From Computer Science to Service Science: Queues with Human Customers and Servers

Hideaki Takagi
Faculty of Engineering, Information and Systems
University of Tsukuba
Tsukuba Science City, Ibaraki 305-8573, Japan
takagi@sk.tsukuba.ac.jp

March 17, 2014

Abstract

Professor Leonard Kleinrock has made key contributions during the initial stage of development in computer communication networks from several aspects. One of them is the ingenious application of queueing theory to the performance evaluation of communication networks. Queueing theory is still useful in contemporary service systems having human customers and servers as precious resources. However, some new theoretical development is needed to cope with human service systems. In this article, the potential of queueing theory is discussed in the scope of emerging *service science*. We begin with a snapshot of Professor Kleinrock's laboratory in the early 1980s. We then review the performance metric called *power* worked out by Kleinrock as it determines the optimal input rate in a service system. As an example of service science for healthcare, we show modeling of obstetric patient flow in a hospital by means of Little's law and a network of M/M/m and M/G/∞ queues.

Keywords and Phrases: Queueing theory, performance evaluation, service science, perception of customers, satisfaction of servers, "power", patient flow in hospital, Little's law, queueing network, time-varying arrival rate.

1 Personal Retrospective and View of Queueing Theory

My career as a researcher of computer and communication science started when I received the Ph.D. in Computer Science from the Computer Science Department of the University of California, Los Angeles (UCLA), in May 1983 [23]. My dissertation was supervised by Professor Leonard Kleinrock, who was then the principal investigator (PI) of the Contract MDA 903-82-C-0064 from the Defense Advanced Research Projects Agency

(DARPA) of the Department of Defense. My doctoral committee consisted of Professors Mario Gerla, James R. Jackson, Steven A. Lippman, Richard R. Muntz, and Leonard Kleinrock (Committee Chair). Two volumes of *Queueing Systems* [13, 14] were textbooks in Professor Kleinrock's classes "CS212A Queueing Systems: Theory and Applications" and "CS212C Computer Communications Network". Later the solutions manuals were published coauthored by Kleinrock and Richard Gail [18, 19, 20].

I spent a great time with Professor Kleinrock's students during that period including Mart Molle, Randy Nelson, Richard Gail, Ken Kung, Hanoch Levy, Yehuda Afek, Fathi Belgith, and Joe Green. I also talked a lot with other professors' students like Mike Molloy, Bruce Walker, Paul Hurley, Edmundo de Souza e Silva, Mary Vernon, Kin Leung, Frank Schaffa, and Luis Filipe de Moraes. All of them later became outstanding figures in academia and industry across the world. Several ex-students such as Ed Coffman, Simon Lam, Fouad Tobagi, Farouk Kamoun, Parviz Kermani, John Silvester, and Yechiam Yemini occasionally dropped by at the weekly students meeting in Professor Kleinrock's office and talked to younger *academic brothers* in a very friendly fashion. I sat in classes of Professors Richard R. Muntz, Mario Gerla, Walter J. Karplus, Wesley W. Chu, and Izhak Rubin (Systems Science) and met many distinguished visitors such as Vinton Cerf, Wim Cohen, Bob Kahn, Hisashi Kobayashi, Alan Konheim, and Steve Lavenberg.

Cheerful support staff members Lillian Larijani, George Ann Hornor, Brenda Ramsey, Verra Morgan, Ruth Pordy, and Terry Peters helped students in administration as well as in preparation for research papers, because typewriting and figure drawing were done manually at that time.

On June 9 and 10, 1994, a two-day symposium entitled *Experts on Networks* was held at the Sequoia room of the UCLA Faculty Center in order to honor Professor Leonard Kleinrock and his contributions to computer science on his 60th birthday and over 30 years at UCLA. On that occasion, a genealogical tree of his students, students of students, and so on, was compiled by Bob Felderman. According to the tree, I am the 21st student amongst those 35 students (at that time) who received doctoral degrees under the supervision of Professor Kleinrock, the first of whom being Ed Coffman. On June 11, participants in the workshop were invited by Mrs. Stella Schuler Kleinrock to the party at Professor Kleinrock's house in the Brentwood area of western Los Angeles.

With such background, I worked on the application of queueing theory to computers and communication networks during my most active research life. However, the use of analytic studies has soon become less important in these fields. I think there are two reasons for this decline. First, the rapid technological development, referred to as *Moore's law*, has brought powerful CPU and abundant memory for computers as well as virtually unlimited bandwidth of optical fiber for telecommunication networks so cheaply that we do not have to concern ourselves with the wise (and stingy) use of these resources any longer. Second, the protocol and control of the operation in these systems have become so complicated and interdependent that theoretical treatment, often based on simplifying fictitious assumptions such as independence of events and exponential distribution, cannot cope with the operation of real systems. Thus it is no wonder that analytic methods based individually on the human brain power have been replaced by algorithmic and event-driven simulation methods which can capitalize on the ever-growing computer power.

My recent interest is the mathematical and statistical treatment of service systems with human customers and servers, for example, queues in airports, call centers, hospitals, etc., in the context of an emerging discipline called *service science*. Notice that human beings still remain to be a precious resource that may not be abused in service systems. In most countries, whether developed or under development, the service sector of industry now takes a predominant portion of the national economy in terms of gross domestic product (GDP) as well as the population share of labor force. Unlike manufacturing, however, not much scientific approach has been exploited so far for the increase of productivity and promotion of innovation in the service industry. The *services science* was advocated as a new academic discipline in the so-called *Palmisano Report* from the Council on Competitiveness in the United States published in December 2004 [2].

According to my view, the study of service systems with human customers and servers is one of a few fields in which the queueing theory can still make prominent impacts on the practical side (another potential field may be the radio communication technology where the channel bandwidth continues to be a physically limited resource [26]). Application of queueing theory to human service systems is not new at all. It was only that the exclusively driving application was computers and communication networks from the 1960s to the 1980s. Science of service is the field of study in the 21st century on top of operations research, data science and computer science. I hope that the basis of our knowledge which was cultivated under the leadership of Professor Kleinrock enables us to pave the way for this new affluent area of research. Some books address the queueing theory for service systems [3, 11, 24].

The rest of this article is organized as follows. In Section 2, we characterize service systems involving human customers and servers by highlighting the difference from those systems in which the service is provided by tireless machines. In Section 3, we review the performance metric named *power* which I think Professor Kleinrock favored, and we confirm one of his conjectures numerically. In Section 4, as an example of queueing theory application to service science, we show a preliminary piece of work on the modeling of obstetric patient flow in a hospital by means of Little's law and a network of M/M/m and M/G/∞ queues. We conclude in Section 5 by discussing several features of queueing models for human service systems that challenge the queueing theorists.

2 Service Systems with Human Servers and Facilities

Service can be defined as the activity to bring value (satisfaction) to not only the recipients (customers) but also the providers (employees) by optimal management of a set of available resources. The service industry is so diverse. Thus the challenge of *service science* is to create the principles and architecture of service operation and to implement it in the system just as the computer science has developed the architecture of computer operation and implemented it in the system since the 1950s. Some service systems do not involve human servers in real time such as reservation of a hotel room, purchase of goods, money transfer etc. over the Internet and the automatic check-in processing in the airport. However, there are still many service systems in which human servers play

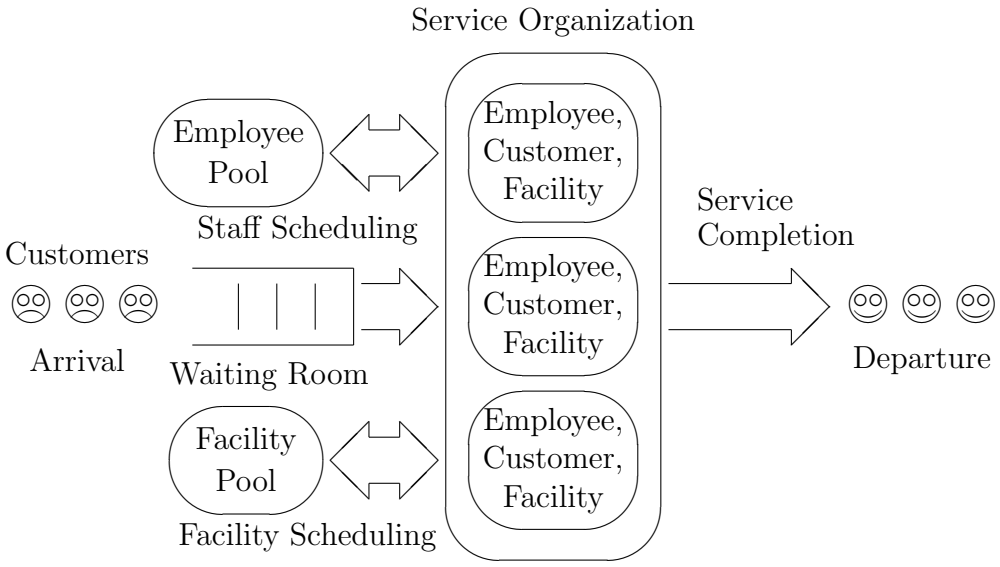

Figure 1: Generic service system involving human servers and facilities.

a major and indispensable role.

Figure 1 shows a generic service system involving human servers and facilities. In a service organization, employees provide service by using some physical facilities as enablers. One of the characteristic features of service that differentiate "service" from "goods" is said to be the *intangibility*, meaning that the service is action or performance that cannot be touched physically in the same manner as goods can be sensed as objects [4, p.20]. However, our experience tells us that a comfortable physical facility and environment are essential for the customer satisfaction. Employees and facilities are the operational resources that are not free but require sizable investment to keep them available for customers. Therefore, the manager of service organization holds a finite number of employees and facilities in the pool to supply them when needed to administer the service.

Customers come to the service organization to get service. A certain set of resources (employees and facilities) must be allocated to each service for a certain duration of time. If the necessary resource is not available when customers arrive, they are forced to wait in the waiting room. The group of customers in the waiting room is called a *queue*. The space of a physical waiting room is usually finite. If there is no vacancy in the waiting room when a customer arrives, he is not admitted to the system. When the required set of employees and service facilities becomes available, one customer (usually at the head position) in the waiting room is selected for the next service. Customers who have completed the service depart from the system. Staff scheduling and facility scheduling are necessary in order to allocate the set of employees and service facilities to customers efficiently.

Many real service systems involving human servers and facilities fit in the model of Figure 1. A simple example is the checkout counter in a supermarket. In a coffee shop, when a customer orders a cup of coffee, an employee handles the order using the counter as a facility. When the customer drinks coffee, his table is the facility while there may be

no human servers afterwards as *self-service*. In a call/contact center, the operator is a server and the telephone line is the facility. Finally, in a hospital, patients are customers while medical equipments, surgery rooms and beds are facilities, while doctors, nurses, and technicians are servers. There are many kinds of service in the hospital such as examination, diagnosis, surgery, clinical treatment, and rehabilitation.

When the service system has sufficient resources (employees and facilities), the above-mentioned service is provided without delay. However, if the organization cannot supply enough resources to cope with the demand of customers, the system enters the state called *congestion*. In this state, as the availability of resources decreases, there is less chance to take in waiting customers for service. Thus there will be a long waiting line, which results in a long waiting time of customers leading to their dissatisfaction. Customers who arrive when the system is full are rejected, or they may balk if they see a long waiting line upon arrival. On the contrary, some people may elect to join a long waiting line at the restaurant in the hope that it means good food. Impatient customers in the waiting room may leave the system without getting service after having waited an unbearably long time. Those customers who have balked at joining the queue or who have abandoned while waiting may come back later. Then both new and retrial customers together rush to the system resulting in even higher arrival rate. In a call center, a customer who has waited a long time may complain before starting an intended conversation with an operator, which leads to an elongated holding time of a communication line and the operator.

We must pay attention to the job satisfaction of employees too. Employees who are tired of long working hours and processing many claims from customers may decrease the efficiency, morale, and quality of service, and eventually quit the job. The satisfaction of servers (employees) is an important feature of human service systems which we have not dealt with in queueing models for computers and communication networks.

3 "Power" of Kleinrock

The manager of a service system is happy if his resource (servers) is fully utilized by accepting as many customers as possible. However, the waiting time of customers increases (customers feel unhappy) if the system admits too many customers. On the other hand, if the system is kept empty, arriving customers should be satisfied because they do not have to wait. Then, however, the manager may not be happy because his resource is wasted without customers. See Table 1 for the trade-off of resource utilization and customer satisfaction in a system modeled with an M/M/m queue. Therefore it is a good idea to control the customer arriving process so that the system may be neither too crowded nor too empty. The situation is similar whether the system is a computer network, a highway road, or a checkout counter of supermarket. Then, what is the optimal level of controlling the customer arrivals?

In order to answer this question mathematically, a metric of system performance called *power* was proposed and studied. Although the notion of power had been introduced earlier by others [7, 27], I believe it is Professor Kleinrock who exploited its significance intensively and extensively [15, 16, 17]. The power was later studied in detail

Table 1: Trade-off of resource utilization and customer satisfaction in the M/M/m queue.

Arrival Rate	Server Utilization	Waiting Time
$\lambda \to 0$	0 (waste of resource)	0 (customer satisfaction)
$\lambda \to m\mu$	1 (efficient use of resource)	∞ (customer dissatisfaction)

in the dissertations of his students Harry Richard Gail, Jr. [5, 6] and Jau-Hsiung Huang [12].

Suppose that the system is in a stable state, which means that the input rate equals the output rate (*throughput*). Given the input rate λ, the power is defined as the ratio of the throughput to the *mean response time* $T(\lambda)$:

$$P(\lambda) := \frac{\lambda}{T(\lambda)}. \tag{1}$$

It is the idea of power that the optimal operating point of the system is determined as the value of λ that makes the power maximum. Yoshioka et al. [27] discuss the analogy to the optimization of an electric circuit.

For example, in the M/M/1 with arrival rate λ and service rate μ, we have [13, p.98]

$$T(\lambda) = \frac{1}{\mu - \lambda},$$

where we assume $\lambda < \mu$ for stability. It follows that

$$P(\lambda) = \lambda(\mu - \lambda),$$

which is a quadratic function in λ that has the maximum at $\lambda = \mu/2$. Very interestingly, this value of λ makes the mean number of customers in the whole system just unity:

$$E[N] = \lambda E[T] = \frac{\lambda}{\mu - \lambda} = 1.$$

That is to say, there should be exactly one customer in the system (who is being served) on average at the optimal operating point.

We get the same result for the M/G/1 queue which has [13, p.190]

$$E[T] = \frac{E[N]}{\lambda} = b\left[1 + \frac{\rho(1 + C_b^2)}{2(1 - \rho)}\right], \tag{2}$$

with $\rho = \lambda b$, where b is the mean service time and C_b^2 is the squared coefficient of variation of the service time. Then we have

$$P(\lambda) = \frac{1}{b^2}\left[\frac{2\rho(1 - \rho)}{2 - (1 - C_b^2)\rho}\right],$$

which becomes maximum at

$$\rho = \frac{1}{1 + \sqrt{(1 + C_b^2)/2}}$$

This value of ρ again makes $E[N] = 1$ [6, 16].

Generally speaking, by differentiating $P(\lambda)$ in Eq. (1) with respect to λ, we can show that the maximum power occurs at the value of λ that satisfies the relation

$$\frac{dT(\lambda)}{d\lambda} = \frac{T(\lambda)}{\lambda}. \tag{3}$$

This means that the power becomes maximum at the value of λ such that the straight line through the origin in the $(\lambda, T(\lambda))$ plane is tangent to the $T(\lambda)$ curve, or at the *knee*, so to speak, of the $T(\lambda)$ curve [6]. This situation is plotted in Figure 2 for the M/M/1 queue.

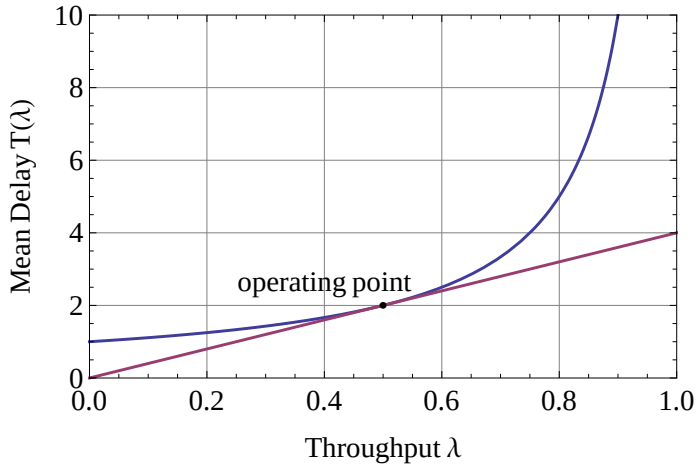

Figure 2: Determination of the optimal throughput in the M/M/1 queue.

In order to study the power for the M/M/m queue, where m denotes the number of servers, let us redefine the power in terms of the physically non-dimensional quantities $\rho := \lambda/\mu$ and $\mu E[T]$ by

$$P(m, \rho) := \frac{\rho/m}{\mu E[T]}. \tag{4}$$

For $m = 1$, we have

$$P(1, \rho) = \rho(1 - \rho),$$

which becomes maximum at $\rho = \frac{1}{2}$ that makes $E[N] = 1$.

For the M/M/m queue, the redefined power is given by

$$P(m, \rho) = \frac{\rho}{m} \Big/ \left[1 + \frac{C(m, \rho)}{m - \rho}\right], \tag{5}$$

where

$$C(m, \rho) = \frac{\dfrac{\rho^m}{m!}}{\left(1 - \dfrac{\rho}{m}\right) \displaystyle\sum_{k=0}^{m-1} \frac{\rho^k}{k!} + \frac{\rho^m}{m!}} \tag{6}$$

is the famous *Erlang's C-formula* for the probability that a customer waits upon arrival [13, p.103]. We also note that

$$E[T] = \frac{E[N]}{\lambda} = \frac{1}{\mu}\left[1 + \frac{C(m,\rho)}{m-\rho}\right]. \tag{7}$$

For $m = 2$, we have

$$P(2,\rho) = \frac{\rho}{8}(4-\rho^2),$$

which becomes maximum at $\rho = 2/\sqrt{3}$. Then we have $E[N] = \sqrt{3} = 1.732\cdots$. In Figure 3, we plot $P(m,\rho)$ for various values of m. By numerical computation, we find the value of ρ, denoted by ρ^*, that maximizes $P(m,\rho)$ for each m, and calculate the mean number of customers in the system $E[N]$ when $\rho = \rho^*$. The results are shown in Table 2. As $m \to \infty$, we observe that

$$\lim_{m\to\infty}\frac{\rho^*}{m} = 1 \quad ; \quad \lim_{m\to\infty}\frac{E[N^*]}{m} = 1 \quad ; \quad \lim_{m\to\infty} P(m,\rho^*) = 1. \tag{8}$$

Thus we can say again that there should be exactly one customer per server on average at the optimal operating point as $m \to \infty$ in the M/M/m queue, although this conclusion has not been derived theoretically.

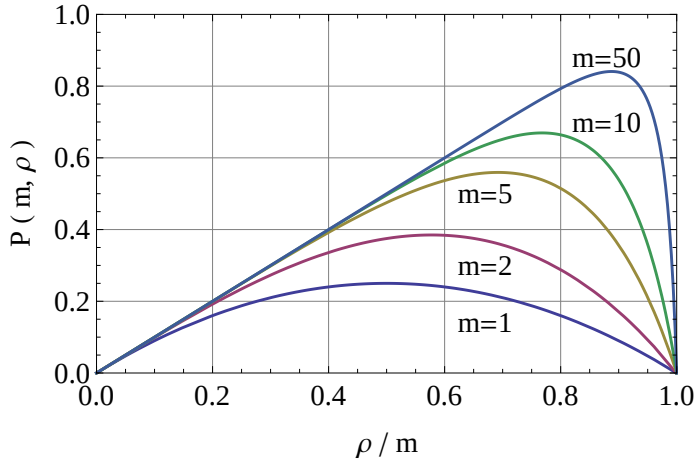

Figure 3: Kleinrock's "power" for the M/M/m queue.

Table 2: Optimal load for maximizing the power in the M/M/m queue.

m	1	2	5	10	50	100	500	1000
ρ^*/m	0.5	0.5774	0.6921	0.7678	0.8408	0.9191	0.9630	0.9737
$E[N^*]/m$	1	0.8660	0.5594	0.8806	0.9371	0.9541	0.9786	0.9847
$P(m,\rho^*)$	0.25	0.3849	0.5594	0.6695	0.8408	0.8855	0.9477	0.9628

4 Patient Flow in a Hospital

As an example of queues with human servers and facilities, we show a queueing network model for the patient flow in a hospital. Basic performance measures in hospital management include the bed utilization and the *length-of-stay* (LOS) of each inpatient. The bed utilization is calculated as the ratio of the mean number of patients staying in the entire hospital or in each clinical ward to the total number of beds that the facility has. Little's law of queueing theory relates the mean number of patients present there to the patient admission rate and the mean LOS. These are simple calculations. Taking one step forward, we propose an approximate queueing network model for the patient flow. Our model makes it possible to predict the probability distribution for the number of patients staying in each ward at the midnight census from the observed data of patient admission rate and the histogram for the duration of each hospitalization in that ward. A more detailed report of the present study is included in the Proceeding of the SRII Global Conference 2014 [25].

There is extensive work on applying the techniques of operations research, in particular that of queueing theory, to the patient flow in hospitals as surveyed in [8, 9, 22]. Application of Little's law is found in [21]. The patient flows in the obstetric and neonatal units are studied in [1, 10].

A complete data set of every movement of all the inpatients from room to room covering two years was provided us by the Medical Information Department of the University of Tsukuba Hospital (UTH) in Japan. The statistical treatment of these log data with the resulting publication of research findings in our project has been approved by the Institutional Review Board (IRB) of UTH. By focusing on the obstetric unit of the hospital in which patients are assumed to be hospitalized at random times, we have analyzed the patient flow by a queueing model. Upon admission, each obstetric patient is assigned to a bed in one of the two wards, one for high-risk delivery and the other for normal delivery, and then she may be transferred between the two wards before discharge.

We confirm Little's law for the flow of obstetric patients in each ward. Then we propose a network model of M/G/∞ and M/M/m queues to represent the transfer of patients over the two wards. Although our model is a very rough and simplistic approximation of the real patient flow, the predicted probability distribution is shown to be in good agreement with the observed data. Our method can be used for dimensioning the capacity of obstetric units if the patient demand is predicted.

4.1 Little's Law Applied to Inpatients

The obstetric unit of UTH is called the Center for Maternal, Fetal and Neonatal Health for treatment of normal as well as high-risk childbirth in the Tsukuba and Southern Ibaraki Prefecture areas. There are two wards, numbered 30M and 300, for the obstetric unit (The wards in UTH moved to a new site in December 2012. All the data in this paper refer to the statistics before this movement.):

- Ward 30M is the maternal and fetal intensive care unit (MFICU), which has 6 beds, for the treatment of high-risk delivery.

Table 3: Statistics on the number of obstetric patients in the University of Tsukuba Hospital during two fiscal years 2010–2011.

	Ward 300	Ward 30M
Number of beds	26	6
Total number of patient-days	12,630	3,600
Mean number of patients staying overnight	17.28	4.92
Bed utilization	66.5%	82.1%
Total number of admitted patients	1,963	338
Patients arrival rate per day	2.685	0.462
Mean length-of-stay (LOS) in days	6.43	10.65

- Ward 300 with 26 beds accommodates patients with normal delivery and also plays a role of backup (waiting room) for Ward 30M.

In Table 3, we show several statistics on the number of patients in the two wards of the obstetric unit during two fiscal years 2010–2011, more precisely, a period from the night of April 1, 2010, to the night of March 31, 2012, both inclusively. In this table, for a patient who was assigned to Ward 300 and then transferred to Ward 30M, we count one admission to Ward 300 and another admission to Ward 30M. For convenience' sake, we assume that all the patients who stayed on April 1, 2010 were admitted on that day, and that all the patients who stayed on March 31, 2012 were discharged on the next day. But these boundary effects are negligible. Also those patients who were admitted and discharged on the same day are counted in the total number of admitted patients but they do not contribute to the count of patient-days. Since the year 2012 was a leap year, there were 731 nights in the period of our study.

Let us make a few comments on the numbers in Table 3. For example, take Ward 300. The number of patients in bed is counted at 12:00 midnight on each night, the sum of which over the two years amounts to 12,630. Therefore, the mean number of patients staying in bed on each night is 12,630/731=17.28, which leads to the bed utilization 17.28/26=0.665. On the other hand, since 1,963 patients were admitted during these two years, the arrival rate was 1,963/731=2.685 patients/day. The mean LOS for each patient was 6.43 days. This may sound much longer than the mean LOS of obstetric patients in the United States. In Japan, however, a mother with normal delivery usually stays 5 days in the hospital after childbirth. Therefore, if a baby is born on the day of admission, the mother stays 5 nights. If a baby is born on the day following the admission, she stays 6 nights in the hospital. Mothers with abnormal delivery may stay a few more days in the hospital. A similar calculation for Ward 30M produces the numbers on the rightmost column in Table 3. The mean LOS in Ward 30M is longer than that in Ward 300, because mothers with high-risk delivery are accommodated in Ward 30M.

Let us confirm that the mean LOS of an arbitrary patient (W) can be calculated from the mean number of patients being hospitalized each day during a given observation period (L) and the number of new patients admitted per day in the same period (λ).

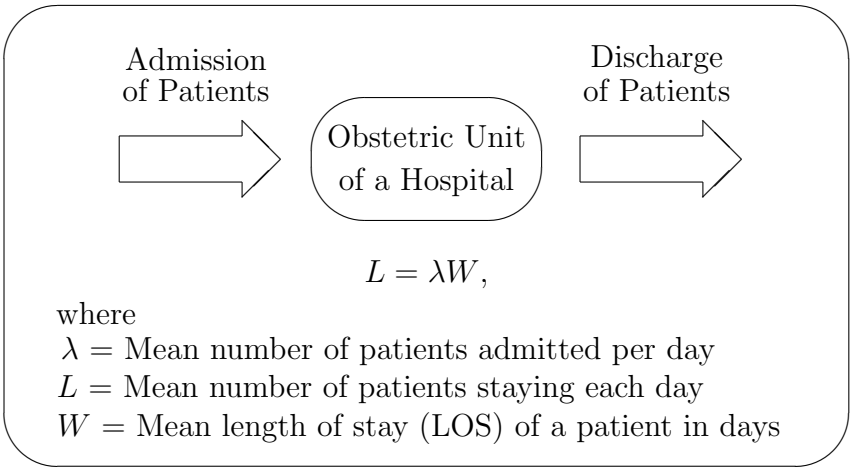

Figure 4: Little's law for obstetric patients in a hospital.

This is done by applying *Little's law* [13, p.17]

$$L = \lambda W \tag{9}$$

to the inpatients in a specified section of the (or the whole) hospital as shown in Figure 4. The only condition for this law is that the system is stable in that the number of patients present in the system does not grow indefinitely.

If we apply Little's law to the obstetric patients in Ward 300 during a period of years 2010–2011 shown in Table 3, we see that the arrival rate is $\lambda = 2.685$ patients/day and the mean number of patients is $L = 17.28$. Then we get the mean LOS as

$$W = L/\lambda = 17.28/2.685 = 6.436 \text{ days},$$

which virtually agrees with the reported value $W = 6.43$ days as it should. A similar calculation can be done for the patients in Ward 30M.

4.2 Queueing Network Model of the Obstetric Patient Flow

From the data analysis of patient flow records, we have found that a major portion of the obstetric patient flow consists of five routes shown in Figure 5. Patients on route 1 are supposed to be those who have normal childbirth. They are just admitted in Ward 300 and simply leave the hospital in about 6 days. Patients on routes 2 through 5 are supposed to be those who have a high-risk delivery. They are to be treated in Ward 30M. Upon arrival they are admitted to Ward 30M if beds are available there. After giving birth, they either leave the hospital (route 2) or move to Ward 300 possibly for after-birth treatment (route 4). If there are no beds available in Ward 30M when they arrive, they are temporarily accommodated in Ward 300 until beds become available in Ward 30M. Then they are transferred to Ward 30M. After giving birth, they either leave the hospital (route 3) or move back to Ward 300 possibly for after-birth treatment (route 5).

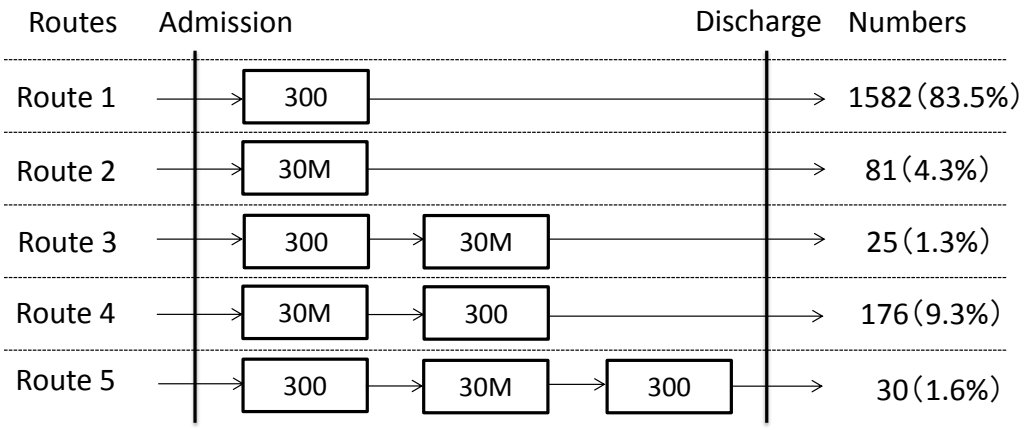

Figure 5: Dominating routes of obstetric patients.

Figure 6 shows a queueing network model we propose for the five dominant routes of obstetric patient flow by using M/G/∞ and M/M/m queues.

Our model is approximate from queueing-theoretic viewpoint in the following sense:

- The stochastic process for the patient flow under study is essentially a discrete-time system in which the patient arrivals and the length of stay are counted in days, while the M/G/∞ and M/M/m queues work in the continuous-time framework.
- The residence of patients in each ward is treated as independent while the entrance of route 5 patients twice into Ward 300 clearly violates this assumption.

Nevertheless, we propose a queueing network model of M/G/∞ and M/M/m systems because of its simplicity in modeling and computation in addition to our belief that the mathematical rigor should not be overly requested for practical purposes. The justification of our method is partly provided from the good agreement of the calculated results with the observations as demonstrated below.

In building a queueing network model of the obstetric patient flow, we pay special attention to routes 3 and 5 on which 55 patients were first admitted to Ward 300 and then transferred to Ward 30M during the two years of observation. There were 48 days on which admissions occurred to Ward 300, out of which Ward 30M was full on 41 days. Therefore we conjecture that Ward 300 was used by patients with high-risk delivery as the "waiting room" for Ward 30M. This conjecture was later confirmed to be the case by consulting with doctors of the UTH.

4.3 Distribution for the Number of Patients in Each Ward

Let us first consider the number of patients in Ward 300, which consist of the following four types:

- Patients going through Ward 300 only (route 1),

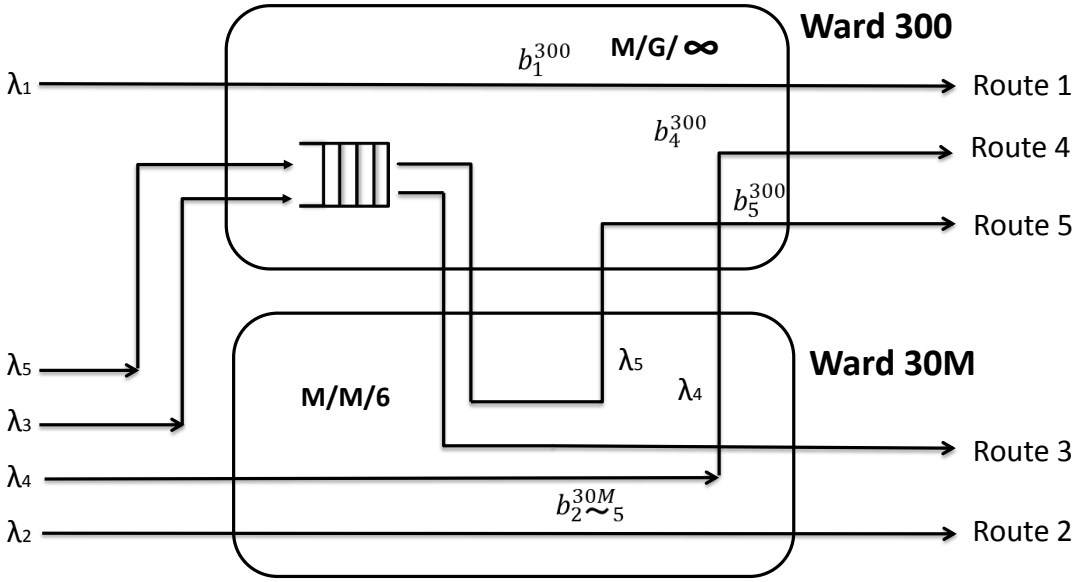

Figure 6: Queueing network model of the obstetric patient flow.

- Patients transferred from Ward 30M (route 4),
- Patients transferred from Ward 30M (route 5), and
- Patients waiting for Ward 30M (routes 3 and 5).

Therefore, we first obtain the probability distribution for the number of patients in Ward 300 for each type. We then obtain the probability distribution for the total number of patients in Ward 300 by the convolution of the four distributions.

We observe that the probability that all the beds are occupied is only about 1% for Ward 300 (see Figure 7 below). Therefore, we may assume that there are a sufficient number of beds in Ward 300 which can accept all patients at any time. Thus we will use an M/G/∞ queue to model the flow of patients on routes 1, 4, and 5 with respectively observed arrival rates and mean LOS. For patients on routes 3 and 5, we assume that the waiting room of an M/M/6 queue virtually exists in Ward 300, while the service facility of the M/M/6 queue exists in Ward 30M.

A model denoted by M/G/∞ is simply a system with sufficiently many servers to which customers arrive in a Poisson process and spend there a random amount of service time which is generally distributed probabilistically [11, p.145] and [13, p.234]. There is no contention for servers among customers. If λ denotes the arrival rate and b denotes the mean service time, the number N of customers present in the M/G/∞ system at an arbitrary time has the Poisson distribution with mean $\rho = \lambda b$:

$$P\{N = k\} = \frac{\rho^k}{k!} e^{-\rho} \qquad k \geq 0. \tag{10}$$

Note that this distribution depends only on the mean of the service time. A useful property in modeling is that the output of an M/G/∞ queue is a Poisson process. Another nice property about the Poisson process is that the superposition of independent Poisson processes forms another Poisson process with added rates.

Now the probability distributions for the number of route 1, 4, and 5 patients in Ward 300 are respectively given by

$$P_k^{(1)} = \frac{\left(\lambda_1 b_1^{300}\right)^k}{k!} e^{-\lambda_1 b_1^{300}} \quad k \geq 0 \quad ; \quad \lambda_1 = 2.164, \quad b_1^{300} = 6.521,$$

$$P_k^{(4)} = \frac{\left(\lambda_4 b_4^{300}\right)^k}{k!} e^{-\lambda_4 b_4^{300}} \quad k \geq 0 \quad ; \quad \lambda_4 = 0.241, \quad b_4^{300} = 7.125,$$

and

$$P_k^{(5)} = \frac{\left(\lambda_5 b_5^{300}\right)^k}{k!} e^{-\lambda_5 b_5^{300}} \quad k \geq 0 \quad ; \quad \lambda_5 = 0.041, \quad b_5^{300} = 6.700.$$

The patients on routes 3 and 5 in Ward 300 have the distribution for the number of customers *in the waiting room* of an M/M/m queue with $m = 6$, which is given by

$$P\{L^{(2-5)} = k\} = \begin{cases} 1 - \dfrac{\rho}{m} C(m,\rho) & k = 0, \\ C(m,\rho)\left(1 - \dfrac{\rho}{m}\right)\left(\dfrac{\rho}{m}\right)^k & k \geq 1, \end{cases} \tag{11}$$

where $\rho = \lambda_{2-5} b_{2-5}^{30\mathrm{M}}$, and $C(m,\rho)$ is given in Eq. (6). We use $\lambda_{2-5} = 0.427$ and $b_{2-5}^{30\mathrm{M}} = 10.452$ so that $\rho = 4.463$.

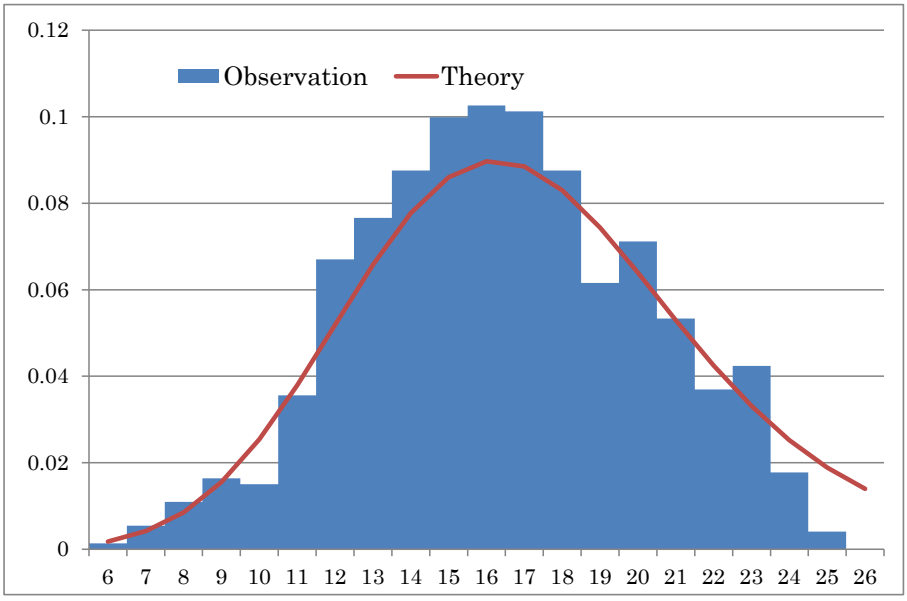

Figure 7: Theoretical versus observed values for the probability distribution of the number of patients in Ward 300.

Finally, the probability distribution for the total number of patients in Ward 300 is obtained by the convolution of the above-mentioned four distributions. Since the first three are independent Poisson distributions, their convolution results in another Poisson distribution with mean

$$\lambda_1 b_1^{300} + \lambda_4 b_4^{300} + \lambda_5 b_5^{300} = 16.103.$$

Therefore, we only have to calculate the convolution of this distribution with the fourth distribution $P\{L^{(2-5)} = k\}$ in Eq. (11). The result is shown in Figure 7, where we see fairly good agreement between the theory and observation except near the capacity 26 of Ward 300. The reason for this discrepancy is that the Poisson distribution has a support until ∞ while the observed distribution must be truncated at the capacity of Ward 300.

We next consider the number of patients in Ward 30M. Its probability distribution is obtained as that for the number of customers *in the service facility* of an M/M/m queue with $m = 6$. It is given by

$$P\{S^{(2-5)} = k\} = \begin{cases} P_0 \dfrac{\rho^k}{k!} & 0 \le k \le m-1, \\ C(m,\rho) & k = m, \end{cases} \tag{12}$$

where

$$\frac{1}{P_0} = \sum_{i=1}^{m-1} \frac{\rho^i}{i!} + \frac{\rho^m}{(m-1)!(m-\rho)} \tag{13}$$

with $\rho = 4.463$. The result is shown in Figure 8. While the theoretical result still captures major characteristics of the observed values, there remains a challenge for better agreement. This completes our modeling and analysis of the obstetric patient flow.

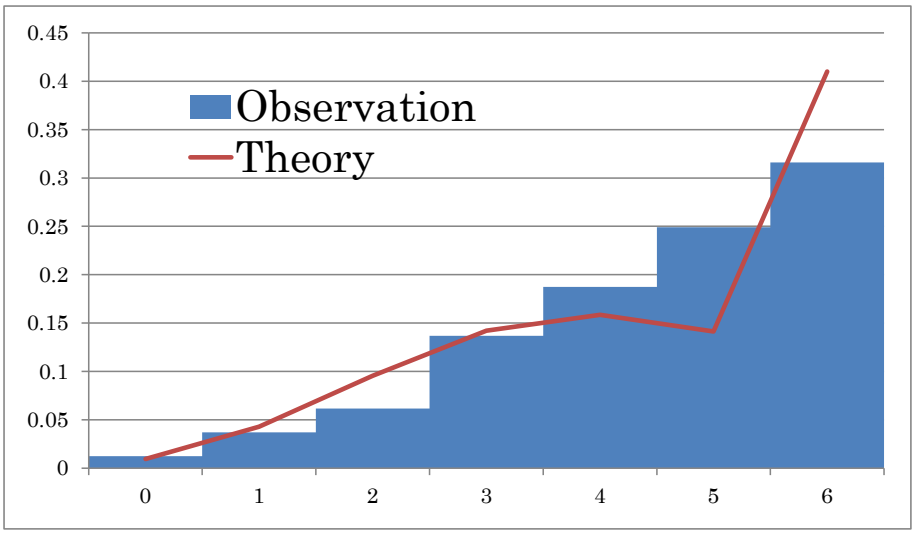

Figure 8: Theoretical versus observed values for the probability distribution of the number of patients in Ward 30M.

5 Research of Queues in Human Service Systems

In the latter half of the 20th century, the queueing theory was vigorously applied to the performance evaluation of computers, communication networks, and manufacturing systems, where the "customers" were jobs, messages, and goods while the "servers" were CPUs, communication lines, and manufacturing machineries, respectively. These servers can work 24 hours a day and 7 days a week without fatigue. At that time the goal of system design was to shorten the processing time by maximizing the efficiency with optimal scheduling of precious service resources. In the 21th century when we aim at the application of queueing theory to service systems with human customers and human servers, we must consider the *perception of customers* when they are waiting as well as before and after receiving the service. Waiting customers are concerned not only with the real waiting time but also with the fairness among the customers, comfort in the waiting room, and something useful to do while waiting. We must also care about the *satisfaction of servers* such as fair and balanced job assignment in the staff scheduling. The latter results from the empirical research that there is no customer satisfaction without employee satisfaction in human service systems. Finally, the manager of the system must be satisfied too.

Below we want to discuss several features of queueing models for human service systems that challenge the queueing theorists.

- Many Servers

  Professor Kleinrock compares the performance of queueing systems with various configurations of queues and servers [14, Section 5.1]. Clearly there is *economy of scale* in the service system with a single big server. An example is the centralized job processing in a company in the 1970s. A "main-frame computer" was installed in a computer room to process all jobs from the entire company. Another example is a "broadband communication network" in the 1980s where many copper cables were multiplexed into a single optical fiber cable for heavy intercity connection. However, the super high-speed service is neither possible with human servers nor liked by human customers. Thus a human service system usually provides a service facility with many servers with mediocre speed. This configuration actually brings less waiting times than a system with a single big server, because each arriving customer waits only if all servers are busy. In fact, it can be analytically shown that the mean waiting time in the M/M/m queue with arrival rate λ and m servers each with service rate μ is less than that in the M/M/1 queue with arrival rate λ and a single server with service rate $m\mu$ [14, p.285]. The probability of waiting for an arriving customer is much less in the M/M/m queue than in the big M/M/1 queue. However, since explicit analytical results are not available for the M/G/m queue with generally distributed service times, we are unable to provide such a discussion in the general setting.

- Time-Varying Customer Arrival Rate

  The amount of service demand from human customers naturally depends on the time of the day according to people's private and working activities in a day.

For example, the customer arrival rate at a restaurant has peaks at meal times. There are few customers around midnight in a supermarket with 24 hour opening. If servers are human, the personnel expenses are proportional to the duration of working hours. Thus the manager tries to control the number of workers depending on the time of day. There are studies of queues with time-varying customer arrival rate and number of servers by means of fluid approximation. At this moment, however, the theory does not seem have developed enough to be easily amenable to practical use.

- Satisfaction of Customers and Servers

  If servers are CPUs and communication lines, we do not care about the "satisfaction" of such physical devices. The time scale of their metal fatigue and failure (days, say) is order of magnitude different from the time scale of their operation (less than microseconds, say) so that failure and operation should be treated by different models. However, the dissatisfaction of employees directly affects the quality of service in the same time scale. Also, employees may feel happy and enhance the quality of service immediately after the appreciation from their customers. Therefore, it is essential to consider the satisfaction of both customers and servers that depend on each other in the same model. In the traditional application of queueing theory, customer satisfaction was discussed in the context of fair resource sharing. But the server satisfaction such as fair and balanced staff scheduling and customer assignment remain to be investigated for human service systems.

- Workforce Management

  A sad reality is that researchers of queueing theory, those of mathematical optimization, and those of statistics and data science are different species of people who do not talk to each other. The academic performance of a researcher is evaluated by the publication of papers in journals with high citation index within their own fields of specialty. However, for the success of scientific approach to service systems, researchers of various specialties must collaborate on the same problem with their respective contributions. An example is the workforce management for human service systems like call centers and hospitals. First, data scientists should analyze the time records of customer arrivals and departures. Then queueing theorists calculate the necessary and sufficient number of employees to meet the satisfaction of customers. Finally, operations researchers provide the good staff scheduling to meet the requests of employees. Only such packaged solution would be accepted by the managers and practitioners of service systems.

Acknowledgment

I would like to thank Professor Parviz Kermani of the City University of New York and Professor John A. Silvester of the University of Southern California for inviting me to contribute the present article to the tribute issue of Professor Leonard Kleinrock's 80th birthday. I am also very grateful to Dr. Richard Gail, my old friend, and Dr. Harry

Rudin, Co-Editor-in-Chief of *Computer Networks*, for their valuable comments on the draft of this article.

References

[1] Asaduzzaman, M., T. H. Chaussalet, and N. J. Robertson, A loss network model with overflow for capacity planning of a neonatal unit, *Annals of Operations Research*, Vol.178, No.1, pp.67–76, July 2010.

[2] Council on Competitiveness, *Innovate America*: *Thriving in a World of Challenges and Change*, National Innovation Initiative Summit and Report, May 2005.

[3] Daskin, M. S., *Service Science*, John Wiley & Sons, 2010.

[4] Fitzsimmons, J. A. and M. J. Fitzsimmons, *Service Management*: *Operations, Strategy, Information Technology*, Sixth edition, McGraw-Hill, 2008.

[5] Gail, H. R., On the Optimization of Computer Network Power, Ph. D. dissertation, Computer Science Department, University of California, Los Angeles, CSD-830922, September 1983.

[6] Gail, R. and L. Kleinrock, An invariant property of computer network power, *International Conference on Communications*, pp.63.1.1–63.1.5, IEEE, June 1981.

[7] Giessler, A., J. Hänle, A. König, and E. Pade, Free buffer allocation - An investigation by simulation, *Computer Networks*, Vol.2, No.3, pp.191–208, July 1978.

[8] Green, L., Capacity planning and management in hospitals. In: *Operations Research and Health Care*: *A Handbook of Methods and Applications*, edited by M. L. Brandeau, F. Sainfort, and W. P. Pierskalla, pp.15–41, Kluwer, 2004.

[9] Green, L., Queueing analysis in healthcare. In: *Patient Flow*: *Reducing Delay in Healthcare Delivery*, edited by R. W. Hall, pp.281–307, Springer, 2006.

[10] Griffin, J., S. Xia, S. Peng, and P. Keskinocak, Improving patient flow in an obstetric unit, *Health Care Management Science*, Vol.15, No.1, pp.1–14, March 2012.

[11] Hall, R. W., *Queueing Systems*: *For Services and Manufacturing*, Prentice-Hall, 1991.

[12] Huang, J.-H., On the Behavior of Algorithms in Multi-Processing Environment, Ph. D. dissertation, Computer Science Department, University of California, Los Angeles, CSD-880085, October 1988.

[13] Kleinrock, L., *Queueing Systems*, Volume I: *Theory*, John Wiley & Sons, 1975.

[14] Kleinrock, L., *Queueing Systems*, Volume II: *Computer Applications*, John Wiley & Sons, 1976.

[15] Kleinrock, L., On flow control in computer networks, *International Conference on Communications*, pp.27.2.1–27.2.5, IEEE, June 1978.

[16] Kleinrock, L., Power and deterministic rules of thumb for probabilistic problems in computer communications, *International Conference on Communications*, pp.43.1.1–43.1.10, IEEE, June 1979.

[17] Kleinrock, L., Performance evaluation of distributed computer-communication systems. In: *Queueing Theory and its Applications, Liber Amicorum for J. W. Cohen*, edited by O. J. Boxma and R. Syski, pp.1–57, North-Holland, 1988.

[18] Kleinrock, L. and R. Gail, *Solutions Manual for Queueing Systems, Volume 1: Theory*. Technology Transfer Institute, Santa Monica, California, 1982.

[19] Kleinrock, L. and R. Gail, *Solutions Manual for Queueing Systems, Volume 2: Computer Applications*. Technology Transfer Institute, Santa Monica, California, 1986.

[20] Kleinrock, L. and R. Gail, *Queueing Systems*: *Problems and Solutions*, John Wiley & Sons, 1996.

[21] Lovejoy, W. S. and J. S. Desmond, Little 's law flow analysis of observation unit impact and sizing, *Academic Emergency Medicine*, Vol.18, No.2, pp.183–189, February 2011.

[22] Palvannan, R. K. and K. L. Teow, Queueing for healthcare, *Journal of Medical Systems*, Vol.36, No.2, pp.541–547, April 2012.

[23] Takagi, H., Analysis of Throughput and Delay for Single- and Multi-Hop Packet Radio Networks, Ph. D. dissertation, Computer Science Department, University of California, Los Angeles, UCLA-ENG-8326, May 1983.

[24] Takagi, H. (editor), *Introduction to Service Science*: *Innovation by Mathematical Models and Data Analysis*, University of Tsukuba Press, to appear in 2014 (in Japanese).

[25] Takagi, H., Y. Kanai, and K. Misue, Queuing network model and visualization for the patient flow in the obstetric unit of the University of Tsukuba Hospital, *SRII Global Conference 2014*, San Jose, California, April 23–25, 2014.

[26] Takagi, H. and B. H. Walke (editors), *Spectrum Requirement Planning in Wireless Communications*: *Model and Methodology for IMT-Advanced*. John Wiley & Sons, 2008.

[27] Yoshioka, Y., T. Nakamura, and R. Sato, An optimum solution of the queuing system, *The Transactions of the Institute of Electronics and Communication Engineers of Japan*, Section B, Vol.J60-B, No.8, pp.590–591, August 1977.

Biography

Hideaki Takagi is currently Provost and Executive Officer at the University of Tsukuba, Japan. He received his B.S. and M.S. degrees in Physics from the University of Tokyo in 1972 and 1974, respectively. In 1974 he joined IBM Japan as a Systems Engineer. From 1979 to 1983, he studied at the University of California, Los Angeles (UCLA), and received his Ph.D. degree in Computer Science. From 1983 to 1993, he was with IBM Research, Tokyo Research Laboratory. He moved to the University of Tsukuba in October 1993 as Professor at the Institute of Policy and Planning Sciences. He was Vice President of the University of Tsukuba during 2002–2003.

His research interests include queueing theory and stochastic processes as applied to the performance evaluation of computer communication networks and human service systems such as call/contact centers and hospitals. He is the author of research monographs *Analysis of Polling Systems* (MIT Press, 1986), and *Queueing Analysis*: *A Foundation of Performance Evaluation*,Volumes 1–3 (Elsevier, 1991–1993). He is also the co-editor of *Spectrum Requirement Planning in Wireless Communications*: *Model and Methodology for IMT-Advanced* (Wiley, 2008). He is IEEE Fellow (1996) and IFIP Silver Core Holder (2001). He served as editors for *IEEE Transactions on Communications* (1986–1993), *IEEE/ACM Transactions on Networking* (1992–1994), *Queueing Systems* (1988–2009), and *Performance Evaluation* (from 1984 onwards).