Queueing Theory

O.J. Boxma

1. Introduction

Queueing phenomena occur in several real-life situations when resources (machines at a factory, elevators, telephone lines, traffic lights) cannot immediately render the amount or the kind of service required by their users. Also, at byte level in modern data-handling technologies (communication systems, computer networks) queueing phenomena may arise; they are typically less visible but their effects at user level are usually not less serious. Quite often such congestion effects may be adequately studied by mathematical methods from *queueing theory*. Adopting the abstract terminology from queueing theory, the object of study is formulated as a network of service units with customers requiring services at those units. The nature of the arrival processes and service requests is usually such that they have to be represented by stochastic processes. Hence the most important performance measures, like waiting times, workloads and queue lengths, are random variables. Accordingly, the main techniques of queueing theory stem from probability theory.

In Section 2 we discuss some elementary phenomena and results from queueing theory. Section 3 contains a brief history of the past 50 years of queueing theory, with some of the applications that guided its development. Section 4 is devoted to polling systems, a class of queueing systems that recently has received much attention in the literature and at CWI.

2. Elementary queueing theory

The influence of randomness on queueing processes is often remarkably strong, and sometimes at first sight counterintuitive. For example, if a hairdresser spends exactly 12 minutes on each customer, and his customers arrive at intervals of exactly 15 minutes, then no customer has to wait. But if these same customers arrive according to a Poisson process with an arrival rate of 4 customers per hour, then they wait 24 minutes on average! And if, moreover, the service time $\mathbf{S}$ spent per customer is negative exponentially distributed with mean 12 minutes (i.e. the probability that $\mathbf{S}$ exceeds x equals $\Pr\{\mathbf{S} > x\} = e^{-x/12}, x \geq 0$), then the mean waiting time of a customer doubles to 48 minutes.

If, in the latter case, an observer enters the shop at a randomly chosen point in time during a service, he might think that the expected time until completion of that service is one half of 12 minutes, i.e. 6 minutes. But it is easily seen that the probability that $\mathbf{S}$ will exceed $x + y$, given that it has already exceeded x, equals $e^{-(x+y)/12}/e^{-x/12} = e^{-y/12}$, which is the probability that $\mathbf{S}$ exceeds y; this is called the memoryless property of the exponential distribution. Hence the residual time until service completion has exactly the same exponential distribution as $\mathbf{S}$ itself, and its mean is 12 minutes instead of 6. For generally distributed service times, the mean residual service time equals $E\mathbf{S}^2/2E\mathbf{S}$, which exceeds $E\mathbf{S}/2$ when service times are not constant. The intuitive explanation of this phenomenon, the

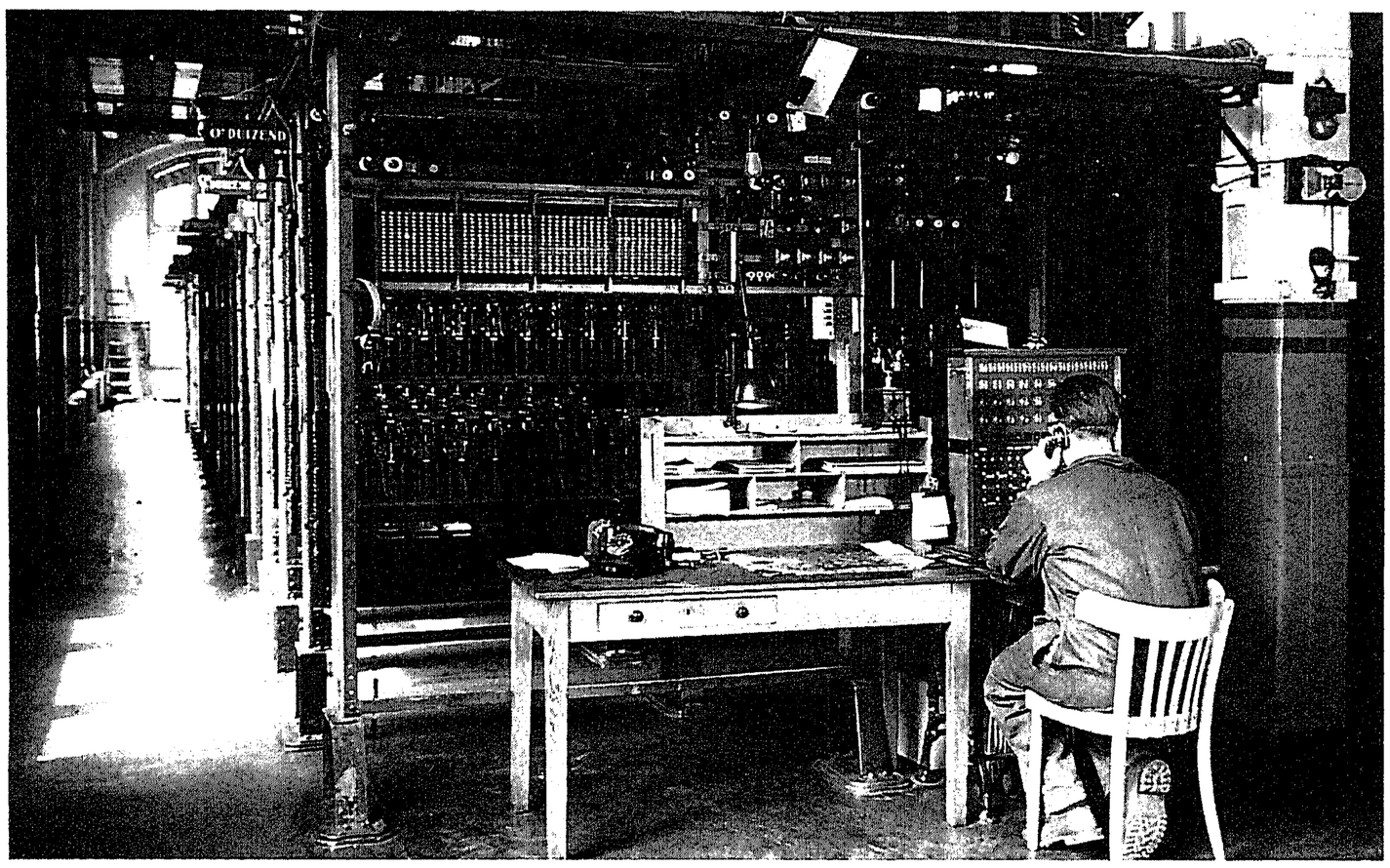

Figure 1. Queueing theory originated early this century in the study of overload in telephone exchanges. (Photo: PTT Telecom.)

'waiting time paradox', is that the observer has a relatively high probability of entering the shop during a relatively long service (the same phenomenon explains why, on average, at a bus stop one has to wait longer than half the interarrival interval indicated by the time table).

The above-mentioned memoryless property is very attractive from a mathematical viewpoint: Given the present, it allows one to disregard the past in studying the future. This is the characteristic of Markov processes, for which a rich literature has been developed in probability theory. At an early stage of queueing theory, in the first half of this century, it allowed the exact analysis of a whole class of 'Markovian' queueing systems. Examples are the loss and delay models developed and analysed by the Danish queueing pioneer A.K. Erlang (1909) with the purpose of dimensioning telephone exchanges (see also figure 1). Another example is the M/M/1 queue (see Box).

The M/M/1 queue

This is a queue with Markovian or memoryless (M) interarrival times and service times, and a single (1) server. The number of customers $\mathbf{X}$ in the M/M/1 queue is also a Markov process, and its steady-state distribution is geometric, i.e., again memoryless: $\Pr\{\mathbf{X} = n\} = (1 - \lambda/\mu)(\lambda/\mu)^n$, $n = 0, 1, \cdots$. Here λ is the arrival rate and μ the service rate. The expected number of *waiting* customers equals $\mathrm{E}\mathbf{X}_w = \mathrm{E}\mathbf{X} - \lambda/\mu = \lambda^2/(\mu(\mu - \lambda))$. Little's formula gives a (very generally valid) relation between the mean number of waiting customers $\mathrm{E}\mathbf{X}_w$ and the mean waiting time $\mathrm{E}\mathbf{W}$:

$$\mathrm{E}\mathbf{X}_w = \lambda \mathrm{E}\mathbf{W}, \tag{2.1}$$

so that for the M/M/1 queue

$$\mathrm{E}\mathbf{W} = \frac{\lambda/\mu^2}{1 - \lambda/\mu}. \tag{2.2}$$

One can now verify that the mean waiting time in the example discussed in the beginning of this section indeed equals 48 minutes.

3. POST-WAR QUEUEING THEORY

3.1. Development of queueing theory

The post-war technological revolution and the resulting new attitude towards pure and applied science had a strong influence on the development of queueing theory. Mathematical modelling in economics and management had always been seriously hampered by the difficulties encountered in the

numerical evaluation of the analytical models. The development of powerful computing machinery lowered this 'numerical' barrier rather abruptly. A new discipline evolved: Operations Research, with queueing theory as one of its fastest growing branches.

Not only as a branch of Operations Research, but also as a branch of Applied Probability, queueing theory attracted the interest of professional mathematicians after the Second World War. Around 1950, the mathematical theory of stochastic processes had reached a certain maturity. Brownian motion and noise phenomena, investigated by physicists in the first quarter of the century, and biological processes such as the development of epidemics and the growth of bacteria populations appeared to be accessible for probabilistic modelling. Around 1950, too, monographs on stochastic processes became available, and the appearance of Feller's famous book *An Introduction to Probability Theory and its Applications* turned out to be a landmark in the development of stochastic modelling. Under the influence of Feller's exposition the techniques required for the analysis of stochastic models were systemized, and were investigated on their merits and on their potential for obtaining numerical results. The influence of these developments on queueing theory was strong, the more so since queueing models turned out to be a gratifying testing ground for many techniques developed in subfields of Probability Theory like Renewal Theory, Birth-and-Death Processes, Branching Processes, Fluctuation Theory and, in particular, (semi-)Markov Processes.

Figure 2. D. van Dantzig.

In The Netherlands, D. van Dantzig (one of the founding fathers of SMC, see figure 2) began, shortly after the Second World War, to teach courses in probability and mathematical statistics. His teaching has been very influential on the development of these fields in The Netherlands. In particular he may be considered to have laid the foundations for the strong international position of Dutch applied probabilists. Van Dantzig's interest in the application of mathematics has also been a stimulus to the development of probabilistic Operations Research in The Netherlands.

In 1953, D.G. Kendall published what was to become one of the most influential papers in queueing theory [4]. In it he analyzed the M/G/1 queue—a single server queue with Poisson arrival process and *generally* (G) distributed service times. The queue length process no longer has the Markov property. Kendall showed an interesting way to circumvent that problem. He observed that the queue length process at the successive epochs (points in time) at which a customer leaves the system is again Markovian—a so-called imbedded Markov chain. He was thus able to determine the steady-state distribution of the queue length process at departure epochs; that distribution could subsequently be shown to be equal to the steady-state queue length distribution at an arbitrary epoch. For future reference we mention the steady-state mean waiting time in the M/G/1 queue:

$$\mathrm{E}\mathbf{W} = \frac{\lambda \mathrm{E}\mathbf{S}^2}{2(1-\rho)}; \tag{3.1}$$

here λ is the arrival rate, $\mathbf{S}$ denotes a service time, and $\rho := \lambda \mathrm{E}\mathbf{S} < 1$ is the offered load. One can now verify the 24-minute result from the hairdresser example in Section 2.

3.2. New stimuli for queueing theory

In the midsixties queueing theory got a new stimulus from the fields of computer engineering and computer-communication networks. At that time computer technology had reached a level of development which required a good insight in the data flow inside the computer as well as in computer networks. For the latter in particular, the classical single service facilities did not suffice; networks of service facilities had to be analyzed. For networks of M/M/1-like queues, J.R. Jackson and also W.J. Gordon and G.F. Newell had shown that the explicit expression for the joint distribution of the queue lengths at the various nodes has a *product form* (in some cases it is a product of the marginal queue length distributions). These product-form results turned out to be very useful for the performance analysis of several basic computer systems, such as the central server system and the computer-terminal system. Furthermore, computer technology created new service disciplines, like processor sharing, that also gave rise to some new product-form results. Landmarks are the beautiful studies of F.P. Kelly and of F. Baskett, K.M. Chandy, R. Muntz and F. Palacios. In The Netherlands researchers like J.W. Cohen, A. Hordijk and N.M. van Dijk also made fundamental contributions.

Towards the end of the sixties it was recognized that the availability of the special resources and capabilities of many separate computer facilities could be extended by resource and load sharing; as a consequence, networks of computer systems, or more generally data communication networks, started to emerge. In line with its tradition, queueing theory has responded very

positively to the challenges posed by these new technological developments. Satellite communication led to queueing models for the phenomenon of colliding transmissions; protocols for message transmission in local area networks led to new priority models (like polling models, cf. Section 4); flexible manufacturing and distributed processing gave rise to queueing models with complicated dependencies between the arrival processes of customer streams at various queues, and between their service processes.

The effectiveness of queueing theory in handling such problems may to a large extent be due to the deep understanding that has been acquired for basic queueing models, such as Erlang's loss model, the M/G/1 queue and product-form networks. Accordingly, queueing theory has established itself as an indispensable tool in the design and performance analysis of computer-communication networks. Presently a new key topic is coming to the front: the performance analysis of Broadband Integrated Services Digital Networks (B-ISDN, see also figure 3). Such networks allow the simultaneous transmission of different traffic types (data, video, voice). Traditional arrival processes are inadequate for describing the sometimes bursty nature and long-ranging dependencies of these traffic streams; again this poses new challenges to performance analysts.

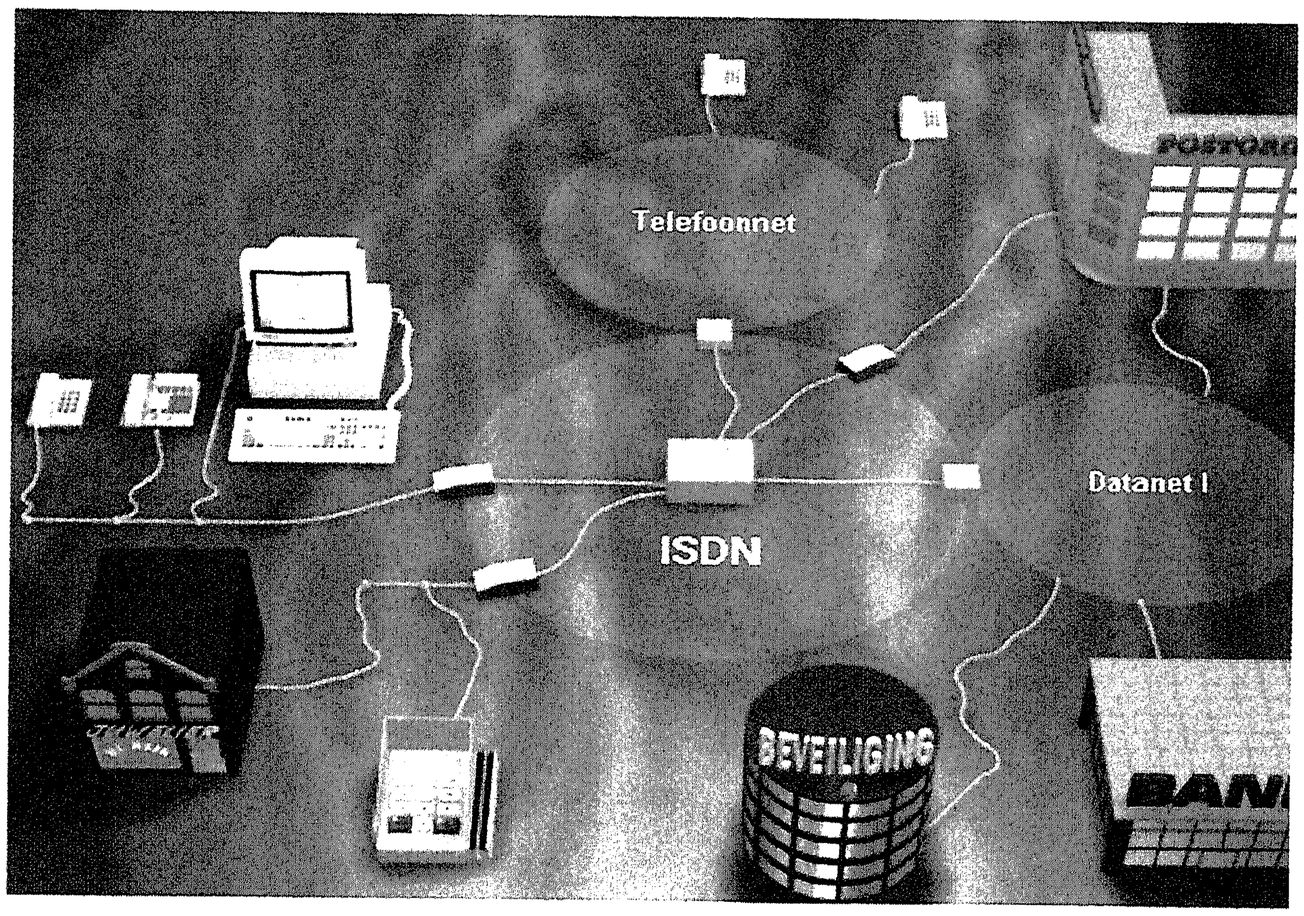

Figure 3. The simultaneous transmission of data, video and voice through broadband networks presents new challenges to queueing theory. (Courtesy PTT Telecom.)

4. CWI-RESEARCH ON POLLING SYSTEMS

4.1. Introduction

While the activities in CWI's research group Analysis and Control of Information Flows in Networks are not restricted to queueing theory (there is also much research in random walks, percolation theory and reliability theory, and in a more distant past many strong contributions have also been made to Markov decision theory), the group's core activity in the last few years has been queueing theory and its application to computer-communication performance analysis. Of the five Ph.D. theses that have been produced in the group in the nineties, two have been devoted to the analysis and optimization of polling systems [1, 3]. Therefore we now pay special attention to this class of queueing systems, starting with a motivating example.

Many communication systems provide a broadcast channel which is shared by all connected stations. When two or more stations wish to transmit simultaneously, a conflict arises. The rules for resolving such conflicts are referred to as 'multi-access protocols'. The token ring protocol is one such protocol, that is being used in many local area networks.

In a token ring local area network, several stations (terminals, file servers, hosts, gateways, etc.) are connected to a common transmission medium in a ring topology. A special bit sequence called the *token* is passed from one station to the next; a station that 'possesses the token' is allowed to transmit messages. After completion of his transmission the station releases the token, giving the next station in turn an opportunity to transmit. This situation can be presented by a so-called *polling model.*

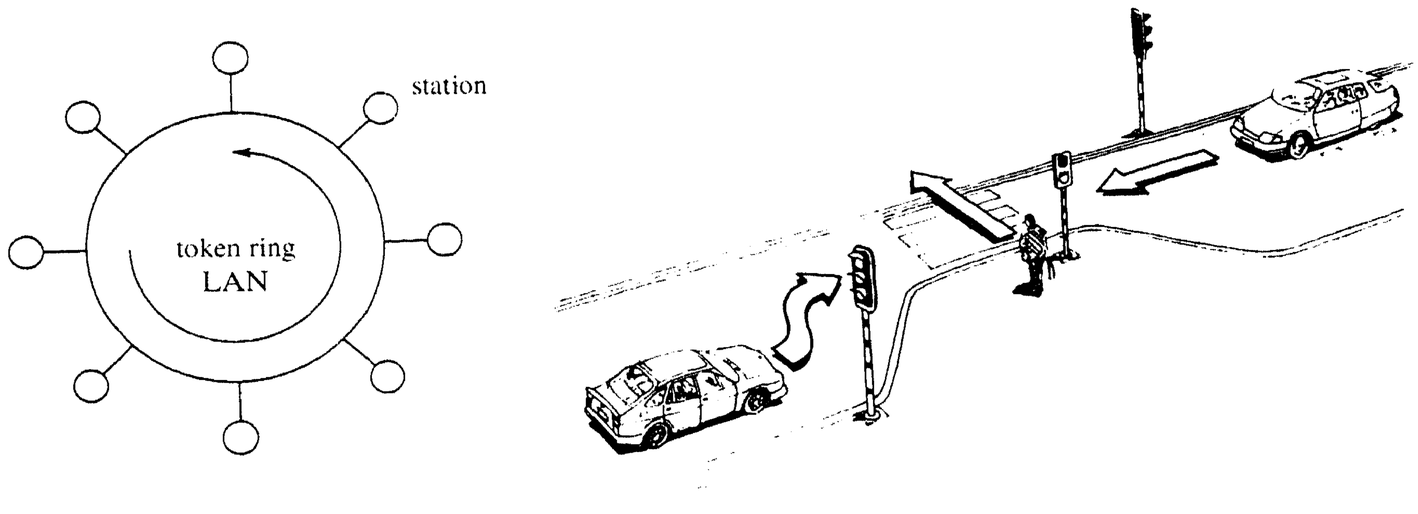

Figure 4. Multi-queueing problems with cyclic service are as common in computer networks as in road traffic situations. CWI has arrived at a conservation law giving insights into average waiting times for all sorts of customers in such systems.

4.2. The polling model

A polling model is a single-server multi-queue model, in which the server attends to the queues in cyclic order. The N queues $Q_1, ..., Q_N$ have infinitely large waiting rooms. Arrival times of customers at the queues are usually assumed to occur according to a Poisson process. Service requirements of customers at a queue are independent, identically distributed stochastic variables; the same holds for the switch-over times of the server between queues. Arrival rates, service time and switch-over time distributions may differ from queue to queue.

A polling model describes the behaviour of a token ring local area network in a natural way. The server represents the token-passing mechanism, and the customers represent messages generated at the stations. Many other situations in which several users compete for access to a common resource can also be described by this polling model. Examples are a repairman patrolling a number of machines which may be subject to breakdown, assembly work on a carousel in a production system, a computer with multi-drop terminals, and a signalized road traffic intersection (see figure 4). Depending on the application, various service disciplines at the queues may be considered. Common disciplines are *exhaustive service* (the server continues to work at a queue until it becomes empty), *gated service* (the server serves exactly those customers who were present when he arrived at the queue) and *1-limited service* (the server serves just one customer—if anyone is present—before moving on to the next queue).

Exact results for waiting time distributions are known when the service discipline at each queue is either exhaustive or gated. In a recent CWI report, J.A.C. Resing has shown that a detailed exact analysis (using the theory of multitype branching processes) is possible for the broader class of polling systems for which the service discipline at each queue has a so-called 'branching property'. Hardly any exact results for individual queue lengths or waiting times are known when this property is violated at one or more

queues. However, even in such a case there exists a simple expression for a certain *weighted sum* of all the steady-state mean waiting times; L. Kleinrock has shown in 1964 that, when all switch-over times between queues are zero:

$$\sum_{i=1}^{N} \rho_i \mathrm{E}\mathbf{W}_i \;=\; \rho \frac{\sum_{i=1}^{N} \lambda_i \mathrm{E}\mathbf{S}_i^2}{2(1-\rho)} \,. \tag{4.1}$$

Here $\mathrm{E}\mathbf{W}_i$ denotes the mean waiting time at Q_i, λ_i the arrival rate, $\mathbf{S}_i$ the generic service time, and $\rho_i := \lambda_i \mathrm{E}\mathbf{S}_i$ the offered traffic load; $\rho = \Sigma_i \rho_i$ denotes the total offered load. This is called a *conservation law*: if the service discipline at a queue is changed, the weighted sum of mean waiting times (the left-hand side of (4.1)) remains the same, although the individual mean waiting times may change. Note that this formula is a generalization of formula (3.1) for the mean waiting time in a single M/G/1 queue.

4.3. Work conservation and work decomposition

The conservation law is a consequence of the 'principle of work conservation'. Suppose that the scheduling policy, i.e., the procedure for deciding at any time which customer(s) should be in service, has the properties that it does not allow the server to be idle when at least one customer is present and does not affect the amount of service given to a customer or the arrival time of any customer. Comparing the sample paths of the 'workload process' for such a system under different scheduling disciplines leads to the observation that *the workload process is independent of the scheduling discipline.*

The principle of work conservation has in the past proven to be very useful. It enables one to analyze the workload process of queueing systems with a highly complicated scheduling discipline as if the scheduling were a relatively simple one, such as the First Come First Served discipline.

For the token ring local area network mentioned above, the time for the token to be passed from station to station is in general not negligible. Correspondingly, in the polling model the time the server needs for switching from station to station has to be taken into account. This fact considerably complicates the analysis: the principle of work conservation is no longer valid, since now the server may be idle (switching), although there is at least one customer in the system. However, under certain conditions there exists a natural modification of the principle of work conservation for polling systems with switch-over times, based on a *decomposition* of the amount of work in the system [2, 3]. This result states that—under certain conditions—the amount of work in the polling system, $\mathbf{V}_{with}$, is in distribution equal to the sum of the amount of work in the simpler 'corresponding' system *without* switch-over times, $\mathbf{V}_{without}$, plus the amount of work, $\mathbf{Y}$, at an arbitrary moment during a period in which the server is switching from one queue to another:

$$\mathbf{V}_{with} =^{(d)} \mathbf{V}_{without} + \mathbf{Y}, \tag{4.2}$$

$=^{(d)}$ denoting equality in distribution.

The work decomposition gives rise to similar expressions for a weighted sum of the mean waiting times as Formula (4.1). We can write

$$\mathrm{E}\mathbf{V}_{with} = \sum_{i=1}^{N} \mathrm{E}\mathbf{S}_i \mathrm{E}\mathbf{X}_{i,w} + \sum_{i=1}^{N} \rho_i \frac{\mathrm{E}\mathbf{S}_i^2}{2\mathrm{E}\mathbf{S}_i} = \sum_{i=1}^{N} \rho_i \mathrm{E}\mathbf{W}_i + \frac{1}{2} \sum_{i=1}^{N} \lambda_i \mathrm{E}\mathbf{S}_i^2.$$

The first relation splits the mean workload into the mean workload of the $\mathbf{X}_{i,w}$ waiting customers and the mean residual workload of the customer in service (if a customer of type i is in service, his mean residual service time equals $E\mathbf{S}_i^2/2E\mathbf{S}_i$, cf. the discussion of the waiting time paradox in the beginning of this essay); the second equality follows from Little's formula $\mathrm{E}\mathbf{X}_{i,w} = \lambda_i \mathrm{E}\mathbf{W}_i$. Taking means in (4.2) now leads to:

$$\sum_{i=1}^{N} \rho_i \mathbf{E}\mathbf{W}_i \;=\; \rho \frac{\sum_{i=1}^{N} \lambda_i \mathbf{E}\mathbf{S}_i^2}{2(1-\rho)} + \mathbf{E}\mathbf{Y}. \tag{4.3}$$

Denote the mean total switch-over time in one cycle of the server by s, and the second moment by $s^{(2)}$. Evaluating $\mathbf{E}\mathbf{Y}$ (cf. [2]) yields:

$$\sum_{i=1}^{N} \rho_i \mathbf{E}\mathbf{W}_i \;=\; \rho \frac{\sum_{i=1}^{N} \lambda_i \mathbf{E}\mathbf{S}_i^2}{2(1-\rho)} + \rho \frac{s^{(2)}}{2s} +$$

$$\frac{s}{2(1-\rho)}[\rho^2 - \sum_{i=1}^{N} \rho_i^2] \;+\; \sum_{i=1}^{N} \mathbf{E}\mathbf{M}_i, \tag{4.4}$$

where $\mathbf{E}\mathbf{M}_i$ denotes the mean amount of work in Q_i left by the server upon its departure from that queue (when $s \to 0$, the fraction of visits to Q_i in which the server finds Q_i empty tends to one, and the right-hand side of (4.4) reduces to the right-hand side of (4.1)). Formula (4.4) has been coined a *pseudo-conservation law*. The main difference with Kleinrock's conservation law is that now the weighted sum of mean waiting times *does* depend on the service discipline at each queue, through $\sum \mathbf{E}\mathbf{M}_i$. For many service disciplines, amongst which are exhaustive, gated and 1-limited service, we are able to determine the right-hand side of (4.4) explicitly.

It is also possible to extend the work decomposition property and pseudo-conservation law to much more general single-server systems with multiple customer classes [2]. Such pseudo-conservation laws often provide the only information available in polling and multiclass systems with nonzero switch-over times. They are therefore of considerable practical importance. One of the main features of the pseudo-conservation laws is that they are very useful for testing *and* developing approximations for the individual mean waiting times [3]. Such approximations have in turn supplied an approach for solving various optimization problems, like (i) determine the optimal visit times in a cyclic polling model [1] (this problem has been studied in a consultancy for PTT Telecom) and (ii) determine the optimal visit order of the stations when non-cyclic polling is allowed.

The work decomposition property (4.2) relates the workload in polling models with and without switch-over times. Such a simple relationship does not in general exist between the joint queue length processes in both models. However, when all service disciplines have the above-mentioned 'branching property', then one can also find [1] a surprisingly simple relation between the joint queue length processes in both models, at particular imbedded points in time (cf. our earlier reference to [4]!).

Other very recent polling research at CWI has led to the first exact results for polling systems with *multiple servers* [1], and to good rules for the NP-complete problem of the optimal (with respect to a weighted sum of

mean waiting times) probabilistic allocation of several customer types to a collection of parallel servers.

4.4. Epilogue

In Section 4 some emphasis has been put on Ph.D. research. In recent years the emphasis in the group has been shifting somewhat towards postdoctoral research: several postdocs have worked in the group, supported by grants from Shell, PTT Research, Esprit and ERCIM.

Finally it deserves to be mentioned that, since the early eighties, the group has strongly benefitted from the advisorship of J.W. Cohen. Through his research he has contributed, more than anyone else, towards establishing queueing theory as a mature mathematical discipline within Applied Probability. Almost ten years after his retirement, his unflagging energy and love for the mathematical modelling and analysis of congestion phenomena continue to stimulate his environment.

REFERENCES

1. S.C. BORST (1994). *Polling Systems*, Ph.D. Thesis, Tilburg University.
2. O.J. BOXMA (1989). Workloads and waiting times in single-server queues with multiple customer classes. *Queueing Systems 5*, 185-214.
3. W.P. GROENENDIJK (1990). *Conservation Laws in Polling Systems*, Ph.D. Thesis, Utrecht University.
4. D.G. KENDALL (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *Ann. Math. Statist. 24*, 338-354.