Queueing Syst (2009) 63: 3–12
DOI 10.1007/s11134-009-9147-4

INVITED PAPER

The first Erlang century—and the next

J.F.C. Kingman

Received: 19 October 2009 / Revised: 26 October 2009 / Published online: 17 November 2009

Abstract The history of queueing theory, particularly over the first sixty years after Erlang's 1909 paper, is summarised and assessed, with particular reference to the influence of Pollaczek and Kendall. The interactions between the world of telephone traffic and that of applied probability and operational research are a significant factor. The history is followed by speculation about the directions in which the theory might now develop, in response to new problems and new possibilities. It is suggested that classical unsolved problems like the queue $\mathrm{M}/\mathrm{G}/k$ might be revisited, and that non-renewal inputs might be handled by martingale techniques.

Keywords History of queueing theory · Multi-server queues · Teletraffic · Poisson arrivals · Martingales

Mathematics Subject Classification (2000) 60K25

The purpose of this lecture, marking the centenary of Erlang's first paper on queueing theory, is to offer a personal perspective on the achievements of the hundred years that have elapsed since then, and to speculate a little on the directions in which the theory might evolve from here. I first encountered queueing theory in 1959, the exact midpoint of the Erlang century, but I was only actively involved for about ten years, and have since been only an admiring spectator of its development.

Erlang's own contributions, and the context in which he worked, have been well described by Brockmeyer, Halstrøm and Jensen in their magisterial edition [2] of his papers. In his 1909 paper, he argued that the number of calls arriving at a telephone exchange in a given time interval would be expected to follow a Poisson distribution,

In honour of Agner Krarup Erlang (1878–1929), and of David George Kendall (1918–2007).

J.F.C. Kingman (✉)
Bristol, UK
e-mail: John.Kingman@bristol.ac.uk

and he showed how this assumption led to expressions for the waiting time distribution if the calls were of fixed equal lengths. In the standard notation introduced in 1953 by Kendall [10], he solved the queue M/D/1.

However, what makes this first paper so remarkable is its clear recognition that congestion in telephone systems is a stochastic phenomenon which can be analysed by probabilistic arguments that go back to Pascal, Laplace and de Moivre. If calls arrived at regular intervals, and took fixed times, there would be no queueing or loss unless the equipment was inadequate for the load. But if they arrive at random, and even more if there is variability in the lengths of calls, no amount of exchange equipment can totally eliminate more or less frequent congestion.

In later papers Erlang analysed more complicated situations; in 1917, for instance, he produced his famous B formula for loss systems, and in 1920 he solved the multiserver queue $M/D/k$ for arbitrary k. His results were picked up by others working in what came to be called *teletraffic* theory, largely in the various national post offices which had claimed responsibility for telephone networks. In the USA this role was played by the Bell System, whose laboratories employed engineers and mathematicians who soon saw the value of Erlang's techniques (see, for instance, [6]).

The next important step, however, was taken in Berlin, where the Reichspost-Zentralamt had recruited in 1923 a young Austrian called Felix Pollaczek. He had studied mathematics in Vienna and electrical engineering in Brno before taking a doctorate in number theory in Berlin, a combination that equipped him well to apply classical mathematical techniques to the problems of teletraffic. His first papers on the subject, in 1930 [20], solved the queue M/G/1 with a general distribution of call length (service time).

It has to be remembered that neither Erlang nor Pollaczek had available a theory of random processes as was developed by Markov and, definitively, by Kolmogorov. Pollaczek throughout his long career (his last paper [24] was published in 1975 when he was 83) translated queueing problems as quickly as he could into integral equations which he attacked using tools of contour integration. Probabilistic methods were foreign to him, and it is truly remarkable that he achieved as much as he did. But his publications make difficult reading, and he came to feel that they did not receive the recognition they deserved. Nor was his personal life easy. He was dismissed from his post in 1933 because of his Jewish origin, and settled in Paris where he made a living as a consulting engineer until he became (temporarily in 1939 and later, from 1944, permanently) a Maître de Recherches with the CNRS.

Thus far the theory had penetrated only the teletraffic community, but Pollaczek's work came to the attention of the great Russian mathematician Aleksandr Yakovlevich Khinchin. Khinchin knew of the work of Markov, and he was in touch with the young Kolmogorov; in 1932 he recast the solution of M/G/1 in more recognisably probabilistic terms (the most accessible reference being the English version [11]). If it is unjust to Pollaczek that we bracket Khinchin's name in the famous formula for this queue, it is also a recognition that Khinchin saw the essentially probabilistic nature of the problem, and made possible the link with the evolving subject of applied probability.

That link however took a long time to be established, and it was only after the 1939–1945 world war that queueing theory became an active interest of the mathematical community. During that war, many mathematicians in different countries found themselves working on unusual problems thrown up by modern warfare, and many of these problems, including those that inspired the new concept of OR (operational research in Europe, operations research in the USA), involved stochastic considerations.

One such mathematician was David Kendall, who had intended to work in astronomy, but had been sidetracked by wartime priorities and by the influence of the statistician Maurice Bartlett. After the war he taught in Oxford, and tried to introduce into a traditional syllabus some elements of probability theory. In 1948 Stalin blockaded West Berlin, and Kendall was fascinated by the way in which communications and supplies were maintained by streams of aircraft in and out of the beleaguered city. He wondered if the problems of avoiding congestion could be approached mathematically and, combing the Oxford libraries, encountered the work of Erlang, Khinchin, Crommelin, Palm, Jensen and others. He decided to draw these together in a paper that he would offer to be read at one of the London meetings of the Royal Statistical Society.

It is the custom at such meetings that the two people who have refereed the paper propose and second a vote of thanks and start a discussion that is then printed *verbatim* with the paper. Kendall's 1951 paper [9] is thus a useful record not only of his views, but also of the reactions of colleagues reading the work. Kendall always regarded the paper as expository (and the reviewer for *Mathematical Reviews*, writing from the Bell Telephone Laboratories, agreed), but the paper introduced the subject to many who had not met it in other ways. One of these was Dennis Lindley, who opened the discussion with his famous recurrence relation

$$W_{n+1} = (W_n + S_n - T_n)^+. \tag{1}$$

Here W_n denotes the waiting time of the nth customer at a single-server queue, S_n his service time, and T_n the time between his arrival and that of the next customer.

Kendall's paper was mainly concerned with $M/G/1$, showing that the results of Pollaczek and Khinchin could be derived by studying (using methods contained in the then newly published book of Feller [5]) an *imbedded Markov chain* obtained by observing the queue length at each service instant. In a later paper [6] he gave a dual argument which solves the queue $GI/M/1$ and even the multi-server extension $GI/M/k$. But Lindley's equation makes it natural to assume that (S_n) and (T_n) are independent sequences of independent, identically distributed random variables; the queue $GI/G/1$ which was to become the subject of an enormous literature (see for instance Cohen's formidable tome [3]).

It is both the strength and the weakness of this approach that it reduces the queue $GI/G/1$ to a problem about sums of independent random variables. It allows to be brought to bear a rich classical theory, from the law of large numbers and the central limit theorem to the law of the iterated logarithm and large deviation theory, but it tempts scientists to restrict attention to those problems that can be analysed in this way, and in particular to single-server queues with renewal arrival processes.

Pollaczek gave in 1957 [21] his own account of $GI/G/1$, in which he used (1) but turned it at once into an integral equation. He claimed that his methods were different from those used by others such as Walter Smith [28] who followed Lindley [18], or those who later used Spitzer's identity [13], but in fact there is a single unifying thread to all these approaches [15].

My own involvement came while I was studying for my first degree in mathematics at Cambridge. In my first summer vacation, I took a temporary job in the small mathematics group which was a part of the Engineering Research Station of the UK General Post Office at Dollis Hill in London. I was given a number of problems, some with a simple probabilistic content which appealed to me. When I returned to the university, I therefore chose to take a course on probability given, as it happened, by Dennis Lindley. He was (and is) a superb lecturer, and used some simple queueing models as examples of the use of probability. When I returned to Dollis Hill in the summer of 1959 I mentioned these to the head of my section, and was told to go to the library and read the Bell System Technical Journal.

In this way I entered queueing theory very much from a teletraffic viewpoint, and I was encouraged to attack problems which were as realistic as possible, and to produce results that might conceivably be of use in the real world. Thus, for instance, many pieces of exchange equipment could hold waiting calls, but not in order of arrival, and so I tried to analyse the queue $M/G/1$ with random order of service. My solution was in the form of an impossibly complicated integral [12], but it was possible (and useful) to derive an approximation valid in *heavy traffic*, when the arrival rate is only just below the critical value at which the queue becomes unstable. This led me to consider robust approximations which could hold for other queues whose arrival processes were much more general than the renewal processes hitherto considered. When I started as a research student, my approach was encouraged by my more senior colleague Bob Loynes, who was proving stability theorems for very general systems [19].

In 1961 I moved from Cambridge to Oxford to work with David Kendall, and found to my surprise that my enthusiasm for queues was not shared. Kendall had seen his two papers used as the basis for the production of many routine PhD theses, giving unusable formal solutions for an arcane variety of implausible queueing models, and he was deeply disillusioned about the future of the subject. I was sufficiently impressed with his attitude to diversify my own research, although in fact its most important outcome, the theory of regenerative phenomena [16], flowed directly from my work at Dollis Hill. But of course, Kendall and I were wrong, for reasons I shall come to shortly.

In 1964, there were two important international conferences, which it is instructive to compare and contrast. The first, in London, was the Fourth International Teletraffic Congress, attended by nearly two hundred delegates from all over the developed world. Some sixty papers were read, and although some of these were of a general theoretical nature, most related to particular technical issues which are now only of historical interest. What came through from that meeting was a strong sense of a teletraffic community, quite separate from the mathematicians who were treating the theory of queues as an autonomous branch of applied probability with little relation to its telephonic origins.

The second meeting, organised in Chapel Hill in the USA by Walter Smith and William Wilkinson [29], was smaller (fewer than thirty participants) and much more mathematical in flavour. Most of the major figures in the post-Kendall development were present, but neither Kendall nor Lindley. Pollaczek (who had attended the London meeting) had accepted an invitation, and submitted a paper [23], but was prevented by ill health from attending. His paper, read and defended by Richard Syski, was particularly interesting as a summary of his analytical method. He had followed up his 1957 monograph on GI/G/1 with another [22] in 1961 on GI/G/k, which was even less penetrable. He now emphasised, as did Syski, that although he could not solve his integral equations in usable explicit form for general service and interarrival time distributions, such solution would in principle be possible if the characteristic functions of those distributions were rational functions.

This observation failed to impress the audience, because the necessary algebra would rapidly become infeasible as the degrees of those rational functions increased. But we should have taken more notice, because Pollaczek's suggestion was a vast extension of an idea of Erlang, that a service time with a gamma distribution (with integer parameter m, say) was equivalent to a sequence of m stages each of which has the negative exponential distribution. Erlang's idea was generalised, first by Marcel Neuts (who was present in Chapel Hill, and was working on semi-Markov queueing models) and more recently by Søren Asmussen and his collaborators [1]. Asmussen models very general distributions as the lifetimes of finite Markov chains, so reducing many queueing problems to matrix calculations well adapted to the modern computer.

And it is that very computer that has made nonsense of the sceptical view that queueing theory had run its course. The literally exponential increase in computing power has had two separate but complementary effects. In the first place, it has greatly widened our ideas of what constitutes a solution to a queueing problem. At the 1959 midpoint there were three ways of dealing with such a problem. If possible, one sought an analytic solution in reasonably explicit form. It might not be exact, but at least good approximations or inequalities could be useful. Failing that, one tried to write down equations that could be solved numerically, but the slow computers then available (I used the Ferranti Pegasus, a valve machine that I had to programme in machine code) greatly limited the feasibility or accuracy of such solutions. The final recourse was simulation, but the small number of replications that could be achieved meant that the results were often misleading.

With modern computers, both numerical solution and simulation have become very much more feasible and reliable. The value of approximate analytic solution remains, because it often gives a deeper understanding which complements and illuminates the results of numerical computation. But there is less temptation to adopt unrealistic assumptions merely for the sake of simple solution.

The other consequence of the silicon revolution has been to throw up many new problems. We all know from everyday experience that both individual computers and networks of computers give rise to frustrating delays that can be traced to congestion at some level of the system. The nature of this congestion may be very specific to the architecture of the hardware or software, and it may be dangerous to carry over insights from different applications, but the Erlang paradigm of regarding the system as a random process is still of value.

In 1964 it was already common to draw an analogy between a computer and a telephone exchange. Indeed, the earliest computers, designed by Alan Turing to break the German codes of the Enigma machine [26], had been built by the telephone engineers of the Post Office. As telephone networks became more global, and computers were more and more networked, the two technologies rapidly converged, and today are almost indistinguishable. In an important sense the Internet is the ultimate telephone exchange.

Although all this was well below the horizon in 1964, there were straws in the wind at the Chapel Hill meeting. Thomas Saaty, who worked then for the US Arms Control and Disarmament Agency, talked about networks of queues [25]. No doubt there was much that he had to leave unsaid, but it was already clear that there were important applications of such networks. Saaty cited results of R.R.P. Jackson and J.R. Jackson, and these were later picked up by Peter Whittle [31] and, with remarkable results, by Frank Kelly [7, 8]. It seems likely that the analysis of congestion in complex networks, up to and including the Internet, will be one of the most important components of the next queueing century.

Another application, also of great difficulty and also of practical importance, is that of road traffic. We are all familiar with the mysterious congestion that can occur, even on fast multi-lane roads with no obvious obstacles. There have been many attempts to explain these, for instance by analogy with statistical mechanics, and scientists of the calibre of Prigogine and Lighthill have proposed equations whose solution might yield explanation. As a recent symposium [27] makes clear, no solution so far carries complete conviction, and it seems likely that both probabilistic insight and fast computation will be needed to make progress.

There will of course be other areas of application not envisaged at Chapel Hill. One of these, which has become important in the UK, is that of waiting lists for elective surgery. Typically, in the British Health Service, there are quite long waiting times which do not fluctuate too much over time, and understanding these is politically important. No explanation from simple queueing theory will work, and it seems that there are feedback mechanisms that prevent finite busy periods.

The coming years will, we must hope, see progress made in understanding and controlling congestion in many different contexts, aided by the computing power of modern information technology. But it is also worth asking whether anything can be done about some of the simple unsolved problems which have, perhaps wisely, been left to one side of the mainstream of research. For example, $\mathrm{M}/\mathrm{G}/1$ was solved by Pollaczek, and $\mathrm{M}/\mathrm{D}/k$ for general k by Erlang, but what about $\mathrm{M}/\mathrm{G}/k$? This is surely an important system, with the Poisson arrivals that are still the most useful input process, and independent service times having a given but arbitrary distribution.

The nearest we have to a solution is that we can approximate the service-time distribution by a phase type distribution which is the lifetime distribution of some finite Markov chain. The process can then be analysed as a Markov chain whose states include that of the subsidiary chain, and the matrix manipulations needed are well adapted to the computer. But although every service-time distribution can be so approximated, the approximation may fail to capture important aspects. For example, every phase type distribution has an exponential tail. However, the work of Whitt [30] has shown that the qualitative behaviour of the queue can be markedly different if the service-time distribution has a heavy tail.

Yet $\mathrm{M}/\mathrm{G}/k$ is really a very simple system. If the nth customer arrives at time τ_n and has service time S_n, the points (τ_n, S_n) form a two-dimensional Poisson process

[17] which is the only random element of the model. It defines the cumulative load up to time t as a compound Poisson process, and the k servers then reduce this load at a rate equal to the number of servers working at the time. The difficulty of analysing the queue comes in relating this number to the (τ_n, S_n), which is a combinatorial rather than a stochastic problem.

This way of looking at the queue has two advantages. Firstly, it directs attention to the queue discipline. The advantage of FIFO 'first come, first served' is clear in single-server queues, but not when there are several servers and service times vary. Any queue discipline is an algorithm for reducing the unserved load, and needs to maximise (the time integral of) the number of busy servers. Which algorithms are admissible depends on such factors as whether the servers can recognise in advance the customers with long service times. If it were practical in any particular application, the 'best' service procedure would be for the k servers to combine to serve the customer at the head of the queue, reducing his service time by a factor of k in a manner illustrated by the pit stop mechanics in Formula One motor racing.

The other advantage is that the mean density of the two-dimensional Poisson process can be allowed to depend explicitly on time, so that secular variations in arrival rate and service-time distribution can be modelled at little extra cost. Indeed, one can allow that density to be itself a random process, leading to inputs which are the 'doubly stochastic Poisson processes' introduced by David Cox [4] and now named *Cox processes* [17].

And this could be an important observation, because the most serious criticism of the classical theory is its insistence on the GI input. The assumption that the interarrival times are independent and identically distributed, so that the arrivals form a renewal process, was made by Pollaczek, Lindley and Kendall because it was less restrictive than that they are a Poisson process, but still allows analytic progress to be made. Yet it is very difficult to produce a plausible example of a queue with a non-Poisson renewal input (except for the trivial case of arrivals at regular intervals). Sensible alternatives to the Poisson process do exist, but they are very far from renewal processes. For example, a Cox process can only be a renewal process if the underlying random rate takes only two values, one of them zero [14].

If a queue has an arrival process which cannot be well modelled by a Poisson process or one of its near relatives, it is likely to be difficult to fit any simple model, still less to analyse it effectively. So why do we insist on regarding the arrival times as random variables, quantities about which we can make sensible probabilistic statements? Would it not be better to accept that the arrivals form an irregular sequence, and carry out our calculations without positing a joint probability distribution over which that sequence can be averaged?

It is likely that the service mechanism, which may be very complicated, involving perhaps a network of queues with many servers, can be plausibly modelled as a random process, probably a (homogeneous) Markov process if enough variables are specified. Then we test that mechanism by subjecting it to an irregular but deterministic sequence of arrivals.

If the nth customer arrives at time τ_n, we assume that the state $Z(t)$ of the service mechanism is a homogeneous Markov process (perhaps on a large state space) on the interval (τ_n, τ_{n+1}). Moreover, the arrival at τ_n itself causes a jump in Z, and we assume that this is Markov, depending on the past only through $Z(\tau_n)$. In this situation, there is an imbedded Markov process

$$Z(\tau_n+); \quad n = 1, 2, 3, \ldots, \tag{2}$$

but it is not homogeneous unless the τ_n are in arithmetic progression. Moreover, even in the simplest cases, and if Z has finitely many states, the transition matrices of (2) do not commute, and this rules out any analytic solution. However, matrix multiplication is one of the strong points of modern computers, and there might be a possibility of numerical solutions in some non-trivial cases.

Such computation would have to be carried out for a range of possible sequences (τ_n). One way of doing this would be to use a large number of actual observed arrival sequences, to evaluate the performance of the queue under typical conditions. This might be more efficient, and correspond better with reality, than using observed sequences to fit a statistical model, and then sampling from the fitted model.

Markov theory makes strong use of time homogeneity, but martingale theory does not. Suppose that we decide arbitrarily to stop the arrivals after the nth customer. The service mechanism will continue to process the earlier customers, but after a random time T the system will become empty, and we can use this random variable to define a martingale on the interval $[\tau_n, \infty)$. Fix a positive number λ, and define a function ϕ on the state space of Z by requiring $\phi\{Z(\tau_n)\}$ to be the conditional expectation of $e^{-\lambda T}$ given the process Z up to τ_n. Then it is easy to see that

$$\Phi(t) = e^{-\lambda t}\phi\big\{Z(t)\big\} \quad (t \le T), \qquad \Phi(t) = \Phi(T) \quad (t > T) \tag{3}$$

defines such a martingale. Now keep the same function ϕ, but restore the arrivals after the nth. It will still be true that (3) defines a martingale Φ_n on the interval (τ_n, τ_{n+1}) between successive arrivals.

This in itself is of little use, but it may be possible to sew these small martingales, or constant multiples of them, together to form a martingale on the whole line, and then all the apparatus of martingale theory—stopping identities, inequalities, central limit theorems, and the like—becomes available. To illustrate what might be done, consider a very simple example. In the queue ID/M/k, where ID denotes an irregular deterministic arrival sequence, the number $N(t)$ present at time t is, between arrivals, a pure death process with death rates

$$\beta_r = \sigma \min(r, k), \tag{4}$$

where σ is the service rate and r a generic value of $N(t)$. There are also queues in which the death rates are not of the form (4), where for instance the servers speed up if the queue gets too long. Take N as the process Z, and compute that

$$\phi(N) = \prod_{r=1}^{N}\left(\frac{\beta_r}{\beta_r + \lambda}\right). \tag{5}$$

Because $N(\tau_n+) = N(\tau_n-) + 1$, (5) implies that

$$\Phi_n(\tau_n+) = \left(\frac{\beta_{N_n}}{\beta_{N_n} + \lambda}\right)\Phi_{n-1}(\tau_n-), \tag{6}$$

where $N_n = N(\tau_n+)$. If therefore

$$A_n = \prod_{m=1}^{n} \left(\frac{\beta_{N_m} + \lambda}{\beta_{N_m}} \right), \tag{7}$$

the process

$$\Psi(t) = A_n \Phi_n(t) \quad (\tau_n \le t < \tau_{n+1};\ n = 1, 2, 3, \ldots) \tag{8}$$

is continuous at each τ_n and is therefore a martingale on $[\tau_1, \infty)$.

This construction depends on the multiplicative property (6), and I do not know which more complex queueing systems have such a property. Nor do I know how far the martingale (8) can be exploited. My purpose is merely to challenge younger queueing theorists to take a broader view, not necessarily accepting without question the formulations of the past.

References

1. Asmussen, S.: Applied Probability and Queues. Springer, New York (2003)
2. Brockmeyer, E., Halstrøm, H.L., Jensen, A.: The Life and Works of A.K. Erlang. Akademiet for de Tekniske Videnskaber, København (1948)
3. Cohen, J.W.: The Single Server Queue. North-Holland, Amsterdam (1969)
4. Cox, D.R.: Some statistical models related to series of events. J. R. Stat. Soc. B **17**, 129–164 (1955)
5. Feller, W.: An Introduction to Probability Theory and Its Applications. Wiley, New York (1950)
6. Fry, T.C.: Probability and Its Engineering Uses. Van Nostrand, New York (1928)
7. Kelly, F.P.: Reversibility and Stochastic Networks. Wiley, London (1979)
8. Kelly, F.P., Zachary, S., Ziedins, I. (eds.): Stochastic Networks: Theory and Applications. Oxford University Press, Oxford (1996)
9. Kendall, D.G.: Some problems in the theory of queues. J. R. Stat. Soc. B **13**, 151–173 (1951); discussion 173–185
10. Kendall, D.G.: Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. Ann. Math. Stat. **24**, 338–354 (1953)
11. Khinchin, A.Y.: Mathematical Methods in the Theory of Queueing. Griffin, London (1960)
12. Kingman, J.F.C.: On queues in which customers are served in random order. Proc. Camb. Philos. Soc. **58**, 79–91 (1962)
13. Kingman, J.F.C.: The use of Spitzer's identity in the investigation of the busy period and other quantities in the queue GI/G/1. J. Aust. Math. Soc. **2**, 345–356 (1962)
14. Kingman, J.F.C.: On doubly stochastic Poisson processes. Proc. Camb. Philos. Soc. **60**, 923–930 (1964)
15. Kingman, J.F.C.: On the algebra of queues. J. Appl. Probab. **3**, 285–326 (1966)
16. Kingman, J.F.C.: Regenerative Phenomena. Wiley, London (1972)
17. Kingman, J.F.C.: Poisson Processes. Oxford University Press, Oxford (1993)
18. Lindley, D.V.: The theory of queues with a single server. Proc. Camb. Philos. Soc. **48**, 277–289 (1952)
19. Loynes, R.M.: The stability of a queue with non-independent interarrival and service times. Proc. Camb. Philos. Soc. **58**, 497–520 (1962)
20. Pollaczek, F.: Ueber eine Aufgabe der Wahrscheinlichkeitstheorie. Math. Z. **32**, 64–100 (1930) 729–750
21. Pollaczek, F.: Problèmes stochastiques posés par le phénomène de formation d'une queue d'attente à un guichet et par des phénomènes apparentés. Mém. Sci. Math. **136**, 1–122 (1957)
22. Pollaczek, F.: Théorie analytique des problèmes stochastiques relatifs à un groupe de lignes téléphoniques avec dispositif d'attente. Mém. Sci. Math. **150**, 1–114 (1961)
23. Pollaczek, F.: Concerning an analytic method for the treatment of queueing problems. In: Smith, W.L., Wilkinson, W.E. (eds.) Congestion Theory, pp. 1–42. University of North Carolina, Chapel Hill (1964)

24. Pollaczek, F.: Order statistics of partial sums of mutually independent random variables. J. Appl. Probab. **12**, 390–395 (1975)
25. Saaty, T.L.: Stochastic network flows: advances in networks of queues. In: Smith, W.L., Wilkinson, W.E. (eds.) Congestion Theory, pp. 86–107. University of North Carolina, Chapel Hill (1964)
26. Sebag-Montefiore, H.: Enigma: The Battle for the Code. Weidenfeld, Nicolson (2000)
27. Smith, M., Briggs, K., Kelly, F.P. (eds.): Networks: modelling and control. Philos. Trans. R. Soc. A 366, 1875–2092 (2008)
28. Smith, W.L.: On the distribution of queueing times. Proc. Camb. Philos. Soc. **49**, 449–461 (1953)
29. Smith, W.L., Wilkinson, W.E. (eds.): Congestion Theory. University of North Carolina, Chapel Hill (1964)
30. Whitt, W.: The impact of a heavy-tailed service-time distribution upon the $\mathrm{M}/\mathrm{GI}/s$ waiting-time distribution. Queueing Syst. Theory Appl. **36**, 71–87 (2000)
31. Whittle, P.: Systems in Stochastic Equilibrium. Wiley, London (1986)