Maximum Entropy Analysis of Queueing Network Models

Demetres Kouvatsos
Computer Systems Modelling Research Group
University of Bradford
Bradford, BD7 1DP
England

Abstract

The principle of Maximum Entropy (ME) provides a consistent method of inference for estimating the form of an unknown discrete-state probability distribution, based on information expressed in terms of true expected values. In this tutorial paper entropy maximisation is used to characterise product-form approximations and resolution algorithms for arbitrary continuous-time and discrete-time Queueing Network Models (QNMs) at equilibrium under Repetitive-Service (RS) blocking and Arrivals First (AF) or Departures First (DF) buffer management policies. An ME application to the performance modelling of a shared-buffer Asynchronous Transfer Mode (ATM) switch architecture is also presented. The ME solutions are implemented subject to Generalised Exponential (GE) and Generalised Geometric (GGeo) queueing theoretic mean value constraints, as appropriate. In this context, single server GE and GGeo type queues in conjunction with associated effective flow streams (departure, splitting, merging) are used as *building blocks* in the solution process. Physical interpretations of the results are given and extensions to the quantitative analysis of more complex queueing networks are discussed.

1. Introduction

Queueing network models (QNMs) are widely recognised as powerful tools for representing discrete flow systems (such as computer, communication and flexible manufacturing systems) as complex networks of queues and servers and analysing their performance. Within this framework the servers represent the active or passive resources of the system such as processors, memory and communication devices and the customers circulating through the servers stand for the jobs, messages or components being processed by and competing for these resources.

Classical queueing theory provides a conventional framework for formulating and solving the QNM. The variability of interarrival and service times of jobs can be modelled by continuous-time or discrete-time probability distributions. Exact and approximate analytical methods have been proposed in the literature for solving equations describing system performance. These techniques lead to efficient computational algorithms for analysing QNMs and over the years a vast amount of

progress has been made worldwide (e.g., [1-19]). However, despite persistant attempts for generalisation some problems still remain without a satisfactory solution.

In the continuous-time domain the accuracy of analytic approximations for general QNMs, particularly those with finite capacity and multiple-job classes, may be adversely affected when based on intuitive heuristics, while very often gross assumptions are made in order to assure exact numerical solutions (e.g., [1-14]). Many theoretical advances on continuous-time queues have been extended to discrete-time queues (e.g., [15-19]). However, there are few standard and unique results that are known for discrete-time queues due to the inherent difficulties associated with the occurrence of simultaneous events at the boundary epochs of a slot including bulk arrivals and departures (c.f., [20]).

Since the mid-60s it has become increasingly evident that classical queueing theory cannot easily handle *by itself* complex queueing systems and networks with many interacting elements. As a consequence, alternative ideas and tools, analogous to those applied in the field of Statistical Mechanics, have been proposed in the literature (e.g., [21-28]). It can be argued that one of the most fundamental requirements in the analysis of complex queueing systems is the provision of a convincing interpretation for a probability assignment free from arbitrary assumptions. In a more general context, this was the motivation behind the principle of Maximum Entropy (ME), originally developed and thoroughly discussed by Jaynes [29-31] in Statistical Physics. The principle provides a self-consistent method of inference for estimating uniquely an unknown but true probability distribution, based on information expressed in terms of known true mean value constraints. It is based on the concept of the entropy functional introduced earlier in Information Theory by Shannon [32]. Over the recent years the principle of ME, subject to queueing theoretic constraints, has inspired a new and powerful analytic framework for the approximate analysis of complex queueing systems and arbitrary queueing networks (c.f., [22, 25-28, 33-64]).

This tutorial paper presents ME product-form approximations and resolution algorithms for both continuous-time and discrete-time First-Come-First-Served (FCFS) QNMs at equilibrium, subject to Repetitive-Service (RS) blocking mechanism with either Fixed (RS-FD) or Random (RS-RD) destination and Arrivals First (AF) or Departures First (DF) buffer management policies. An ME application to the performance analysis of a shared buffer Aynchronous Transfer Mode (ATM) switch architecture is also described. The ME solutions are implemented subject to queueing theoretic Generalised Exponential (GE) and Generalised Geometric (GGeo) mean value constraints. In this context, single server GE and GGeo-type queues in conjunction with associated flow streams (departure, splitting, merging) play the role of *building blocks* in the solution process.

The principle of ME is introduced in Section 2. The GE and GGeo distributions and related formulae for merging of flow streams are described in Section 3. The ME analysis of GE and GGeo type queues in conjunction with interdeparture-time flow formulae are presented in Section 4. The ME product form approximations and resolution algorithms for arbitrary QNMs are reviewed in Section 5. A ME application to a shared buffer ATM switch architecture is carried out in Section 6. Conclusions are given in the last Section, followed by an annotated bibliography.

Remarks: RS blocking occurs when a job upon service completion at queue i attempts to join a destination queue j whose capacity is full. Consequently, the job is rejected by queue j and immediately receives another service at queue i. In the case of RS-FD blocking this is repeated until the job completes service at a moment where the destination queue j is not full. In the RS-RD case each time the job completes service at queue i, a downstream queue is selected independently of the previously chosen destination queue j. Moreover, AF and DF policies for discrete-time queues stipulate how the buffer is filled or emplied in case of simultaneous arrivals and departures at a boundary epoch of a slot (or unit time interval). Under AF arrivals take precedence over departures while the reverse takes place under DF (c.f., Fig. 1).

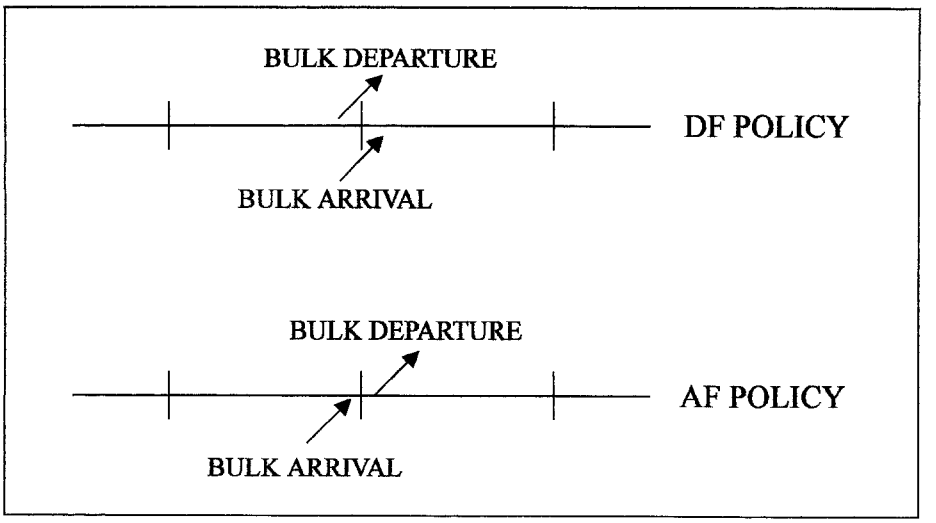

Fig. 1. AF and DF buffer management policies per slot

2. The Principle of ME

2.1 Formalism

Consider a system Q that has a set of possible discrete states $S = \{S_0, S_1, S_2, ...\}$ which may be finite or countable infinite and state S_n, n=0,1,2,... may be specified arbitrarily. Suppose the available information about Q places a number of constraints on $p(S_n)$, the probability distribution that the system Q is at state S_n. Without loss of generality, it is assumed that these take the form of mean values of several suitable functions $\{f_1(S_n), f_2(S_n), ..., f_m(S_n)\}$, where m is less than the number of possible states. The principle of maximum entropy (ME) [29-31] states that, of all distributions satisfying the constraints supplied by the given information, the minimally prejudiced distribution $p(S_n)$ is the one that maximises the system's entropy function

$$H(p) = -\sum_{S_n \in S} p(S_n) \ln\{p(S_n)\}, \tag{2.1}$$

subject to the constraints

$$\sum_{S_n \in S} p(S_n) = 1, \tag{2.2}$$

$$\sum_{S_n \in S} f_k(S_n) p(S_n) = \langle f_k \rangle, \quad k=1,2,...,m, \tag{2.3}$$

where $\{\langle f_k \rangle\}$ are the prescribed mean values defined on the set of functions $\{f_k(S_n)\}$, k=1,2,...,m. Note that in a stochastic context, for example, these functions may be defined on the state space S of a Markov process with states $\{S_n\}$, $n \geq 0$, and $p(S_n)$ can be interpreted as the asymptotic probability distribution of state S_n at equilibrium.

The maximisation of H(p), subject to constraints (2.2)-(2.3), can be carried out using Lagrange's Method of Undetermined Multipliers leading to the solution

$$p(S_n) = \frac{1}{Z} \exp\left\{ -\sum_{k=1}^{m} \beta_k f_k(S_n) \right\}, \tag{2.4}$$

where $\exp\{\beta_k\}$, k=1,2,...,m are the Lagrangian coefficients determined from the set of constraints $<f_k>$, and Z, known in statistical physics as the *partition function* (or normalising constant), is given by

$$Z = \exp\{\beta_0\} = \sum_{S_n \in S} \exp\left\{ - \sum_{k=1}^{m} \beta_k f_k(S_n) \right\}, \tag{2.5}$$

where β_0 is a Lagrangian multiplier specified by the normalisation constraint. It can be verified that the Lagrangian multipliers $\{\beta_k\}$, k=1,2,...,m satisfy relations:

$$- \frac{\partial \beta_0}{\partial \beta_k} = <f_k>, \quad k=1,2,...,m, \tag{2.6}$$

while the ME functional can be expressed by

$$\max_p H(p) = \beta_0 + \sum_{k=1}^{m} \beta_k <f_k>. \tag{2.7}$$

Although it is not generally possible to solve (2.6) for $\{\beta_k\}$ explicitly in terms of $\{<f_k>\}$, numerical methods for obtaining approximate solutions are available. When system Q has a countable infinite set of states, S, the entropy function H(p) is an infinite series having no upper limit, even under the normalisation constraint. However, the added expected values $\{<f_k>\}$ of (2.3) introduce the upper bound (2.7) and the ME solution $\{p(S_n)\}$ exists.

The characterisation of a closed-form ME solution requires the priori estimation of the above multipliers in terms of constraints $\{<c_k>\}$. Note that these constraints may not all be known a priori; but it may be known that these constraints exist. This information, therefore, can be incorporated into the ME formalism in order to characterise the form of the state probability (2.4). As a result, the mean value constraints may become explicit parameters of the ME solution. The analytic implementation of this solution, however, clearly requires the priori calculation of these constraints via queueing theoretic (or even operational) exact of approximate formulae expressed in terms of basic system parameters.

2.2 Justification of ME Principle

The principle of ME has its roots in Bernoulli's principle of *Insufficient Reason (IR)* implying that *the outcomes of an event should be considered initially equally probable unless there is evidence to make us think otherwise.* The entropy functional H(p) may be informally interpreted as the expected amount of uncertainty that exists prior to the system occupying anyone of its states. For a finite set of states $\{S_n\}$, H(p) reaches its maximum (2.7) when all outcomes of an event are equally probable (i.e., prior to the execution of an experiment one is faced with maximum uncertainty on which outcome will be realised). To this end, one should initially start with the distribution of *IF* (i.e., a uniform type distribution) and then *adjust* this distribution to maximise the entropy if prior constraint information is known. In this context, the principle of ME may be stated as *Given the propositions of an event and any information relating to them, the best estimate for the corresponding probabilities is the distribution that maximised the entropy subject to the available information.*

In an information theoretic context [29], the ME solution corresponds to the maximum disorder to system states, and thus is considered to be the least biased distribution estimate of all solutions that satisfy the system's constraints. In sampling terms, Jaynes [30] has shown that, given the imposed constraints, the ME solution can be experimentally realised in overwhelmingly more ways than any other distribution. Major discrepancies between the ME distribution and the experimentally observed distribution indicate that important physical constraints have been overlooked. Conversely, experimental agreement with the ME solution represents evidence that the constraints of the system have been properly identified.

More details on the principle of ME can be found in Tribus [65]. A generalisation to the principle of Minimum Relative Entropy (MRE) - requiring, in addition, a prior estimate of the unknown distribution - can be seen in Shore and Johnson [66].

2.3 ME Analysis in Systems Modelling

In the field of systems modelling expected values of various performance distributions of interest, such as the number of jobs in each resource queue concerned, are often known, or may be explicitly derived, in terms of moments of interarrival and service time distributions. Note that the determination of the distributions themselves, via classical queueing theory, may prove an infeasible task even for systems of queues with moderate complexity. Hence, the methodology of entropy maximisation may be applied to characterise useful information theoretic approximations of performance distributions of queueing systems and networks.

Focusing on a general QNM, the ME solution (2.4) may be interpreted as a product-form approximation, subject to the set of mean values $\{<f_k>\}$, k=1,2,...,m, viewed as marginal type constraints per queue. Thus, for an open QNM, entropy maximisation suggests a decomposition into individual queues with revised interarrival and service times. The marginal ME solutions of these queues, in conjunction with related formulae for the first two moments of the effective flow, can play the role of building blocks towards the computation of the performance metrics (c.f., [27]). For a closed QNM, the implementation of ME solution (2.4) clearly requires the a priori estimation of the Lagrangian coefficients $\exp\{\beta_k\}$, k=1,2,...,m. To this end, a modified algorithm of an open network satisfying the principles of flow and population conservation (*pseudo* open network) may be used in conjunction with a convolution type procedure for the estimation of the performance metrics (c.f., [28]).

3. The GE and GGeo Distributional Models

3.1 The GE Distribution

The GE distribution is of the form

$$F(t) = P(W \le t) = 1 - \tau e^{-\sigma t}, t \ge 0, \quad (3.1)$$

where

$$\tau = 2/(C^2 + 1), \quad (3.2)$$

$$\sigma = \tau\nu, \quad (3.3)$$

W is a mixed-time random variable (rv) of the interevent-time, while $1/\nu$ is the mean and C^2 is the squared coefficient of variation (SCV) of rv W(c.f., Fig 2).

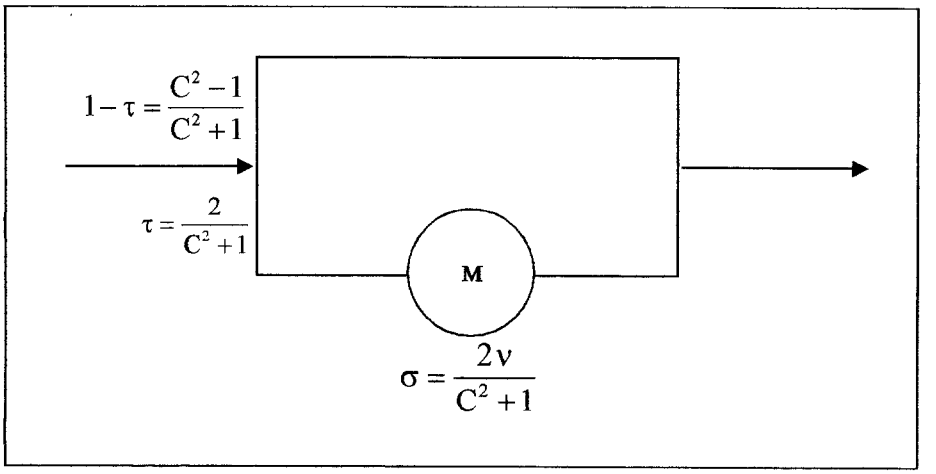

Fig. 2. The GE distribution with parameters τ and σ.

For $C^2 > 1$, the GE model (3.1) is a mixed-time probability distribution and it can be interpreted as either

(i) an extremal case of the family of two-phase exponential distributions (e.g., Hyperexponential-2 (H_2)) having the same ν and C^2, where one of the two phases has zero service time; or

(ii) a bulk type distribution with an underlying counting process equivalent to a Compound Poisson Process (CPP) with parameter $2\nu/C^2 + 1)$ and geometrically destributed bulk sizes with mean $= (C^2 + 1)/2$ and SCV $= (C^2 - 1)/(C^2 + 1))$ given by

$$P(N_{cp} = n) = \begin{cases} \sum_{i=1}^{n} \frac{\sigma^i}{i!} e^{-\sigma} \binom{n-1}{i-1} \tau^i (1-\tau)^{n-i}, & \text{if } n \geq 1, \\ e^{-\sigma}, & \text{if } n = 0, \end{cases} \tag{3.4}$$

where N_{cp} is a Compound Poisson (CP) rv of the number of events per unit time corresponding to a stationary GE-type intervent rv.

By using the bulk interpretation of the GE-type distribution and applying the Law of Total Probability it can be shown that merging M GE-type streams with parameters (ν_i, C_i^2), i=1,2,...,M results in a GE-type overall stream. This stream, however, generally corresponds to a non-renewal CPP process (with non-geometric bulk sizes) - unless all C_i^2's are equal - with parameters (τ, C^2) determined by [27, 28].

$$\nu = \sum_{i=1}^{M} \nu_i, \tag{3.5}$$

$$C^2 = -1 + \left[\sum_{i=1}^{M} \frac{\nu_i}{\nu} (C_i^2 + 1)^{-1} \right]^{-1}. \tag{3.6}$$

3.2 The GGeo Distribution

The GGeo distribution is of the form

$$f_n = P(Y = n) = \begin{cases} 1 - \tau, & \text{if } n = 0, \\ \tau\sigma(1-\sigma)^{n-1}, & \text{if } n \geq 1, \end{cases} \tag{3.7}$$

where

$$\tau = 2/(C^2 + 1 + \nu)\,, \tag{3.8}$$

$$\sigma = \tau\nu\,, \tag{3.9}$$

Y is a discrete-time rv of the interevent-time, while $1/\nu$ and C^2 are the mean and SCV of rv Y (c.f., Fig. 3).

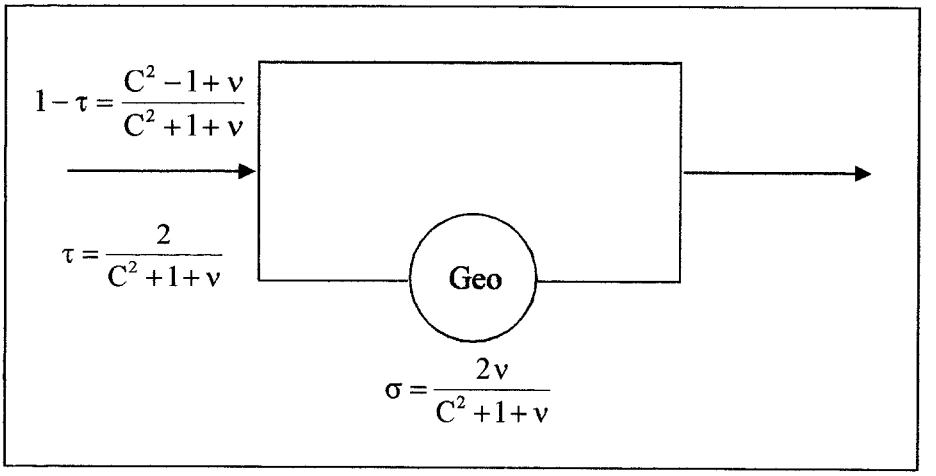

Fig. 3. The GGeo distribution with parameters τ and σ.

For $C^2 \geq |\,1 - \nu\,|$, the GGeo model (3.7) is a discrete-time probability distribution implying a bulk intervent pattern according to a Bulk Bernoulli Process (BBP) with a rate σ, while the number of events (e.g, arrivals or departures) at the boundary epochs of a slot is geometrically distributed with parameter τ. Thus, the GGeo distribution is generated by a sequence of bulk Bernoulli independent and identically distributed non-negative integer valued rv.'s $\{Y_k\}$, where Y_k is the number of events in a slot, with a probability distribution given by

$$g_\ell = \begin{cases} P(Y_k = 0) = 1 - \sigma\,, & \text{if } \ell = 0\,, \\ P(Y_k = \ell) = \sigma\tau(1-\tau)^{\ell-1}\,, & \text{if } \ell \geq 1\,. \end{cases} \tag{3.10}$$

It can be verified via the Law of Total Probability that merging M GGeo-type streams with parameters (ν_i, C_i^2), i=1,2,...,M, the corresponding overall parameters (ν, C^2) are given by (3.5) and

$$C^2 = -1 - \nu + \left[\sum_{i=1}^{M} \frac{\nu_i}{\nu} (C_i^2 + \nu_i + 1)^{-1} \right]^{-1}, \qquad (3.11)$$

respectively [60, 64]. Note, however, that the resulting distribution is not of GGeo-type.

Remarks: The GE and GGeo distributions are versatile, possessing pseudo-memoryless properties which make the solution of many queueing systems and networks analytically tractable. The GE and GGeo cannot be physically interpreted as stochastic models outside the ranges $C^2 \geq 1$ and $C^2 \geq |1 - \nu|$, respectively. However, they can be meaningfully considered as pseudo-distributions of flow model approximations of stochastic models, in which negative branching pseudo-probabilities (or weights) are permitted. The utility of other pseudo-distributions in systems modelling has been pointed out in [67, 68].

4. The ME Building Blocks: GE and GGeo Type Single Server Queues

This Section determines the form of ME solutions for stable G/G/1 and G/G/1/N queues, subject to queueing theoretic GE and GGeo type constraints. Moreover, it presents closed-form expressions for the SCV of the interdeparture-time distribution of the GE/GE/1 and GGeo/GGeo/1 queues at equilibrium.

4.1 A Stable ME G/G/1 Queue

Consider a stable FCFS G/G/1 queue with infinite capacity where jobs belong to a single class and arrive according to an arbitrary interarrival-time distribution with mean $1/\lambda$ and SCV, Ca^2. Moreover, they are served by a single server having a general service-time distribution with mean $1/\mu$ and SCV, Cs^2. Let at any given time the state of the system be described by the number of jobs present (waiting or receiving service) and p(n) be the steady-state probability that the G/G/1 queue is at state n, n=0,1,2,... .

The form of a universal ME solution, p(n), for either a continuous-time or discrete-time G/G/1 queue at equilibrium can be obtained by maximising the entropy functional, H(p), subject to the following constraints:

(a) Normalisation

$$\sum_{n=0}^{\infty} p(n) = 1 .$$

(b) Server utilisation (UTIL), $\rho (0 < \rho = \lambda/\mu < 1)$,

$$\sum_{n=1}^{\infty} h(n)\, p(n) = \rho ,$$

where $h(n) = 0$, if $n = 0$, or 1, if $n \geq 1$.

(c) Mean Queue Length (MQL), $L(\rho < L < +\infty)$,

$$\sum_{n=1}^{\infty} n\, p(n) = L .$$

The maximisation of H(p), subject to constraints (a)-(c), can be obtained via Langrange's Method of Undetermined Mulitipliers leading to the following solution:

$$p(n) = \begin{cases} 1 - \rho , & \text{if } n = 0 , \\ (1 - \rho) g\, x^{n} , & \text{if } n \geq 1 , \end{cases} \tag{4.1}$$

where g and x are the Lagrangian coefficients corresponding to constraints ρ and L, respectively, and are given by

$$g = \frac{(1 - x)\, \rho}{(1 - \rho)\, x} = \frac{\rho^2}{(L - \rho)\, (1 - \rho)} , \tag{4.2}$$

and

$$x = \frac{L - \rho}{L} . \tag{4.3}$$

The ME solution (4.1) can be rewritten, via (4.2), as a GGeo state probability distribution with parameters ρ and $1 - x$, namely

$$p(n) = \begin{cases} 1 - \rho , & \text{if } n = 0 , \\ \rho (1 - x) x^{n-1} , & \text{if } n \geq 1 . \end{cases} \tag{4.4}$$

The implementation of the ME solution (4.4) depends on the analytic determination of the MQL, L. This is achieved in the next Section by focusing on the GE/G/1 and GGeo/G/1 queues at equilibrium.

4.1.1 The GE and GGeo Type Information Constraints

Consider GE/G/1 and GGeo/G/1 queues at equilibrium. By applying the generalised Laplace and z-transform equations, respectively, it can be shown [39, 60] that for either queue the random observer's MQL is given by the same formula, namely

$$L = \frac{\rho}{2}\left[1 + \frac{Ca^2 + \rho\ Cs^2}{1 - \rho}\right], \qquad (4.5)$$

where for $L \geq \rho$ it follows that $\rho \geq (1 - Ca^2)/(1 + Cs^2)$. It is, therefore, implied that the ME solution of a G/G/1 queue is insensitive with respect to GE and GGeo interarrival-time distributions. Moreover, it can be established, via the generalised embedded Markov chain approach for continuous-time and discrete-time queues, that the ME solution, p(n) n=0,1,2,... becomes exact if the underlying service time distributions of the GE/G/1 and GGeo/G/1 queues at equilibrium are also of $GE(\mu, Cs^2)$ and $GGeo(\mu, Cs^2)$ types, respectively (c.f., [39, 60]).

In the next Section closed-form expressions for the SCV of the interdeparture-time process are established.

4.1.2 The GE and GGeo Type Parameters of the Interdeparture Process

Consider the GE/GE/1 and GGeo/GGeo/1 queues at equilibrium. By applying the Laplace transform and z-transform equations of the interdeparture-time distribution (within continuous-time and discrete-time domains), respectively, it can be shown that the mean departure rate is given by λ, while the SCV, Cd^2, is determined by

(i) $$Cd^2 = \rho(1 - \rho) + (1 - \rho)\,Ca^2 + \rho^2\,Cs^2, \qquad (4.6)$$

for a stable GE/GE/1 queue [27] and

(ii) $$Cd^2 = \rho^2(Cs^2 + \mu - 1) - \rho(Ca^2 + \lambda - 1) + Ca^2, \qquad (4.7)$$

for a stable GGeo/GGeo/1 queue under both AF and DF buffer management policies [60].

4.2 A Stable ME G/G/1/N Queue

Consider a single class FCFS G/G/1/N censored queue at equilibrium with finite capacity, N and general interarrival and service time distributions with known first two moments. The notation of Section 4.1 applies, as appropriate. It is assumed that

the arrival process is *censored*, i.e., arriving jobs are turned away when the buffer is full.

The form of a universal ME solution p(n), n=0,1,2,...,N, applicable to either a continuous-time or discrete-time G/G/1/N queue at equilibrium, can be established, subject to the constraints of

(a) Normalisation

$$\sum_{n=0}^{N} p(n) = 1 ,$$

(b) UTIL, υ $(0 < \upsilon < 1)$,

$$\sum_{n=1}^{N} h(n)\, p(n) = \upsilon ,$$

(c) MQL, Ω $(\upsilon \le \Omega < N)$,

$$\sum_{n=1}^{N} n\, p(n) = \Omega$$

(d) Full buffer state probability, $\varphi = p(N)$ $(0 < \varphi < 1)$,

$$\sum_{n=0}^{N} f(n)\, p(n) = \varphi ,$$

where f(n) = 0 , if $0 \le n \le N-1$, or 1, if n = N, satisfying the flow-balance condition

$$\lambda(1 - \pi) = \mu\, \upsilon , \qquad (4.8)$$

where π is the blocking probability that a tagged job within an arriving bulk will find a full buffer.

By applying the Method of Lagrange's Undetermined Mulitipliers the ME solution, p(n), subject to constraints (a)-(d), is expressed by

$$p(n) = \begin{cases} 1/Z\,, & \text{if } n=0\,, \\ (1/Z) g\, x^{n}, & \text{if } n=1,2,\ldots,N-1\,, \\ (1/Z) g\, x^{n}\, y, & \text{if } n=N\,, \end{cases} \tag{4.9}$$

where Z is the normalising constant given by

$$Z = 1 + g\, x\, \frac{1 - x^{N-1}}{1 - x} + g\, y\, x^{N}\,, \tag{4.10}$$

and {g, x, y} are the Lagrangian coefficients corresponding to constraints υ, Ω and φ, respectively. By making asymptotic connections to infinite capacity G/G/1 queues at equilibrium as $N \rightarrow +\infty$, the Lagrangian coefficients g and x are assumed invariant with respect to N and are given by (4.2) and (4.3), respectively, while Ω reduces to L and is given by (4.5). The determination of Lagrangian coefficient y, however, depends on the blocking probability of a tagged job, π and the probability of an arriving bulk to find a n jobs in the system, $p_a(n)$, n=1,2,...,N. These probabilities are determined in the next Section by considering GE/GE/1/N and GGeo/GGeo/1/N queues at equilibrium.

4.2.1 The Blocking Probability π

Consider a stable GE/GE/1/N queue or a GGEo/GGeo/1/N queue under AF or DF buffer management policy. An universal analytic expression for blocking probability π can be established in terms of Lagrangian coefficients {x, g, y} by applying probabilistic arguments focusing on a tagged job within an arriving bulk. Clearly, the blocking probability π is conditioned on the position of a tagged job within the bulk and the number of jobs a bulk finds on arrival in the system. The following two distinct and exhaustive blocking cases are considered:

(i) The arriving bulk finds the queue empty with probability $p_a(0)$ and the tagged job is blocked. This event occurs with probability

$$\pi(N_a=0, N_t=N) = p_a(0) \sum_{j=N+1}^{\infty} j \left[\frac{\tau_a (1-\tau_a)^{j-1}}{1/\tau_a} \right] \sum_{k=0}^{j-(N+1)} \tau_s (1-\tau_s)^{k} \frac{j-N-k}{j}\,, \tag{4.11}$$

where N_a, N_t are rv.'s representing the number of jobs seen in the queue by an arriving bulk and its tagged job, respectively, $j\,\tau_a^2(1-\tau_a)^{j-1}$ is the probability that the bulk containing the tagged job has size j, $\tau_s(1-\tau_s)^k$ is the probability that the first k jobs depart through the zero service branch of $GE(\tau_s,\sigma_s)$ or GGeo (τ_s,σ_s) distribution and (j-N-k)/j is the probability that the tagged job occupies one of the positions N+k+1, N+k+2, ..., j within the bulk (i.e., it is blocked). Expression (4.11) can be written simplified to the following compact form:

$$\pi(N_a = 0, N_t = N) = \delta\, p_a(0)\,(1-\tau_a)^N, \tag{4.12}$$

where $\delta = \tau_s / (\tau_s(1-\tau_a) + \tau_a)$.

(ii) An arriving bulk finds on arrival n jobs in the queue, n=1,2,...,N (including the one in the Exponential or Geo service branch of a $GE(\tau_s,\sigma_s)$ or GGeo (τ_s,σ_s) distribution, respectively) and the tagged job is blocked. By applying similar arguments to those of case (i), this event takes place with probability

$$\pi\,(N_a \geq 1, N_t = N) = \sum_{n=1}^{N} (1-\tau_a)^{N-n}\, p_a(n), \tag{4.13}$$

where $p_a(n)$ is the probability that the arriving bulk finds n($\geq$1) jobs in the queue.

Thus, a general form of the blocking probability π of a GE/GE/1/N or GGeo/GGeo/1/N queue (under AF or DF policies) can be expressed as the sum of the probabilities (4.12) and (4.13), namely,

$$\pi = \delta\, p_a(0)\,(1-\tau_a)^N + \sum_{n=1}^{N} (1-\tau_a)^{N-n}\, p_a(n). \tag{4.14}$$

4.2.2 The Arriver's Probability, $p_a(n)$, $n \geq 1$

The arriving bulk of a GE/GE/1/N queue or GGeo/GGeo/1/N queue under AF policy has the same steady state probability as that of the random observer's ME solution

(4.4), i.e., $p_a(n) = p(n)\ \forall n$ (c.f., [20, 54]). However, for a censored GGeo/GGeo/1/N queue under DF buffer management policy, the following mutually exclusive events are distinguished:

(i) $\{N_a = 0\}$

An arriving bulk finds an empty system (with probability $p_a(0)$). Under DF policy the state of the queue before any departure(s) take place (i.e., outside observer's view point) can only be in one of the following cases:

(i_1) The queue was idle with probability p(0);

(i_2) The queue was in state n ($1 \le n \le N$) with probability p(n), n = 1,2,...,N and, therefore, all cells of the queue departed at the end of the slot together with the one in Geo service. This even occurs with probability

$$\sigma_s(1 - \tau_s)^{n-1} .$$

Thus,

$$p_a(0) = p(0) + \sigma_s \sum_{n=1}^{N} (1 - \tau_s)^{n-1} p(n), \qquad (4.15)$$

(ii) $\{N_a = n \ge 1\}$

An arriving bulk finds n(n = 1,2,...,N) jobs in the queue. In a similar fashion to case (i_2), just before any departure(s) take place (under DF policy) there were k($k \ge n$) cells in the queue with random observer's probability p(k). Two mutually exclusive cases are encountered in this situation:

(ii_1) $\{k = n\}$

The cell in Geo service stays there for another service slot with probability

$$1 - \sigma_s$$

(ii_2) $\{k \ge n + 1\}$

The cell in Geo service departs (with probability σ_s) together with k - (n + 1) the cells who leave through the zero service branch of the GGeo distribution with probability

$$(1 - \tau_s)^{k-n-1} \tau_s .$$

Combining cases (ii_1) and (ii_2) - and applying the Law of Total Probability, it follows that

$$p_a(n) = (1 - \sigma_s)\, p(n) + \sigma_s \sum_{k=n+1}^{N} \tau_s (1 - \tau_s)^{k-n-1} p(k). \qquad (4.16)$$

Substituting (4.15) and (4.16) into (4.14) and after some algebraic manipulation, it follows that

$$\pi = \delta\, p(0) (1 - \tau_a)^N + (1 - \sigma_s) \sum_{k=1}^{N} p(k) (1 - \tau_a)^{N-\kappa}$$
$$+ \sigma_s\, \delta (1 - \tau_a) \sum_{k=1}^{N} p(k) (1 - \tau_a)^{N-k} . \qquad (4.17)$$

4.2.3 The Lagrangian Coefficient, y

The Lagrangian coefficient y corresponding to the full buffer state probability can be determined by substituting π into the flow balance condition (4.8) and solving with respect y. After some manipulation it can be verified that for a GE/GE/1/N queue and GGeo/GGeo/1/N queue under a DF policy, y is given by a universal form, namely,

$$y = \frac{1}{1 - (1 - \tau_s)x} . \qquad (4.18)$$

For a GGeo/GGeo/1/N queue under AF policy it can be shown that y is given by

$$y = \frac{1-\rho}{1-x} + (1-\tau_a) \left[\frac{\rho}{1-\tau_a - x} - \left(\frac{1-\rho}{1-x}\, \delta + \frac{\rho}{1-\tau_a - x} \right) \left(\frac{1-\tau_a}{x} \right)^{N-1} \right] . \qquad (4.19)$$

Remarks: It can be shown that the ME solution (4.9) satisfies the global balance equations of the censored GE/GE/1/N and GGeo/GGeo/1/N queues at equilibrium (c.f., 53, 54]). Moreover, these ME solutions can be used as building blocks in the performance analysis of some multi-buffered ATM switch architectures (e.g., [63]).

5. Entropy Maximisation and Arbitrary QNMs

5.1 Open QNMs with RS Blocking

Consider an arbitrary open queueing network under RS-RD or RS-FD blocking mechanisms consisting of M FCFS single server queues with general external interarrival and service times. The notation of Section 4 is adopted with the incorporation of subscript i. At any given time the state of the network is described by a vector $\underline{n} = (n_1, n_2, ..., n_M)$, where n_i denotes the number of jobs at queue i, i=1,2,...,M, such that $0 \le n_i \le N_i$. Let $p(\underline{n})$ be the equilibrium probability that the queueing network is in state $\underline{n}$.

It can be shown that the ME solution, $p(\underline{n})$, subject to normalisation and marginal constraints of the type (b)-(d), namely υ_i, Ω_i and φ_i (i=1,2,...,M), is given by the product-form approximation (c.f., [53, 54, 58])

$$p(\underline{n}) = \prod_{i=1}^{M} p_i(n_i), \tag{5.1}$$

where $p_i(n_i)$ is the marginal ME solution of a censored G/G/1/N_i queue i, i=1,2,...,M (c.f., Section 4.2). Thus, entropy maximisation suggests a decomposition of the original open network into individual censored G/G/1/N_i queues with revised interarrival-time and service-time distributions.

Assuming that all flow processes (i.e., merge, split, departure) are renewal and their interevent-time distributions being of GE or GGeo type, each queue i, i=1,2,...,M, can be seen as a censored GE/GE/1/N_i or GGeo/GGeo/1/N_i queue with (a) GE or GGeo overall interarrival process (including *rejected* jobs), formed by the merging of departing streams towards queue i, generated by queues $\{j\}$, $\forall j \in A_i$, where A_i is the set of upstream queues of queue i (which may include outside world *queue 0*) and (b) an effective service-time distribution reflecting the total time during which a server of queue i is occupied by a particular job.

The rate and SCV of the effective service-time depend on (a) the type of RS blocking mechanism enforced, and (b) all blocking probabilities π_{ij}, $i \neq j$, $\forall j \in D_i$ (i.e., the probabilities that a completer from queue i is blocked by queue j ($\neq$ i)), where D_j is the set of all downstream queues of queue i. Using the properties of the GE or GGeo distribution it can be verified that the rate $\hat{\mu}_i$, and SCV, $\hat{Cs}_i^2$, of the effective service time are determined for all i=1,2,...,M by [54].

$$\hat{\mu}_i = \begin{cases} \mu_i(1 - \pi_{ci}) , & \text{if RS-RD} , \\ \mu_i \left[\sum_{j \in D_i} \frac{\alpha_{ij}}{1 - \pi_{ij}} \right]^{-1} , & \text{if RS-FD} , \end{cases} \tag{5.2}$$

$$\hat{Cs}_i^2 = \begin{cases} \pi_{ci} + Cs_i^2 (1 - \pi_{ci}) , & \text{if RS-RD,} \\ -1 + \frac{Cs_i^2}{\sum_{j \in D_i} \frac{\alpha_{ij}}{(1 - \pi_{ij})}} + \frac{\sum_{j \in D_i} \frac{\alpha_{ij}(1 + \pi_{ij})}{(1 - \pi_{ij})^2}}{\left[\sum_{j \in D_i} \frac{\alpha_{ij}}{(1 - \pi_{ij})} \right]^2} , & \text{if RS-FD,} \end{cases} \tag{5.3}$$

where π_{ci} is the blocking probability that a completer from queue i is blocked under RS-RD blocking mechanism, i.e., $\pi_{ci} = \sum_{j \in D_i} \alpha_{ij} \pi_{ij}$, with $\{\alpha_{ij}\}$, i=1,2,...,M ($i \neq j$) being the associated transition probabilities. Note that only in the case of RS-RD blocking the effective service-time is represented exactly by a GE or GGeo distribution with parameters $\hat{\mu}_i$ and $\hat{Cs}_i^2$.

The effective arrival stream of jobs at the GE/GE/1/N_i or GGeo/GGeo/1/N_i queue can be seen as the result of a two-way splitting of overall merging stream with parameter π_i (i.e., the blocking probability of either an external arriver or a completer in the

network is blocked by queue i). Denoting $\hat{\lambda}_i$ and $\hat{Ca}_i^2$ as the parameters of the effective interarrival process, it follows that the corresponding parameters of the overall interarrival process are given by

$$\lambda_i = \frac{\hat{\lambda}_i}{1 - \pi_i}, \quad Ca_i^2 = \frac{\hat{Ca}_i^2 - \pi_i}{1 - \pi_i}, \quad i=1,2,...,M. \tag{5.4}$$

Moreover, $\{\hat{\lambda}_i\}$, i=1,2,...,M, must satisfy the effective job flow rate equations, i.e.,

$$\hat{\lambda}_i = \hat{\lambda}_{0i} + \sum_{j=1}^{M} \hat{\alpha}_{ji} \hat{\lambda}_j, \quad i=1,2,...,M, \tag{5.5}$$

where $\hat{\lambda}_{0i}$ is the effective external arrival rate and $\hat{\alpha}_{ji}$ is the effective transition probability [58] given by

$$\hat{\alpha}_{ji} = \begin{cases} \dfrac{\alpha_{ji} (1 - \pi_{ji})}{(1 - \pi_{cj})}, & \text{if RS-RD}, \\ \alpha_{ji}, & \text{if RS-FD}, \end{cases} \tag{5.6}$$

j=1,2,...,M, $i \in D_j$ (n.b., $\pi_{j0} \equiv 0$).

The SCV of the effective interarrival process of each queue i, $\hat{Ca}_i^2$, i=1,2,...,M, can be approximated within a continuous-time or a discrete-time domain by applying the merging GE-type or GGeo-type flow formulae (3.6) and (3.11), respectively and is given by

$$\hat{Ca}_i^2 = \begin{cases} -1 + \left[\dfrac{\hat{\lambda}_{0i}}{\hat{\lambda}_i} \left(\hat{Cd}_{0i}^2 + 1 \right)^{-1} \right. \\ \qquad \left. + \sum_{j=1}^{M} \dfrac{\hat{\lambda}_j \, \hat{\alpha}_{ji}}{\hat{\lambda}_i} \left(\hat{Cd}_{j\,i}^2 + 1 \right)^{-1} \right]^{-1}, \text{ for GE streams,} \\ \\ -1 - \hat{\lambda}_i + \left[\dfrac{\hat{\lambda}_{0i}}{\hat{\lambda}_i} \left(\hat{Cd}_{0i}^2 + 1 + \hat{\lambda}_{0i} \right)^{-1} \right. \\ \qquad \left. + \sum_{j=1}^{M} \dfrac{\hat{\lambda}_j \, \hat{\alpha}_{ji}}{\hat{\lambda}_i} \left(\hat{Cd}_{j\,i}^2 + 1 + \hat{\lambda}_j \, \hat{\alpha}_{ji} \right)^{-1} \right]^{-1}, \text{ for GGeo streams,} \end{cases} \tag{5.7}$$

where for i=1,2,...,M

$$\hat{\lambda}_{0i} = \lambda_{0i} (1 - \pi_{0i}), \tag{5.8}$$

$$\hat{Cd}_{0i}^2 = \pi_{0i} + (1 - \pi_{0i}) \, Ca_{0i}^2, \; 0 \in A_i, \tag{5.9}$$

$$\hat{Cd}_{j\,i}^2 = 1 - \hat{\alpha}_{ji} + \hat{\alpha}_{ji} \, \hat{Cd}_j^2, \; j \in A_i, \tag{5.10}$$

λ_{0i} and Ca_{0i}^2 are the overall rate and SCV of the external interarrival-times, while $\hat{Cd}_j^2$ is the SCV of the effective interdeparture time of queue j, j=1,2,...,M, represented by either a GE/GE/1/N_j or GGeo/GGeo/1/N_j queueing models, namely

$$\hat{Cd}_j^2 = \begin{cases} \hat{\rho}_j (1 - \hat{\rho}_j) + (1 - \hat{\rho}_j) \, \hat{Ca}_j^2 + \hat{\rho}_j^2 \, \hat{Cs}_j^2, & \text{for GE/GE/1/}N_j, \\ \hat{\rho}_j^2 (\hat{Cs}_j^2 + \hat{\mu}_i - 1) - \hat{\rho}_j (\hat{Ca}_j^2 + \hat{\lambda}_j - 1) + \hat{Ca}_j^2, & \text{for GGeo/GGeo/1/}N_j, \end{cases} \tag{5.11}$$

with $\hat{\rho}_j = \hat{\lambda}_j / \hat{\mu}_j$. Moreover by applying the Law of Total Probability π_i is clearly given by

$$\pi_i = \sum_{j \in A_i} \lambda_{ji} \pi_{ji} / \sum_{j \in A_i} \lambda_{ji}, \tag{5.12}$$

where $\lambda_{ji} = \hat{\lambda}_j \hat{\alpha}_{ji}/(1 - \pi_{ji})$, $j \in A_i - \{0\}$. The behaviour of a single queueing station i as a building block within an open QNM can be observed in Fig. 4, while the revised GE/GE/1/N_i or GGeo/GGeo/1/N_i queue i in isolation can be seen in Fig. 5.

Finally, queue 0, $0 \in A_j$, and each queue i, $i \in A_j$, generate an overall arriving stream to queue j, j=1,2,...,M, seen as a censored GE/GE/1/N_j queue or GGeo/GGeo/1/N_j queue under AF or DF policies with rate $\hat{\lambda}_i \hat{\alpha}_{ij}/(1 - \pi_{ij})$ and SCV equal to $Cd^2_{ij} = (\hat{Cd}^2_{ij} - \pi_{ij})/(1 - \pi_{ij})$, $i \in A_j$. Thus, if each of these streams is considered in isolation to be the only arriving stream to queue j, then π_{ij} ($j \in D_i$ and $N_j < +\infty$) can be determined by an expressions analogous to (4.14) and (4.17), as appropriate. For example, for a GE/GE/1/N_i queue or GGeo/GGeo/1/N_i queue under AF policy, if clearly follows that

$$\pi_{ij} = (1 - \tau_{aij})^{N_j} \left(\frac{\hat{\tau}_{sj}}{\hat{\tau}_{sj}(1 - \tau_{aij}) + \tau_{aij}} \right) p_j(n_j) + \sum_{n_j=1}^{N_j} (1 - \tau_{aij})^{N_j - n_j} p_j(n_j), \quad i=0,1,...,M, \quad j=1,2,...,M \quad (i \neq j), \tag{5.13}$$

where

$$\tau_{aij} = \begin{cases} \dfrac{2}{Cd^2_{ij} + 1}, & \text{if } i \neq 0, \\ \dfrac{2}{Ca^2_{0i} + 1}, & \text{if } i = 0, \end{cases} \tag{5.14}$$

and

$$\hat{\tau}_{sj} = \frac{2}{\hat{Cs}^2_j + 1}, \quad j=1,2,...,M. \tag{5.15}$$

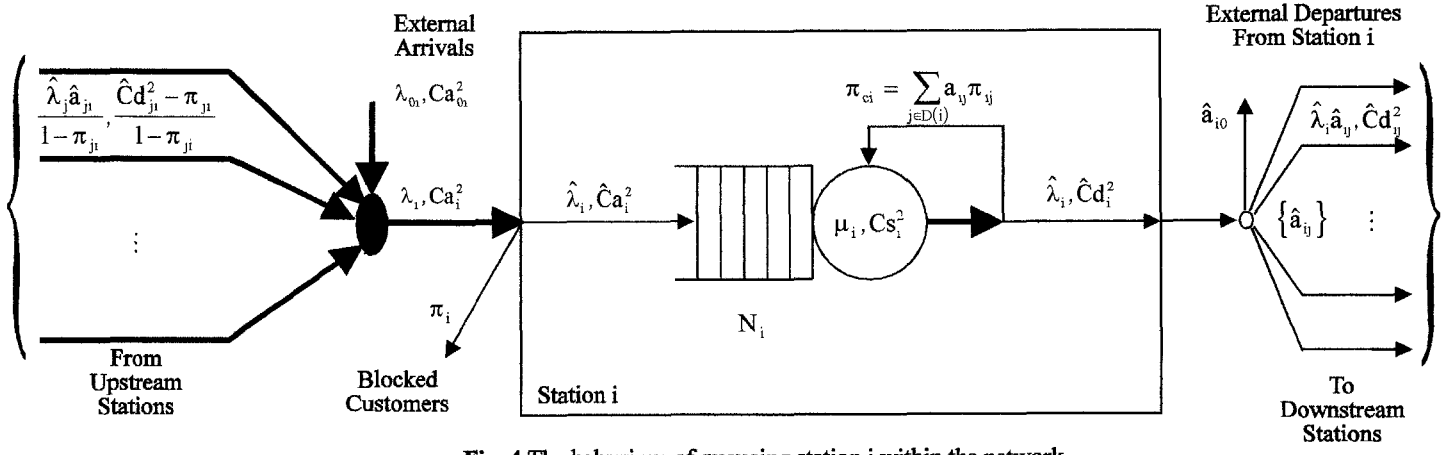

Fig. 4 The behaviour of queueing station i within the network.

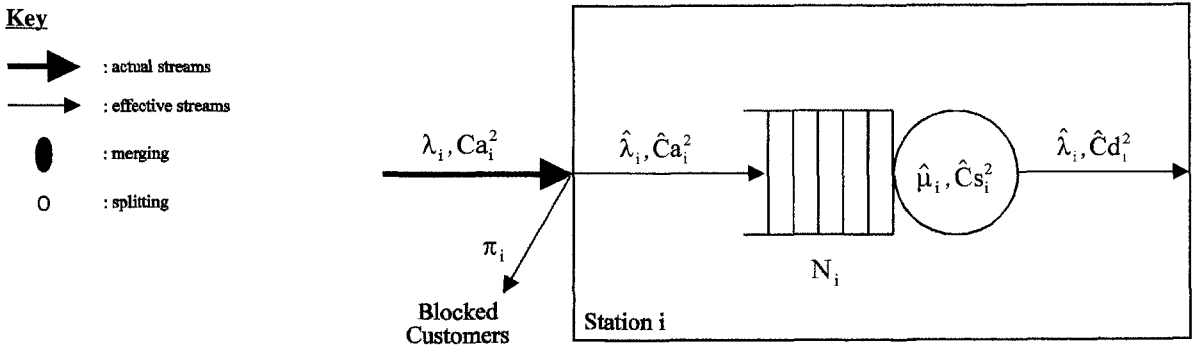

Fig. 5 The revised censored queue i in isolation.

A similar expression to that of (4.17) holds for a GGeo/GGeo/1/N_j queue under DF policy.

The ME approximation algorithm involves the solution of the system of non-linear equations for the blocking probabilities $\{\pi_{ij}\}$. It starts with some initial value for $\hat{Cd}_i^2$ and then an iterative procedure is applied involving all queues until a convergence for $\hat{Cd}_i^2$, i=1,2,...,R, is achieved. A diagramatic representation of the solution process can be seen in Fig. 6. The main steps of a revised ME algorithm (c.f., [58]) are given below:

ME Algorithm
Begin
Inputs: M, N_i, μ_i, Cs_i^2, λ_{0i}, Ca_{0i}^2, i=1,2,...,M,
$\{\alpha_{ij}\}$, i=1,2,...,M; j=0,1,...,M;

Step 1: Feedback correction (if $\alpha_{ii} > 0$, i=1,2,...,M);

Step 2: Initialisation: $\hat{Cd}_i^2$, i=1,2,...,M;
$\{\pi_{ij}\}$, i=0,1,...,M; j=1,2,...,M;

Step 3: Solve the system of non-linear equations $\{\pi_{ij}\}$, i=0,1,...,M; j=1,2,...,M;

Step 3.1: Calculate:

$$\{\hat{\alpha}_{ij}\},\ i=1,2,...,M;\ j=0,1,...,M;$$

Step 3.2: Compute:

$$\{\hat{\lambda}_i, \hat{\lambda}_{0i}, \pi_i, \hat{Ca}_i^2, \lambda_i, Ca_i^2, \pi_{ci}, \hat{\mu}_i, \hat{Cs}_i^2\},\ i=1,2,...,M;$$

Step 3.3: Use the Newton-Raphson method to find new values for $\{\pi_{ij}\}$, i=0,1,...,M; j=1,2,...,M;

Step 3.4: Return to Step 3.1 until convergence of $\{\pi_{ij}\}$, $\forall$ i, j;

Step 4: Fine new values for $\{\hat{Cd}_i^2\}$, i=1,2,...,M;

Step 5: Return to Step 3 until convergence of $\{\hat{Cd}_i^2\}$, $\forall$i;

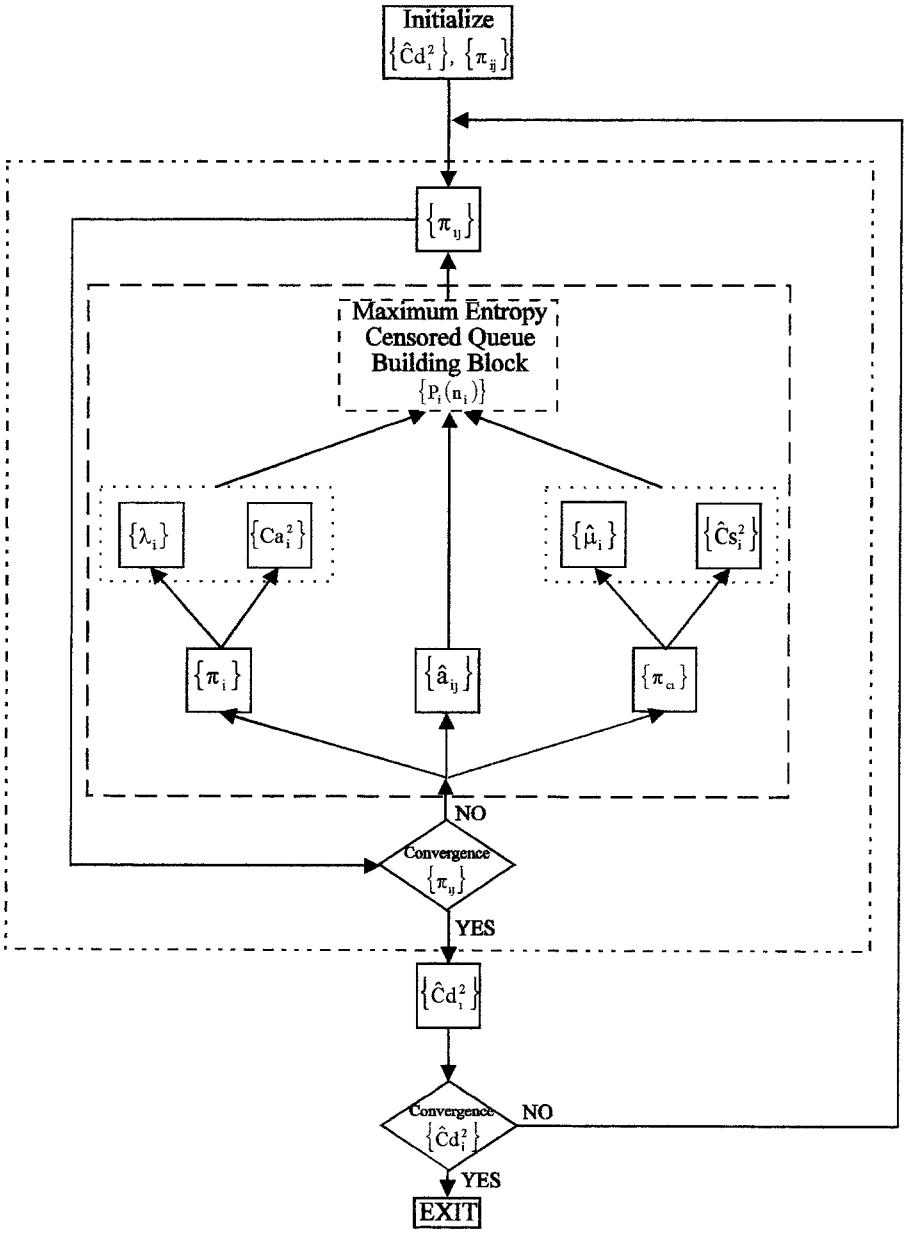

Fig. 6 A diagramatic representation of the two nested recursive schemes.

Step 6: Apply entropy maximisation to evaluate each queue i, i=1,2,...,M, as a censored $GE(\lambda_i, Ca_i^2)/GE(\hat{\mu}_i, \hat{Cs}_i^2)/1/N_i$ queue or $GGeo(\lambda_i, Ca_i^2/GE(\hat{\mu}_i, \hat{Cs}_i^2)/1/N_i$ queue, if $N_i < +\infty$, and as corresponding infinite capacity queues at equilibrium, if $N_i \to +\infty$;

Step 7: Obtain the performance metrics of interest;

End.

Remarks: The main computational effort of the ME algorithm is at every iteration between Steps 3 and 5. The non-linear system of equations $\{\pi_{ij}\}$, i=0,1,...,M, j=1,2,...,M, can be rewritten in the form $\pi = F(\pi)$, where π and F are column vectors of dimension Ψ, where Ψ is the cardinality of the set $\{\pi_{ij}\}$, $i \neq j$ and $N_j < +\infty$. It can be verified that the computational cost of the ME algorithm for open networks is $o(k\Psi^3)$, where k is the number of iterations between steps 3 and 5 and Ψ^3 is the number of manipulations for inverting the Jacobian matrix of F with respect to π.

Notably, the existence and unicity for the solution of the system of non-linear equations $\pi = F(\pi)$ cannot be proved analytically due to the complexity of the expression of the blocking probabilities $\{\pi_{ij}\}$. Furthermore, no strict mathematical justification can be given for the convergence of $\{\hat{Cd}_i^2\}$, i=1,2,...,M; nevertheless, numerical instabilities are very rarely observed in many experiments that have been carried out, even when the probabilities π_{ij} are relatively close to 1. Only in networks with high interarrival-time and service-time variability, in conjunction with high server utilisations and feedback streams, the solutions of the non-linear system $\pi = F(\pi)$ goes towards the trivial solution of $\pi_{ij} = 1$. This can be attributed to the fact that the effective utilisations $\{\hat{\lambda}_i/\hat{\mu}_i\}$ attain the value 1 and thus the system becomes unstable.

5.2 Closed QNMs with RS Blocking

Consider an arbitrary closed queueing network under RS-RD or RS-FD blocking mechanisms consisting of M FCFS single server queues with finite capacity N_i, i=1,2,...,M, general service-times and a fixed number of jobs K such that

$N_1 + N_2 + ... + N_M > K$. For each queue i, i=1,2,...,M the notation of the previous section apply and, in addition, let $\hat{N}_i$ be the virtual capacity of queue i, i.e., $\hat{N}_i = \min (N_i, K)$, κ_i be the minimum number of jobs always present, i.e., $\kappa_i = \max \{0, K - \sum_{i \neq j} \hat{N}_j\}$ and $\tilde{N}_i$ be the rv of the number of jobs in queue i as seen by a random observer.

By analogy to Sections 4 and 5.1, the form of the ME state probability, $p(\underline{n})$, $\underline{n} \in S(K, M)$, where $S(K, M) = \{\underline{n} ; \sum_i n_i = K, \kappa_i \leq n_i \leq \hat{N}_i, i=1,2,...,M\}$, can be determined subject to constraints of normalisation, *active* UTIL, $\upsilon_i(K)$, i.e., $\upsilon_i(K) = P(\tilde{N}_i > \kappa_i)$, *active* MQL, $\Omega_i(K)$, i.e., $\Omega_i(K) = \sum_{n_i > \kappa_i} (n_i - \kappa_i) p_i(n_i)$ and full buffer probability, $\varphi_i(K)$, $\varphi_i(K) = p_i(N_i)$, satisfying the job flow rate equations.

By applying Lagrange's Method of Undertermined Multipliers, a ME product-form aproximation is characterised, subject to the above constraints, and is given by

$$p(\underline{n}) = \frac{1}{Z(L, M)} \prod_{i=1}^{M} \hat{g}_i(n_i) x_i^{n_i - \kappa_i} y_i^{f(n_i)}, \qquad (5.16)$$

where Z(L, M) is the normalising constant,

$$\hat{g}_i(n_i) = \begin{cases} 1, & \text{if } n_i = \kappa_i, \\ g_i, & \text{if } \kappa_i < n_i \leq \hat{N}_i, \end{cases} \qquad (5.17)$$

$f(n_i) = \max \{0, n_i - \hat{N}_i + 1\}$, and g_i, x_i and y_i are the Lagrangian coefficients corresponding to constraints $\upsilon_i(K)$, $\Omega_i(K)$ and $\varphi_i(K)$, i=1,2,...,M respectively.
The implementation of ME solution (5.16) proceeds in two stages:

Stage 1: Determine the ME product-form approximation, $p(\underline{n})$, for a pseudo open network with RS blocking, subject to normalisation and constraints of the type υ_i, Ω_i and φ_i, i=1,2,...,M (c.f., Section 5.1). Incorporate the conservation of flow rate equations (i.e. $\hat{\lambda}_i = \sum_j \alpha_{ji} \hat{\lambda}_j$, i=1,2,...,M) and the fixed population mean constraints (i.e.,

$K = \sum_j \Omega_j$) within the ME algorithm for open QNMs with RS blocking (c.f., Section 5.1). Decompose the pseudo open network into individual censored GE/GE/1/κ_i;N_i or GGeo/GGeo/1/κ_i;N_i queues, if $N_i < K$ or stable GE/GE/1/κ_i;$+\infty$ or GGeo/GGeo/1/κ_i;$+\infty$ queues, if $N_i \geq K$, with modified arrival and service processes within a continuous-time or discrete-time domain, respectively (c.f., [53, 54]). (n.b., constraint $K = \sum_j \Omega_j(K)$ replaces, in a mathematical sense, the unknown external arrival process of the flow balance rate equations.)

Remarks: Focusing on state κ_i, $1 \leq \kappa_i \leq K - 1$, of GE/GE/1/$\kappa_i$;$N_i$ and GGeo/GGeo/1/κ_i;N_i censored queues with $N_i < +\infty$ or $N_i \to +\infty$, a job on service completion is forced to repeat service so that there will always be a minimum number of κ_i jobs present. These queues can be analysed via entropy maximisation in a similar fashion to that of Section 4.2 (c.f., [53, 54]).

Stage 2: Associate the analytic estimates of Lagrangian coefficients g_i, x_i and y_i, i=1,2,...,M, from Stage 1 with the ME state probability $p(\underline{n})$ of the original closed network (c.f., (5.16)). For the general case $\{\kappa_i \geq 0$ and $N_i \leq K\}$, use an efficient convolution type technique to compute ME solution iteratively until the job flow rate equations as applied to the original closed network, are satisfied.

The overall computational requirements of the ME algorithm are of $o(\gamma_1 \Psi^3)$ for Stage 1 and of $o(\gamma_2 MK)$ for Stage 2, where γ_1, γ_2 are the numbers of the corresponding iterations at each stage, respectively. Note that the recursive calculation of the normalising constant Z(M, K) in Stage 2 can be carried out under the most general type of product-form approximation.

As in the case of open networks (c.f., Section 5.1), no mathematical proof of convergence is given due to the complexity of the expressions involved in Stages 1 and 2. However, numerous experiments have always converged under stability conditions $(\hat{\lambda}_i/\hat{\mu}_i) < 1$, i=1,2,...,M.

Remarks: For the case of *reversible* networks with finite capacity, the ME algorithm captures their exact solution (c.f., [8]). However, the ME solutions for these networks may be viewed directly as the truncated ME solutions of the corresponding reversible networks with infinite capacity, reducing, clearly, to the exact solutions.

6. ME Application to a Queueing Model of a Shared Buffer ATM Switch

In the field of high speed networks the ATM switch architecture incorporating a single memory of fixed size which is shared by all output ports is of particular importance. An incoming cell (or packet) is stored in a shared buffer of finite capacity while its address is kept in the address buffer. The cells destined for the same output port can be linked by an address chain pointer or their addresses can be stored into a FCFS buffer which relates to a particular output port. A cell will be lost if on arrival it finds either the shared buffer or the address buffer full. An example of such a switch architecture is the Prelude architecture proposed by CNET, France [69]. Some specific types of queueing models and analytic approximations for a shared buffer ATM switch architecture can be seen in [70-72].

In this Section the principle of ME is applied to characterise product-form approximations for the performance analysis of a general FCFS queueing model of a shared buffer ATM switch architecture with bursty arrivals (c.f., Fig. 7). At the implementation stage of the ME solution the arrival process at each output port of the ATM switch is modelled by a CPP (c.f., continuous-time domain) or a BBP (c.f., discrete-time domain) with geometrically distributed bulk sizes. These processes are most appropriate to model simultaneous arrivals and departures at the output ports generated by different bursty sources under a given buffer management policy (c.f., AF or DF).

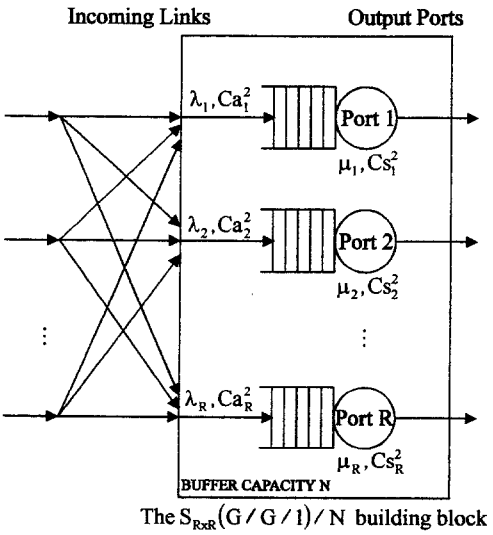

Fig. 7 A queueing model of the shared buffer ATM switch

6.1 Shared Buffer Model Formulation and Notation

Consider a general queueing model of a shared buffer switch with bursty arrivals depicted in Fig. 7. The queueing model consists of R parallel single server queues, where R is the number of the input (or output) ports. Each server represents an output port, while each queue stands for the address queue of the shared buffer switch. There are R bursty and heterogeneous arrival streams of cells, belonging to a single class - one per input port - each of which is generally distributed. A cell upon arrival at the input port i joins the queue of the output port j with transition probability α_{ij}, i, j=1,2,...,R such that $\sum_j \alpha_{ij} = 1$, i=1,2,...,R. Assuming that a merging of the arrival streams at each output port queue j, j=1,2,...,R, has been applied, let $1/\lambda_j$, Ca_j^2, j=1,2,...,R be, respectively, the mean and SCV of the interarrival process. The transmission (or service)-time of a cell at each output port j follows a general distribution with mean and SCV $1/\mu_j$ and Cs_j^2, j=1,2,...,R, respectively.

Remark: For an ATM switch, due to the fixed cell size and the nature of the associated outgoing links, the transmission times are assumed to be Deterministic (D) (i.e., $Cs_j^2 = 0$, j=1,2,...,R).

The buffer management scheme adopted here is the so-called *complete sharing* with the buffer capacity of each queue, N_i being equal to N. In general, the queueing model of the shared buffer switch can be denoted by S_{RxR}(G/G/1/)/N such that

(a) the interarrival and service times at an R x R switch are heterogeneous and generally distributed,
(b) each output port queue has a single server,
(c) shared buffer capacity of the switch is N.

Let at any given time the state of the system be represented by a state vector $\underline{n} = (n_1, n_2, ..., n_R)$, where n_i is the number of cells at each queue i, i=1,2,..,R. Moreover, let sets S(N, R) and A(η, R) be defined by

$$S(N, R) = \{\underline{n} = (n_1, n_2, ..., n_R) / \sum_{j=1}^{R} n_j \le N, 0 \le n_j \le N, j=1,2,...,R\},$$

$$A(\eta, R) = \{\underline{n} = (n_1, n_2, ..., n_R) / \sum_{j=1}^{R} n_j \leq \eta,\ 0 \leq n_j \leq \eta,\ j=1,2,...,R\},\ \eta=0,1,...,N.$$

Moreover, let $\{p(\underline{n})\}$, $\underline{n} \in S(N, R)$, $\{p(\eta)\}$, $\eta=0,1,....,N$, and $p_i\{n_i\}$, $i=1,2,...,R$; $n_i=0,1,...,N$, be the joint, aggregate and marginal state probabilities, respectively. Finally, let $\{p_a(\underline{n})\}$, $\underline{n} \in S(N, R)$ and $\{p_a(n)\}$, $n=0,1,...,N$ be the joint and aggregate state probabilities of an arriving bulk, respectively.

6.2 A ME solution for an S_{RxR}(G/G/1)/N Queueing System

Consider a general S_{RxR}(G/G/1)/N queueing model of a shared buffer switch architecture depicted in Fig. 7. Motivated by earlier ME analysis of simpler types of queues and networks (c.f., Sections 4 and 5), it is assumed that the following constraints about the joint state probabilities $\{p(\underline{n}\}, \underline{n} \in S(N, R)\}$ are known to exist: The normalisation, and for all i=1,2,...,R, the UTIL, $U_i (0 < U_i < 1)$, MQL, $Q_i (U_i \leq Q_i < N)$ and conditional state probability of an aggregate full buffer state with $n_i > 0$, ζ_i $(0 < \zeta_i < 1)$.

The form of the ME solution $p(\underline{n})$, $\underline{n} \in S(N, R)$, can be established by maximising the entropy function H(p), subject to normalisation, U_i, Q_i and φ_i, i=1,2,..,R, constraints. The maximisation of H(p) can be carried out by using Lagrange's Method of Undetermined Multipliers leading to the product-form solution

$$p(\underline{n}) = \frac{1}{Z} \prod_{j=1}^{R} g_j^{s_j(\underline{n})} x_j^{n_j} y_j^{f_j(\underline{n})}, \quad \forall \underline{n} \in S(N, R), \tag{6.1}$$

where Z is the normalising constant,

$s_j(\underline{n}) = 1$, if $n_i > 0$ or 0, otherwise,

$f_j(\underline{n}) = 1$, if $\{\sum_i n_i = N$ and $s_j(\underline{n}) = 1\}$ or 0, otherwise,

and $\{g_j, x_j, y_j\}$ are the Lagrangian coefficients corresponding to constraints $\{U_j, Q_j, \zeta_j\}$, j=1,2,...,R, respectively.

Suitable formulae for computing the complex sums of Z and U_j, j=1,2,...,R can be determined by following the Generating Function Approach (c.f. [73]). By defining appropriate z-transforms involving the Lagrangian coefficients, the normalising constant can be expressed by

$$Z = \sum_{\eta=0}^{N-1} C_1(\eta) + C_2(N), \tag{6.2}$$

where $\{C_1(\eta), \eta=0,1,...,N-1\}$ and $C_2(N)$ are determined via the following recursive and refined formulae:

$$C_k(\eta) = \begin{cases} C_{1,R}(\eta), & k = 1; \eta=0,1,...,N-1, \\ C_{2,R}(N), & k = 2; \eta = N, \end{cases} \tag{6.3}$$

where

$$C_{k,r}(\eta) = C_{k,r-1}(\eta) - (1 - \beta_r) x_r C_{k,r-1}(\eta - 1) + x_r C_{k,r}(\eta - 1), \tag{6.4}$$

$$\eta=1,2,...,N-2 + k; \quad k=1,2; \quad r=2,...,R,$$

with initial condition (for r = 1):

$$C_{k,1}(\eta) = \begin{cases} 1, & \text{if } \eta=0, \\ \beta_1 x_1^{\eta}, & \text{if } \eta=1,2,...,N-2+k, \end{cases}$$

$$C_{k,r}(0) = 1, \qquad k=1,2; \quad r=2,...,R,$$

where $\beta_r = g_r$, if k = 1, or $g_r y_r$, if k = 2.

Similarly, the utilisation U_i can be expressed as

$$U_i = \frac{1}{Z} \left\{ \sum_{\eta=1}^{N-1} C_1^{(i)}(\eta) + C_2^{(i)}(N) \right\}, \quad i=1,2,...,R, \tag{6.5}$$

where

$$C_k^{(i)}(\eta) = (1 - \beta_i)\, x_i\, C_k^{(i)}(\eta - 1) + \beta_i\, x_i\, C_k(\eta - 1), \qquad (6.6)$$
$$\eta=2,3,...,N-2+k;\ k=1,2;\ i=1,2,...,R,$$

with initial condition (for $\eta = 1$): $C_k^{(i)}(1) = \beta_i\, x_i$.

The Lagrangian coefficients $\{g_i, x_i, i=1,2,...,R\}$ can be approximated by making asymptotic connections to infinite capacity queues. Assuming that x_i is invariant with respect to the capacity N, as $N \to +\infty$, it can be verified that (c.f., (4.2), (4.3))

$$g_i = \frac{\rho_i(1 - x_i)}{x_i(1 - \rho_i)}, \quad x_i = \frac{L_i - \rho_i}{L_i}, \quad i=1,2,...,R, \qquad (6.7)$$

where L_i is the asymptotic MQL of queue i and $\rho_i = \lambda_i/\mu_i$. Moreover, the Lagrangian coefficients $\{y_i\}$, i=1,2,...,R, can be approximated by making use of the flow balance condition

$$\lambda_i(1 - \pi_i) = \mu_i U_i,\ i=1,2,...,R, \qquad (6.8)$$

where π_i is the blocking probability (or cell-loss) that an arriving cell destined for queue i, i=1,2,...,R, finds the shared buffer of the switch full (i.e., $\Sigma_i n_i = N$).

6.3 Marginal State Probabilities $\{p_i(\ell_i), \ell_i=0,1,...,N;\ i=1,2,...,R\}$

Let $\tilde{N}^{(i)}$ be the rv of the number of cells seen by a random observer at queue i, i=1,2,..,R. Using ME solution (6.1) and recursive expressions (6.6), it follows that

$$p_i(\ell_i) = P_i[\tilde{N}^{(i)} \geq \ell_i] - P_i[\tilde{N}^{(i)} \geq \ell_i + 1], \qquad (6.9)$$

with $P_i[\tilde{N}^{(i)} \geq N + 1] = 0$, where

$$P_i[\tilde{N}^{(i)} \geq \ell_i] = \frac{x_i^{\ell_i - 1}}{Z} \left\{ \sum_{\eta=\ell_i}^{N-1} C_1^{(i)}(\eta - \ell_i + 1) + C_2^{(i)}(N - \ell_i + 1) \right\}, \qquad (6.10)$$

$$\ell_i = 1,2,...,N; \quad i=1,2,...,R.$$

Remarks: Appropriate recursive expressions can also be defined (c.f., [59]) for marginal MQLs $\{Q_i, i=1,2,...,R\}$ and aggregate state probabilities $\{p(\eta), \eta=0,1,...,N\}$, while the aggregate cell-loss probability, π_a, is expressed by $\pi_a = \Sigma_i(\lambda_i/\lambda)\pi_i$, where $\lambda = \Sigma_i \lambda_i$.

6.4 The GE and GGeo Type Blocking Probabilities

Analytic expressions for the blocking probabilities $\{\pi_i\}$, i=1,2,...,R, can be established within a continuous-time or discrete-time domain by considering the S_{RxR}(GE/GE/1)/N or S_{RxR}(GGeo/GGeo/1)/N queueing models, respectively, and focusing on a tagged cell of an arriving bulk that is blocked. In this context, the blocking probability π_i is conditional on the position of a tagged cell within the bulk and also on the number of cells that its bulk finds in the system irrespective of its destination queue. Without loss of generality, it is assumed that the arriving bulk of the tagged cell is destined for output port queue i, i=1,2,...,R.

The following distinct and exhaustive blocking cases of a tagged cell are considered:

(i) The bulk on arrival finds the system in state $\{A(\eta, R)$ and $n_i = 0\}$, $\eta=0,1,...,N-1$.

In this case the size of the arriving bulk must be at least $N - \eta + 1$ and the tagged cell is one of those members of the bulk that is blocked. By following the same probabilistic arguments as those used for a single queue with capacity $N - \eta$ (c.f., Section 4.2.1), it is implied that

Prob {A tagged cell is blocked and its bulk on arrival finds the system at state $[A(\eta, R)$ and $n_i = 0]\}$

$$= \delta_i \left(\sum_{\substack{\underline{n} \in A(\eta, R) \\ \wedge\ n_i = 0}} P_a(\underline{n}) \right) (1 - \tau_{ai})^{N-\eta}, \quad \eta = 0,1,...,N-1, \tag{6.11}$$

where $$\delta_i = \tau_{si}/[\tau_{si}(1 - \tau_{ai}) + \tau_{ai}],$$

$$\tau_{ai} = \begin{cases} 2/(Ca_i^2 + 1), & \text{if GE}, \\ 2/(Ca_i^2 + \lambda_i + 1), & \text{if GGeo}, \end{cases} \text{ and, } \tau_{si} = \begin{cases} 2/Cs_i^2 + 1), & \text{if GE}, \\ 2/(Cs_i^2 + \mu_i + 1), & \text{if GGeo}. \end{cases}$$

(ii) The bulk on arrival finds the system in state $\{A(\eta, R) \text{ and } n_i > 0\}$, $\eta = 1,2,...,N-1$

In this case the bulk finds server i busy and there are in total η cells queueing or being transmitted in the entire system. Thus only $N - \eta$ buffer places are available and the tagged cell occupies any position within its bulk greater or equal to $N - \eta + 1$. It is therefore, implied that

Prob {A tagged cell is blocked and its bulk on arrival finds the system at state $[A(\eta, R) \text{ and } n_i > 0]$}

$$= \left(\sum_{\substack{\underline{n} \in A(\eta, R) \\ \wedge\ n_i > 0}} P_a(\underline{n}) \right) (1 - \tau_{ai})^{N-\eta}, \quad \eta = 1,2,...,N-1. \tag{6.12}$$

(iii) The bulk on arrival finds the entire system full with aggregate probability $p_a(N)$.

Applying the Law of Total Probability it follows that

$$\pi_i = \delta_i \sum_{\eta=0}^{N-1} \left(\sum_{\substack{\underline{n} \in A(\eta, R) \\ \wedge\ n_i = 0}} P_a(\underline{n}) \right) (1 - \tau_{ai})^{N-\eta} + \sum_{\eta=1}^{N-1} \left(\sum_{\substack{\underline{n} \in A(\eta, R) \\ \wedge\ n_i > 0}} P_a(\underline{n}) \right) (1 - \tau_{ai})^{N-\eta} + p_a(N). \tag{6.13}$$

To determine a computational formula for π_i, the following two sets of queueing models are considered:

Case I: S_{RxR}(GE/GE/1)/N and S_{RxR}(GGeo/GGeo/1)/N Queueing Models under AF policy.

In this case, $p_a(\underline{n}) = p(\underline{n})$, $\forall\ \underline{n} \in S(N, R)$ and $p_a(\eta) = p(\eta)$, η=0,1,...,N (c.f., [20, 53, 54]). Therefore, using z-transforms, ME solution $p(\underline{n})$, a corresponding expression for $p(\eta)$ (c.f., [59, 63]), η=0,1,...,N and coefficients $C_1(\eta)$, η=0,1,...,N-1, $C_2(N)$ and $C_1^{(i)}(\eta)$, η=1,2,...,N-1, it can be verified after some manipulation that

$$\pi_i = \frac{1}{Z}\ \{F_i(N) + C_2(N)\}\ , \tag{6.14}$$

where

$$F_i(N) = \delta_i \sum_{\eta=0}^{N-1} C_1(\eta)\,(1 - \tau_{ai})^{N-\eta} + (1 - \delta_i) \sum_{\eta=1}^{N-1} C_1^{(i)}(\eta)\,(1 - \tau_{ai})^{N-\eta}\ . \tag{6.15}$$

Case II: S_{RxR}(GGeo/GGeo/1)/N Queueing Model under DF policy.

In this case $p_a(\underline{n})$, $\underline{n} \in S(N, R)$ can be determined analytically in a similar fashion to that of Section 4 under the assumption (made for mathematical tractability) that *only simultaneous arrivals and departures relating to the same queue i, i=1,2,...,R, can actually take place.*

To this end, the cell-loss probability, π_i, i=1,2,...,R can be expressed by [63]

$$\pi_i = \frac{1}{Z}\ \left\{ \hat{F}_i(N) + C_2(N) - \mu_i\,\tau_{ai}\,\delta_i\,C_2^{(i)}(N) \right\}, \tag{6.16}$$

where

$$\hat{F}_i(N) = (1 - \sigma_{si}) \left\{ \delta_i \sum_{\eta=0}^{N-1} C_1(\eta) (1 - \tau_{ai})^{N-\eta} + (1 - \delta_i) \sum_{\eta=1}^{N-1} C_1^{(i)}(\eta) (1 - \tau_{ai})^{N-\eta} \right\}$$

$$+ \sigma_{si} \delta_i \left\{ \sum_{\eta=0}^{N-1} C_1(\eta) (1 - \tau_{ai})^{N-\eta} - \tau_{ai} \sum_{\eta=1}^{N-1} C_1^{(i)}(\eta) (1 - \tau_{ai})^{N-\eta} \right\}, \quad (6.17)$$

and $\sigma_{si} = \mu_i \tau_{si}$.

Remark: In the case of R = 1, formulae (6.14) and (6.16) reduce to those of the corresponding GE/GE/1/N and GGeo/GGeo/1/N_i censored queues of Section 4 (c.f., (4.14) and (4.17)).

6.5 Estimating Lagrangian Coefficients $\{g_i, x_i, y_i, i=1,2,...,R\}$

The Lagrangian coefficients $\{g_i, x_i\}$, i=1,2,...,R can be evaluated directly via expressions (6.7) by making use of the asymptotic MQL formula (4.5).

The Lagrangian coefficients $\{y_i\}$, i=1,2,...,R, can be determined numerically by substituting π_i of (6.14) or (6.16) and equations (6.2) and (6.5) into the flow-balance condition (6.8) and solving the resulting system of R non-linear equations with R unknowns $\{y_i, i=1,2,...,R\}$, namely

Case 1:

$$C_2^{(i)}(N) = \rho_i \left\{ \sum_{\eta=0}^{N-1} C_1(\eta) - F_i(N) \right\} - \sum_{\eta=1}^{N-1} C_1^{(i)}(\eta), \quad (6.18)$$

Case 2:

$$(1 - \mu_i \tau_{ai} \delta_i \rho_i) C_2^{(i)}(N) = \rho_i \left\{ \sum_{\eta=0}^{N-1} C_1(\eta) - F_i(N) \right\} - \sum_{\eta=1}^{N-1} C_1^{(i)}(\eta), \quad (6.19)$$

i=1,2,...,R and $N \geq 2$.

Systems (6.18) and (6.19) can be solved by applying the numerical algorithm of Newton-Raphson in conjunction with an efficient recursive scheme (c.f., [59]).

Conclusions

The principle of ME, subject to queueing theoretic mean value constraints, provides a powerful methodology for the analysis of complex queueing systems and networks. In this tutorial paper, the ME principle is used to characterise product-form approximations and resolution algorithms for arbitrary FCFS continuous-time and discrete-time QNMs at equilibrium under RS-RD and RS-FD blocking mechanisms and AF or DF buffer management policies. An ME application to the performance modelling of a shared-buffer ATM switch architecture with bursty arrivals is also presented. The ME solutions are implemented computationally subject to GE and GGeo type mean value constraints, as appropriate. In this context, the ME state probabilities of the GE/GE/1, GE/GE/1/N, GGeo/GGeo/1 and GGeo/GGeo/1/N queues at equilibrium and associated formulae for the first two moments of the effective flow streams (departure, splitting, merging) are used as *building blocks* in the solution process.

The ME algorithms capture the exact solution of reversible QNMs. Moreover, extensive validation studies (c.f., [27, 28, 36, 39, 40, 42-64] indicate that the ME approximations for arbitrary GE and GGeo type QNMs are generally very comparable in accuracy to that obtained by simulation models. Moreover, it has been conjectured that typical performance measures (such as MQLs for open QNMs and system throughputs and mean response-times for closed QNMs) obtained by GE or GGeo type ME approximations, given the first two moments of the external interarrival-times and/or service-times, define optimistic or pessimistic performance bounds - depending on the parameterisation of the QNM (c.f., [51]) - on the same quantities derived from simulation models when representing corresponding interevent-times by a family of two-phase distributions, as appropriate (e.g., Hyperexponential-2), having the same given first two moments.

The analytic methodology of entropy maximisation and its generalisations are versatile and can be applied within both *queue-by-queue decomposition* and *hierarchical multilevel aggregation* schemes to study the performance of other types of queueing systems and networks, particularly in the discrete-time domain (e.g., Banyan interconnection networks with blocking), representing new and more complex digitised structures of high speed networks. Work of this kind is the subject of current studies.

Acknowledgements

The author wishes to thank the Science and Engineering Research Council (SERC), UK, for supporting research work on entropy maximisation and QNMs with grants GR/D/12422, GR/F/29271 and GR/H/18609. Thanks are also extended to Spiros Denazis for drawing diagrams of Figs. 1-7.

Bibliography

In the continuous-time domain exact solutions and efficient computational algorithms for arbitrary QNMs with or without blocking have been presented, subject to reversibility type conditions (e.g, [1, 2, 8, 12]. More general, and therefore more realistic, QNMs have been analysed via approximate methods (e.g, [3-7, 9-11]). These include diffusion approximations for general QNMs (e.g., [4, 5]) and heuristic MVA for Markovian QNMs with priority classes (e.g, [9]). The accuracy of approximate methods, however, may be adversely affected in certain cases when based on either gross assumptions or intuitive heuristics. Comprehensive surveys on open and closed QNMs with finite capacity can be seen in [13] and [14], respectively.

In the discrete-time domain analytical results are mainly based on reinterpretation of earlier theoretical advances on continuous-time queues (e.g., [15-19]). There are inherent difficulties and open issues associated with the analysis of discrete-time queues due to the occurrence of simultaneous events (e.g, arrivals and departures) at the boundary epochs of a slot.

To overcome some of the drawbacks of earlier techniques, alternative ideas and tools from Statistical Mechanics have been proposed in the literature (e.g., 21-28]. In this context, the principle of ME (c.f., Jaynes [29-31], based on the concept of entropy functional [32], has inspired over the recent years a novel analytic methodology for the approximate solution of complex queues and QNMs. Note that Tribus [65] used the principle to derive a number of probability distributions. The mathematical foundations of the method and its generalisation to the principle of Minimum Relative Entropy (MRE) can be found in Shore and Johnson [66].

Entropy maximisation was first applied to the analysis of Markovian queueing systems at equilibrium by Ferdinand [22] who determined the stationary ME queue length distribution (qld) of an M/M/1/N queue by analogy to Statistical Mechanics. It was also shown that the ME steady-state probability of an open QNM, subject to marginal MQL constraints, corresponds to the solution of an open Jacksonian queueing network. Shore [25, 33] applied the principle of ME to investigate the stationary qlds of the M/M/∞ and M/M/∞/N queues [33] and also to propose ME

approximations for stable M/G/1 and G/G/1 queues, subject to mql and utilisation constraints, respectively [25]. The latter ME solution turns out to be identical to the qld of a stable M/M/1 queue. Independently, El-Affendi and Kouvatsos [26] derived a new ME qld for a stable M/G/1 queue, subject to both utilisation and MQL constraints. This ME solution becomes exact when the underlying service-time distribution is represented by the GE distribution. The ME principle was also applied to the analysis of a stable G/M/1 queue [26]. Cantor et al. [34] used the principle of MRE to characterise an equivalent stationary state probability to that of an M/M/c/N/K machine repair model and also to analyse two queues in tandem. Lutton et al. [35] applied the ME principle for solving under exponential assumptions queueing problems encountered in telephone exchanges. Guiasu [37] examined the ME condition for a stable M/M/1 queue. Rego and Szpankowski [38] investigated the presence of exponentiality in entropy maximised M/G/1 queues. Kouvatsos [39, 40] determined GE-type approximations for the ME qlds of the stable G/G/1 and G/G/1/N queues, respectively. Arizono et al. [41] analysed M/M/c queues at equilibrium, while Kouvatsos and Almond [42] and Wu and Chan [43] investigated stable G/G/c/N/K and G/M/c queues, respectively. Strelen [47] applied entropy maximisation to obtain a piecewise approximation for the waiting-time densities of G/G/1 queues. A new ME priority approximation for a stable G/G/1 queue was proposed by Kouvatsos and Tabet-Aouel [45]. This work was also extended to the ME analysis of a stable multiple server G/G/c/PR queue under preemptive-resume (PR) rule [46, 47]. Moreover, Kouvatsos and Georgatsos [48] presented an ME solution for a stable G/G/c/SQJ queueing system with c parallel serve queues operating under the shortest queue with jockeying (SQJ) routing criterion. A stable M/G/1 retrial queue was analysed via entropy maximisation by Falin et al. [49]. More recently, ME advances have been made towards the analysis of GGeo-type discrete-time queues (c.f., [60, 62, 73-74]).

The first ME approximation for arbitrary GE-type FCFS open closed QNMs with infinite capacity, subject to normalisation and marginal constraints of UTIL and MQL, was proposed in Kouvatsos [27]. Subsequent reformulation of the ME problem for closed QNMs incorporates the condition of flow conservation expressed by the job flow balance equations [28], or (equivalently) the work rate theorem [36]. Extensions of the ME methodology to the analysis of arbitrary GE-type QNMs with infinite capacity and mixed (priority/non-priority) service disciplines have been reported in [50, 51] (single server stations) and [52] (multiple server stations). ME has also been applied in the approximate analysis of arbitrary FCFS open and closed GE-type QNMs with finite capacity and RS-RD or RS-FD blocking mechanisms with single [53, 58] or multiple [54, 58] server stations. An ME extension to arbitrary open QNMs with RS-RD blocking and multiple-job classes with various non-priority based service disciplines has been recently presented by Kouvatsos and Denazis [56].

Furthermore, an ME analysis of arbitrary QNMs with GE-type stations and SQJ routing has appeared in [57]. The ME principle has also been used to study arbitrary FCFS open and closed discrete-time QNMs of GGeo-type with RS-RD blocking [64]. Finally, MRE applications, given fully decomposable subset and aggregate constraints, towards a hierarchical multilevel aggregation of arbitrary closed QNMs can be found in [74, 75].

In the performance modelling field of high speed networks entropy maximisation has been used in both continuous-time (GE-type) and discrete-time (GGeo-type) domains for the analysis of multi-buffered and shared buffer ATM switch architectures [59, 61, 63] as well as Distributed Queue Dual Bus (DQDB) networks with multiple-job classes [76-78].

A comprehensive review of entropy and relative entropy optimisation and arbitrary continuous-time QNMs with finite capacity, multiple-job classes, multiple-server stations and mixed service disciplines together with some computer and communication systems applications can be seen in Kouvatsos [79].

[1] Baskett, F., K.M., Muntz, R.R. and Palacios, F.G., Open, Closed and Mixed Networks with Different Classes of Customers, JACM 22, 2 (1975) 248-260.

[2] Reiser, M. and Kobayashi, H., Queueing Networks with Mulitple Closed Chains: Theory and Computational Algorithms, IBM J. Res. & Dev. 19 (1975) 283-294.

[3] Marie, R., An Approximate Analytical Method for General Queueing Networks, IEEE Trans. on Software Eng. SE-5, No. 5 (1979) 530-538.

[4] Reiser, M. and Kobayashi, H., Accuracy of the Diffusion Approximation for Some Queueing Systems, IBM J. Res. & Dev. 18 (1974) 110-124.

[5] Gelenbe, E. and Pujolle, G., The Behaviour of a Single Queue in a General Queueing Network, Acta Info. 7 (1976) 123-160.

[6] Courtois, P.J., Decomposability: Queueing and Computer Systems Applications, Academic Press Inc., New York (1977).

[7] Chandy, K.M., Herzog, U. and Woo, L., Approximate Analysis of General Queueing Networks, IBM J. Res. & Dev. 19 (1975) 43-49.

[8] Kelly, F.P., Reversibility and Stochastic networks, Wiley, New York (1979).

[9] Bryant, R.M., Krzesinski, A.E., Lasmi, M.S. and Chandy, K.M., The MVA Priority Approximation, T.O.C.S. 2(4) (1984) 335-359.

[10] Altiok, T. and Perros, H.G., Approximate Analysis of Arbitrary Configurations of Queueing Networks with Blocking, Annals of OR (1987) 481-509.

[11] Van Dijk, N.M., A Simple Bounding Methodology for Non-Product-Form Queueing Networks with Blocking, Queueing Networks with Blocking, Perros, H.G. and Altiok, T. (Eds.), North-Holland (1980) 3-18.

[12] Akyildiz, I.F. and Von Brand, H., Exact Solutions for Open, Closed and Mixed Queueing Networks with Rejection Blocking, Theoret. Comput. Sci. 64 (1989) 203-219.

[13] Perros, H.G., Approximation Algorithms for Open Queueing Networks with Blocking, Stochastic Analysis of Computer and Communication Systems, Takagi (Ed.), North-Holland (1990) 451-494.

[14] Onvural, R.O., Survey of Closed Queueing Networks with Blocking, ACM Computing Surveys 22, 2 (1990), 83-121.

[15] Hsu, J. and Burke, P.J., Behaviour of Tandem Buffers with Geometric Input and Markovian Output, IEEE Trans. Commun. 25, (1976) 2-29.

[16] Bharath-Kumar, K., Discrete-Time Queueing Systems and their Networks, IEEE Trans. on Commun. 28, 2, (1980) 260-263.

[17] Hunter, J.J., Mathematical Techniques of Applied Probability, Vol. 2, Discrete-time Models: Techniques and Applications, Academic Press (1983).

[18] Kobayashi, H., Discrete-Time Queueing Systems, Chapter 4 of Probability Theory and Computer Science, Louchard, G. and Latouche, G. (Eds.), Academic Press (1983).

[19] Pujolle, G., Discrete-time Queueing Systems for Data Networks Performance Evaluation, Queueing, Performance and Control in ATM, Cohen, J.W. and Pack, C.D. (Eds.), North-Holland, (1991) 239-244.

[20] Gravey, A. and Hebuterne, G., Simultaneity in Discrete-time Single Server Queues with Bernoulli Inputs, Performance Evaluation, 14, (1992) 123-131.

[21] Benes, V.E., Mathematical Theory of Connecting Networks and Telephone Traffic, Academic Press, New York (1965).

[22] Ferdinand, A.E., A Statistical Mechanical Approach to Systems Analysis, IBM J. Res. Dev. 14 (1970) 539-547.

[23] Pinsky, E. and Yemini, Y., A Statistical Mechnics of Some Interconnection Networks, Performance '84, Gelenbe, E. (Ed.), North-Holland (1984) 147-158.

[24] Pinsky, E and Yemini, Y., The Canonical Approximation in Performance Analysis, Computer Networking and Performance Evaluation, Hasegawa, T. et al. (Eds.), North-Holland (1986) 125-137.

[25] Shore, J.E., Information Theoretic Approximations for M/G/1 and G/G/1 Queueing Systems, Acta Info. 17 (1982) 43-61.

[26] El-Effendi, M.A. and Kouvatsos, D.D., A Maximum Entropy Analysis of the M/G/1 and G/M/1 Queueing Systems at Equilibrium, Acta Info. 19 (1983) 339-355.

[27] Kouvatsos, D.D., Maximum Entropy Methods for General Queueing Networks, Modelling Techniques and Tools for Performance Analysis, Potier, D. (Ed.), North-Holland (1985) 589-609.

[28] Kouvatsos, D.D., A Universal Maximum Entropy Algorithm for the Analysis of General Closed Networks, Computing Networking and Performance Evaluation, Hasegawa, T. et al. (Eds.), North-Holland (1986) 113-124.

[29] Jaynes, E.T., Information Theory and Statistical Mechanics, Phys, Rev. 106, 4 (1957) 620-630.

[30] Jaynes, E.T., Information Theory and Statistical Mechanics II, Phys. Rev. 108, 2 (1957) 171-190.

[31] Jaynes, E.T., Prior Probabilities, IEEE Trans. Systems Sci. Cybern. SSC-4 (1968) 227-241.

[32] Shannon, C.E., A Mathematical Theory of Communication, Bell Syst. Tech. J. 27 (1948) 379-423, 623-656.

[33] Shore, J.E., Derivation of Equilibrium and Time-Dependent Solutions of M/M/∞/N and M/M/∞ Queueing Systems Using Entropy Maximisation, Proc. 1978 National Computer Conference, AFIPS (1978) 483-487.

[34] Cantor, J., Ephremides, A. and Horton, D., Information Theoretic Analysis for a General Queueing System at Equilibrium with Application to Queues in Tandem, Acta Info. 23 (1986) 657-678.

[35] Lutton, J.L., Bonomi, E. and Felix, M.R., Information Theory and Statistical Mechanics Approach to Solving a Queueing Problem Encountered in Telephone Exchanges, Research Report PA/1199, Division ATR - Centre National d'Etudes des Telecommunications, France (1984).

[36] Walsta, R., Iterative Analysis of Networks of Queues, PhD Thesis, Technical Report CSE1-166, University of Toronto, Canada (1984).

[37] Guiasu, S., Maximum Entropy Condition, in Queueing Theory, J. Opl. Res. Soc. 37, 3 (1986) 293-301.

[38] Rego, V. and Szpankowski, W., The Presence of Exponentiality in Entropy Maximised M/GI/1 Queues, Computers Opns. Res. 16, 5 (1989) 441-449.

[39] Kouvatsos, D.D., A Maximum Entropy Analysis of the G/G/1 Queue at Equilibrium, J. Opl. Res. Soc. 39, 2 (1989) 183-200.

[40] Kouvatsos, D.D., Maximum Entropy and the G/G/1/N Queue. Acta Info. 23 (1986) 545-565.

[41] Arizono, I., Cui, Y. and Ohta, H., An Analysis of M/M/s Queueing Systems Based on the Maximum Entropy Principle, J. Opl. Res. Soc. 42, 1 (1991) 69-73.

[42] Kouvatsos, D.D. and Almond, J., Maximum Entropy Two-Station Cyclic Queues with Multiple General Servers, Acta info. 26 (1988) 241-267.

[43] Wu, J.S. and Chan, W.C., Maximum Entropy Analysis of Multiple-Server Queueing Systems, J. Opl. Res. Soc. 40, 9 (1989) 815-825.

[44] Strelen, C., Piecewise Approximation of Densities Applying the Principle of Maximum Entropy: Waiting Time in G/G/1-Systems, Modelling Techniques and Tools for Computer Performance Evaluation, Puigjaner, R. and Potier, D. (Eds.), Plenum Press (1989) 421-438.

[45] Kouvatsos, D.D. and Tabet-Aouel, N.M., A Maximum Entropy Priority Approximation for a Stable G/G/1 Queue, Acta Info. 27 (1989) 247-286.

[46] Tabet-Aouel, N.M. and Kouvatsos, D.D., On an Approximation to the Mean Response Times of Priority Classes in a Stable G/G/c/PR Queue, J. Opl. Res. Soc., 43, 3 (1992) 227-239.

[47] Kouvatsos, D.D. and Tabet-Aouel, N.M., An ME-Based Approximation for Multi-Server Queues with Preemptive Priority, To appear in the European Journal of Operational Research (1993).

[48] Kouvatsos, D.D. and Georgatsos, P.H., General Queueing Networks with SQJ Routing, Proc. 5th UK Perf. Eng. Workshop, University of Edinburgh, September 1988.

[49] Falin, G., Martin, G.I. and Artaleo, J., Information Theoretic Approximations for the M/G/1 Retrial Queue, To appear in Acta Infomatica.

[50] Kouvatsos, D.D., Georgatsos, P.H. and Tabet-Aouel, N.M., A Universal Maximum Entropy Algorithm for General Multiple Class Open Networks with Mixed Service Disciplines, Modelling Techniques and Tools for Computer Performance Evaluation, Puigjaner, R. and Potier, D. (Eds.), Plenuum Publishing Corporation (1989) 397-419.

[51] Kouvatsos, D.D. and Tabet-Aouel, N.M., Product-Form Approximations for an Extended Class of General Closed Queueing Networks, Performance '90, King, P.J.B. et al. (Eds.), North-Holland (1990) 301-315.

[52] Kouvatsos, D.D. and Tabet-Aouel, N.M., Approximate Solutions for Networks of Queues with Multiple Server Stations under PR Discipline, Proc. 8th UK Perf. Eng. Workshop, Harrison, P. and Field, T. (Eds.), Imperial College, London, 1992.

[53] Kouvatsos, D.D. and Xenios, N.P., Maximum Entropy Analysis of General Queueing Networks with Blocking, Queueing Networks with Blocking, Perros, H.G. and Altiok, T. (Eds.), North-Holland, Amsterdam (1989) 281-309.

[54] Kouvatsos, D.D. and Xenios, N.P., MEM for Arbitrary Queueing Networks with Multiple General Servers and Repetitive-Service Blocking, Perf. Eval. 10 (1989) 169-195.

[55] Kouvatsos, D.D., Denazis, S.G. and Georgatsos, P.H., MEM for Arbitrary Exponential Open Networks with Blocking and Multiple Job Classes, Proc. 7th Perf. Eng. Workshop, Hillston, J. et al. (Eds.), Springer-Verlag (1991) 163-178.

[56] Kouvatsos, D.D. and Denazis, S.G., Entropy Maximised Queueing Networks with Blocking and Multiple Job Classes, Perf. Eval. 17 (1993) 189-205.

[57] Kouvatsos, D.D. and Georgatsos, P.H., Queueing Models of Packet-Switched Networks with Locally Adaptive Routing, Teletraffic and Datatraffic in Period of Change, ITC-13, Jensen, A. and Iversen, V.B. (Eds.), North-Holland (1991) 303-308.

[58] Kouvatsos, D.D. and Denazis, S.G., Comments on and Tuning to: MEM for Arbitrary Queueing Networks with Repetitive-Service Blocking under Random Routing, Technical Report CS-18-91, University of Bradford, (May 1991).

[59] Kouvatsos, D.D. and Denazis, S.G., A Universal Building Block for the Approximate Analysis of a Shared Buffer ATM Switch Architecture, Annals of Operations Research (to appear).

[60] Kouvatsos, D.D. and Tabet-Aouel, N.M., GGeo-type Approximations for General Discrete-Time Queueing Systems, Proc. IFIP Workshop TC6 on Modelling and Performance Evaluation of ATM Technology, Pujolle, G. et al. (Eds.), La Martinique, North-Holland (1993) (to appear).

[61] Kouvatsos, D.D., Tabet-Aouel, N.M. and Denazis, S.G., A Discrete-Time Queueing Model of a Shared Buffer ATM Switch Architecture with Bursty Arrivals, Proc. of 10th UK Teletraffic Symposium, IEE, BT Labs., Ipswich (1993) 19/1-19/9.

[62] Kouvatsos, D.D. and Tabet-Aouel, N.M., The GGeo/G/1 Queue under AF and DF Buffer Management Policies, Proc. of 1st UK Workshop on Performance Modelling and Evaluation of ATM Networks, Kouvatsos, D.D. (Ed.), University of Bradford (1993) 7/1-7/11.

[63] Kouvatsos, D.D., Tabet-Aouel, N.M. and Denazis, S.G., ME-Based Approximations for General Discrete-Time Queueing Models, Res. Rep. CS-11-93, Department of Computing, University of Bradford (1993).

[64] Kouvatsos, D.D., Tabet-Aouel, N.M. and Denazis, S.G., Approximate Analysis of Discrete-Time Networks with or without Blocking, Proc. of 5th International Conference on Data Communication Systems and their Performances, Viniotis, Y. and Perros, H. (Eds.), (to appear).

[65] Tribus, M., Rational Description, Decisions and Designs, Pergamon, New York (1969).

[66] Shore, J.E. and Johnson, R.W., Axiomatic Derivation of the Principle of ME and the Principle of Minimum Cross-Eentropy, IEEE Trans. Info. Theory TI-26 (1980) 26-37.

[67] Sauer, C., Configuration of Computing Systems: An Approach Using Queueing Network Models, PhD Thesis, University of Texas (1975).

[68] Nojo, S. and Watanabe, H., A New Stage Method Getting Arbitrary Coefficient of Variation by Two Stages, Trans. IEICE 70 (1987) 33-36.

[69] Devault, M., Cochennec, J.Y. and Servel, M., The *Prelude* ATD Experiment: Assessments and Future Prospects, IEEE J. Sac, 6(9) (1988) 1528-1537.

[70] Iliadis, I., Performance of a Packet Switch with Shared Buffer and Input Queueing, Teletraffic and Datatraffic in a Period of Change, ITC-13, Jensen, A. and Iversen, V.B. (Eds), North-Holland (1991), 911-916.

[71] Yamashita, H., Perros, H.G. and Hong, S., Performance Modelling of a Shared Buffer ATM Switch Architecture, Teletraffic and Datatraffic in a Period of Change, ITC-13, Jensen, A. and Iversen, V.B. (Eds), North-Holland (1991), 993-998.

[72] Hong, S., Perros, H.G. and Yamashita, H., A Discrete-Time Queueing Model of the Shared Buffer ATM Switch with Bursty Arrivals, Res. Rep., Computer Science Department, North Carolina State University (1992).

[73] Williams, A.C. and Bhandiwad, R.A., A Generating Function Approach to Queueing Network Analysis of Multiprogrammed Computers, Networks, 6 (1976) 1-22.

[74] Kouvatsos, D.D. and Tomaras, P.J., Multilevel Aggregation of Central Server Models: A Minimum Relative Entropy Approach, Int. J. Systems Science, 23, 5 (1992) 713-739.

[75] Tomaras, P.J. and Kouvatsos, D.D., MRE Hierarchical Decomposition of General Queueing Network Models, Acta Info. 28 (1991) 265-295.

[76] Tabet-Aouel, N.M. and Kouvatsos, D.D., A Discrete-Time Maximum Entropy Solution for the Performance Analysis of DQDB Architecture, Proc. of 9th UK Perf. Eng. Workshop, Woodward, M. (Ed.), Pentech Press (1993) (to appear).

[77] Tabet-Aouel, N.M. and Kouvatsos, D.D., Entropy Maximisation and Unfinished Work Analysis of the Logical DQDB Queue, Res. Rep. TAN/DDK 13, Department of Computing, University of Bradford (1993), 16/1-16/7.

[78] Tabet-Aouel, N.M. and Kouvatsos, D.D., Performance Analysis of a DQDB under a Priority MAC Protocol, Proc. of 1st UK Workshop on Performance Modelling and Evaluation of ATM Networks, Kouvatsos, D.D. (Ed.), University of Bradford (1993) 24/1-24/12.

[79] Kouvatsos, D.D., Entropy Maximisation and Queueing Network Models, Annals of Operations Research (to appear).