Journal of the Operational Research Society (2009) 60, S83–S92

www.palgrave-journals.com/jors/

Reflections on queue modelling from the last 50 years

D Worthington*

Lancaster University, Lancaster, UK

Queueing theory continues to be one of the most researched areas of operational research, and has generated numerous review papers over the years. The phrase 'queue modelling' is used in the title to indicate a more practical emphasis. This paper uses work taken predominantly from the last 50 years of pages of the *Operational Research Quarterly* and the *Journal of the Operational Research Society* to offer a commentary on attempts of operational researchers to tackle real queueing problems, and on research foci past and future. A new discipline of 'queue modelling' is proposed, drawing upon the combined strengths of analytic and simulation approaches with the responsibility to derive meaningful insights for managers.

Journal of the Operational Research Society (2009) **60,** S83–S92. doi:10.1057/jors.2008.178
Published online 11 February 2009

Keywords: queueing theory; queueing models; simulation

1. Introduction

Queueing theory is one of the longest established modelling approaches available to operational researchers, with its origins in the early telecommunications work of AK Erlang, see Brockmeyer *et al* (1948). It also continues to be one of the most researched areas of operational research, so much so that it has generated numerous review papers over the years, for example Saaty (1966); Bhat (1969); Bhat *et al* (1979); Koenigsberg (1982); Neuts (1984); Bitran and Dasu (1992) and Preater (2001).

This paper takes the occasion of the 50th Operational Research Society Conference to add to this process, and by using the pages of the *Operational Research Quarterly* and the *Journal of the Operational Research Society* to provide much of the evidence, attempts to emphasize the interest of the Operational Research Society (ORS) in the application of operational research.

In the title of this paper, and in what follows, the phrase 'queueing models' is used rather than the more traditional 'queueing theory'. This is partly to emphasize the application rather than just the theory. Additionally 'queueing models' is used to indicate modelling approaches in general which are appropriate for tackling queueing problems, not exclusively just those based on queueing theory.

A more accurate definition of a 'queueing system' is in fact a 'service system', as the crucial feature is not the queue but rather a service, and customers who wish to receive the service. If the customers are lucky, either by system design or by chance, they will receive the service without queueing; if they are unlucky they will need to queue. Given this more descriptive definition, it is clear that we can include a very wide set of systems. For example, in a healthcare setting there are emergency ambulance systems where an important objective is to avoid the occurrence of queues if at all possible, outpatient clinics where the aim of an appointment system is to provide the medical staff with a smooth flow of work without causing undue waiting for patients, and hospital waiting lists where the waiting list can act as a rationing mechanism. Similar ranges of queueing systems exist in very different application areas. For example, in call centre work Whitt (2006) distinguishes between revenue generating and service-orientated call centres. The former usually require very low chances of delay, whereas the latter (with the exception of emergency services) do not specify a delay probability, but require answering $x\%$ (eg 80%) of calls within y (eg 30) seconds. Such relatively long waits can again act as a rationing mechanism by deterring customers, or at least by encouraging them to ring back at quieter times of day.

Queue modelling is potentially very valuable, as many real situations in important industries can be formulated as queueing/service systems. Over the last 50 years or so in the *Operational Research Quarterly* and the *Journal of the Operational Research Society* the most frequently reported application areas have been healthcare and transport. In healthcare, for example Bailey (1957), when writing generally about hospital planning and design, highlighted the scope for operational research methods to extract the maximum amount of benefit for the community out of restricted resources, citing appointment systems and emergency beds as examples. Queue modelling studies of outpatient appointment systems have provided a rich vein of work, see for example Jackson *et al* (1964) and Brahimi and Worthington (1991). Other examples in healthcare queue modelling include emergency ambulance services—Groom (1977), hospital waiting

*Correspondence: D Worthington, Department of Management Science, Lancaster University, Lancaster LA1 4YX, UK.
E-mail: d.worthington@lancaster.ac.uk

lists—Worthington (1987), intensive care—Griffiths *et al* (2005), NHS Direct—Royston *et al* (2003), and NHS walk-in centres—Ashton *et al* (2005).

Applications from transport over the same period include airline passenger check-in—Lee and Longton (1959), control policies for signalized road intersections—Green and Hartley (1966), port investment in berths for containers—Edmond and Maggs (1978), queues and delays at roundabouts and priority junctions—Kimber *et al* (1986), queues for freight services through the Channel Tunnel—Griffiths *et al* (1991), and balancing delays and throughput for the Suez Canal Authority—Griffiths (1995).

Other important queue modelling application areas that have been reported in the *Operational Research Quarterly* and the *Journal of the Operational Research Society* include coal mining (Faulkner, 1968), telecommunications (Lawrie, 1980), call centres (Chassioti and Worthington, 2004), office processing (Mayhew, 1987), and retail (Parkan, 1987).

Because of the level of mathematics that is sometimes involved in queueing theory, queue modelling is often seen as a challenging undertaking. However, the challenge is also in part a reflection of the nature of the real queueing problems, which typically incorporate random variation in arrival and service processes, exhibit stochastic behaviour which is often time-dependent, and are often parts of more complex systems, for example networks of queues. These same features mean that the real-life management of these systems is also often challenging, requiring the manager to have a good feel for the impact of stochastic variation, and for performance measures in use to be probabilistic rather than deterministic.

The focus of the remainder of this paper is encapsulated well in two quotes from Bhat's review of 40 years ago (Bhat, 1969):

> The main purpose of the study of queueing systems is to understand real life queueing situations as well as possible.

> Most of the results used have been relatively simple and it is quite clear that the applied researcher has not benefited from the advancement of knowledge on the basic model.

Both these quotes emphasize that the eventual purpose of modelling work is to enable us to better deal with real-life situations. The second also highlights the clear challenge to link the (academic) advancement of knowledge to identifiable benefits in applied research. Both issues remain pertinent today, particularly for organizations such as the ORS.

The aim of this review is therefore to consider a range of queue modelling contributions that have been made over the last 50 years, and to reflect upon them in the light of the needs of managers of real-life queueing systems. The next section draws on queue modelling work to propose five dimensions of queues and queue modelling, and the third section uses the same body of work to distinguish three approaches to queue modelling. The penultimate section then offers a queue manager's perspective. In the final section, the five dimensions, three approaches, and manager's perspectives are combined to reflect upon past queue modelling research and development, and to propose possible future foci.

A final point of motivation is to recognize that queueing theory is a classic modelling approach, in the sense that it makes its assumptions very clear at the outset. The modeller is more aware than in many other approaches that their model is a simplification of reality. This can either lead to the divergence of theory and practice identified by Bhat in 1969, or can serve to heighten the modeller's ability to extract practical insights from their necessarily simplified models. Lessons learned for queue modelling may therefore also offer relevant lessons for other areas of modelling.

2. Five dimensions of queues and queue modelling

In this section five dimensions are proposed to describe the range of real-life queues and the queueing models that have been developed to represent them.

2.1. Single server versus multi-server versus infinite server

While the number of servers is clearly a key parameter in practice, the fact that possible numbers of servers are grouped as $S=1$, $1 < S < \infty$, and $S=\infty$ is a consequence of the mathematics rather than reality. In particular, while single-server systems exist, they are only treated separately from $S > 1$ because the mathematics usually turns out to be simpler. For example, the expressions for the mean number in the system ($E(Q)$) for M/M/1 and M/M/S queues (using Kendall's notation, see eg Gross and Harris, 1985) with arrival rate λ and mean service time $1/\mu$ are:

$$M/M/1: \quad E(Q) = \frac{\lambda}{(\mu - \lambda)} \tag{1}$$

$$M/M/S: \quad E(Q) = \frac{\lambda}{\mu} + \frac{(\lambda/\mu)^S \lambda\mu}{(S-1)!(S\mu - \lambda)^2} P_0 \tag{2}$$

where

$$P_0 = \left[\sum_{k=0}^{S-1} \frac{(\lambda/\mu)^k}{k!} + \frac{(\lambda/\mu)^S}{S!} \frac{S\mu}{S\mu - \lambda} \right]^{-1} \tag{3}$$

The special case of $S = \infty$ is not because systems with infinite numbers of servers exist, but because the mathematics is often simplified by assuming $S \to \infty$ and because some systems can be well represented by assuming $S \to \infty$. For example, the simple expression for the mean number in M/M/∞ queueing systems is

$$M/M/\infty: \quad E(Q) = \frac{\lambda}{\mu} \tag{4}$$

that is, simpler than for M/M/S and even than for M/M/1 systems. Furthermore, the distribution about this mean is simply Poisson.

2.2. *Exponential versus non-exponential*

This second dimension is also mathematically driven, with important implications for the types of queueing systems that can be studied. The exponential assumption implies that inter-arrival times and service times are exponentially distributed. Exponential inter-arrival times imply random arrivals, which may well be realistic for many queueing problems; however, exponentially distributed service times are few and far between. The non-exponential assumption usually implies other distributions of service times, and sometimes other distributions of inter-arrival times.

When systems are exponential, for example M/M/1, M/M/S, and M/M/∞, mathematical formulation is as a Markov process, and solution for steady-state behaviour follows easily from the associated birth–death equations. Simple formulae for the full steady-state probability distributions are obtained easily in addition to Equations (1)–(4) for their means.

However, when systems are non-exponential, for example M/G/1, M/G/S, and M/G/∞, analysis is usually more difficult. In the case of the single-server system M/G/1, mathematical formulation is now as an imbedded Markov chain. A closed-form mathematical solution for the mean number in the system results in the extremely insightful Pollaczek–Khintchine formula:

$$M/G/1: \quad E(Q) = \frac{\lambda}{\mu} + \frac{\lambda^2(1/\mu^2 + Var(\text{service time}))}{2(1-\lambda/\mu)} \qquad (5)$$

Obtaining the distribution of number in the system is, however, more difficult as it requires the numerical solution of the steady-state equations associated with the imbedded Markov chain.

The multi-server version (M/G/S) is more demanding again, this time requiring approximation of the service time distribution by a phase type distribution, see for example Neuts (1975) or by a discrete time distribution, see for example Wall and Worthington (1994). In both approaches the approximate service time distribution is normally chosen to match the mean and variance of the true service time distribution as closely as possible. Both approaches result in Markov chain formulations, the former in continuous time and the latter in discrete time. Solution of the resulting steady-state equations then involve the use of generating functions and/or numerical methods.

In contrast the infinite-server version (M/G/∞) is less demanding. In fact, the steady-state behaviours of numbers of customers in M/G/∞ and M/M/∞ queues are identical, see for example Gross and Harris (1985), so that the distribution is again Poisson with mean given by Equation (4).

2.3. *Steady state versus time-dependent*

Most real queueing systems exhibit time-dependent behaviour to a greater or lesser extent, either while settling to steady-state or because parameters (often the arrival rate) change over time. In many cases this time dependence is relatively unimportant and can be ignored by the operational researcher; however, in many others it is important to somehow reflect the time dependence. In the main, time-dependent behaviour is tougher to analyse than steady-state behaviour.

For example, taking the simplest case in terms of our previous two dimensions, time-dependent solutions for M(t)/M/1 systems require the solution of sets of differential equations. Solutions can be obtained numerically, see for example Grassmann (1977), or more elegantly by using transforms and/or special functions, see for example Sharma (1990) and Bunday (1996).

Moving away from assumptions of exponential service times and single-server systems increases the challenge for the analyst substantially. Analysis of M(t)/G/S systems necessarily requires phase-type or discrete approximations for the service time distribution, as for the equivalent steady-state systems. However the resulting time-inhomogeneous Markov chains, in either continuous or discrete time, also require time-dependent solutions. Progress is sometimes possible analytically, see for example Griffiths *et al* (2006), but in the main numerical solution is required. In this respect, the discrete time approach performs very well, see for example Worthington and Wall (1999). The discrete time step not only allows matching of mean and variance of service time, but also keeps track with the passage of real time, whereas phases in the continuous time formulations do not.

A third important approach for modelling M(t)/G/S is to use steady-state results to approximate time-dependent behaviour. There is a considerable body of work on the use of stationary approximation-based methods (eg simple stationary approximation—SSA, pointwise stationary approximation—PSA, modified offered load—MOL) in which time-dependent queueing systems are assumed to settle instantaneously to the steady-state behaviour associated with the current arrival rate, see Green *et al* (2007) for an extensive account. In the main these approximations have been found to work well when service levels are high (ie probabilities of delay are small), or when fluctuations in arrival rate are small, or when cycle lengths are large relative to the mean service time.

As before the infinite-server version (M(t)/G/∞) is again less demanding. As shown by Eick *et al* (1993) the time-dependent number of customers in the system is again Poisson, but in this case the mean is given by the formula:

$$M(t)/G/\infty: \quad \text{mean}(t) = \int_{-\infty}^{t} \lambda(u) F(t-u) \mathrm{d}u \qquad (6)$$

where $\lambda(u)$ is the arrival rate at time u and $F(t-u) =$ prob(service time $> t-u$).

This approach is particularly easy to use if service time is represented by a discrete distribution, in which case the integral expression is replaced by a summation expression.

2.4. Single node versus tandem queues versus networks of queues

The pioneering work on queueing networks was the research of RRP Jackson (1954, 1956) and JR Jackson (1957, 1963). JR Jackson considered *open queueing networks* in which there are external arrivals to and departures from the system at one or more of the nodes of the system. By assuming random arrivals and exponential service times he reported that a fairly general class of networks can be exactly analysed as individual queues, and each queue can be formulated as an M/M/S system. The solution method is known as the *product-form* solution, which means that the steady-state joint probabilities can be expressed as the product of the marginal probabilities of each independent queue. The process of modelling the overall system as a set of independent subsystems is referred to as *decomposition*.

The concept of *decomposition* has underpinned much subsequent research to develop approximations for non-exponential systems. The Queueing Network Analyzer was developed at Bell Laboratories, Whitt (1983), to model open queueing networks in which external arrivals do not need to be Poisson and service times do not need to be exponential. As with phase-type approximations for single-node systems, arrival processes and service times are each characterized by two parameters, one to represent the rate and one to represent variability. Each service node is analysed as a standard GI/G/S queue, and performance measures for the whole network are approximated by assuming that the nodes are stochastically independent and are evaluated numerically. Hasslinger and Rieger (1996), among others, showed some advantages of refining this decomposition approach using discrete time distributions to represent service times, rather than the two-parameter characterization of Whitt.

In practice there are often physical limitations on the buffer capacities in queueing networks. For example in a manufacturing context *blocking* occurs if the flow of jobs through a workstation is stopped when the buffer of a destination workstation has reached its maximum capacity. Such queueing networks are generally difficult to analyse and, except in a few special cases, do not have *product-form* solutions, see for instance Perros (1994). Even with exponential assumptions, analysis of these systems involves the numerical solution of Markov chain models. Furthermore because the number of states for exact numerical analysis grows dramatically with the number of workstations and the size of the buffer capacities, exact analysis becomes computationally impossible and most analyses are based on analytical approximations using partial decomposition methods, see for example Lee *et al* (1998), or on simulation techniques.

As soon as time dependence of network behaviour is considered analytic work to date is very limited and the main modelling approach is simulation, with one important exception. When networks are uncapacitated, that is, infinite servers are assumed at each node, the time-dependent number of customers at each node has a Poisson distribution with its own time-dependent mean. The equations for the means take a similar form to Equation (6), but now need to consider all the possible routes by which a customer could have arrived at the node of interest by time t, see Massey and Whitt (1993) for details.

2.5. Extra features (eg queueing system rules and behavioural factors)

Special features such as priority systems, server vacations, batch arrivals, bulk services, customer balking, and reneging are well represented in the queueing literature, be they motivated by particular real-life problems or intended to improve understanding of a class of problems. For example: Kimber *et al* (1986) used simulation to model delays at priority junctions, whereas Williams (1980) studied a class of multi-server priority queues using queueing theory; Samanta *et al* (2007) used a discrete time formulation to model single-server systems with multiple vacations motivated by potential applications in high-speed computer networks; Alfa (1979) also used a discrete time approach to develop a numerical method for the time-dependent behaviour of single-server queues with batch arrivals; Griffiths *et al* (1991) applied bulk-service queueing theory to estimate queues at the Channel Tunnel, whereas Cromie and Chaudhry (1976) produced tables and charts for the steady-state behaviour of exponential bulk service queues; van Ackere and Ninios (1993) used simulation to model a single-server queue with advertising and balking; and Ke and Wang (1999) tackled an exponential machine repair problem with balking, reneging, and server breakdowns.

In the main the introduction of extra features makes the modelling task more challenging than it would otherwise be. However there are exceptions. One classic example is the Erlang Loss formula, see for example Gross and Harris (1985), where the introduction of balking whenever the servers are all busy means that the analytically unsolvable M/G/S queue becomes the easily solvable M/G/S/S queue.

3. Three approaches to queue modelling

The examples of queueing models and their applications in the previous sections are now used to propose three distinct approaches to queue modelling, ranging from the classic mathematical formulae of queueing theory to simulation modelling. While these two extremes of our set of three approaches are clearly very different, both are entirely consistent with the focus of this paper, that is (Bhat, 1969) 'The main purpose of the study of queueing systems is to understand real life queueing situations as well as possible'.

It is also interesting to note that while simulation modelling has only become easily available and user friendly in recent years, Bailey (1952), in one of the earliest studies of hospital outpatient clinics, phrased the problem in the

language of queueing theory but chose to tackle it by using a pre-computer simulation to model the consequences of different appointment systems.

3.1. Analytic formulations and formulae

Classic queueing theory often appears in terms of analytic or mathematical formulations, and at its simplest results in easy-to-use formulae. The formulae themselves can be found by relatively simple methods, for example the solution of birth–death equations, or by much more challenging use of mathematical transforms and generating function methods. Furthermore these formulae can bring with them important insights that apply to many real queueing situations. At the simplest level, Equation (1) for the M/M/1 queue is trivial to evaluate and immediately shows the importance of traffic intensity (λ/μ) as the sole determinant of the level of congestion. Similarly Equations (2) and (3) are very easy to evaluate in a spreadsheet for M/M/S queues, and the mathematics behind them show the crucial role of traffic intensity ($\lambda/S\mu$) in determining whether or not steady state exists, and the level of congestion that results.

The Pollaczek–Khintchine formula (Equation (4)) has probably generated even more insights into queue behaviour. For the M/G/1 queue Equation (4) again demonstrates the primary importance of traffic intensity in determining the level of congestion, but adds the additional insight that increasing variability of service time also increases congestion. Furthermore, for M/G/1 systems no other moments of service time matter at all. This insight has led to the hypothesis that something similar may be true for multi-server systems, and hence to the successful approach of using phase-type and discrete time approximations for M/G/S systems referred to earlier, Neuts (1975) and Worthington and Wall (1999).

3.2. Analytic formulations & numerical solutions

Many queueing problems can be formulated analytically, but the resulting equations do not lead to easy-to-use formulae. For example, the time-dependent behaviour of many queueing systems can be expressed in terms of Chapman–Kolmogorov equations. For the simple M(t)/M(t)/S system they take the form:

$$\begin{aligned}
\frac{\mathrm{d}p_0(t)}{\mathrm{d}t} &= -\lambda(t)p_0(t) + \mu(t)p_1(t) \\
\frac{\mathrm{d}p_n(t)}{\mathrm{d}t} &= -(\lambda(t) + n\mu(t))p_n(t) + (n+1)\mu(t)p_{n+1}(t) \\
&\quad + \lambda(t)p_{n-1}(t) \qquad \text{for } (1 \leqslant n < S) \\
\frac{\mathrm{d}p_n(t)}{\mathrm{d}t} &= -(\lambda(t) + S\mu(t))p_n(t) + S\mu(t)p_{n+1}(t) \\
&\quad + \lambda(t)p_{n-1}(t) \qquad \text{for } (n \geqslant S)
\end{aligned}$$

where

$\lambda(t)$ time-dependent mean arrival rate
$\mu(t)$ time-dependent mean service rate
$p_n(t)$ probability of n customers in the system at time t.

As noted earlier, for special cases, for example when $S = 1$, solutions can be found in terms of transforms and/or special functions. However, in the main these require significant numerical work to evaluate them. Alternatively, numerical methods developed to solve sets of differential equations can be used to solve the Chapman–Kolmogorov equations, see for example Grassmann (1977). With the increasing power of computers both these approaches are becoming increasingly feasible to use in research and indeed in practice.

Similarly, when a discrete time modelling approach is adopted, time-dependent behaviour can be formulated as a time-inhomogeneous Markov chain, requiring solution of the time-dependent equations (Worthington and Wall, 1999):

$$\boldsymbol{\pi}(n) = \boldsymbol{\pi}(n-1)\mathrm{P}(n) \quad \text{for } n = 1, 2, \ldots$$

where $\boldsymbol{\pi}(n)$ is the vector of state probabilities at time n, and $\mathrm{P}(n)$ is the associated time-dependent transition matrix. Despite the potentially very large size of the state space associated with these formulations, modern computer power again means that these problems can be solved numerically and hence that these methods are also becoming popular as research tools and feasible to use in practice.

Thus in terms of providing performance measures for a particular queueing system with particular parameter values, these numerical methods can increasingly be viewed 'like formulae' that we might evaluate in a spreadsheet to obtain performance measures. Furthermore, while understanding and insights into the behaviour of these systems are not as obvious as when seen in formulae, the extensive theories associated with differential equations and Markov chains can be used to provide understanding and insights into the behaviour of queueing systems that are modelled in this way.

3.3. Simulation models

As noted earlier, simulation modelling has long been part of queue modelling, and many of the applications and some of the research described earlier in this paper have made some use of simulation models. Drawing a parallel with 'analytic formulations and numerical solutions', simulation models are also capable of providing performance measures for a particular queueing system with particular parameter values, and hence can similarly also be viewed 'like formulae'. An added caveat is that the results will have confidence intervals, although these can be very narrow by having sufficient runs of the model.

Unlike the numerical methods described earlier, simulation cannot rely on underlying theory to provide understanding and insights into the behaviour of the queueing systems being modelled, but instead needs to rely on the weight of

empirical evidence. This however can be very effective, especially in a practical situation where managers might react better to empirical evidence than to mathematically derived insights. For example, in a wholly simulation based study of queueing problems in UK accident and emergency (A&E) departments, Fletcher *et al* (2007) achieved general recognition that the NHS target requiring 98% of patients to complete their stay in A&E within 4 hours was reasonable, and general understanding of what needed to be done in individual departments for that to be possible, for example matching staffing levels to underlying arrival rates.

In a research setting the flexibility of simulation modelling is a major asset; it can provide a way of investigating the accuracy of many analytic formulations and solutions that have simplified a queueing problem, in some respect, to make the mathematical analysis feasible.

In a practical setting, the flexibility is also potentially very valuable in its ability to cater for particular features of the problem of interest. However, this also brings with it the serious risk of over-fitting, in that the inclusion of particular features combined with the necessarily superficial level of validation possible in most studies (eg does the model reproduce current performance measures with current parameter values?) can lead to misplaced faith in such models.

3.4. Mapping the three approaches onto the five dimensions of queues and queue modelling

In summary, Tables 1A and 1B overlay the three approaches to queue modelling onto the five dimensions of queues and queue modelling to provide a structured overview of queue modelling research and development to date. The two tables are used to distinguish the 'steady state *versus* time-dependent' dimension, the columns distinguish 'exponential *versus* non-exponential', the rows distinguish 'single node *versus* networks', and the three levels of number of servers are distinguished within the cells. In each of the cells defined by these four dimensions the modelling approaches used to tackle these problem types are identified, using the notation:

AF&F: *Analytic formulations and formulae*
AF&N: *Analytic formulations and numerical solution*
SIM: *Simulation Modelling*

Where the results obtained are approximations this is indicated. The fifth (catch-all) dimension of extra features is then added in the commentary.

It is immediately clear from Tables 1A and 1B that the most easy-to-use and insightful approaches (AF&F) are mainly restricted to the steady-state behaviour of exponential systems. The major exceptions to this pattern are the infinite server systems for which analytic formulations and formulae are available in all cells. However, unless there is an infinite number of servers, once we move away from the top two cells in Table 1A, that is, into cells more likely to match the requirements of real-life queues, analytic formulations lead to mainly approximate results, requiring numerical solution methods to be used. As noted before, but not repeated in every cell, simulation is also an option in all cases. Finally we note that the impact of the fifth dimension, that is extra features, is generally (but not always) that their introduction makes the modelling task more challenging than it would otherwise be.

Table 1A Modelling approaches available for steady-state behaviour of different types of queueing problems

	Exponential	*Non-exponential*
Single node	M/M/1: AF&F M/M/S: AF&F M/M/∞: AF&F	M/G/1: AF&F, AF&N M/G/S: AF&N (approx.) M/G/∞: AF&F
Network	Jackson: AF&F Other: AF&N (approx.) ∞ server: AF&F	Any: AF&N (approx.) ∞ server: AF&F

Table 1B Modelling approaches available for time-dependent behaviour of different types of queueing problems

	Exponential	*Non-exponential*
Single node	M(t)/M/1: AF&N M(t)/M/S: AF&N M(t)/M/∞: AF&F	M(t)/G/1: AF&N (approx.) M(t)/G/S: AF&N (approx.) M(t)/G/∞: AF&F
Network	Any: SIM ∞ server: AF&F	Any: SIM ∞ server: AF&F

4. Decision-making contexts for queue models

Before endeavouring to offer comment, in the final section of this paper, on queue modelling research and development to date and in the future, it is relevant to consider the types of decision-making contexts that require understanding and analysis of queueing systems, together with the needs of managers in those contexts. We first of all consider queue management decision making with respect to the usual strategic *versus* operational dimension.

4.1. Strategic versus operational

In terms of queueing systems, *strategic decisions* often concern the level of resources to provide in a new facility or existing facility, or the design or redesign of a facility. For example Edmond and Maggs (1978) discuss the use of queue models in decisions about investment in berth construction and cargo handling equipment, whereas Ashton *et al* (2005) were investigating possible designs for a new NHS walk-in centre. Such work usually requires estimates of current

system parameters (eg arrival rates and service times) or forecasts of their future levels.

Operational decisions for queueing systems often concern staffing levels for a next time period, quite possibly with a time of day or day of week component. Such decisions can be static, as in the case of the Griffiths *et al* (2005) work on rostering of supplementary nurses for an intensive care unit and Ding and Glazebrook (2005) who study the problem of warranty outsourcing; or *dynamic* as in the case of Ansell *et al* (2003) who investigate the allocation of different job types among a choice of service stations taking into account the degree of congestion at each service station. Such work, whether static or dynamic, again requires estimates of current system parameters or forecasts of their future levels.

Our second consideration in respect of the decision-making context for queue modelling is the distinction between optimizing models and decision supporting models.

4.2. Optimizing versus decision supporting

As recognized by Bailey (1957) and many others, queueing system decision making usually requires some balance between service level and cost. Sometimes costs can be associated with service levels and resources, in which case the queueing results can perhaps feed into some optimizing algorithm, as is the assumption made by Ansell *et al* (2003). Alternatively, resources can be treated as a constraint with the objective to maximize service level, as is the case in much call centre rostering work, see for example Green *et al* (2007).

In many real circumstances, however, queue models are required to support the decision-making process by indicating the likely consequences of particular options. In these situations, queue models may simply help identify good and bad ways of organizing a facility, as in much of the work on hospital appointment systems. Alternatively, as in the case of the Ashton *et al* (2005) investigation of possible designs for a new NHS walk-in centre, predicted performance levels were needed to be weighed alongside other decision criteria.

Our final consideration with respect to the decision-making context for queue modelling is to try to identify what managers actually want from the modeller.

4.3. What do managers want from queue modelling?

It is thought provoking to recognize that in practice, and perhaps in the manager's mind, nearly all queueing problems have already been solved—in the sense that the queueing system functions. So an initial issue for much queue modelling work could well be to interest the manager in the possibility of an improved solution for those cases where the current solution is not very good.

Once a manager of a queueing system, or of a bigger system which includes queues, is interested in the possibility of system improvements, 'models' will normally be used as some sort of evidence in support of decision making. As such 'models' can take a variety of forms:

- a general insight, for example in almost all queueing systems reducing random variation will lead to reduced congestion;
- a specific insight, for example in appointment systems it is generally good practice to reduce doctor lateness and double-book the first appointment time;
- a specific model, for example predictions of patient waiting times and doctor idleness under alternative appointment systems.

While much queue modelling in practice is undertaken with the third of these forms of evidence in mind, the actual value of the model may often be closer to the first two forms, in that it serves to highlight insights or underpin arguments that the decision makers need to use.

5. Past and future foci for queue modelling research and development

In this final section the previous dimensions, approaches, and perspectives are used to identify some past and possible future research and development themes. Some coincide with the dimensions of queues and queue modelling described earlier, others are themes that span a number of dimensions, and some suggest a different perspective.

5.1. Moving beyond exponential assumptions

This important move towards more realistic assumptions, particularly about the distribution of service times, has led, in particular, to research into the use of phase-type distributions described earlier. See Neuts (1995) for an account of many of the theoretical developments and van Ackere and Ninios (1993) for an example of their use in practice.

5.2. Time-dependent behaviour

As outlined earlier there has been considerable progress in modelling the time-dependent behaviour of single-node systems using PSA-based approximations, and using discrete time modelling. A major area of application for the PSA-based methods has been call centre staffing problems. A recent special issue of *Management Science* edited by Koole (2008) concentrates on call centre management and refers to much of the PSA-based work. Chassioti and Worthington (2004) outline the use of the discrete time approach in the same context.

The infinite server results for time-dependent behaviour of single-node and multi-node systems are remarkable for their simple formulaic form.

5.3. Queueing networks

As outlined earlier, research in this area has developed exact and approximate product-form solutions for a range of

networks at steady state. In some cases full decomposition of the network into independent queues is appropriate, in other cases the decomposition is partial. Apart from the simple infinite-server formulae, time-dependent analysis currently requires simulation.

5.4. Extra features

Research and development to incorporate special features into queueing systems are wide ranging, and examples that incorporate queueing system rules and customer behaviour have been described earlier. Because special features often complicate the formulation of such systems, exponential assumptions predominate.

5.5. Discrete time modelling

While not one of the previously identified dimensions, discrete time modelling has been a significant research focus across a number of the dimensions, for three main reasons. Some real-life systems are more correctly formulated in discrete time than in continuous time, particularly the performance of computer and communication systems. See Samanta *et al* (2007) for a particular example, and Bruneel and Kim (1993) for a fuller account. Secondly, discrete time models can be easier to analyse than continuous time models and can therefore provide a way to understand queueing issues that are too difficult to investigate in continuous time, see for example Atkinson (1995). Thirdly, not only are discrete time queues easier to analyse but, as noted earlier, they can be used to go beyond exponential distributions, for example by matching mean and variance, and are very suitable for modelling time-dependent behaviour.

5.6. Design and control of queueing systems

Research and development in this area are designed to reflect the decision-making context in which some queues need to be analysed, that is, where the aim is to optimize some overall system in which queues are an important component. Because of the needs of the associated optimizing algorithms, queueing results in these situations are often required to be formulaic and hence exponential assumptions are often made. However, given the current speed of numerical and simulation methods, recent work is starting to use these approaches as a practical and more accurate representation of the queue performance. See for example Feldman *et al* (2008) who tackle the problem of allocating staff in call centres and other facilities with time-varying demands.

In other cases, even when exponential assumptions are made, the stochastic systems are very complex to analyse and lead into other challenging mathematical areas, see for example Ding *et al* (2008) who investigate allocation models and heuristics for outsourcing of repairs under warranty.

5.7. Improving accessibility of models and results

This type of work is a longstanding concern of queue modellers and is more evident in the early years in the *Operational Research Quarterly* than in the pages of the *Journal of the Operational Research Society*. This work includes papers aimed more towards managers explaining the virtues of queue modelling in various contexts; for example, Taylor and Jackson (1954) on the provision of spare machines in the context of the airline industry; and the creation of queueing tables to overcome the practical difficulties of using some of the queueing theory results, for example, Cromie and Chaudhry (1976).

In more recent years some of this work has transferred to journals aimed at specific industries, or outlets such as *OR Insight* where the focus is on writing in a style more accessible to managers. Other progress relevant to accessibility includes the increasing availability of queueing and simulation software.

5.8. Better understanding and use of 'order of importance' ideas in model building

The Pollaczek–Khintchine formula shows that the only two aspects of a service time distribution that matter are its first two moments in respect of the mean number of customers in an M/G/1 queue. It also led to the hypothesis that for M/G/S and M(t)/G/S systems, good quality approximations would be obtained by matching the first two moments of service time. As noted earlier, subsequent empirical work has demonstrated that the hypothesis is valid for many practical queueing systems.

More recently research by Chassioti (2005) has shown empirical evidence that when a balking parameter is introduced into time-dependent queues the variance of service time becomes less important; and Green *et al* (2007) reported that performance measures such as probability of delay tend to be insensitive to the service time distribution beyond the mean in time-varying systems. Hence, there is some empirical evidence to support the idea that as the queueing systems under study become more complex, it may become more and more reasonable to represent service times more and more simply.

More complex queueing situations, and hence possible candidates for research along these lines, include queueing networks and the optimization of queueing systems.

5.9. Better understanding of robustness/sensitivity of results to uncertainty

The earlier discussion of the decision-making context for queueing models drew attention to the fact that many real applications have to use estimates of current system parameters or forecasts of their future levels. Given the obvious uncertainty associated with these values, the predicted performance of alternative queueing systems will clearly be inaccurate and the relative merits of alternative decisions may

well be wrong. Very little work is apparent in the literature to date that attempts to understand the implications for queue modelling of this type of uncertainty.

5.10. Drawing out insights for managers

The earlier analysis of decision-making contexts in which queue models are potentially useful, highlighted that managers often require models to provide insights as well as, or indeed rather than, detailed predictions of queue performance. One implication of this is that queue modellers should make more effort to produce insights for queue management from their research.

Some researchers do indeed attempt to draw insights, but little is known about what might constitute a useful and accessible insight for managers. This is clearly an issue that spans more than just queue modelling, and the *Journal Management Science* has recently started to require 'management insights' for each paper that it publishes. The recent special issue on call centres (Koole, 2008) therefore provides some examples in the queue modelling area.

5.11. Queue modelling as a discipline

There is much evidence throughout this review that analytic approaches (using both formulaic and numerical methods) and simulation approaches to queue modelling are complementary. With the ever increasing power of computers there is increasing scope for numerical methods and simulation models to be used alongside traditional queueing theory to help 'understand real life queueing systems as well as possible'.

In research, analytic queue modellers often use simulation to validate their approximations, and queueing theorists could use simulation to produce counter examples when trying to establish viable theorems. The possibility that numerical methods and simulation can each be used 'like formulae' in some circumstances opens the door for them to be incorporated into research involving more complicated formulations and algorithms.

In practice, queue modelling often goes either the analytic route or the simulation route because of the background of the modeller. Each can be inadequate on its own, with the analytic approach requiring unconvincing assumptions to be made, or the simulation approach risking over fitting or development of a sledgehammer to crack a nut. The possibility that numerical methods and simulation can be used 'like formulae' is also clearly very attractive in a practical situation.

The importance of integrating analytic and simulation approaches in research and in applications makes a strong case for a new discipline of 'queue modelling'. It could draw upon the combined strengths of analytic and simulation approaches to tackle queueing problems in research and in practice. Alongside these modelling approaches it could take up the challenge of understanding these systems from both the modelling and management perspectives, including the responsibility to draw out managerial insights, all with the purpose 'to understand real-life queueing systems as well as possible'.

References

Alfa AS (1979). A numerical method for evaluating delay to a customer in a time-inhomogeneous, single server queue with batch arrivals. *J Opl Res Soc* **30**: 665–667.

Ansell PS, Glazebrook KD and Kirkbride C (2003). Generalised 'join the shortest queue' policies for the dynamic routing of jobs to multi-class queues. *J Opl Res Soc* **54**: 379–389.

Ashton R, Hague L, Brandreth M, Worthington D and Cropper S (2005). A simulation-based study of a NHS walk-in centre. *J Opl Res Soc* **56**: 153–161.

Atkinson JB (1995). The general two-server queueing loss system: Discrete-time analysis and numerical approximation of continuous-time systems. *J Opl Res Soc* **46**: 386–397.

Bailey NTJ (1952). A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. *J Roy Stat Soc Ser B* **14**: 185–199.

Bailey NTJ (1957). Operational research in hospital planning and design. *Opl Res Q* **8**: 149–157.

Bhat UN (1969). Sixty years of queueing theory. *Mngt Sci* **15**: B280–B294.

Bhat UN, Shalaby M and Fischer MJ (1979). Approximation techniques in the solution of queueing problems. *Nav Res Log Q* **26**: 311–326.

Bitran GR and Dasu S (1992). A review of open queueing network models of manufacturing systems. *Queueing Syst Theory Appl* **12**: 95–134.

Brahimi M and Worthington DJ (1991). Queueing models for out-patient appointment systems – a case study. *J Opl Res Soc* **42**: 733–746.

Brockmeyer E, Halstrom HL and Jensen A (1948). *The Life and Works of A K Erlang*. J Jorgenson & Co.: Copenhagen.

Bruneel H and Kim BG (1993). *Discrete-time Models for Communication Systems Including ATM*. Kluwer Academic: Boston.

Bunday BD (1996). *An Introduction to Queueing Theory*. Arnold: London.

Chassioti E (2005). *Queueing models for call centres*. PhD Thesis, Management Science, Lancaster University.

Chassioti E and Worthington DJ (2004). A new model for call centre queue management. *J Opl Res Soc* **55**: 1352–1357.

Cromie MV and Chaudhry ML (1976). Analytically explicit results for the queueing system $M/M^x/c$ with charts and tables for certain measures of efficiency. *Opl Res Q* **27**: 733–745.

Ding L and Glazebrook KD (2005). A static allocation model for the outsourcing of warranty repairs. *J Opns Res Soc* **56**: 825–835.

Ding L, Glazebrook KD and Kirkbride C (2008). Allocation models and heuristics for the outsourcing of repairs for a dynamic warranty population. *Mngt Sci* **54**: 594–607.

Edmond ED and Maggs RP (1978). How useful are queue models in port investment decisions for container berths? *J Opl Res Soc* **29**: 741–750.

Eick SG, Massey WA and Whitt W (1993). The physics of the $Mt/G/\infty$ queue. *Opns Res* **41**: 731–742.

Faulkner JA (1968). The use of closed queues in the deployment of coal-face machinery. *Opl Res Q* **19**: 15–23.

Feldman Z, Mandelbaum A, Massey WA and Whitt W (2008). Staffing of time-varying queues to achieve time-stable performance. *Mngt Sci* **54**: 324–338.

Fletcher A, Halsall D, Huxham S and Worthington D (2007). The DH accident and emergency department model: A national generic model used locally. *J Opl Res Soc* **58**: 1554–1562.

Grassmann W (1977). Transient solutions in markovian queues: An algorithm for finding them and determining their waiting-time distributions. *Eur J Opl Res* **1**: 396–402.

Green DH and Hartley MG (1966). The simulation of some simple control policies for a signalized intersection. *Opl Res Q* **17**: 263–277.

Green LV, Kolesar PJ and Whitt W (2007). Coping with time-varying demand when setting staffing requirements for a service system. *Product Opn Mngt* **16**: 13–39.

Griffiths JD (1995). Queueing at the Suez Canal. *J Opl Res Soc* **46**: 1299–1309.

Griffiths JD, Holland W and Williams JE (1991). Estimation of queues at the Channel Tunnel. *J Opl Res Soc* **42**: 365–373.

Griffiths JD, Leonenko GM and Williams JE (2006). The transient solution to $M/E_k/1$ queue. *Opns Res Lett* **34**: 349–354.

Griffiths JD, Price-Lloyd N, Smithies M and Williams JE (2005). Modelling the requirement for supplementary nurses in an intensive care unit. *J Opl Res Soc* **56**: 126–133.

Groom KN (1977). Planning emergency ambulance services. *Opl Res Q* **28**: 641–651.

Gross D and Harris CM (1985). *Fundamentals of Queueing Theory*. John Wiley & Sons Inc.: New York.

Hasslinger G and Rieger ES (1996). Analysis of open discrete time queueing networks: A refined decomposition approach. *J Opl Res Soc* **47**: 640–653.

Jackson JR (1957). Networks of waiting lines. *Opns Res* **5**: 518–521.

Jackson JR (1963). Jobshop-like queueing systems. *Mngt Sci* **10**: 131–142.

Jackson RRP (1954). Queueing systems with phase type service. *Opl Res Q* **5**: 109–120.

Jackson RRP (1956). Random queueing processes with phase-type service. *J Roy Stat Soc Ser B* **18**: 129–132.

Jackson RRP, Welch JD and Fry J (1964). Appointment systems in hospitals and general practice: Design of an appointments system. *Opl Res Q* **15**: 219–237.

Ke JC and Wang KH (1999). Cost analysis of the M/M/R machine repair problem with balking, reneging, and server breakdowns. *J Opl Res Soc* **50**: 275–282.

Kimber RM, Daly P, Barton J and Giokas C (1986). Predicting time-dependent distributions of queues and delays for road traffic at roundabouts and priority junctions. *J Opl Res Soc* **37**: 87–97.

Koenigsberg E (1982). Twenty five years of cyclic queues and closed queue networks: A review. *J Opl Res Soc* **33**: 605–619.

Koole G (2008). Special issue on call center management. *Mngt Sci* **54**: 237.

Lawrie NL (1980). An application of queueing theory to a teletraffic problem. *J Opl Res Soc* **31**: 975–981.

Lee AM and Longton PA (1959). Queueing processes associated with airline passenger check-in. *Opl Res Q* **10**: 56–71.

Lee HS, Bouhchouch A, Dallery Y and Frein Y (1998). Performance evaluation of open queueing networks with arbitrary configuration and finite buffers. *Ann Opns Res* **79**: 181–206.

Massey WA and Whitt W (1993). Networks of infinite-server queues with nonstationary Poisson input. *Queueing Syst* **13**: 183–250.

Mayhew LD (1987). Resource inputs and performance outputs in social security offices. *J Opl Res Soc* **38**: 913–928.

Neuts MF (1975). Computational uses of the method of phases in the theory of queues. *Comput Math Appl* **1**: 151–166.

Neuts MF (1984). Matrix-analytic methods in queueing theory. *Eur J Opl Res* **15**: 2–12.

Neuts MF (1995). *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. John Hopkins University Press: Baltimore.

Parkan C (1987). Simulation of a fast-food operation where dissatisfied customers renege. *J Opl Res Soc* **38**: 137–148.

Perros HG (1994). *Queueing Networks with Blocking: Exact and Approximate Solutions*. Oxford University Press: New York.

Preater J (2001). *A bibliography of queues in health and medicine*. Keele University: UK.

Royston G, Halsall J, Halsall D and Braithwaite C (2003). Operational research for informed innovation: NHS Direct as a case study in the design, implementation and evaluation of a new public service. *J Opl Res Soc* **54**: 1022–1028.

Saaty TL (1966). Seven more years of queues. A lament and a bibliography. *Nav Res Log Q* **13**: 447–476.

Samanta SK, Gupta UC and Sharma RK (2007). Analysis of finite capacity discrete-time GI/Geo/1 queueing system with multiple vacations. *J Opl Res Soc* **58**: 368–377.

Sharma OP (1990). *Markovian Queues*. Ellis Horwood: Chichester.

Taylor J and Jackson RRP (1954). An application of the birth and death process to the provision of spare machines. *Opl Res Q* **5**: 95–108.

van Ackere A and Ninios P (1993). Simulation and queueing theory applied to a single-server queue with advertising and balking. *J Opl Res Soc* **44**: 407–414.

Wall AD and Worthington DJ (1994). Using discrete distributions to approximate general service time distributions in queueing models. *J Opl Res Soc* **45**: 1398–1404.

Whitt W (1983). The main paper: The queueing network analyzer. *Bell Syst Tech J* **92**: 2779–2815.

Whitt W (2006). Fluid models for multiserver queues with abandonments. *Opns Res* **54**: 37–54.

Williams TM (1980). Nonpreemptive multi-server priority queues. *J Opl Res Soc* **31**: 1105–1107.

Worthington DJ (1987). Queueing models for hospital waiting lists. *J Opl Res Soc* **38**: 413–422.

Worthington D and Wall A (1999). Using the discrete time modelling approach to evaluate the time-dependent behaviour of queueing systems. *J Opl Res Soc* **50**: 777–788.

Received September 2008;
accepted December 2008 after one revision