for Stochastic Models*, A. S. Alfa and S. R. Chakravarthy, Editors, Neshanic Station, NJ, Notable Publications, Inc., 429–443.

[243] Newell G. F. (1971) *Applications of Queueing Theory.* Chapman and Hall, London.

[244] Ou J., Wein L. M. (1992) Performance bounds for scheduling queueing networks. *Ann. Appl. Probab.* **2** 460–480.

[245] Ou J., Wein L. M. (1992) On the improvement from scheduling a two-station queueing network in heavy traffic. *Oper. Res. Lett.* **11** 225–232.

[246] Perros H. G., Altiok T. (1989) *Queueing Networks with Blocking.* North-Holland, Amsterdam.

[247] Perros H. G. (1994) *Queueing Networks with Blocking.* Oxford Press, New York.

[248] Peterson W. P. (1991) A heavy traffic limit theorem for networks of queues with multiple customer types. *Math. Oper. Res.* **16** 90–118.

[249] Pliska S. R. (1997) *Introduction to Mathematical Finance: Discrete Time Models.* Blackwell Publishers Inc., Malden, MA.

[250] Prabhu N. U. (1965) *Queues and Inventories: A Study of Their Basic Stochastic Processes.* Wiley, New York.

[251] Prabhu N. U. (1970) Ladder variables for a continuous time stochasic process. *Z. Wahrscheinlichkeitstheorie verw. Geb.* **16** 157–164.

[252] Prabhu N. U. (1980) *Stochastic Storage Processes: Queues, Insurance Risk, and Dams.* Springer-Verlag, New York.

[253] Prabhu N. U. (1982) Conferences on stochastic processes and their applications: a brief history. *Stochastic Process. Appl.* **12** 115–116.

[254] Prabhu N. U. (1988) Stochastic processes and their applications. In *Encyclopedia of Statistical Sciences.* **8** S. Kotz and N. L. Johnson (eds.), John Wiley, New York.

[255] Prohorov Y. V. (1963) Transient phenomena in queueing processes. *Liet. Mat. Rink.* **3** 199–206 (in Russian).

[256] Puterman M. L., Shin M. C. (1978) Modified policy iteration algorithms for discounted Markov decision processes. *Management Sci.* **24** 1127–1137.

[257] Puterman M. L., Shin M. C. (1982) Action elimination procedures for modified policy iteration algorithms. *Operations Research.* **30** 301–318.

[258] Puterman M. L. (1994) *Markov Decision Proceses: Discrete Stochastic Dynamic Programming.* Wiley, New York.

[259] Reich E. (1957) Waiting times when queues are in tandem. *Ann. Math. Statist.* **28** 768–773.

[260] Reich E. (1963) Note on queues in tandem. *Ann. Math. Statist.* **34** 338–341.

[261] Reich E. (1965) Departure processes (with discussion). *Proc. Sympos. Congestion Theory (Chapel Hill, N.C., 1964).* Univ. North Carolina Press, Chapel Hill, N.C., 439–457.

[262] Reiman M. I. (1978) Open queueing networks in heavy traffic. PhD Dissertation, Department of Operations Research, Stanford University, Stanford, CA.

[263] Reiman M. I. (1984) Open queueing networks in heavy traffic. *Math. Oper. Res.* **9** 441–458.

[264] Reiman M. I., Wein L. M. (1998) Dynamic scheduling of a two-class queue with setups. *Oper. Res.* **46** 532–547.

[265] Reiser M., Lavenberg S. S. (1980) Mean value analysis of closed multichain queueing networks. *J. Assoc. Comput. Mach.* **27** 313–322.

[266] Rendleman R. J., Bartter B. J. (1979) Two-state option pricing. *J. Finance* **34** 1093–1110.

[267] Ross S. M. (1968) Non-discounted denumerable Markovian decision models. *Ann. Math. Statist.* **39** 412–423.

[268] Rybko A. N., Stolyar A. L. (1992) On the ergodicity of random processes that describe the functioning of open queueing networks. (Russian) *Problemy Peredachi Informatsii* **28** 3–26; translation in *Problems Inform. Transmission* **28** 199–220.

[269] Saaty T. (1961) *Elements of Queueing Theory.* McGraw-Hill, New York.

[270] Samorodnitsky G., Taqqu, M. S. (1994) *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance.* Chapman & Hall, New York.

[271] Scarf H. (2002) Inventory theory. *Operations Research* **000** 000–000.

[272] Schäl M. (1975) Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal. *Z. Wahrscheinlichkeitstheorie verw. Gerb.* **32** 179–196.

[273] Schassberger R. (1977) Insensitivity of steady-state distributions of generalized semi-Markov processes: Part I. *Ann. Prob.* **5** 87–89.

[274] Schassberger R. (1978) Insensitivity of stationary probabilities in a network of queues. *Adv. in Appl. Prob.* **10** 906–912.

[275] Schassberger R. (1978) Insensitivity of steady-state distributions of generalized semi-Markov processes: Part II. *Ann. Prob.* **6** 85–93.

[276] Schassberger R. (1978) Insensitivity of steady-state distributions of generalized semi-Markov processes with speeds. *Adv. in Appl. Prob.* **10** 836–851.

[277] Sennott L. (1999) *Stochastic Dynamic Programming and the Control of Queueing Systems.* Wiley, New York.

[278] Serfozo R. F. (1979) An equivalence between continuous and discrete time Markov decision processes. *Oper. Res.* **27** 616–620.

[279] Serfozo R. F. (1993) Queueing networks with dependent nodes and concurrent movements. *Queueing Systems: Theory and Applications* **13** 143–182.

[280] Serfozo R. F. (1994) Little laws for utility processes and waiting times in queues. *Queueing Systems: Theory and Applications* **17** 137–181.

[281] Serfozo R. F. (1999) *Introduction to Stochastic Networks.* Springer-Verlag, New York.

[282] Shaked M., Shanthikumar J. G. (1994) *Stochastic Orders and Their Applications.* Probability and Mathematical Statistics, Academic Press, Inc., Boston, MA.

[283] Shwartz A., Weiss A. (1995) *Large Deviations for Performance Analysis: Queues, Communications, and Computing.* Chapman & Hall, New York.

[284] Sigman K. (1995) *Stationary Marked Point Processes: An Intuitive Approach.* Chapman & Hall, New York.

[285] Smith W. L. (1953) On the distribution of queueing times. *Proc. Camb. Phil. Soc.* **49** 449–461.

[286] Smith W. L. (1955) Regenerative stochastic processes. *Proc. Roy. Soc.* **A 232** 6–31.

[287] Syski R. (1960) *Introduction to Congestion Theory in Telephone Systems.* Oliver and Boyd, Edinburgh and London.

[288] Shanthikumar J. G., Yao D. D. (1992) Multiclass queueing systems: polymatroidal structure and optimal scheduling control. *Operations Research* **40** S293–S299.

[289] Spitzer F. (1957) The Wiener-Hopf equation whose kernel is a probability density. *Duke Math. J.* **24** 327–344.

[290] Stewart W. J. (ed.) (1991) *Numerical Solution of Markov Chains.* Marcel Dekker, Inc., New York.

[291] Stewart W. J. (1994) *Introduction to the Numerical Solution of Markov Chains.* Princeton University Press, Princeton, NJ.

[292] Stewart W. J. (ed.) (1995) *Computations with Markov Chains.* Kluwer Academic Publishers, Boston.

[293] Stidham S. (1970) On the optimality of single-server queueing systems. *Operations Research* **18** 708–732.

[294] Stidham S. (1972) $L = \lambda W$: a discounted analogue and a new proof. *Operations Research* **20** 1115–1126.

[295] Stidham S. (1972) Regenerative processes in the theory of ueues, ith applications to the alternating-priority queue. *Adv. in Appl. Prob.* **4** 542–577.

[296] Stidham S. (1974) A last word on $L = \lambda W$. *Operations Research* **22** 417–421.

[297] Stidham S., Prabhu N. U. (1974) Optimal control of queueing systems. In *Mathematical Methods in Queueing Theory*, B. Clarke, Editor, Lecture Notes in Economics and Mathematical Systems **98** Springer-Verlag, Berlin, 263–294.

[298] Stidham S. (1978) Socially and individually optimal control of arrivals to a $GI/M/1$ queue. *Management Science* **24** 1598–1610.

[299] Stidham S. (1979) On the relation between time averages and customer averages in stationary random marked point processes. Technical Report, Department of Industrial Engineering, N. C. State University, Raleigh.

[300] Stidham S. (1981) On the convergence of successive approximations in dynamic programming with non-zero terminal reward. *Z. Op. Res.* **25** 57–77.

[301] Stidham S., van Nunen J. (1983) The shift-function approach for Markov decision processes with unbounded returns. Technical Report, Program in Operations Research, N. C. State University, Raleigh.

[302] Stidham S. (1994) Successive approximations for Markovian decision processes with unbounded rewards: a review. In *Probability, Statistics and Optimization: a Tribute to Peter Whittle*, F.P. Kelly, Editor, Wiley, Chichester, England, 467–484.

[303] Stidham S. (2000) Optimal control of Markov chains. In *Computational Probability.* W. K. Grassmann, Editor, Kluwer Academic Publishers, Boston, 325–365.

[304] Stoyan D. (1983) *Comparison methods for Queues and Other Stochasic Proceses.* Wiley, New York.

[305] Strauch R. (1966) Negative dynamic programming. *Ann. Math. Statist.* **37** 871–890.

[306] Symposium on Stochastic Processes. (1969) *J. Roy. Statist. Soc. B* **11** 150–264.

[307] Takács L. (1962) *Introduction to the Theory of Queues.* University Texts in the Mathematical Sciences, Oxford University Press, New York.

[308] Takagi H. (1986) *Analysis of Polling Systems.* MIT Press, Cambridge, MA.

[309] Takagi H. (1988) Queueing analysis of polling models. *ACM Comput. Surveys* **20** 5–28.

[310] Takagi H. (1990) Queueing analysis of polling models: an update. *Stochastic Analysis of Computer and Communication Systems.* North-Holland, Amsterdam, 267–318.

[311] Takagi H. (1993) *Queueing Analysis: A Foundation of Performance Evaluation. Vol. 3. Discrete-time systems.* North-Holland Publishing Co., Amsterdam.

[312] Takagi H. (1997) Queueing analysis of polling models: progress in 1990–1994. In *Frontiers in Queueing.* J. Dshalalow, Editor, Probab. Stochastics Ser., CRC, Boca Raton, FL, 119–146.

[313] Takine T. (2001) Distributional form of Little's law for *FIFO* queues with multiple Markovian arrival streams and its application to queues with vacations. *Queueing Systems Theory Appl.* **37** 31–63.

[314] Taylor H. M. (1965) Markovian sequential replacement processes. *Ann. Math. Statist.* **36** 1677–1694.

[315] Taylor L. M., Williams R. J. (1993) Existence and uniqueness of semimartingale reflecting Brownian motions in an orthant. *Probab. Theory Related Fields* **96** 283–317.

[316] Tsoucas P. (1991) The region of achievable performance in a model of Klimov. Research Report RC16543, IBM T. J. Watson Research Center, Yorktown Heights, NY.

[317] van der Wal J. (1981) *Stochastic Dynamic Programming.* Mathematical Centre Tract **139** Amsterdam.

[318] van Hee K., Hordijk A., van der Wal J. (1977) Successive approximations for convergent dynamic programming. In *Markov Decision Theory,* H. Tijms and J. Wessels, Editors, Mathematical Centre Tract **93**, Amsterdam, 183–211.

[319] van Nunen J. (1976) A set of successive approximation methods for discounted Markov decision processes. *Z. Op. Res.* **20** 203–208.

[320] van Nunen J. (1976) *Contracting Markov Decision Processes.* Mathematical Centre Tract **71** Amsterdam.

[321] van Nunen J., Wessels J. (1978) A note on dynamic programming with unbounded rewards. *Management Sci.* **24** 576–580.

[322] Veinott A. F. (1969) On discrete dynamic programming with sensitive optimality criteria. *Ann. Math. Statist.* **40** 1635–1660.

[323] Weber R. R., Stidham, S. (1987) Optimal control of service rates in networks of queues. *Adv. Appl. Prob.* **19** 202–218.

[324] Wein L. M. (1990) Optimal control of a two-station Brownian network. *Math. Oper. Res.* **15** 215–242.

[325] Wein L. M. (1992) Dynamic scheduling of a multiclass make-to-stock queue. *Oper. Res.* **40** 724–735.

[326] Wagner H. (2002) And then there were none. *Operations Research* **000** 000–000.

[327] Wald A. (1947) *Sequential Analysis.* Wiley, New York.

[328] Walrand J. (1988) *An Introduction to Queueing Networks.* Prentice-Hall, Englewood Cliffs, NJ.

[329] Welsh D. J. (1976) *Matroid Theory.* Academic Press, London.

[330] Wessels J. (1977) Markov programming by successive approximations with respect to weighted supremum norms. *J. Math. Anal. Appl.* **58** 326–335.

[331] Whitt W. (1974) Heavy traffic limit theorems for queues: a survey. In *Mathematical methods in queueing theory* (Proc. Conf., Western Michigan Univ., Kalamazoo, Mich., 1973). *Lecture Notes in Econom. and Math. Systems*, Vol. 98, Springer, Berlin, 307–350.

[332] Whitt W. (1991) A review of $L = \lambda W$ and extensions. *Queueing Systems: Theory and Applications.* **9** 235–268.

[333] Whitt W. (1992) $H = \lambda G$ and Palm transformation. *Adv. in Appl. Probab.* **24** 755–758.

[334] Whitt W. (2000) An overview of Brownian and non-Brownian *FCLT*'s for the single-server queue. *Queueing Systems: Theory and Applications.* **36** 39–70.

[335] Whitt W. (2000) The impact of a heavy-tailed service-time distribution upon the *M/GI/s* waiting-time distribution. *Queueing Systems Theory Appl.* **36** 71–87.

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

[337] Whittle P. (1968) Equilibrium distribution for an open migration process. *J. Appl. Prob.* **5** 567–571.

[338] Whittle P. (1979) A simple condition for regularity in negative programming. *J. Appl. Prob.* **16** 305–318.

[339] Whittle P. (1980) Stability and characterizations in negative programming. *J. Appl. Prob.* **17** 635–645.

[340] Whittle P. (1982) *Optimization over Time: Dynamic Programming and Stochastic Control.* Wiley, New York.

[341] Whittle P. (2001) Applied probability in Great Britain. *Operations Research* **000** 000–000.

[342] Williams R. J. (1998) Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Systems Theory Appl.* **30** 27–88.

[343] Williams R. J. (2000) On dynamic scheduling of a parallel server system with complete resource pooling. In *Analysis of Communication Networks: Call Centres, Traffic and Performance,* Fields Inst. Commun. **28** Amer. Math. Soc., Providence, RI, 49–71.

[344] Willinger W. (1995) Traffic modeling for high-speed networks: theory versus practice. In *Stochastic Networks,* F.P. Kelly and R.J. Williams, Eds. IMA Volumes in Mathematics and its Applications **71**, Springer-Verlag, New York, 169–186.

[345] Wolff R. (1970) Work-conserving priorities. *J. Appl. Probab.* **7** 327–337.

[346] Wolff R. (1982) Poisson arrivals see time averages. *Operations Research* **30** 223–231.

[347] Yao D. D. (1994) *Stochastic Modeling and Analysis of Manufacturing Systems.* Springer-Verlag, New York.

[348] Yechiali U. (1971) On optimal balking rules and toll charges in the *GI/M/1* queuing process. *Operations Res.* **19** 349–370.

[349] YECHIALI U. (1972) Customers' optimal joining rules for the *GI/M/s* queue. *Management Sci.* **18** 434–443.