Proceedings of the 8th International Conference on Applied Informatics
Eger, Hungary, January 27–30, 2010. Vol. 1. pp. 9–30.

Queueing Theory and its Applications, A Personal View*

János Sztrik

University of Debrecen, Faculty of Informatics, Hungary
sztrik.janos@inf.unideb.hu

Abstract

This paper deals with the Queueing Theory and some mathematical models of queueing systems. Starting with the historical backgrounds it gives an overview of different solution methods and tools. Then the basic laws and formulas are introduced. It highlights several recent advances and developments of the theory and new applications fields are listed. It briefly summarizes the achievements of Hungarian researchers to the Queueing Theory and the most cited result of the author is mentioned. It ends with the References of the most important sources.

Keywords: queueing theory, applications, finite-source models, telecommunication systems, operational research, manufacturing systems, teletraffic engineering

MSC: 60K25[Queueing Theory], 68M20, 90B22

1. Introduction

As the reader will quickly discover, this article is a short survey - from my personal perspective - of 32 years of research, teaching on the modeling, analysis, and applications of queueing systems. My choice of topics is far from exhaustive; I have focused on those research achievements that I believe have been some of the most significant in their contributions to queueing theory and to its applications. They are the contributions that I have admired and appreciated the most over the course of my teaching and research activities. Another author would undoubtedly have made different choices, as they did in several survey papers on queueing theory. The selection has not been easy at all since there are so many nice results. My

*Research is partially supported by Hungarian Scientific Research Fund-OTKA K 60698/2006. The work is supported by the TAMOP 4.2.1./B-09/1/KONV-2010-0007 project. The project is implemented through the New Hungary Development Plan, co-financed by the European Social Fund and the European Regional Development Fund.

aim is very simple, I would like to draw the attention of readers to a very unpleasant activity, namely waiting. I have collected some sayings or Murphy's Laws on Queueing. Here you are:

- *"If you change queues, the one you have left will start to move faster than the one you are in now.*
- *Your queue always goes the slowest.*
- *Whatever queue you join, no matter how short it looks, will always take the longest for you to get served."*

A queue is a waiting line (like customers waiting at a supermarket checkout counter); queueing theory is the mathematical theory of waiting lines. More generally, queueing theory is concerned with the mathematical modeling and analysis of systems that provide service to random demands.

A queueing model of a system is an abstract representation whose purpose is to isolate those factors that relate to the system's ability to meet service demands whose occurrences and durations are random. Typically, simple queueing models are specified in terms of the arrival process the service mechanism and the queue discipline. The arrival process specifies the probabilistic structure of the way the demands for service occur in time; the service mechanism specifies the number of servers and the probabilistic structure of the duration of time required to serve a customer, and the queue discipline specifies the order in which waiting customers are selected from the queue for service. Selecting or constructing a queueing model that is rich enough to reflect the complexity of the real system, yet simple enough to permit mathematical analysis) is an art. The ultimate objective of the analysis of queueing systems is to understand the behavior of their underlying processes so that informed and intelligent decisions can be made in their management.

Then, the mathematical analysis of the models would yield formulas that presumably relate the physical and stochastic parameters to certain *performance measures*, such as average response/ waiting time, server utilization, throughput, probability of buffer overflow, distribution function of response/waiting time, busy period of server, etc. The art of applied queueing theory is to construct a model that is simple enough so that it yields to mathematical analysis, yet contains sufficient detail so that its performance measures reflect the behavior of the real system. In the course of modeling one could use analytical, numerical, asymptotic, and simulation methods integrated into performance evaluation tools.

In the course of modeling we make several assumptions regarding the basic elements of the model. Naturally, there should be a mechanism by which these assumptions could be verified. Starting with testing the goodness of fit for the arrival and service distributions, one would need to estimate the parameters of the model and/or test hypotheses concerning the parameters or behavior of the system. Other important questions where statistical procedures play a part are in the determination of the inherent dependencies among elements and dependence of the system on time.

Starting with a congestion problem in teletraffic the range of applications has grown to include not only telecommunications and computer science, but also manufacturing, air traffic control, military logistics, design of theme parks, call centers, supermarkets, inventories, dams, hospitals, and many other areas that involve service systems whose demands are random. Queueing theory is considered to be one of the standard methodologies (together with linear programming, simulation, etc.) of operations research and management science, and is standard fare in academic programs in industrial engineering, manufacturing engineering, etc., as well as in programs in telecommunications, computer engineering, and computer science. There are dozens of books and thousands of papers on queueing theory, and they continue to be published at an ever-increasing rate. Searching the Google for "Queueing Theory" I have found 1880 hits.

This tremendous push for new results forced more and more academic journals to publish articles in queueing and even open new sections. In 1986, Baltzer Verlag, AG launched a new academic journal entitled *Queueing Systems* (edited by N.U. Prabhu), which is devoted entirely to queueing. Many other journals, in the field of probability, operational research, telecommunication, industrial engineering, computer science, management science publish articles on queueing extensively. The flow of new theories and methodologies in queueing has become very hard to keep up with. Surveys on the hottest topics in queueing and related areas are scattered over a large variety of scientific magazines. A sort of manual would be desirable, indicating where to find the hottest topics and where to concentrate one's efforts should queueing become one's interest. A careful elaboration of major themes would take many years of work after which the results would then be outdated!

After all these considerations I have decided to give a view of my personal perspectives of Queueing Theory and its Applications. I have been teaching and doing research on this topic for 32 years. I have realized that the applications of the theory became more and more important. In the following I would like to mention some of my favorite books on theory: *Allen [1], Artalejo-Gomez-Corral [2], Bolch-Greiner-De Meer-Trivedi [4], Cooper [5], Gnedenko-Kovalenko [14], Gross-Shortle-Thompson-Harris [15], Khintchine [20], Kleinrock [21], Kobayashi-Mark [23], Takagi [30], Takács [32], Trivedi [35].* Some recent books on applications, mainly computer networks and telecommunications, due to web-based research are the following: *Daigle [6], Giambene [13], Haghighi [17], Kleinrock [21].* Finally, I would like to mention some materials written by Hungarian authors: *Györfi [16], Lakatos-Szeidl-Telek [24], Sztrik [29]*, and the Hungarian translation of the famous book by *Kleinrock* [22].

Due to the huge interest of theory and application the queueing community led by Professor Hlynka, created a home page

`http://web2.uwindsor.ca/math/hlynka/queue.html`

where the most important information can be found concerning materials, conference, softwares, etc.

Just to imagine how important this topic I visited the Google Scholar for citations of the most popular books on queueing theory, not mentioning the ones which

Agner Krarup Erlang, 1878–1929

use the theory. I found the following number of citations, Allen [1]: 715, Cooper [5]: 1166, Gnedenko-Kovalenko [14]: 314, Gross-Harris [15]: 2610 (for earlier editions), Kleinrock [21], for Vol.1: 5813, Vol.2:1736, Manual: 2500, Kobayashi [23]: 350 (for earlier edition), Takagi [30]: 1190, Takács [32]: 888. I must point out that many of the Russian contributions are not refereed and not cited in spite of their importance due to the lack of information for western researchers.

2. Origin and Developments

The history of Queueing Theory goes back nearly 100 years. It was born with the work of A. K. Erlang who published in 1909 his paper, *The Theory of Probabilities and Telephone Conversations*, [10]. His most important work, *Solutions of Some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges* , [11] was published in 1917, which contained formulas for loss and waiting probabilities which are now known as Erlang's loss formula (or Erlang B-formula) and delay formula (or Erlang C-formula), respectively. Erlang's loss model assumes Poisson arrivals of telephone calls, namely, the number of sources or subscribers is sufficiently large. If the number of sources is finite and not so large, then a more accurate loss formula is provided by the Engset'loss formula, which was published by the Norwegian mathematician Engset. We should mention that the Erlang and the Engset loss model and their loss formulas remained the most widely used results in telephone engineering.

Erlang laid the foundation for the place of Poisson (and hence, exponential)

distribution in queueing theory. His papers written in the next 20 years contain some of the most important concepts and techniques; the notion of statistical equilibrium and the method of writing down balance of state equations (later called Chapman-Kolmogorov equations) are two such examples. Erlang's motivation was to develop tools for the analysis and design of telephone systems) an application that continues to the present day to motivate research in queueing theory.

It should be noted that in Erlang's work, as well as the work done by others in the twenties and thirties, the motivation has been the practical problem of congestion. The trend toward the analytical study of the basic stochastic processes of the system continued, and queueing theory proved to be a fertile field for researchers who wanted to do fundamental research on stochastic processes involving mathematical models.

Mathematical modeling is a process of approximation. A probabilistic model brings it a little bit closer to reality; nevertheless it cannot completely represent the real world phenomenon because of involved uncertainties. Therefore, it is a matter of convenience where one can draw the line between the simplicity of the model and the closeness of the representation.

Renewed interest in queueing theory and its potential applications outside of telecommunications came with the codification of the field of operations research in the early 1950's. And in the early 1960' s queueing theory was rediscovered by researchers interested in the performance analysis of time-shared computer systems. The use of queueing theory as a tool for the performance evaluation of computer systems and components of computer systems is now well established with the result that most undergraduate computer science majors receive at least some exposure to its models and methodology.

By the mid-70's researchers interested in modeling 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. 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.

Recently web-based research dominates the main directions. It is difficult to list all the main trends, but in my opinion the followings certainly belong to them: *long-range dependence, numerical problems of stochastic processes, time-dependent solutions, modeling tools, retrial systems, approximations, simulations, statistical inference.* For a more detailed discussions about ongoing research, see Dshalalow [8, 9]. Readers interested in the history of queueing theory are referred to, among others Dshalalow [8], Takagi [30] where extensive Bibliographies were collected involving numerous contributions of Russian experts lead by Khintcine, Gnedenko, Kovalenko, Bocharov, Rykov, too.

Boris Vladimirovich Gnedenko, 1912–1995

Leonard Kleinrock, 1934–

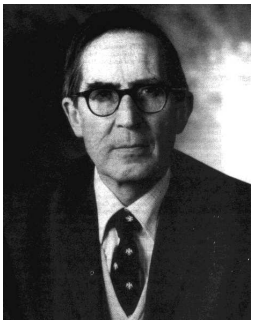
Research on queueing theory and its applications is very active, year-by-year different conferences, workshops are held. Just for illustration I would like to mention two of them
Third Madrid Conference on Queueing Conference, June 28 - July 1, 2010 , and *22nd International Teletraffic Congress, September 7-9, 2010, Amsterdam.* Some topics are

- Performance of wireless/wired networks
- Business models for QoS
- Performance and reliability tradeoffs

- Performance models for voice, video, data and P2P applications
- Scheduling algorithms
- Simulation methods and tools

It is my feeling that at present queueing theory is divided into two directions. One is highly abstract and the other highly practical. It seems that this split will continue to grow wider and wider. Progress in the theory of stochastic processes (especially point, regenerative, and stationary processes) will influence new approaches to queueing theory. This may be in the form of new methods, new interpretations, and the development of new theories with wide applicability. Researchers in abstract probability usually do not have queueing theory in mind; different talents are required to find applicability of their results. Other examples, are diffusion approximation, the large deviations technique, and random fields. One may hope that the near future will bring applications of superprocesses, the object of current research in stochastic processes. Progress in technical developments of systems involving various forms of traffic created the need for mathematical analysis of performance of individual systems. This brings new problems which require new tools, and the search for these tools is of great practical importance. This is clearly visible not only in teletraffic theory, but also in other disciplines where queueing methods are used (biological and health studies, computers). As already mentioned, simulation and numerical analysis are frequently the only way to obtain approximate results. It is therefore hoped that the gap between these two directions may eventually be diminished. Idealistically, this could be achieved when theoreticians learn about practical problems and practitioners learn about theory. In present times of great specialization, this is highly unrealistic. Nevertheless, one could try to work in this direction, at least with our students in universities, by stressing the importance of theory and applications. Otherwise, researchers could not find a common language.

3. Kendall's Notation



David G. Kendall, 1918–2007

3.1. Components of a Queueing System

While analyzing a queueing system we can identify some basic elements of it. Namely,

Input process: if the occurrence of arrivals and the offer of service are strictly according to schedule, a queue can be avoided. But in practice this does not happen. In most cases the arrivals are the product of external factors. Therefore, the best one can do is to describe the input process in terms of random variables which can represent either the number arriving during a time interval or the time interval between successive arrivals. If customers arrive in groups, their size can be a random variable as well.

Service mechanism: the uncertainties involved in the service mechanism are the number of servers, the number of customers getting served at any time, and the duration and mode of service. Networks of queues consist of more than one servers arranged in series and/or parallel. Random variables are used to represent service times, and the number of servers, when appropriate. If service is provided for customers in groups, their size can also be a random variable.

System capacity: at most how many customers can wait at a time in a queueing system is a significant factor for consideration. If the waiting room is large, one can assume that for all practical purposes, it is infinite. But our everyday experience with the telephone systems tells us that the size of the buffer that accommodates our call while waiting to get a free line is important as well.

Service discipline: all other factors regarding the rules of conduct of the queue can be pooled under this heading. One of these is the rule followed by the server in accepting customers for service. In this context, the rules such as First-Come, First-Served (FCFS), Last-Come, First-Served (LCFS), and Random Selection for Service (RS) are self-explanatory. In many situations customers in some classes get priority in service over others. There are many other queue disciplines which have been introduced for the efficient operation of computers and communication systems. Also, there are other factors of customer behavior such as balking, reneging, and jockeying, that require consideration as well.

3.2. Classification of Systems

The following notation, *known as Kendall's notation* , is widely used to describe elementary queueing systems:

$$\text{A/B/m/K/N /Z},$$

where

- A indicates the distribution of the interarrival times,
- B denotes the distribution of the service times,
- m is the number of servers,

- K is the capacity of the system, that is the maximum number of customers staying at the facility (sometimes in the queue),
- N denotes the number of sources,
- Z refers to the service discipline.

As an example of Kendall's notation, the expression

$$\text{M/G/1 - LCFS preemptive resume (PR)}$$

describes an elementary queueing system with exponentially distributed inter-arrival times, arbitrarily distributed service times, and a single server. The queueing discipline is LCFS where a newly arriving job interrupts the job currently being processed and replaces it in the server. The servicing of the job that was interrupted is resumed only after all jobs that arrived after it have completed service.

$$\text{M/G/1/K/N}$$

describes a finite-source queueing system with exponentially distributed source times, arbitrarily distributed service times, and a single server. There are N request in the system and they are accepted for service iff the number of requests staying at the server is less than K. The rejected customers return to the source and start a new source time with the same distribution. It should be noted that as a special case of this situation the $M/G/1/N/N$ system could be considered. However, in this case we use the traditional $M/G/1//N$ notation, that is the missing letter, as usual in this framework, means infinite capacity, and FCFS service rule.

It is natural to extend this notation to heterogeneous requests, too. The case when we have different customers is denoted by $\rightarrow$. So, the

$$\vec{M}/\vec{G}/1/K/N$$

denotes the above system with different arrival rates and service times.

4. Basic Formulas

This section is devoted to the most well-known formulas of queueing theory. The selection is subjective, but I think these are the ones from which many others have been derived.

4.1. Erlang's Formulas

As we mentioned in the earlier section the whole theory started with a practical problem. Erlang's task can be formulated as follows: What fraction of the incoming calls is lost because of the busy line at the telephone exchange office. Of course, the answer is not so simple, since we first should know the inter-arrival and service time distributions. After collecting data Erlang verified that the Poisson-process arrival

and exponentially distributed service were appropriate mathematical assumptions. He considered the $M/M/n/n$ and $M/M/n$ cases, that is the system where the arriving calls are lost because all the servers are busy, and where the calls have to wait for service, respectively. Assuming that the arrival intensity is λ, service rate is μ he derived the famous formulas for loss and delay systems, called Erlang B and C ones, respectively.

Denoting $\rho = \lambda/\mu$, the steady state probability that an arriving call is lost can be obtained in the following way

$$P_n = \frac{\frac{\rho^n}{n!}}{\sum_{k=0}^{n} \frac{\rho^k}{k!}} = B(n, \rho),$$

where $B(n, \rho)$ is the well-known **Erlang B- formula**, or loss formula. It can easily be seen that the following recurrence relation is valid

$$B(n, \rho) = \frac{\rho B(n-1, \rho)}{n + \rho B(m-1, \rho)} \qquad n = 2, 3, ...$$

$$B(1, \rho) = \frac{\rho}{1+\rho}.$$

Similarly, by using the B-formula the steady state probability that an arriving customer has to wait can be written as

$$C(n, \rho) = \frac{nB(n, \rho)}{n - \rho(1 - B(n, \rho)},$$

which is called **Erlang C-formula, or Erlang's delay formula**.

It should be mentioned that the B-formula is insensitive to the service time distribution, in other words it remains valid for any service time distribution with mean $1/\mu$.

4.2. Little's Law

Little's law, Little's result, or Little's theorem is perhaps the most widely used formula in queueing theory was published by J. Little [25] in 1961. It is simple to state and intuitive, widely applicable, and depends only on weak assumptions about the properties of the queueing system.

It says that the average number of customers in the system is equal to the average arrival rate of customer to the system multiplied by the average system time per customer.

Historically, Little's law has been written as

$$L = \lambda W$$

and in this usage it must be remembered that W is defined as *mean response time*, the mean time spent in the queue and at the server, and not just simply as the mean time spent waiting to be served, L refers to the *average number of customers in the system* and λ is *the mean arrival rate*. Little's law can be applied when we relate L to the average number of customers waiting to receive service, L_q and W to the mean time spent waiting for service, W_q, that is another well-known form is

$$L_q = \lambda W_q.$$

The same applies also to the servicing aspect itself. In other words, Little's law may be applied individually to the different parts of a queueing facility, namely the queue and the server.

It may be applied even more generally than we have shown here. For example, it may be applied to separate parts of much larger queueing systems, such as subsystems in a queueing network. In such a case, L should be defined with respect to the number of customers in a subsystem and W with respect to the total time in that subsystem. Little's law may also refer to a specific class of customer in a queueing system, or to subgroups of customers, and so on. Its range of applicability is very wide indeed.

Finally, we comment on the amazing fact that the proof of Little's law turns out to be independent of

- specific assumptions regarding the arrival distribution $A(t)$
- specific assumptions regarding the service time distribution $B(t)$
- number of servers
- particular queueing discipline

Little's law is important for three reasons:

- because it is so widely applicable (it requires only very weak assumptions), it will be valuable to us in checking the consistency of measurement data
- for example, in studying computer systems we frequently will find that we know two of the quantities related by Little's law (say, the average number of requests in a system and the throughput of that system) and desire to know the third (the average system residence time, in this case)
- it is central to the algorithms for evaluating several queueing network models

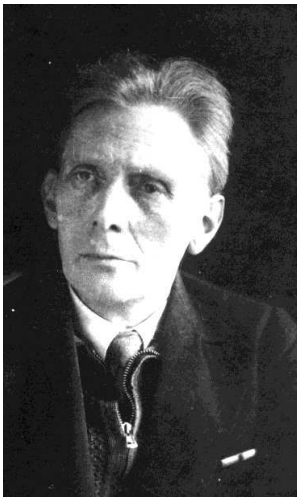
Given a computer system, Little's law can be applied at many different levels: to a single resource, to a subsystem, or to the system as a whole. The key to success is consistency: the definitions of population, throughput, and residence time must be compatible with one another.

Over the past few years, it has become increasingly important in many fields of applications. For more discussions on this topic one may read, for example Jewel [19], Ramalhota-Amaral-Cochito [26], Wolf [36].

4.3. Pollaczek-Khintchine Formulas

This section deals with formulas of an $M/G/1$ queueing system at which the customers arrive according to a Poisson-process with parameter λ, the service time is arbitrarily distributed, there is no restriction to the number of customers staying in the system, and they are serviced according to the order of their arrivals, that is the service discipline is FCFS. These formulas are treated almost every book on queueing theory but the notation is quite different. Each author prefers his own designation and as a consequence it is very difficult to find the proper form. We make difference between the type of formulas, the *mean value* and *transform* ones. Of course, the first ones are much easier to obtain. Independently of each other, Pollaczek and Khintchine derived them in the period 1930-50.

Felix Pollaczek, 1892–1981



Alexander Y. Khintchine, 1894–1959

As usual we need some notations. In steady state let us denote by

- N number of customers in the system,
- $P(z)$ the generating function of N, that is $P(z) = E(z^N)$,
- $B^*(s) = \int_0^\infty e^{-st} dB(t)$ the Laplace-Stieltjes transform of the service time S,
- $W^*(s)$ the Laplace-Stieltjes transform of the waiting time in the system, or response time T,
- $W_q^*(s)$ the Laplace-Stieltjes transform of the waiting time in the queue T_q,
- $C_S^2 = \frac{Var(S)}{E^2(S)}$ squared coefficient of variation of service time S,
- $L = E(N), \quad L_q = E(N_q)$ average number of customers in the system, queue, respectively,

- $W = E(T)$, $W_q = E(T_q)$ mean waiting time in the system, in the queue, respectively,
- $\rho = \lambda E(S)$.

Hence, the *mean value formulas* are as follows:

$$L = \rho + \rho^2 \frac{1 + C_S^2}{2(1-\rho)},$$

$$L_q = \rho^2 \frac{1 + C_S^2}{2(1-\rho)},$$

or by using the Little's law we have

$$W = E(S)(1 + \rho \frac{1 + C_S^2}{2(1-\rho)}),$$

$$W_q = E(S)\rho \frac{1 + C_S^2}{2(1-\rho)}.$$

The *transform formulas or equations* can be written as

$$P(z) = \frac{(1-\rho)(z-1)B^*(\lambda - \lambda z)}{z - B^*(\lambda - \lambda z)},$$

$$P(z) = W^*(\lambda - \lambda z),$$

$$W^*(s) = \frac{(1-\rho)sB^*(s)}{s - \lambda(1 - B^*(s))},$$

$$W_q^*(s) = \frac{(1-\rho)s}{s - \lambda(1 - B^*(s))}.$$

Hence, by applying the well-known properties of generating functions and Laplace-Stieltjes transforms using repeated differentiation for the higher moments we get

$$E(N(N-1)\cdots(N-k+1)) = \lambda^k E(T^k),$$

which is nice generalization of Little's formula for higher moments in FCFS case.

5. Hungarian Contributions to the Theory of Queues

In this section I would like to mention several Hungarian researchers without discussing their main results in details. The selection is subjective and it is based on my information collected during my research activities.

5.1. Lajos Takács and his Work

There is no doubt that he is the most well-known,reputed and celebrated Hungarian in this field. In the second half of 1994, many scientific institutions (including the Institute of Mathematical Statistics, Operations Research Society of America, The Institute of Management Sciences, and Hungarian Academy of Sciences) celebrated his 70th birthday, which took place on August 21, 1994. In addition, a special volume, *Studies in Applied Probability (31A of Journal of Applied Probability, edited by J. Galambos and J. Gani)*, [12], appeared in the first half of 1994 honoring Professor Takács. Another paper, [7] was devoted to his 70th birthday written by Dshalalow and Syski where we can read
" Because of his extraordinary accomplishments, Professor Takács is one of the most celebrated contemporary probabilists. He has published over 200 papers and books, many of which have had a huge impact on the contemporary theory of probability and stochastic processes. His numerous works are yet to be explored. Although some people view Takács as a queueing theorist, it is just one of many areas of his remarkable influence. This opinion about him is also because queueing theory (or, as they say, just queueing) has become so overwhelmingly popular, and because Takács is indisputably one of the greatest contributors to the theory who ever lived. This may outshine his other contributions to the theory of probability, stochastic processes, combinatorial analysis, and even physics. For instance, it is not widely known that Takács was the first to introduce and study semi-Markov processes in the early fifties and which he had been using even before 1954, perhaps one of the most extraordinary achievements in the theory of stochastic processes in the second half of the twentieth century.

In 1962 Takács published his **Introduction to the Theory of Queues**, *a masterpiece and, at the same time, one of the most widely cited monographs in queueing. His other masterpiece,* **On Fluctuations of Sums of Random Variables,** *published in 1978 is less popular, but it undoubtedly deserves more attention. Due to his phenomenal diversity, Takács left traces in many areas of mathematics and probability, such as: queueing theory, combinatorial analysis, point processes, random walks, branching processes, Markov processes, semi-Markov processes, probability on groups, signal processes, statistics, fluctuation theory, sojourn time problems in stochastic processes, ballot theorems, and random graphs. In spite of his numerous commitments, Takács is still amazingly productive. In 1993 he was elected a Foreign Member of the highly renowned Hungarian Academy of Sciences, and recently he was awarded the prestigious John von Neumann Theory Prize by the Operations Research Society of America and The Institute of Management Sciences. Currently, he is a professor emeritus at Case Western Reserve University in Cleveland, Ohio. "*

Besides these I would like to draw your attention to his other 2 books *Stochastic Processes, Problems and Solutions* [31] and *Combinatorical Methods in the Theory of Stochastic Processes* [33]. Readers interested in his results and Bibliography are referred to [7, 12]. Even the *Wikipedia* [38] devotes some pages to his works.

Reading the works of Takács we must admit that they require sound knowlegde

Lajos Takács, 1924–

in mathematics. The methods he used in the proofs are too sophisticated for our students in their BSc or MSc studies. In the following I would like to list some of his formulas and equations I found very useful in teaching queueing theory.

As I wrote earlier the $M/G/1$ system is basic for other more complicated queueing situations. In many problems occuring in operational research or computer science, the higher moments of the involved random variables are very important, for example, waiting time in the system T, or in the queue T_q, busy period of the server B just to mention some. The Pollaczek-Khintchine transform formulas enables us to get them by standard methods, even if the calculations are not easy. I found the following Theorem in [1] pp. 202 stated

Takács Recurrence Theorem: Consider an $M/G/1$ queueing system in which the $(j+1)$th moments of service time S, that is $E(S^{j+1})$ exitst.
Then $E(T_q), ..., E(T_q^j)$ also exist and

$$E(T_q^k) = \frac{\lambda}{1-\rho} \sum_{i=1}^{k} \binom{k}{i} \frac{E(S^{i+1})}{(i+1)} E(T_q^{k-i}), \quad k = 1, 2, ...j,$$

where $E(T_q^0) = 1$.

Hence, the moments $E(T), E(T^2), ..., E(T^j)$ exit and

$$E(T^k) = \sum_{i=0}^{k} \binom{k}{i} E(S^i) E(T_q^{k-i}), \quad k = 1, 2, ...j.$$

As a consequence we get

$$E(T_q) = \frac{\lambda E(S^2)}{2(1-\rho)},$$

$$E(T_q^2) = 2E^2(T_q) + \frac{\lambda E(S^3)}{3(1-\rho)},$$

$$E(T) = E(T_q) + E(S),$$

$$E(T^2) = E(T_q^2) + \frac{E(S^2)}{1-\rho},$$

from which the $Var(T_q), Var(T)$ can be obtained.

Another useful equation is the so-called **Takács' equation** for the Laplace-Stieljes transform $G^*(s)$ of the busy period B. Takács proved that

$$G^*(s) = B^*(s + \lambda - \lambda G^*(s)).$$

By differentiating this equation with respect to s and as s tends to zero we have

$$E(B) = \frac{E(S)}{1-\rho},$$

$$E(B^2) = \frac{E(S^3)}{(1-\rho)^3}.$$

5.2. Present Contributors

To the best of my knowledge the following universities are the centre of education and research on queueing related topics

- Eötvös Loránd University (A. Benczúr, L. Lakatos, L. Szeidl)
- Budapest University of Technology and Economics (L. Györfi, M. Telek, S. Molnár)
- University of Debrecen (J. Tomkó, M. Arató, B. Almási, A. Kuki, J. Sztrik, and several PhD students)

At Debrecen University queueing theory is a mandatory part of BSc curriculum in software engineering, computer engineering and business information management. It is optional part of MSc in software engineering and the Doctoral School of Informatics has a programme in this subject. Several printed and digital lecture notes help students and web-based Java applets can be run to get the main performance measures of the systems. Please visit to following link

`http://irh.inf.unideb.hu/user/jsztrik/`

6. The $\vec{G}/M/r//N/FCFS$ system

Over the years I have been published around 140 papers on topics related to queueing systems. In the following I would like to introduce the results of my paper *On the finite-source $\vec{G}/M/r/$ queue* [28] to which I have received 22 citations.

Requests arrive from a finite source of size N and are served by one of $r(r \leq N)$ servers at a service facility according to a First-Come-First-Served (FCFS) discipline. The service times of the requests are supposed to be identically and exponentially distributed random variables with means $1/\mu$. After completing service, request i returns to the source and stays there for a random time having general distribution function $F_i(x)$ with density $f_i(x)$. All of these random variables are assumed to be independent of each other.

6.1. The mathematical model

Let the random variable $v(t)$ denote the number of requests staying in the source at time t and $(\alpha_1(t), \ldots, \alpha_{v(t)}(t))$ indicate their indices ordered lexicographically. Let us denote by $(\beta_1(t), \ldots, \beta_{N-v(t)}(t))$ the indices of the requests waiting for the service facility in the order of their arrival. Clearly the sets $\{\alpha_1(t), \ldots, \alpha_{v(t)}(t)\}$ and $\{\beta_1(t), \ldots, \beta_{N-v(t)}(t)\}$ are disjoint.

Introduce the process

$$\underline{Y}(t) = (\alpha_1(t), \ldots, \alpha_{v(t)}(t); \beta_1(t), \ldots, \beta_{N-v(t)}(t)).$$

The stochastic process $(\underline{Y}(t), t \geq 0)$ is not Markovian unless the distribution functions $F_i(x)$ are exponential, $i = 1, \ldots, N$.

Let us also introduce the supplementary variables $\xi_{\alpha_l(t)}$ to denote the random time that request $\alpha_l(t)$ has been spending in the source until time t, $l = 1, \ldots, N$. Define

$$\underline{X}(t) = (\alpha_1(t), \ldots, \alpha_{v(t)}(t); \xi_{\alpha_1(t)}, \ldots, \xi_{\alpha_{v(t)}}(t); \beta_1(t), \ldots, \beta_{N-v(t)}(t)).$$

Then process $(\underline{X}(t), t \geq 0)$ exhibits the Markov property.

Let V_k^N and C_k^N denote the set of all variations and combinations of order k of the integers $1, 2, \ldots, N$ respectively, ordered lexicographically. Then the state space of the process $\underline{X}(t)$ consist of the sets

$$(i_1, \ldots, i_k; x_1, \ldots, x_k; j_1, \ldots, j_{N-k}), \qquad (i_1, \ldots, i_k) \in C_k^N,$$
$$(j_1, \ldots, j_{N-k}) \in V_{N-k}^n, \qquad x_i \in \mathbf{R}_+, \qquad i = 1, \ldots, k, \quad k = 0, \ldots, N$$

Let $Q_{i_1,\ldots,i_k;j_1,\ldots,j_{N-k}}(x_1, \ldots, x_k; t)$ denote the probability that at time t the process is in state $(i_1, \ldots, i_k; x_1, \ldots, x_k; j_1, \ldots, j_{N-k})$ if k requests with indices $(i_1, \ldots, i_k)$ have been staying in the source for times $(x_1, \ldots, x_k)$, respectively, while the rest need service and their indices in order of arrival are $j_1, \ldots, j_{N-k})$.

Let λ_i defined by $1/\lambda_i = \int_0^\infty x dF_i(x)$. Then we have:

Theorem 6.1. *If* $1/\lambda_i < \infty, i = 1, \ldots, N$, *then the process* $(\underline{X}(t), t \geq 0)$ *possesses a unique limiting (stationary) ergodic distribution independent of the initial conditions, namely*

$$Q_{0;j_1,\ldots,j_N} = \lim_{t\to\infty} Q_{0;j_1,\ldots,j_N}(t),$$
$$Q_{i_1,\ldots,i_k;j_1,\ldots,j_{N-k}}(x_1,\ldots,x_k) = \lim_{t\to\infty} Q_{i_1,\ldots,i_k;j_1,\ldots,j_{N-k}}(x_1,\ldots,x_k;t). \tag{6.1}$$

Notice that $\underline{X}(t)$ belongs to the class of piecewise-linear Markov processes, subject to discontinuous changes treated by [14] in detail. Our statement follows from the theorem on page 211 of that monograph.

Let $Q_{i_1,\ldots,i_k,j_1,\ldots,j_{N-k}}$ denote the steady state probability that requests with indices $(i_1,\ldots,i_k)$ are in the source and the order of arrival of the rest to the service facility is $(j_1,\ldots,j_{N-k})$. Furthermore, denote by $Q_{i_1,\ldots,i_k}$ the steady state probability that requests with indices $(i_1,\ldots,i_k)$ are staying at the source.
As it was proved in [28] that these probabilities can be expressed in the following form

$$Q_{i_1,\ldots,i_k} = \frac{(N-k)!}{r!r^{N-r-k}\mu^{N-k}\lambda_{i_1},\ldots,\lambda_{i_k}} C_N, \tag{6.2}$$
$$(i_1,\ldots,i_k) \in C_k^N,\ k = 0,1,\ldots,N-r.$$

Similarly,

$$Q_{i_1,\ldots,i_k} = \frac{1}{\mu^{N-k}\lambda_{i_1}\ldots\lambda_{i_k}} C_N \tag{6.3}$$
$$(i_1,\ldots,i_k) \in C_k^N,\ k = N-r,\ldots,N. \tag{6.4}$$

Let $\hat{Q}_k$ and $\hat{P}_l$ denote the steady state probabilities that k requests are staying in the source and l requests are at the service facility, respectively. Clearly

$$Q_{i_1,\ldots,i_N} = Q_{1,\ldots,N} = \hat{Q}_N = \hat{P}_0 \qquad \hat{Q}_k = \hat{P}_{N-k}.$$

It is easy to see that

$$C_n = \hat{Q}_n\lambda_1\ldots\lambda_n \qquad \text{and} \qquad \hat{Q}_k = \sum_{(i_1,\ldots,i_k)\in C_k^N} Q_{i_1,\ldots,i_k},$$

where $\hat{Q}_N$ can be obtained with the aid of the norming condition

$$\sum_{k=0}^{N} \hat{Q}_k = 1.$$

In the homogeneous case, when $\lambda_i = \lambda, \qquad i = 1,\ldots,N$ relations (6.2) and (6.1) yield

$$\hat{Q}_k = \frac{N!}{k!r!r^{N-r-k}}\left(\frac{\lambda}{\mu}\right)^{N-k} \hat{Q}_N, \qquad \text{for } 0 \leq k \leq N-r,$$

$$\hat{Q}_k = \binom{N}{k}\left(\frac{\lambda}{\mu}\right)^{N-k}\hat{Q}_n, \qquad \text{for } N-r \le k \le N.$$

Thus, the probability that k requests are not in the source is

$$\hat{P}_k = \binom{N}{k}\left(\frac{\lambda}{\mu}\right)^{k}\hat{P}_0, \qquad \text{for } 0 \le k \le r,$$

$$\hat{P}_k = \frac{N!}{(N-k)!r!r^{k-r}}\left(\frac{\lambda}{\mu}\right)^{k}\hat{P}_0, \qquad \text{for } r \le k \le N.$$

Before determining the main characteristics of the system we need one more theorem. In order to formulate it, we introduce some further notations. Let $Q^{(i)}(P^{(i)})$ denote the steady state probability that request i is in the source (at the service facility) for $i = 1, \ldots, N$. It is clear that the process $(\underline{Y}(t), t \ge 0)$ is a Markov-regenerative process with state space

$$\bigcup_{\substack{(i_1,\ldots,i_k)\in C_k^N,\quad (j_1,\ldots,j_{N-k})\in V_{N-k}^N,\\ (i_1,\ldots,i_k)\cap(j_1,\ldots,j_{N-k})=0,\\ k=0,1,\ldots,N}} \{(i_1,\ldots,i_k;j_1,\ldots,j_{N-k})\}.$$

Let H_i be the event that request i is in the source and $Z_{H_i}(t)$ its characteristic function, that is

$$Z_{H_i}(t) = \begin{cases} 1 & \text{if } \underline{Y}(t) \in H_i \\ 0 & \text{otherwise} \end{cases}$$

Then we have

Theorem 6.2.

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T Z_{H_i}(t)dt = \frac{1/\lambda_i}{1/\lambda_i + \overline{W}_i + 1/\mu} = Q^{(i)} = 1 - P^{(i)},$$

where $\overline{W}_i$ denotes the mean waiting time of request i.

The statement is a special case of a theorem concerning the expected sojourn time for semi-Markov processes, (see [34]).

Sometimes we need the long-run fraction of time the request i spends in the source. This happens e.g., in the *machine interference model*. In that case for the utilization of machine i we have

$$U_i = Q^{(i)} = \sum_{k=1}^{n}\sum_{i\in(i_1,\ldots,i_k)\in C_k^N} Q_{i_1,\ldots,i_k}.$$

6.2. The main performance measures

(i) **Utilizations**

Utilizations can now be considered for individual servers or for the system as a whole. The process $(\underline{X}(t), t \geq 0)$ is assumed to be in equilibrium. Considering the system as the whole, it will be empty only when there are no requests at the service facility and will be busy at other times. As usual, using renewal-theoretic arguments for the system utilization, that is the long-run fraction of time when at least one server is busy, we have

$$U = 1 - \hat{Q}_N \qquad \text{and} \qquad \hat{Q}_N = \frac{E_{\eta^*}}{E_{\eta^*} + E\delta}$$

where $\eta^* = min(\eta_1, \ldots, \eta_N)$, random variable η_i denotes the source time of request i, $i = 1, \ldots, N$, and $N\delta$ denotes the average busy period of the system.

Thus the expected length of the busy period is given by

$$E\delta = E_{\eta^*} \frac{1 - \hat{Q}_N}{\hat{Q}_N}.$$

In particular, if $F_i(x) = 1 - exp(-\lambda_i x),\ i = 1, \ldots, N$, we get

$$E\delta = \frac{1 - \hat{Q}_N}{\hat{Q}_N} \frac{1}{\sum \lambda_i}.$$

It is also easy to see that for the utilization of a given server, which is called utilization in general, the following relation holds:

$$U_s = \frac{1}{r}\left(\sum_{k=1}^{N} k\hat{P}_k + r \sum_{k=r+1}^{N} \hat{P}_k \right) = \frac{\overline{r}}{r},$$

where $\overline{r}$ denotes the mean number of busy servers.

(ii) **Mean waiting times**

By the virtue of Theorem 6.1 we obtain $Q^{(i)} = (1 + \lambda_i \overline{W}_i + \lambda_i/\mu)^{-1}$. Consequently, the average waiting time of request i is

$$\overline{W}_i = (1 - Q^{(i)})(\lambda_i Q^{(i)})^{-1} - 1/\mu.$$

It follows that the mean sojourn time of request i, that is, the sum of waiting and service times , can be obtained by

$$\overline{T}_i = \overline{W}_i + 1/\mu = (1 - Q^{(i)})(\lambda_i Q^{(i)})^{-1}, \qquad \text{for } i = 1, \ldots, N. \tag{6.5}$$

Since $\sum_{i=1}^{N}(1 - Q^{(i)}) = \overline{N}$, where $\overline{N}$ denotes the mean number of requests staying at the service facility we have, by reordering and adding (6.5)

$$\sum_{i=1}^{N} \lambda_i \overline{T}_i Q^{(i)} = \overline{N}. \tag{6.6}$$

This is the **Little's formula** for the finite source $\vec{G}/M/r$ queue. In particular, if $F_i(x) = F(x),\ i = 1, \ldots, N$, (6.6) can be written as $\lambda(N - \overline{N})\overline{T} = \overline{N}$, where $(N - \overline{N})$ is the expected number of requests staying in the source.

References

[1] Allen, A.: *Probability, statistics, and queueing theory : with computer science applications,* Academic Press, 1990

[2] Artalejo, J.R – Gomez-Corral, A.: *Retrial queueing systems : a computational approach,* Springer, 2008

[3] Bingham, N.H.: The work of Lajos Takács on Probability Theory *Journal of Applied Probability* 31A(1994) 29-39

[4] Bolch, G. – Greiner, S. – de Meer, H. – Trivedi, K.S.: *Queueing networks and Markov chains : modeling and performance evaluation with computer science applications,* John Wiley and Sons, 1998, 2006

[5] Cooper, R.: *Introduction to queueing theory,* CEEPress Books, 1990

[6] Daigle, J.: *Queuing theory for telecommunications,* Addison-Wesley Publisher, 1992

[7] Dshalalow, J.H. – Syski, R.: Lajos Takács and his work, *Journal of Applied Mathematics and Stochastic Analysis* 7(1994) 215-237

[8] Dshalalow, J.: *Advances in queueing : theory, methods, and open problems,* CRC Press, 1995

[9] Dshalalow, J.: *Frontiers in queueing : models and applications in science and engineering ,* CRC Press, 1996

[10] Erlang, A.K.: The Theory of Probabilities and Telephone Conversations, *Nyt Tidsskrift for Matematik B* 20(1909)

[11] Erlang, A.K.: Solutions of Some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges, *Post Office Electrical Engineering Journal* 10(1917) 189-197

[12] Galambos, J. – Gani, J.: Studies in Applied Probability, Papers in honour of Lajos Takács, *Journal of Applied Probability* 31A(1994)

[13] Giambene, G.: *Queuing theory and telecommunications : networks and applications*, Springer, 2005

[14] Gnedenko, B.V. – Kovalenko, I.N.: *Introduction to Queueing Theory,* Birkhauser, 1968, 1989

[15] Gross, D. – Shortle, J.F. – Thompson, J.M. – Harris, C.M.: *Fundamentals of Queueing Theory*, Wiley, 2008

[16] Györfi, L.: *Tömegkiszolgálás informatikai rendszerekben*, Műegyetemi Kiadó, 1996, 2007

[17] Haghighi, A.M.: *Queueing models in industry and business,* Nova Science Publisher, 2008

[18] Haribaskaran, G.: *Probability, queueing theory and reliability engineering,* Laxmi, 2006

[19] Jewel, W.S.: A Simple Proof of $L = \lambda W$, *Operations Research* 15(1967) 1109-1116

[20] Khintchine, A.Y.: *Mathematical Methods in the Theory of Queueing*, Griffin, 1960

[21] Kleinrock, L.: *Queueing Systems I-II,* John Wiley, 1975,1976

[22] Kleinrock, L.: *Sorbanállás, kiszolgálás : Bevezetés a tömegkiszolgálási rendszerek elméletébe*, Műszaki Kiadó, 1979

[23] Kobayashi, H. – Mark, B.L.: *System Modeling and Analysis*, Pearson International Edition, 2009

[24] Lakatos, L. – Szeidl, L. – Telek, M.: Tömegkiszolgálás, *Informatikai Algoritmusok, Edited by Iványi, A. ELTE Eötvös Kiadó* 1298-1347, 2005

[25] Little, J.D.C: A Proof for the Queuing Formula $L = \lambda W$, *Operations Research* 9(1961) 383-387

[26] Ramalhoto, M.F. – Amaral, J.A. – Cochito, M.T.: A survey of J. Little's formula, *International Statistical Review* 51(1983) 255-278

[27] Stidham, S.: A Last Word on $L = \lambda W$, *Operations Research* 22(1974) 417-421

[28] Sztrik, J.: On the finite-source $\vec{G}/M/r$ queu, *European Journal of Operational Research* 20(1985), 261-268

[29] Sztrik, J.: *Bevezetés a sorbanállási elméletbe és alkalmazásaiba,* Debreceni Egyetemi Kiadó, 1994

[30] Takagi, H.: *Queuing analysis : a foundation of performance evaluation, I-III*, North-Holland, 1991-1993

[31] Takács, L.: *Stochastic Processes, Problems and Solutions* John Wiley and Sons, 1960

[32] Takács, L.: *Introduction to the Theory of Queues,* Oxford University Press, 1962

[33] Takács, L.: *Combinatorial Methods in the Theory of Stochastic Processes,* John Wiley and Sons, 1977

[34] Tomkó, J.: On sojourn times for semi-Markov processes, *Proceedings of 14th Meeting of Statisticians, Poland,* 399-402, 1981

[35] Trivedi, K.S.: *Probability and statistics with reliability, queuing, and computer scence applications,* Prentice-Hall, 1982, 2002

[36] Wolf, R.W.: *Stochastic Modeling and the Theory of Queues,* Prentice-Hall, 1989

[37] `http://web2.uwindsor.ca/math/hlynka/queue.html`

[38] `http://en.wikipedia.org/wiki/Lajos Takács`

[39] `http://irh.inf.unideb.hu/user/jsztrik/`

[40] `http://irh.inf.unideb.hu/user/jsztrik/education/09/index.html`

[41] `http://irh.inf.unideb.hu/user/jsztrik/education/05/index.html`