

Generalized QBD Processes, Spectral Expansion and Performance Modeling Applications

Tien Van Do¹ and Ram Chakka²

¹ Department of Telecommunications
Budapest University of Technology and Economics
H-1117, Magyar tudósok körútja 2., Budapest, Hungary
do@hit.bme.hu

² Meerut Institute of Engineering and Technology (MIET),
Meerut, India 250005.
ramchakka@yahoo.com

Abstract. This paper suggests new queuing models, in the Markovian framework, which can tackle the presence of burstiness in the traffic and autocorrelations among the inter-arrival times of packets in the performance evaluation of next generation networks. These models are essentially based on certain generalizations of the Quasi Birth-Death (QBD) processes. Efficient steady state solution of these new queuing models, along with some illustrative applications, is presented. The proposed models and their further evolutions have the potential to be useful tools for the performance evaluation of modern telecommunication networks.

1 Introduction

The concept of Quasi Birth-Death (QBD) processes, as a generalization of the classical birth and death processes (e. g. the $M/M/1$ queue) was first introduced in the late sixties by [55] and [26]. A QBD process is a Markov process on a two-dimensional lattice, finite in one dimension (finite or infinite in the other). A state is described by two integer-valued random variables: the one in the finite dimension is the phase and the other is the level [37, 39, 41]. Transitions in a QBD process are possible within the same level or between adjacent levels. It is observed that QBD processes create a useful framework for the performability analysis pertaining to many problems occurring in telecommunications and computer networks [4, 8, 18, 20, 23, 36, 40, 42, 45, 56, 57].

In a QBD process, if the nonzero jumps in levels are not accompanied with changes in a phase, then these processes can be known as Markov modulated Birth and Death processes. The large or infinite number of states involved makes the solution of these models nontrivial. There are several methods of solving these models, either the whole class of models or some of the subclasses.

Seelen has analyzed a Ph/Ph/c queue in the QBD frame work [46]. Seelen's method is an approximate one where the Markov chain is first truncated to a

finite state Markov chain. By exploiting the structure an efficient iterative solution algorithm can be applied. The second method is to reduce the infinite-states problem to a linear equation involving vector generating function and some unknown probabilities. The latter are then determined with the aid of the singularities of the coefficient matrix. A comprehensive treatment of that approach, in the context of a discrete-time process with a general M/G/1 type structure, is presented in [30]. The third way of solving these models is the well known matrix geometric method, first proposed by Evans [26, 41]. In this method a nonlinear matrix equation is first formed from the system parameters and the minimal non-negative solution R of this equation is computed by an iterative method. The invariant vector is then expressed in terms of the powers of R . Neuts claims this method has a probabilistic interpretation for the steps in computation. However, this method suffers from the fact that there is no way of knowing how many iterations are needed to compute R to a given accuracy. It can also be shown that for certain parameter values the computation requirements are uncertain and formidably large. The fourth method is known as spectral expansion method [5, 6, 39]. It is based on expressing the invariant vector of the process in terms of eigenvalues and left eigenvectors of a certain matrix polynomial. The generating function approach and the spectral expansion method are closely related. However, the latter computes steady state probabilities directly using an algebraic expansion while the former provides them through a transform.

It is confirmed by a number of works that the spectral expansion method is better than the matrix geometric method in a number of aspects [5, 32, 33, 39]. It is observed that the spectral expansion method is proved to be a mature technique for the performance analysis of various problems [5–17, 20–25, 27, 28, 31, 32, 38, 39, 49–51, 53, 54, 52, 58].

Due to heterogeneous requirements concerning network technology and services that next generation networks (NGN) [34] are required to support, the issue of modeling the packet traffic and nodes in modern communication networks has become complicated because of the existence of burstiness (time varying arrival or service rates, arrivals or services of packets in *batches*) and important correlations among inter-arrival times [44]. In addition the traffic arriving at a node is often the superposition of traffic from a number of sources (homogeneous or heterogeneous), which further complicates the analysis of the system. Self-similar traffic models such as the FBM [43] can represent both burstiness and auto-correlations, but they are not analytically tractable in a queuing context.

The CPP³, defined in [19] and employed in [14, 29] and the $\sum_{k=1}^K CPP_k$ (superposition of K independent CPPs) traffic models often give a good representation of the burstiness (batch size distribution) of the traffic from one or more sources (along with mathematical tractability), but not the auto-correlations of the inter-arrival times (of batches) observed in real traffic. The usefulness and applicability of these models has been validated by measurements for example in [8, 23].

³ Throughout this paper when we refer to CPP we mean the compound Poisson process with independent and geometric batch-sizes, this is for convenience of referring.

Recently, we have proposed two new queuing models, the MM $\sum_{k=1}^K CPP_k/GE/c/L$ G-queue [10] with homogenous servers (the *Sigma* queue) and with heterogeneous servers (the *HetSigma* queue [8]). We have also developed some transformations which, when applied to the steady state balance equations, result in QBD-M type computable form. These models do provide a large flexibility to accommodate geometric as well as non-geometric batch sizes in both arrivals and services, and hence are capable of emerging as generalized Markovian node models. In these queues, the GE service time distribution is widely used, which is motivated by the fact that only the mean and variance may be computed reliably from the measurements. Therefore a choice of the distributions which implies least bias is that of the GE distribution [35, 47]. The parameters of the GE distribution are estimated from the real traffic trace in our numerical study. Moreover, the accommodation of *large* or *unbounded* batch-sizes, with efficient steady state queuing solutions, is a definitive advantage of our models besides the ability to accommodate negative customers.

In this paper, we present the spectral expansion methodology for the QBM-M queue. Then we give the short overview of the *HetSigma* queuing model, efficient computation of its steady state performance, possible extensions, along with some non-trivial applications to the performance evaluation of some problems in telecommunications networks.

The rest of the paper is organized as follows. Section 2 gives a brief overview of the important stochastic processes and distributions used in the new queuing model. It also presents the spectral expansion for QBD-M processes. The *HetSigma* queue is described in Section 3. The required steady state solution and an application are given in Sections 4 and 5, respectively. Some extensions are discussed in Section 6. Future directions to this research and conclusions are dealt in Section 7.

2 A Brief Overview of the Stochastic Processes and Distributions Involved

2.1 The QBD-M process

The QBD-M (Quasi Simultaneous-Bounded-Multiple Births and Simultaneous-Bounded-Multiple Deaths) process is a two-dimensional Markov process on a finite or semi-infinite lattice strip [5, 6, 39]. The state at any time t is denoted by two integer valued random variables, $I(t)$ and $J(t)$. $I(t)$ takes a finite set of values (*phases*) $\{1, 2, \dots, N\}$, and $J(t)$ takes a set of values (*levels*) $\{0, 1, \dots, L\}$, where L can be finite or infinite. We assume that the Markov process, Y if L is infinite, and \bar{Y} if L is finite, is denoted by $\{[I(t), J(t)]; t \leq 0\}$ and is irreducible.

The possible transitions underlying this Markov process are given by the following transition rate matrices, each of size $N \times N$:

A_j : purely lateral (phase) transitions – $A_j(i, k)$ is the transition rate from state (i, j) to state (k, j) ($i \neq k; 0 \leq i, k \leq N; j = 0, 1, \dots, L$).

- $B_{j,j+s}$: bounded s -step upward transitions – $B_{j,j+s}(i, k)$ is the transition rate from the state (i, j) to state $(k, j+s)$ ($0 \leq i, k \leq N; 1 \leq s \leq y_1; j = 0, 1, \dots, L-1$).
 $B_{j,j+s} = 0$ if $j + s > L$.
- $C_{j,j-s}$: bounded s -step downward transitions – $C_{j,j-s}(i, k)$ is the transition rate from state (i, j) to state $(k, j-s)$ ($0 \leq i, k \leq N; 1 \leq s \leq y_2; j = 1, 2, \dots, L$).
 $C_{j,j-s} = 0$ if $j - s < 0$.

There is a threshold T such that, $A_j = A$ ($j \geq T$), $B_{j,j+s} = B_s$ ($j \geq T - y_1$), $C_{j,j-s} = C_s$ ($j \geq T$), thus these matrices are independent of j .

The spectral expansion solution of the QBD-M process is based on the observation that the steady state balance equations can be written in the form

$$\sum_{k=0}^y \mathbf{v}_{j+k} Q_k = 0 \quad (T - y_1 \leq j \leq L - y - 1; y = y_1 + y_2), \quad (1)$$

where the coefficient matrices Q_k can be obtained from system parameters, following the methodology in [6, 8, 14].

Therefore, when L is finite, the probability invariant vector \mathbf{v}_j is given by [5, 6, 39]

$$\mathbf{v}_j = \sum_{l=1}^{y_1 N} a_l \psi_l \lambda_l^{j-T+y_1} + \sum_{l=1}^{y_2 N} b_l \gamma_l \xi_l^{L-1-j} \quad (T - y_1 \leq j \leq L - 1), \quad (2)$$

where (λ_k, ψ_k) are the left-eigenvalue and eigenvector pairs of the characteristic, quadratic matrix-polynomial $Q(l) = \sum_{k=0}^y Q_k \lambda^k$ pertaining to the Markov process Y , and (ξ_k, γ_k) are the left-eigenvalue and eigenvector pairs of the characteristic, quadratic matrix polynomial $\bar{Q}(l) = \sum_{k=0}^y Q_{y-k} \lambda^k$. a_l and b_l are the constants, which can be determined with the aid of the state-dependent balance equations [5].

When L is infinite (unbounded) and the ergodicity condition is satisfied, then the above solution reduces to

$$\mathbf{v}_j = \sum_{l=1}^{Ny_1} a_l \psi_l \lambda_l^{j-T+y_1} \quad (j = T - y_1, T - y_1 + 1, \dots). \quad (3)$$

2.2 The QBD-U process

In a QBD-M process, if y_1 or y_2 is *unbounded*, it becomes a QBD-U process (Quasi Simultaneous-Unbounded-Multiple Births and Simultaneous-Unbounded-Multiple Deaths). QBD-U processes are very useful in performance modeling of NGN as we shall see in the rest of this paper. Only in certain special cases of the QBD-U processes, there have been efficient, exact steady state solution methods [14, 15].

2.3 The Generalized Exponential (GE) distribution

Excellent treatment of the GE distribution, its usefulness and applications are available in [35, 47]. The GE distribution is versatile, possessing pseudo-memoryless properties. This makes the solution of many queuing systems and networks employing GE distribution analytically tractable [35]. The GE distribution is given in the following form:

$$F(t) = P(W \leq t) = 1 - (1 - \phi)e^{-\mu t} \quad (t \geq 0), \quad (4)$$

where W is the GE random variable with parameters μ, ϕ . Thus, the GE parameter estimation can be obtained by $1/\nu$, the mean, and C_{coeff}^2 , the squared coefficient of variation of the inter-event time of the sample as

$$1 - \phi = 2/(C_{coeff}^2 + 1) \quad ; \quad \mu = \nu(1 - \phi). \quad (5)$$

For $C_{coeff}^2 > 1$, the GE model is a mixed-type probability distribution, e.g. Hyperexponential-2 having the same mean and coefficient of variation, and with one of the two phases having zero service time, or, a bulk type distribution with an underlying counting process equivalent to a Batch (or Bulk) Poisson Process (BPP) with batch-arrival rate μ and geometrically distributed batch size with mean $1/(1 - \phi)$ and SCV $(C_{coeff}^2 - 1)/(1 + C_{coeff}^2)$ (c.f. [47]). It can be observed that there is an infinite family of BPPs with the same GE-type inter-event time distribution. It is shown that, among them, the BPP with geometrically distributed bulk sizes (referred as the CPP through this paper) is the only one that constitutes a renewal process (the zero inter-event times within a bulk/batch are *independent* if the bulk size distribution is geometric [35]).

The choice of the GE distribution is often motivated by the fact that measurements of actual inter-arrival or service times may be generally limited in accuracy, and so only a few parameters (for example the mean and variance) can be computed reliably. Typically, when only the mean and variance can be relied upon, a choice of a distribution which implies least bias (bias means, introduction of arbitrary and false assumptions) is that of GE-type distribution [35, 47].

2.4 The CPP, MMCP and $MM\sum_{k=1}^K CPP_k$ processes

Though BPP and CPP are synonymous, when we refer to a CPP in this paper, we actually refer to a CPP with independent and geometric batch-sizes, for convenience of referring.

When the parameters of a CPP are modulated by an external Markov chain, we obtain the MMCP (Markov modulated CPP) process. Let the generator matrix of the modulating CTMC (continuous time Markov chain) be given by,

$$Q = \begin{bmatrix} -q_1 & q_{1,2} & \dots & q_{1,N} \\ q_{2,1} & -q_2 & \dots & q_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ q_{N,1} & q_{N,2} & \dots & -q_N \end{bmatrix},$$

where $q_{i,k}(i \neq k)$ is the instantaneous transition rate from phase i to phase k , $q_{i,i} = 0 \forall i$, and $q_i = \sum_{j=1}^N q_{i,j}$, ($i = 1, \dots, N$). Let $\mathbf{r} = (r_1, r_2, \dots, r_N)$ be the vector of equilibrium probabilities of the modulating phases. Then, \mathbf{r} is uniquely determined by the equations, $\mathbf{r}Q = 0$; $\mathbf{r}\mathbf{e}_N = 1$, where \mathbf{e}_N stands for the column vector with N elements, each of which is unity. In the MMCPP arrival process, the inter-arrival time distribution of customers, in phase i , is GE with parameters (σ_i, θ_i) .

The MM $\sum_{k=1}^K CPP_k$ is obtained by Markov modulation of the parameters of the superposition of K independent CPP streams. That is, all the K independent CPPs are jointly Markov modulated.

3 The *HetSigma* Queuing Model

We introduce the terminology *HetSigma* to denote the MM $\sum_{k=1}^K CPP_k/GE/c/L$ G-queue with *heterogenous* servers. In the queue, the effective customer arrival process is MM $\sum_{k=1}^K CPP_k$ in which the superposed K CPPs are independent and their parameters are jointly Markov modulated. The same modulating process also modulates the parameters of the service time and those of the CPP of the negative customers, as we shall see below in detail.

3.1 The arrival process

The arrival and service processes are modulated by the same continuous time, irreducible Markov phase process with N states. Let Q be the generator matrix of this process. The arrival process, in any given modulating phase, is the superposition of K independent CPP arrival streams of customers (or packets, in packet-switched networks) and an independent CPP of *negative* customers. Customers of different arrival streams are not distinguishable. The parameters of the GE inter-arrival time distribution of the k^{th} ($1 \leq k \leq K$) customer arrival stream, in modulating phase i , are $(\sigma_{i,k}, \theta_{i,k})$, and (ρ_i, δ_i) are those of the negative customers. That is, the inter-arrival time probability distribution function is $1 - (1 - \theta_{i,k})e^{-\sigma_{i,k}t}$, in phase i , for the k^{th} stream of customers, and $1 - (1 - \delta_i)e^{-\rho_i t}$ for the negative customers. Thus, in a given phase, all the $K + 1$ arrival *point*-processes are Compound Poisson, with batches arriving at each point having geometric size distribution. Specifically, in phase i , the probability that a batch is of size s is $(1 - \theta_{i,k})\theta_{i,k}^{s-1}$ for the k^{th} stream of customers, and $(1 - \delta_i)\delta_i^{s-1}$ for the negative customers. Strictly during a given phase i , the effective arrival process is $\sum_{k=1}^K CPP_{i,k}$, where $CPP_{i,k}$ is the k^{th} CPP arrival process in the modulating phase i .

Let $\sigma_{i,\cdot}, \overline{\sigma_{i,\cdot}}$ be the average arrival rate of customer batches and customers in phase i respectively. Let $\sigma, \overline{\sigma}$ be the overall average arrival rate of batches and

customers respectively. Then, it can be written

$$\sigma_{i,\cdot} = \sum_{k=1}^K \sigma_{i,k}, \quad \bar{\sigma}_{i,\cdot} = \sum_{k=1}^K \frac{\sigma_{i,k}}{(1 - \theta_{i,k})}, \quad \sigma = \sum_{i=1}^N \sigma_{i,\cdot} r_i, \quad \bar{\sigma} = \sum_{i=1}^N \bar{\sigma}_{i,\cdot} r_i \quad (6)$$

3.2 The GE multi-server

The *HetSigma* queue is the extension of the *Sigma* queue in [10], where the service facility has c *heterogeneous* servers in parallel. The servers are numbered just as their service priorities, *i.e.* $1, 2, \dots, c$, without loss of generality. The GE-distributed service time parameters of server n , in phase i , are $\mu_{i,n}, \phi_{i,n}$ ($n = 1, 2, \dots, c$). A number of scheduling policies can be thought of. Though, in principle, a number of scheduling policies can indeed be modeled by following our methodology, the one that we have adopted in this paper, for illustration and detailed study, is as follows. A set of service priorities is chosen by giving each server a unique service priority, 1 is the highest and c is the lowest. This set can be chosen arbitrarily from the $c!$ different possible ways. However, the impact of choosing service priorities can be very high on the performance measures, whose study is not in the scope of this paper. The optimal allocation of service priorities can be an interesting research item for investigation.

The service discipline is FCFS (First Come First Scheduled, for service) and each server serves at most one positive customer at any given time. Customers, on their completion of service, leave the system. When the number of customers in the system, j , (including those in service if any) is $\geq c$, then only c customers are served with the rest $(j - c)$ waiting for service. When $j < c$, only the first j servers, (*i.e.*, servers numbered $1, 2, \dots, j$), are occupied and the rest are idle. This is made possible by what is known as customer switching. Thus, when server n becomes idle, an awaiting customer would be taken up for service. If there is no awaiting customer, then a customer that is being served by the lowest possible priority server (*i.e.*, among servers $(c, c - 1, \dots, n + 1)$) switches to server n . In such a switching, the (batch) service time is governed by either *resume or repeat with resampling*, thus preserving the Markovian property. The switching is instantaneous or the switching time is treated negligible. Negative customers neither wait in the queue, nor are served.

The operation of the GE server is similar to that described above in the *Sigma* case [10]. The batch size associated with a service completion is bounded by one more than the number of customers waiting to commence service at the departure instant. When $c \leq j < L + 1$, the maximum batch size at a departure instant obviously is $j - c + 1$, only one server being able to complete a service period at any one instant under the assumption of exponentially distributed batch-service times. Thus, in phase i , the probability that a departing batch is of size s can be shown as, $\sum_{n=1}^c \frac{\mu_{i,n}(1-\phi_{i,n})\phi_{i,n}^{s-1}}{\mu_{i,\cdot}}$ for $1 \leq s \leq j - c$ and $\sum_{n=1}^c \frac{\mu_{i,n}\phi_{i,n}^{j-c}}{\mu_{i,\cdot}}$ for $s = j - c + 1$, where $\mu_{i,\cdot} = \sum_{n=1}^c \mu_{i,n}$. However, when $1 \leq j \leq c$, the departing batch has size 1 since each customer is already engaged by a server and there are no customers waiting to commence service.

3.3 Negative customer semantics

A negative customer removes a positive customer in the queue, according to a specified *killing discipline*. A number of different killing disciplines are indeed possible, suitable in different contexts.

The RCE killing discipline We consider here a variant of the RCE killing discipline (removal of customers from the end of the queue), where the most recent positive arrival is removed, but which does *not* allow a customer actually in service to be removed: a negative customer that arrives when there are no positive customers waiting to start service has no effect. We may say that customers in service are immune to negative customers or that the service itself is *immune servicing*. Such a killing discipline is suitable for modeling e.g. load balancing, where work is transferred from overloaded queues but never work, that is, actually in progress.

When a batch of negative customers of size l ($1 \leq l < j - c$) arrives, l positive customers are removed from the end of the queue leaving the remaining $j - l$ positive customers in the system. If $l \geq j - c \geq 1$, then $j - c$ positive customers are removed, leaving none waiting to commence service (queue length becomes c). If $j \leq c$, the negative arrivals have no effect since all customers are in service.

$\bar{\rho}_i$, the average arrival rate of negative customers in phase i and $\bar{\rho}$, the overall average arrival rate of negative customers are given by

$$\bar{\rho}_i = \frac{\rho_i}{1 - \delta_i} ; \quad \bar{\rho} = \sum_{i=1}^N r_i \bar{\rho}_i. \quad (7)$$

Other killing disciplines Apart from the RCE with *immune servicing*, there are two other popular killing disciplines, the RCE-*inimmune servicing* and the RCH killing disciplines. The applicability of the killing disciplines rather depends on the situation and the purpose, and hence depending on these, many more killing disciplines are theoretically possible. Our methodology can easily be extended to many other killing disciplines also, this is explained briefly in this section.

The RCE-inimmune servicing- In this, the negative customer removes the most recent positive arrival regardless of whether it is in service or waiting. This is the traditional killing discipline suited to the modeling of killing signals in speculative parallelism. It can also be used to model cell losses caused by the arrival of a corrupted cell or one encountering a full buffer, when the preceding cells of a packet would be discarded.

The RCH discipline- Another popular killing discipline is the RCH (Removal of customers from the head of the queue) killing discipline. This is appropriate for modeling server breakdowns, where a customer in service will be lost for sure and may be also a portion of queue of waiting customers. The RCH killing discipline is already applied to the case of the MM CPP/GE/c/1 G-queue in [14].

3.4 Condition for stability

When L is finite, the system is ergodic since the representing CTMC is irreducible. Otherwise, i.e. when the queuing capacity is unbounded, the overall average departure rate increases with the queue length, and its maximum (the overall average departure rate when the queue length tends to ∞) can be determined as,

$$\bar{\mu} = \sum_{n=1}^c \sum_{i=1}^N \frac{r_i \mu_{i,n}}{1 - \phi_{i,n}}. \quad (8)$$

Hence, the necessary and sufficient condition for the existence of steady state probabilities is, $\bar{\sigma} < \bar{\rho} + \bar{\mu}$.

3.5 The Markov model

The state of the system at any time t can be specified completely by two integer-valued random variables, $I(t)$ and $J(t)$. $I(t)$ varies from 1 to N , representing the phase of the modulating Markov chain, and $0 \leq J(t) < L + 1$ represents the number of positive customers in the system at time t , including any in service. The system is now modeled by a CTMC \bar{Y} (Y if L is infinite), on a rectangular lattice strip. Let $I(t)$, the phase, vary in the horizontal direction and $J(t)$, the queue length or *level*, in the vertical direction. We denote the steady state probabilities by $\{p_{i,j}\}$, where $p_{i,j} = \lim_{t \rightarrow \infty} \text{Prob}(I(t) = i, J(t) = j)$, and let $\mathbf{v}_j = (p_{1,j}, \dots, p_{N,j})$. The process \bar{Y} evolves due to the following instantaneous transition rates:

- (a) $q_{i,k}$ – purely lateral transition rate – from state (i, j) to state (k, j) , for all $j \geq 0$ and $1 \leq i, k \leq N$ ($i \neq k$), caused by a phase transition in the modulating Markov process ($q_{i,i} = 0$);
- (b) $B_{i,j,j+s}$ – s -step upward transition rate – from state (i, j) to state $(i, j + s)$, caused by a new batch arrival of size s of positive customers in phase i . For a given j , s can be seen as bounded when L is finite and unbounded when L is infinite;
- (c) $C_{i,j,j-s}$ – s -step downward transition rate – from state (i, j) to state $(i, j - s)$, ($j - s \geq c + 1$), caused by either a batch service completion of size s or a batch arrival of negative customers of size s , in phase i ;
- (d) $C_{i,c+s,c}$ – s -step downward transition rate – from state $(i, c + s)$ to state (i, c) , caused by a batch arrival of negative customers of size $\geq s$ or a batch service completion of size s ($1 \leq s \leq L - c$), in phase i ;
- (e) $C_{i,c-1+s,c-1}$ – s -step downward transition rate, from state $(i, c - 1 + s)$ to state $(i, c - 1)$, caused by a batch departure of size s ($1 \leq s \leq L - c + 1$), in phase i ;
- (f) $C_{i,j+1,j}$ – 1-step downward transition rate, from state $(i, j + 1)$ to state (i, j) , ($c \geq 2$; $0 \leq j \leq c - 2$), caused by a single departure, in phase i .

Notice that \bar{Y} and Y (i.e., when $L = \infty$) are essentially QBD-U processes.

3.6 The transition rate matrices

The transition rate matrices and parameters can be obtained as [8].

$$\begin{aligned}
B_{i,j-s,j} &= \sum_{k=1}^K (1 - \theta_{i,k}) \theta_{i,k}^{s-1} \sigma_{i,k} \quad (\forall i; 0 \leq j-s \leq L-2; j-s < j < L); \\
B_{i,j,L} &= \sum_{k=1}^K \sum_{s=L-j}^{\infty} (1 - \theta_{i,k}) \theta_{i,k}^{s-1} \sigma_{i,k} = \sum_{k=1}^K \theta_{i,k}^{L-j-1} \sigma_{i,k} \quad (\forall i; j \leq L-1); \\
C_{i,j+s,j} &= \sum_{n=1}^c \mu_{i,n} (1 - \phi_{i,n}) \phi_{i,n}^{s-1} + (1 - \delta_i) \delta_i^{s-1} \rho_i \\
&\quad (\forall i; c+1 \leq j \leq L-1; 1 \leq s \leq L-j) \\
&= \sum_{n=1}^c \mu_{i,n} (1 - \phi_{i,n}) \phi_{i,n}^{s-1} + \delta_i^{s-1} \rho_i \quad (\forall i; j = c; 1 \leq s \leq L-c) \\
&= \sum_{n=1}^c \phi_{i,n}^{s-1} \mu_{i,n} \quad (\forall i; j = c-1; 1 \leq s \leq L-c+1) \\
&= 0 \quad (\forall i; c \geq 2; 0 \leq j \leq c-2; s \geq 2) \\
&= \sum_{n=1}^{j+1} \mu_{i,n} \quad (\forall i; c \geq 2; 0 \leq j \leq c-2; s = 1);
\end{aligned}$$

Define,

$$\begin{aligned}
B_{j-s,j} &= \text{Diag} [B_{0,j-s,j}, B_{1,j-s,j}, \dots, B_{N,j-s,j}] \quad (j-s < j \leq L); \\
B_s &= B_{j-s,j} \quad (j < L)
\end{aligned}$$

$$= \text{Diag} \left[\sum_{k=1}^K \sigma_{0,k} (1 - \theta_{0,k}) \theta_{0,k}^{s-1}, \dots, \sum_{k=1}^K \sigma_{N,k} (1 - \theta_{N,k}) \theta_{N,k}^{s-1} \right];$$

$$\Sigma_k = \text{Diag} [\sigma_{0,k}, \sigma_{1,k}, \dots, \sigma_{N,k}] \quad (k = 1, 2, \dots, K);$$

$$\Theta_k = \text{Diag} [\theta_{0,k}, \theta_{1,k}, \dots, \theta_{N,k}] \quad (k = 1, 2, \dots, K);$$

$$\Sigma = \sum_{k=1}^K \Sigma_k;$$

$$R = \text{Diag} [\rho_0, \rho_1, \dots, \rho_N]; \quad \Delta = \text{Diag} [\delta_0, \delta_1, \dots, \delta_N];$$

$$M_n = \text{Diag} [\mu_{0,n}, \mu_{1,n}, \dots, \mu_{N,n}] \quad (n = 1, 2, \dots, c);$$

$$\Phi_n = \text{Diag} [\phi_{0,n}, \phi_{1,n}, \dots, \phi_{N,n}] \quad (n = 1, 2, \dots, c);$$

$$\begin{aligned}
C_j &= \sum_{n=1}^j M_n \quad (1 \leq j \leq c); \\
&= \sum_{n=1}^c M_n = C \quad (j \geq c); \\
C_{j+s,j} &= \text{Diag}[C_{0,j+s,j}, C_{1,j+s,j}, \dots, C_{N,j+s,j}]; \\
E &= \text{Diag}(\mathbf{e}'_N).
\end{aligned}$$

Then, we get,

$$\begin{aligned}
B_s &= \sum_{k=1}^K \Theta_k^{s-1} (E - \Theta_k) \Sigma_k; \quad B_1 = B = \sum_{k=1}^K (E - \Theta_k) \Sigma_k; \\
B_{L-s,L} &= \sum_{k=1}^K \Theta_k^{s-1} \Sigma_k; \\
C_{j+s,j} &= \sum_{n=1}^c M_n (E - \Phi_n) \Phi_n^{s-1} + R(E - \Delta) \Delta^{s-1} \\
&\quad (c+1 \leq j \leq L-1; s = 1, 2, \dots, L-j); \\
&= \sum_{n=1}^c M_n (E - \Phi_n) \Phi_n^{s-1} + R \Delta^{s-1} \quad (j = c; s = 1, 2, \dots, L-c); \\
&= \sum_{n=1}^c M_n \Phi_n^{s-1} \quad (j = c-1; s = 1, 2, \dots, L-c+1); \\
&= 0 \quad (c \geq 2; 0 \leq j \leq c-2; s \geq 2); \\
&= C_{j+1} \quad (c \geq 2; 0 \leq j \leq c-2; s = 1).
\end{aligned}$$

4 Steady State Solution

The method presented in this Section is done for sufficiently large L such that $L \geq 2c + K + 3$. When $L < 2c + K + 3$, then the Markov process \bar{Y} can be solved by traditional methods [48].

First, the steady state balance equations are obtained [8]. Let the term $\langle j \rangle$ denote the vector balance equation for level j . A novel methodology is developed to solve these equations *exactly and efficiently*. First these complicated equations are *transformed* to a computable form by using mathematically oriented transformations. The resulting transformed equations are of the QBD-M type and hence can be solved.

Define the functions, $F_{K,l}$ ($l = 1, 2, \dots, K$) and $H_{c,n}$ ($n = 1, 2, \dots, c$) using their properties and recursions as given below.

$$\begin{aligned}
F_{k,0} &= E, \quad F_{k,k} = \prod_{i=1}^k \Theta_i \quad (k = 1, 2, \dots, K); \\
F_{k,l} &= 0 \quad (k = 1, 2, \dots, K; l < 0); \quad F_{k,l} = 0 \quad (k = 1, 2, \dots, K; l > k); \\
F_{1,0} &= E; \quad F_{1,1} = \Theta_1; \\
F_{k,l} &= F_{k-1,l} + \Theta_k F_{k-1,l-1} \quad (2 \leq k \leq K, 1 \leq l \leq k-1);
\end{aligned} \tag{9}$$

$$\begin{aligned}
H_{m,0} &= E, \quad H_{m,m} = \prod_{i=1}^m \Phi_i \quad (m = 1, 2, \dots, c); \\
H_{m,n} &= 0 \quad (m = 1, 2, \dots, c; n < 0); \quad H_{m,n} = 0 \quad (m = 1, 2, \dots, c; n > m); \\
H_{1,0} &= E; \quad H_{1,1} = \Phi_1; \\
H_{m,n} &= H_{m-1,n} + \Phi_m H_{m-1,n-1} \quad (2 \leq m \leq c, 1 \leq n \leq m-1).
\end{aligned} \tag{10}$$

Please note E is the Identity matrix of size $N \times N$. The parameters Θ_i are the same as in [8].

Transformation 1 Modify simultaneously the balance equations for levels j ($L-2-c \geq j \geq c+K+1$), by the transformation:

$$\begin{aligned}
\langle \mathbf{j} \rangle^{(1)} &\leftarrow \langle \mathbf{j} \rangle + \sum_{l=1}^K (-1)^l \langle \mathbf{j}-\mathbf{1} \rangle F_{K,l} \quad (c+K+1 \leq j \leq L-2-c); \\
\langle \mathbf{j} \rangle^{(1)} &\leftarrow \langle \mathbf{j} \rangle \quad (j > L-2-c \text{ or } j < c+K+1).
\end{aligned}$$

Apply the second transformation to the resulting equations.

Transformation 2 Modify simultaneously the balance equations for levels j ($L-2-c \geq j \geq c+K+1$), by the transformation:

$$\begin{aligned}
\langle \mathbf{j} \rangle^{(2)} &\leftarrow \langle \mathbf{j} \rangle^{(1)} + \sum_{n=1}^c (-1)^n \langle \mathbf{j} + \mathbf{n} \rangle^{(1)} H_{c,n} \\
&\quad (c+K+1 \leq j \leq L-2-c); \\
\langle \mathbf{j} \rangle^{(2)} &\leftarrow \langle \mathbf{j} \rangle^{(1)} \quad (j > L-2-c \text{ or } j < c+K+1).
\end{aligned}$$

Apply the third and final transformation to the resulting equations.

Transformation 3 Modify simultaneously the balance equations for levels j ($L-2-c \geq j \geq c+K+1$), by the transformation:

$$\begin{aligned}
\langle \mathbf{j} \rangle^{(3)} &\leftarrow \langle \mathbf{j} \rangle^{(2)} - \langle \mathbf{j} + \mathbf{1} \rangle^{(2)} \Delta \quad (c+K+1 \leq j \leq L-2-c); \\
\langle \mathbf{j} \rangle^{(3)} &\leftarrow \langle \mathbf{j} \rangle^{(2)} \quad (j > L-2-c \text{ or } j < c+K+1).
\end{aligned}$$

Theorem 1. *With these above three transformations, the transformed balance equation, $\langle \mathbf{j} \rangle^{(3)}$'s, for the rows $(c + K + 1 \leq j \leq L - 2 - c)$, will be of the form:*

$$\mathbf{v}_{j-K}Q_0 + \mathbf{v}_{j-K+1}Q_1 + \dots + \mathbf{v}_{j+c+1}Q_{K+c+1} = 0$$

$$(j = L - 2 - c, L - 1 - c, \dots, c + K + 1), \quad (11)$$

where $Q_0, Q_1, \dots, Q_{K+c+1}$ are $K+c+2$ number of j -independent matrices which can be derived algebraically from the system parameters

Proof. With Transformation 1, we get

$$\langle \mathbf{j} \rangle^{(1)} \leftarrow \langle \mathbf{j} \rangle + \sum_{l=1}^K (-1)^l \langle \mathbf{j} - \mathbf{l} \rangle F_{K,l}. \quad (12)$$

Applying Transformation 2 to the j^{th} row, from the above (12), we get

$$\langle \mathbf{j} \rangle^{(2)} \leftarrow \langle \mathbf{j} \rangle^{(1)} + \sum_{n=1}^c (-1)^n \langle \mathbf{j} + \mathbf{n} \rangle^{(1)} H_{c,n}. \quad (13)$$

Expanding the terms, equation (13) can be written as

$$\begin{aligned} \langle \mathbf{j} \rangle^{(2)} \leftarrow \langle \mathbf{j} \rangle + \sum_{l=1}^K (-1)^l \langle \mathbf{j} - \mathbf{l} \rangle F_{K,l} \\ + \sum_{n=1}^c (-1)^n \left[\langle \mathbf{j} + \mathbf{n} \rangle + \sum_{l=1}^K (-1)^l \langle \mathbf{j} - \mathbf{l} + \mathbf{n} \rangle F_{K,l} \right] H_{c,n}. \end{aligned} \quad (14)$$

Applying Transformation 3 to the j^{th} row, and substituting from the above (14), for $\langle \mathbf{j} + \mathbf{1} \rangle^{(2)}$

$$\begin{aligned} \langle \mathbf{j} \rangle^{(3)} \leftarrow \langle \mathbf{j} \rangle + \sum_{l=1}^K (-1)^l \langle \mathbf{j} - \mathbf{l} \rangle F_{K,l} \\ + \sum_{n=1}^c (-1)^n \left[\langle \mathbf{j} + \mathbf{n} \rangle + \sum_{l=1}^K (-1)^l \langle \mathbf{j} - \mathbf{l} + \mathbf{n} \rangle F_{K,l} \right] H_{c,n} \\ - \left[\langle \mathbf{j} + \mathbf{1} \rangle + \sum_{l=1}^K (-1)^l \langle \mathbf{j} + \mathbf{1} - \mathbf{l} \rangle F_{K,l} \right] \Delta \\ - \sum_{n=1}^c (-1)^n \left[\langle \mathbf{j} + \mathbf{1} + \mathbf{n} \rangle + \sum_{l=1}^K (-1)^l \langle \mathbf{j} + \mathbf{1} - \mathbf{l} + \mathbf{n} \rangle F_{K,l} \right] H_{c,n} \Delta. \end{aligned} \quad (15)$$

Expanding and grouping the terms together, equation (15) can be written as

$$\langle \mathbf{j} \rangle^{(3)} \leftarrow \sum_{m=-c-1}^K \langle \mathbf{j} - \mathbf{m} \rangle G_{K,c,m}, \quad (16)$$

where

$$\begin{aligned} G_{K,c,m} &= \sum_{\substack{l=-1,\dots,K \\ n=0,\dots,c}}^{l-n=m} (-1)^{l+n} [F_{K,l} H_{c,n} + F_{K,l+1} H_{c,n} \Delta] \\ &= \sum_{n=0}^c (-1)^{m+2n} [F_{K,m+n} + F_{K,m+n+1} \Delta] H_{c,n} \\ &= (-1)^m \sum_{n=0}^c [F_{K,m+n} + F_{K,m+n+1} \Delta] H_{c,n} \quad (m = -1 - c, \dots, K). \end{aligned} \quad (17)$$

The balance equations $\langle \mathbf{j} + \mathbf{c} + \mathbf{1} \rangle, \dots, \langle \mathbf{j} \rangle, \dots, \langle \mathbf{j} - \mathbf{1} \rangle, \dots, \langle \mathbf{j} - \mathbf{K} \rangle$, respectively are given by,

$$\begin{aligned} \sum_{s=1}^{j+c+1} \sum_{k=1}^K \mathbf{v}_{j+c+1-s} \Theta_k^{s-1} (E - \Theta_k) \Sigma_k + \mathbf{v}_{j+c+1} [Q - \Sigma - C_{j+c} - R] \\ + \sum_{s=1}^{L-j-c-1} \mathbf{v}_{j+c+1+s} C_{j+c+1+s, j+c+1} = 0; \\ \vdots \end{aligned}$$

$$\begin{aligned} \sum_{s=1}^j \sum_{k=1}^K \mathbf{v}_{j-s} \Theta_k^{s-1} (E - \Theta_k) \Sigma_k + \mathbf{v}_j [Q - \Sigma - C_j - R] \\ + \sum_{s=1}^{L-j} \mathbf{v}_{j+s} C_{j+s, j} = 0; \\ \vdots \end{aligned}$$

$$\begin{aligned} \sum_{s=1}^{j-l} \sum_{k=1}^K \mathbf{v}_{j-l-s} \Theta_k^{s-1} (E - \Theta_k) \Sigma_k + \mathbf{v}_{j-l} [Q - \Sigma - C_{j-l} - R] \\ + \sum_{s=1}^{L-j+l} \mathbf{v}_{j-l+s} C_{j-l+s, j-l} = 0; \\ \vdots \end{aligned}$$

$$\begin{aligned} \sum_{s=1}^{j-K} \sum_{k=1}^K \mathbf{v}_{j-K-s} \Theta_k^{s-1} (E - \Theta_k) \Sigma_k + \mathbf{v}_{j-K} [Q - \Sigma - C_{j-K} - R] \\ + \sum_{s=1}^{L-j+K} \mathbf{v}_{j-K+s} C_{j-K+s, j-K} = 0. \end{aligned}$$

Substituting or applying the above to (16), for the coefficients (Q_{K-m}) of \mathbf{v}_{j-m} in $\langle \mathbf{j} \rangle^{(3)}$, we get

$$\begin{aligned} Q_{K-m} &= \sum_{l=-1-c}^{m-1} \left[\sum_{n=1}^K \Theta_n^{m-l-1} (E - \Theta_n) \Sigma_n \right] G_{K,c,l} + [Q - \Sigma - C_{j-m} - R] G_{K,c,m} \\ &\quad + \sum_{l=m+1}^K [C_{j-m, j-l}] G_{K,c,l} \\ &\quad (m = j - L, \dots, -2, -1, 0, \dots, K, \dots, j). \end{aligned} \quad (18)$$

Also, for $m = -1 - c, 0, \dots, K$, substituting $C_{j-m} = C$ and $C_{j-m, j-l} = C_{j-l+l-m, j-l} = \sum_{n=1}^c M_n (E - \Phi_n) \Phi_n^{l-m-1} + R (E - \Delta) \Delta^{l-m-1}$ in (18), we get

$$\begin{aligned} Q_{K-m} &= \sum_{l=-1-c}^{m-1} \left[\sum_{n=1}^K \Theta_n^{m-l-1} (E - \Theta_n) \Sigma_n \right] G_{K,c,l} + [Q - \Sigma - C - R] G_{K,c,m} \\ &\quad + \sum_{l=m+1}^K \left[\sum_{n=1}^c M_n (E - \Phi_n) \Phi_n^{l-m-1} + R (E - \Delta) \Delta^{l-m-1} \right] G_{K,c,l} \\ &\quad (m = -1 - c, \dots, 0, \dots, K). \end{aligned} \quad (19)$$

Using the above, the required Q_l 's can be computed easily. Notice the above Q_l 's in equation (19) are j -independent. The other coefficients, *i.e.* those of $\mathbf{v}_{j-K-1}, \mathbf{v}_{j-K-2}, \dots, \mathbf{v}_0$ and of $\mathbf{v}_{j+c+2}, \mathbf{v}_{j+c+3}, \dots$, can be shown to be zero, case wise, by using computer programs in *Mathematica* or other symbolic manipulation languages. A rigorous proof is indeed possible, but it is beyond the scope of the present paper \square

It is observed that the solution of the HetSigma queue can be performed within the framework of the QBD-M processes with the following threshold parameters $y_1 = K; y = K + c + 1; T_1 = c + K + 1$.

After obtaining $F_{K,l}$'s and $H_{c,n}$'s thus, $G_{K,c,k}$, ($k = -1 - c, \dots, K$) can be computed from (17). Then, using them directly in (19), the required Q_l ($l = 0, 1, \dots, K + c + 1$) can be computed. An alternate way of computing the $G_{K,c,l}$'s is by the following properties and recursion which are obtained from (9), (10) and (17) as

$$\begin{aligned} G_{k,n,l} &= G_{k,n-1,l} - \Phi_n G_{k,n-1,l+1} \\ &\quad (2 \leq k \leq K, -1 \leq l + c \leq k + n \leq k + c), \\ G_{k,c,l} &= G_{k-1,c,l} - \Theta_k G_{k-1,c,l-1} \quad (2 \leq k \leq K, -1 \leq l \leq k + c). \end{aligned} \quad (20)$$

Thus, the resulting equations (11) corresponding to the rows from $c + K + 1$ to $L - 2 - c$, are of the same form as those of the QBD-M processes and hence have an efficient solution by several alternative methods such as the spectral expansion method, Bini-Meini's method [3] and the matrix-geometric methods with folding and block-size enlargement. In our implementation, we have used the spectral expansion method. The required unknowns in the spectral expansion solution and the steady state probabilities can be computed using the remaining balance equations and the normalisation equation.

5 Applications

The *HetSigma* queue has been applied to evaluate performance problems in telecommunication systems [8, 23]. In this section, we present the application of the *HetSigma* queue to the performance analysis of MPLS networks which are being used in the backbone of IP networks.

5.1 A model for a multipath routing in MPLS networks

An example of an MPLS domain with routers and links is illustrated in Figure 1. Traffic demands traversing the MPLS domain are conveyed along pipes called Label Switched Paths (LSPs). When a packet arrives at the ingress router – a Label Edge Router (LER) of the MPLS domain, the LER classifies incoming IP (or other packets for example Ethernet or even MPLS) packets to the appropriate FEC (Forward Equivalence Class) and encapsulates the packets within MPLS packets. Labels are automatically assigned with the use of appropriate protocols, though labeling can also be allocated manually by the network administrator. A routing table (as the result of assignments of labels) in the LER is used to switch packets in MPLS networks.

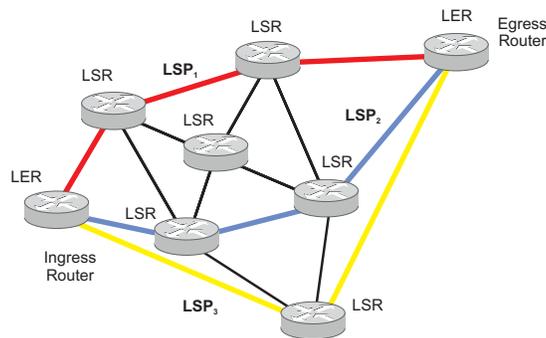


Fig. 1. MPLS domain

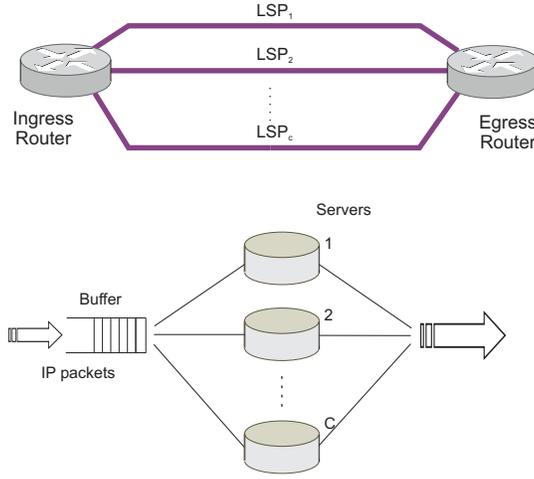


Fig. 2. Model of an ingress node performing load balancing

In what follows, we describe the proposed model for an ingress-egress router pair illustrated in Figure 2. Several paths can be defined and determined between a given ingress-egress (IE) node pair in the MPLS network according to some predefined criteria of the MPLS traffic engineering that is applied (e.g.: paths with disjoint edges) for a single service class. Assume that there are c paths with different bandwidths to be established between the IE node pair for a specific service class. We model the c distinct paths (LSPs) in the system as the c distinct heterogeneous servers in parallel in the corresponding queuing model that we are introducing here for modeling the communication between the IE pair. The LSPs in the network and the corresponding servers in the model are numbered in the same order. The GE-distributed service time parameters of the n^{th} server ($n = 1, 2, \dots, c$) are denoted by (μ_n, ϕ_n) when the server is functional. L is the queuing capacity (finite or infinite), in all phases, including the customers in service, if any.

The packet arrival stream is represented by the $MM \sum_{k=1}^K CPP_k$ arrival process, this can accommodate traffic-burstiness, correlations among inter-arrival times and correlations among batch sizes. It has been shown in the recent work [21, 23] that the CPP is accurate enough to model real traffic (when CPP parameters are estimated from the captured traffic) and can be used for the performance evaluation of real systems. The arrival process (the $MM \sum_{k=1}^K CPP_k$) is inherently modulated by a CTMC X with N_1 states, with generator matrix Q_X .

LSPs are prone to failures because of various reasons (e.g.: unreliable equipment, hardware failures, software bugs or cable cut). Such faults and failures may affect the operation of LSPs and cause packet losses and delays. In case of failures, the load balancing mechanism can move packets that are queued for the affected LSPs to the unaffected LSPs.

When the failed link is repaired, the repaired LSP can be used again. Repair strategy defines the order of repairing failed links, when there are more failed links than repairmen (it is assumed simultaneous repairs can happen). Repair strategies can be preemptive or non-preemptive. However, it is not very realistic to apply a preemptive-priority repair strategy for the operation of networks in the case of link cuts because of the travel cost of a maintenance team. In this paper, we assume that one maintenance team is available to repair failures in the network. However, the analysis can be easily extended to the case of multiple maintenance facilities as well.

We consider four repair strategies, FCFS (First-Come-First-Served), LCFS (Last-Come-First-Served non-preemptive) and two based on link-priorities. In the strategies based on link-priorities, links with higher priorities should be set into repair sooner, even if these failures have occurred later. The priority list of links may then be constructed in a greedy way, with a view to repair earlier the link which would fetch larger gain in performance.

We assume the LSP states (or LSP configurations) that arise due to failures and repairs of the LSPs can be described by a CTMC called Z . These LSP configurations would indeed correspond to all possible multi-server configurations (also termed, operative states of the multi-server) with functional as well as failed servers, in the corresponding queuing model. This can well be so when exponential or phase-type failures and repairs are assumed. Let N_2, Q_Z denote the number of server-configurations and the generator matrix of the arising CTMC Z . Indeed N_2 and Q_Z would depend on the parameters of the parallel servers, number of repairmen and the repair strategy.

In the presence of one repair team, we would need $N_2 = \sum_{l=0}^c Pe(c, l) = 2^{c+1}$ operative states for FCFS and LCFS repair strategies, where $Pe(c, l)$ is the number of permutations of c distinct elements taken l at a time. $Pe(c, l) = \frac{c!}{(c-l)!}$. Let ξ_k and η_k be the failure and repair rates respectively, of the k^{th} LSP (that is, k^{th} server in the multi-server model). Then Q_Z can be determined (illustration for $c = 3$ is presented below). Many other repair strategies can also be modeled, however, that is not in the scope of the present paper.

Both the arrival and the service processes can be thought of being jointly modulated by the same continuous time, irreducible Markov process, with N states where $N = N_1 \cdot N_2$. That generator matrix of this joint modulating process is denoted by Q where Q is determined as

$$Q = Q_Z \oplus Q_X. \quad (21)$$

Let the random variable $I_1(t)$ ($1 \leq I_1(t) \leq N_1$) represent the phase of the modulating process X at any time t . We introduce $I_2(t)$ – an integer-valued random variable – to describe the server configuration of the model (which corresponds to the network state) at time t . We define the following function,

$$\gamma(I_2(t), n) = \begin{cases} 1 & \text{the } n^{th} \text{ server is functional} \\ 0 & \text{otherwise} \end{cases}. \quad (22)$$

The state $I(t)$ ($1 \leq I(t) \leq N$; $N = N_1 \cdot N_2$) of the joint modulating process is constructed by lexicographically sorting the two variables $(I_1(t), I_2(t))$ as illustrated in Table 1.

Table 1. Order of the phase variable

I	1	2	...	N_1	$N_1 + 1$...	$2N_1$...	$N_1N_2 - N_1 + 1$...	N_1N_2
(I_1, I_2)	$(1, 1)$	$(2, 1)$...	$(N_1, 1)$	$(1, 2)$...	$(N_1, 2)$...	$(1, N_2)$...	(N_1, N_2)

From table 1, we can write the following equations,

$$\begin{aligned}
 I(t) &= I_1(t) + (I_2(t) - 1)N_1, I_2(t) = f_2(I(t)) = \left\lfloor \frac{I(t) - 1}{N_1} \right\rfloor + 1, \\
 I_1(t) &= f_1(I(t)) = [(I(t) - 1) \bmod N_1] + 1.
 \end{aligned} \tag{23}$$

Consequently, the parameters of the arrival process are mapped as follows. The parameters of the GE inter-arrival time distribution of the k^{th} ($1 \leq k \leq K$) customer arrival stream in phase i ($i = 1, \dots, N$) are $(\sigma_{f_1(i),k}, \theta_{f_1(i),k})$. The service time parameters of the n^{th} server ($n = 1, 2, \dots, c$) in phase i ($1 \leq i \leq N$), denoted by $(\mu_{i,n}, \phi_{i,n})$ can be determined as,

$$\mu_{i,n} = \gamma(f_2(i), n)\mu_n, \phi_{i,n} = \gamma(f_2(i), n)\phi_n. \tag{24}$$

5.2 Numerical results

An elaborate case study is carried out to determine the performance of a specific ingress-egress node pair in an European Optical Network topology [1]. The network in Figure 3 contains 19 nodes and 78 optical links.

Traffic to the ingress node to be carried to the egress node, shown in the figure, is assumed follow the ON- OFF process with two states ($ON - OFF$). The ON and OFF periods are exponentially distributed with mean 0.4 and 0.5 s respectively. The distribution of inter-arrival times in state ON is GE with parameters ($\sigma = 150.6698990$ (1/s), $\theta = 0.526471$). In state OFF no packets arrive. These parameters of the arrival stream are indeed derived from the inter-arrival times of the recorded samples of the real Bellcore traffic trace BC-pAug89 [2], by matching the mean and variance.

Assume that there are three LSPs established between node 11 (ingress) and node 4 (egress) of the IP/MPLS network as seen in Figure 3. The first LSP is routed through nodes 10 and 5. The second LSP goes through nodes 17 and 16, while the third one is through node 9. Thus, we have, in the model, $c = 3$ servers with GE service time parameters $(\mu_i, \phi_i$ ($i = 1, 2, 3$)) given by, $\mu_1 = 96$ (1/s), $\mu_2 = 128$ (1/s), $\mu_3 = 160$ (1/s); $\phi_1 = \phi_2 = \phi_3 = 0.109929$. Note that

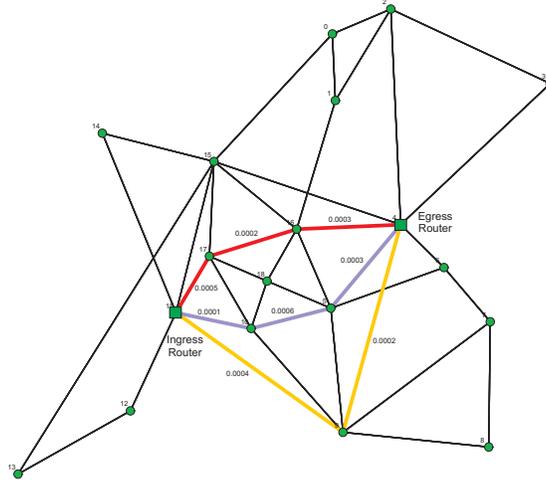


Fig. 3. Network topology

these parameters are obtained, based on the service capacity of the LSPs and the packet lengths from the trace.

A specific LSP fails when a link through which the LSP is routed becomes inoperative (e.g.: cable cut). Normally, the failure rate of each link can be assumed to depend on its length. Using simple calculations, the failure rates are obtained as $\xi_1 = 0.001$, $\xi_2 = 0.001$ and $\xi_3 = 0.0006$. The repair rate values (η_1 , η_2 and η_3) are chosen in such a way to have the availability of the connectivity ensured by the LSPs between the ingress and egress node to be 99.9%.

Four repair strategies are compared in this section as follows. Note that after the identifications of LSP configurations in each strategy, the corresponding Q_Z matrices can be easily determined.

FCFS repairs It can be seen that there would be 16 LSP configurations or operative states represented as: $(0, 0, 0)_{1,2,3}$, $(0, 0, 0)_{1,3,2}$, $(0, 0, 0)_{2,1,3}$, $(0, 0, 0)_{2,3,1}$, $(0, 0, 0)_{3,1,2}$, $(0, 0, 0)_{3,2,1}$, $(0, 0, 1)_{1,2}$, $(0, 0, 1)_{2,1}$, $(0, 1, 0)_{1,3}$, $(0, 1, 0)_{3,1}$, $(1, 0, 0)_{2,3}$, $(1, 0, 0)_{3,2}$, $(0, 1, 1)$, $(1, 0, 1)$, $(1, 1, 0)$ and $(1, 1, 1)$. Here, the k^{th} bit from left within the brackets, is 0 when LSP k is broken, 1 when operative. Also, the suffix indicates the order in which the failed LSPs are to be repaired, if they are greater than one. If the above order is numbered from 0 to 15, then the nonzero and off-diagonal elements of matrix Q_Z can be given by, $Q_Z(0, 10) = Q_Z(1, 11) = Q_Z(6, 13) = \eta_1$, $Q_Z(8, 14) = Q_Z(12, 15) = \eta_1$, $Q_Z(2, 8) = Q_Z(3, 9) = Q_Z(7, 12) = \eta_2$, $Q_Z(10, 14) = Q_Z(13, 15) = \eta_2$, $Q_Z(4, 6) = Q_Z(5, 7) = Q_Z(9, 12) = \eta_3$, $Q_Z(11, 13) = Q_Z(14, 15) = \eta_3$, $Q_Z(10, 3) = Q_Z(11, 5) = Q_Z(13, 7) = \xi_1$, $Q_Z(14, 9) = Q_Z(15, 12) = \xi_1$, $Q_Z(8, 1) = Q_Z(9, 4) = Q_Z(12, 6) = \xi_2$, $Q_Z(14, 11) = Q_Z(15, 13) = \xi_2$, $Q_Z(6, 0) = Q_Z(7, 2) = Q_Z(12, 8) = \xi_3$, and $Q_Z(13, 10) = Q_Z(15, 14) = \xi_3$.

In the above, for example, the state $(1, 0, 0)_{2,3}$ means that LSP 1 is functional, LSPs 2 and 3 are broken, the single repairman is working on LSP 2. From this state, the possible next states are $(0, 0, 0)_{2,3,1}$ which happens by the failure of LSP 1, or, $(1, 1, 0)$ which happens by the repair of LSP 2. These are shown in Q_Z as, $Q_Z(10, 3) = \xi_1, Q_Z(10, 14) = \eta_2$. All other elements of Q_Z can be explained in a similar way.

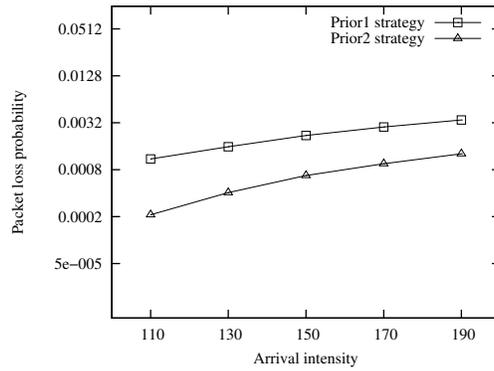


Fig. 4. Packet loss probability versus arrival rate

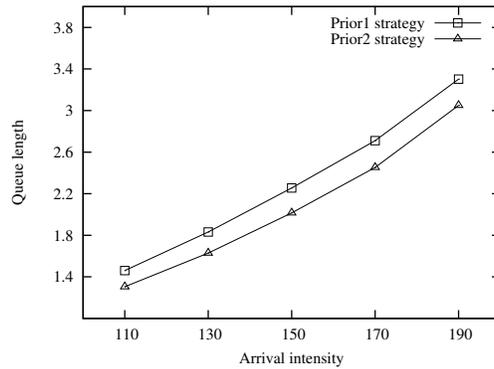


Fig. 5. Mean queue length versus arrival intensity

LCFS In this LCFS non-preemptive repair strategy, the repairman chooses the failed server according to the LCFS discipline, and a repair is never preempted

by a new breakdown. Following the same representation and order as in FCFS strategy, the non zero and off-diagonal elements of matrix Q_Z , in this case, can be obtained as, $Q_Z(0, 10) = Q_Z(1, 11) = Q_Z(6, 13) = \eta_1$, $Q_Z(8, 14) = Q_Z(12, 15) = \eta_1$, $Q_Z(2, 8) = Q_Z(3, 9) = Q_Z(7, 12) = \eta_2$, $Q_Z(10, 14) = Q_Z(13, 15) = \eta_2$, $Q_Z(4, 6) = Q_Z(5, 7) = Q_Z(9, 12) = \eta_3$, $Q_Z(11, 13) = Q_Z(14, 15) = \eta_3$, $Q_Z(10, 2) = Q_Z(11, 4) = Q_Z(13, 7) = \xi_1$, $Q_Z(14, 9) = Q_Z(15, 12) = \xi_1$, $Q_Z(8, 0) = Q_Z(9, 5) = Q_Z(12, 6) = \xi_2$, $Q_Z(14, 11) = Q_Z(15, 13) = \xi_2$, $Q_Z(6, 1) = Q_Z(7, 3) = Q_Z(12, 8) = \xi_3$, and $Q_Z(13, 10) = Q_Z(15, 14) = \xi_3$.

Priority-based strategy In a LSP-priority based repair strategy, it can be crucial to determine the priority of links according to which failed links are to be repaired. In this section, we try to derive a decision function based on system parameters to decide upon the priorities of the failed links.

Let f_i and r_i be the failure and repair rates of link i respectively and b_i , its bandwidth. The bandwidth of a link is computed from the summation of the LSPs bandwidths passing through this particular link. This approach ensures that links used by several LSPs may get higher priority, over the links used only by one LSP, when the former has larger bandwidth contribution. The w_i , the weight of link i is set to

$$w_i = \frac{r_i}{f_i} * b_i. \quad (25)$$

A repair strategy called *Prior1* is based entirely on the failure rate of the links. That is, when there are several failed links, the link with a smallest failure rate is set into repair by the single repair facility available. Another strategy called *Prior2* is based on the decision function (w_i). If $w_i > w_k$, then link i is given higher repair priority than link k in *Prior2*.

To compare *Prior1* and *Prior2*, we plot the curves concerning packet loss and average number of customers waiting in the queue are versus the arrival rate of packets in Figure 4 and 5, respectively. We can observe that applying *Prior2* repair strategy results in a better performance characteristics than *Prior1*. The *Prior2* strategy takes more information into account about the network links towards the decision of repair order which can explain the better results. It balances not only failure rates, but also the repair rates and the bandwidth of links used by the passing-through LSPs. Therefore, we will only investigate strategies FCFS, LCFS and *Prior2* (will be referred as priority strategy) in what follows.

We now examine the effect of the repair rate of optical links on the performance of the system. In order to ensure that the availability of service between the ingress and egress node to lie between 99.9% and 99.999%, it can be shown that the repair rate should be in the interval [0.0397703,0.405355]. We investigate the effect of the repair rate of link within that range, on the performance. The expected queue length was reduced only by a small extent by the increase of repair rate (see Figure 6). The packet-loss probability varied nearly in inverse-proportion to the repair rate (see Figure 7). Of course, the results would actually

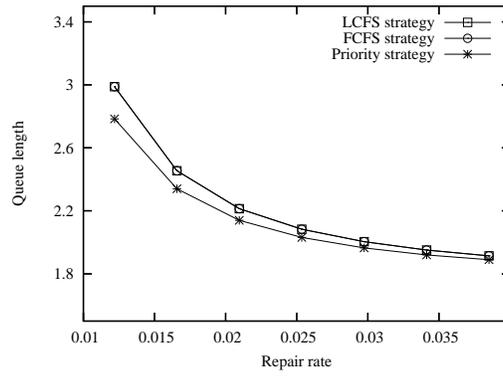


Fig. 6. Mean queue length versus the repair rate

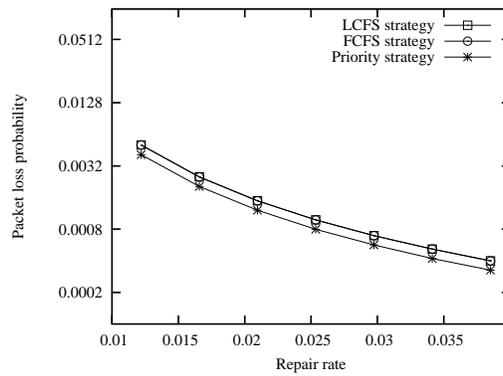


Fig. 7. Packet loss probability versus the repair rate

depend on many other factors too including the load in the network. Accurate estimation of the returns on investment that is used to increase repair rate can be made, only if the cost-patterns involved in increasing the repair rate are known. It may also be observed, in the considered range of experimentation, the various repair strategies considered do not have significant impact on the performance and the availability of the service in the network (in the range of repair rates that are considered).

Figures 8 and 9, where the strategies are compared versus the arrival rate also confirm an observation that the priority based repair outperforms the LCFS and FCFS ones.

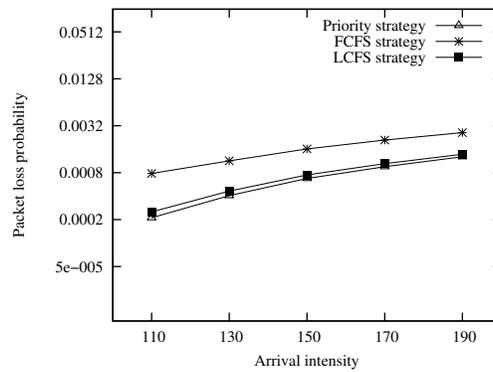


Fig. 8. Packet loss probability versus arrival rate

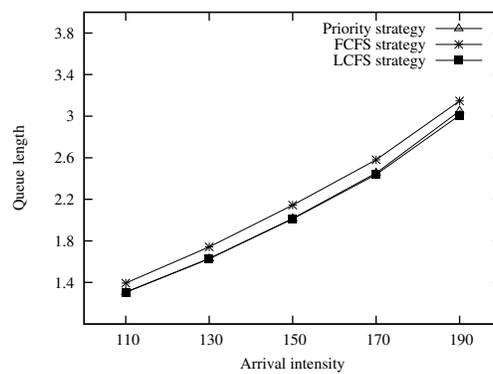


Fig. 9. Mean queue length versus arrival intensity

6 Extensions

6.1 Other killing disciplines

Apart from the killing discipline that was used above, there are two other popular killing disciplines, the RCE-*inimmune servicing* and the RCH killing disciplines. The applicability of the killing disciplines rather depends on the situation and the purpose, and hence depending on these, many more killing disciplines are theoretically possible. Our methodology can easily be extended to many other killing disciplines also, this is explained briefly in this section.

The RCE-*inimmune discipline* In the RCE-*inimmune servicing*, the negative customer removes the most recent positive arrival regardless of whether it is in service or waiting; thus a negative arrival has no effect only when it encounters an empty queue and all servers idle. This is the traditional killing discipline, suited to the modeling of killing signals in speculative parallelism, for example. It can also be used to model cell losses caused by the arrival of a corrupted cell or one encountering a full buffer, when the preceding cells of a packet would be discarded. In this case, the later and the upward transitions remain as before, but the downward transition rates become [13]

$$\begin{aligned}
C_{j+s,j} &= C(E - \Phi)\Phi^{s-1} + R(E - \Delta)\Delta^{s-1} \\
&\quad (c \leq j \leq L - 1; s = 1, 2, \dots, L - j) \\
&= C\Phi^{s-1} + R(E - \Delta)\Delta^{s-1} \quad (c > 1; j = c - 1; 1 \leq s \leq L - c + 1) \\
&= C\Phi^{s-1} + R\Delta^{s-1} \quad (c = 1; j = c - 1 = 0; 1 \leq s \leq L) \\
&= C_{j+1} + R(E - \Delta) \quad (1 \leq j \leq c - 2; s = 1) \\
&= R(E - \Delta)\Delta^{s-1} \quad (1 \leq j \leq c - 2; 2 \leq s \leq L - j) \\
&= R\Delta^{s-1} \quad (c > 1; j = 0; 2 \leq s \leq L) \\
&= C_1 + R \quad (j = 0; s = 1).
\end{aligned}$$

Thus, after obtaining the required transition rates, e.g. lateral, upward and downward, and the balance equations in this case, the same procedure can be used, perhaps with slight appropriate modifications, to transform the balance equations to a suitable form (QBD-M type equations) for solving by the existing methods.

The RCH discipline Another popular killing discipline is the RCH (Removal of customers from the head of the queue) killing discipline. This is appropriate for modeling server breakdowns, where a customer in service will be lost for sure and may be also a portion of queue of waiting customers. Here too, the A matrices and the B matrices remain unchanged, however the C matrices would be different and can be determined. For example, to obtain $C_{j+s,j}$, consider the system (say, a single server system with $c = 1$) in state $(i, j + s)$, where $j + s \leq L$. The rate at which a batch of l negative customers arrives is $(1 - \delta_i)\delta_i^{l-1}\rho_i$ ($l = 1, 2, \dots$).

If $l \geq j + s$, then all the jobs will be removed by the negative customers. If $l < j + s$, then the job *in service* plus $l - 1$ jobs waiting for service would be removed, leaving only $j + s - l$ jobs. However, due to the nature of the generalized exponential service times, a certain number of customers may be 'leaked out', i.e. serviced *instantaneously* just after the killing takes place. Alternatively, if we redefine the operation of the system such that this leakage does not occur, i.e. immediately after a negative arrival, the next customer in the queue (if any) *cannot* skip service, then the equilibrium state probabilities would be same as in the case of the RCE-inimmune servicing.

Assuming leakage, the rate matrices $C_{j+s,j}$ can be derived as [13],

$$\begin{aligned} C_{j+s,j} &= (E - \Phi)\Phi^{s-1}M + (E - \Delta)R(E - \Phi)\{\Phi\Delta\}_{s-1} \\ &\quad (1 \leq j \leq L - 1 ; s \geq 1) ; \\ &= \Phi^{s-1}M + (E - \Delta)R\{\Phi\Delta\}_{s-1} + R\Delta^s \quad (j = 0 ; s \geq 2) ; \\ &= M + R \quad (j = 0 ; s = 1) ; \end{aligned}$$

where,

$$\begin{aligned} \{\Phi\Delta\}_{s-1} &= \sum_{k=0}^{s-1} \Phi^{s-1-k} \Delta^k \quad (s \geq 2) ; \\ &= E \quad (s = 1) . \end{aligned}$$

This case also can be modeled following the previous procedures, perhaps with minor modifications.

6.2 truncated-CPP(t-CPP)

By incorporating several CPP's and superposing them, non-geometric batch size arrivals could be modeled, to certain extent and with certain limitations. That flexibility can be increased, in fact greatly, by replacing the CPP's by truncated-CPP's (or t-CPP's). The method of transforming the balance equation for efficient solution may be applicable in principle, but with some modifications, to this case as well and hence can be extended with some effort.

Let the parameters of the k^{th} t-CPP in the i^{th} phase be $(\sigma_{i,k}, \theta_{i,k}, b_{i,k}, d_{i,k})$, which means the batch size is geometrically distributed and *bounded* by $b_{i,k}$ as the minimum batch size and $d_{i,k}$ as the maximum. Also, the inter-arrival time between successive batches, during that modulating phase, is exponential with parameter $\sigma_{i,k}$. The probability of batch size being s of this t-CPP is then given by $\frac{(1-\theta_{i,k})\theta_{i,k}^{s-1}}{\theta_{i,k}^{b_{i,k}-1} - \theta_{i,k}^{d_{i,k}}}$ if $b_{i,k} \leq s \leq d_{i,k}$, 0 otherwise. Hence, the overall batch size distribution in phase i arrivals is then modified as

$$\pi_{l/i} = \sum_{k=1}^K \frac{\sigma_{i,k} (1 - \theta_{i,k}) \theta_{i,k}^{l-1}}{\sigma_{i,k} \cdot (\theta_{i,k}^{b_{i,k}-1} - \theta_{i,k}^{d_{i,k}})} f_{i,k,l}, \quad (26)$$

where $f_{i,k,l} = 1$ if $b_{i,k} \leq l \leq d_{i,k}$, else 0. And, the overall batch size distribution of arrivals is

$$\pi_{l/} = \sum_{i=1}^N r_i \pi_{i,l}. \quad (27)$$

Clearly, the superposition of t-CPP's (26, 27), because of larger number of parameters and batch size bounds, offers much more flexibility than the superposition of CPP's in order to generate/model certain given non-geometric batch size distributions, by parameter tuning.

6.3 GE with batch size truncation (GE-t)

The service time distribution, GE-t (General Exponential with batch size truncation), is essentially a batch-exponential service, with geometric and bounded batch size distribution. This is quite analogous to the t-CPP. In the case of homogeneous servers the parameters of the GE-t, in phase i , are of the form $(\mu_i, \phi_i, g_i, h_i)$ where μ_i is the batch-service rate, g_i, h_i respectively are lower and upper bounds of the batch size which is geometrically distributed with parameter ϕ_i . However, as far as the service is concerned, the actual batch size served also depends on c, j , where j is the number of jobs in the system. If $j < g_i$, then there will not be any services. If $j \geq h_i$, then the probability that the size of the batch being served is s would be $\frac{(1-\phi_i)\phi_i^{s-1}}{\phi_i^{g_i-1} - \phi_i^{h_i}}$ where $g_i \leq s \leq h_i$. If $g_i < j < h_i$, then the batch size distribution of the served batch can be estimated if the service is clearly defined in such a case. Using these expressions and the steady state probabilities in this case, it is possible to estimate the effective batch size distribution of the served batches. This extension can be incorporated into the main model with a viable solution following same procedures as before, perhaps with minor modifications. This modification can result in, by appropriate parameter tuning, incorporating certain given non-geometric batch size services.

That flexibility can be enhanced even further if we use heterogeneous servers. In such a case, let the parameters of server n ($n = 1, 2, \dots, c$) in phase i be, $(\mu_{i,n}, \phi_{i,n}, g_{i,n}, h_{i,n})$. Extrapolating the earlier analysis to this case, there are no services in phase i if $j < \text{Min}(g_{i,1}, g_{i,2}, \dots, g_{i,c})$. When $j \geq \text{Max}(h_{i,1}, h_{i,2}, \dots, h_{i,c})$ in phase i , then the probability that the batch size is s for the next service can be derived as $\sum_{n=1}^c \frac{\mu_{i,n}(1-\phi_{i,n})\phi_{i,n}^{s-1}}{\mu_{i,n}(\phi_{i,n}^{g_{i,n}-1} - \phi_{i,n}^{h_{i,n}})}$ where $\text{Min}(g_{i,1}, g_{i,2}, \dots, g_{i,c}) \leq s \leq \text{Max}(h_{i,1}, h_{i,2}, \dots, h_{i,c})$, and $\mu_{i.} = \sum_{n=1}^c \mu_{i,n}$. For all other ranges or values of s , the expressions for the probability distribution can be obtained.

6.4 Towards arbitrary batch size distributions

The consideration of arbitrary batch size distributions is most desirable in the performance evaluation of modern telecommunication networks. To achieve this now, is not a very far off thing, owing to the above suggested viable extensions.

And thus, we arrive at a much more general and useful model, that is, *the MM $\sum_{k=1}^K t - CPP_k/GE-t/c/L$ G-queue with heterogeneous servers*. Further work is being carried out on this model.

7 Conclusions

The *HetSigma* queuing model is developed as a generalization of the QBD processes. In this queue, it is possible to accommodate inter-arrival time correlations, service time correlations, batch-size correlations, large and unbounded batch sizes. All these aspects are useful in order to model emerging communication systems. Certain non-trivial transformations are conceived on the balance equations. The transformed balance equations are of the QBD-M type, hence there is a fast solution using the spectral expansion method. The queue is applied for the performance evaluation of MPLS networks with unreliable nodes and can be used to model various problems in telecommunication network as well. It is possible to extend or further generalize these models to develop highly generalized Markovian node models for the emerging Next Generation Networks. That work is underway.

Acknowledgement Ram Chakka thanks the Chancellor of Sri Sathya Sai University, Prasanthi Nilayam, India, for constant encouragement, guidance and inspiration throughout this work.

References

1. Network Research Topologies: The European optical network (EON). <http://www.optical-network.com/topology.php>.
2. The Internet traffic archive. <http://ita.ee.lbl.gov/index.html>.
3. D. Bini and B. Meini. On the solution of a nonlinear matrix equation arising in queueing problems. *SIAM Journal on Matrix Analysis and Applications*, 17(4):906–926, 1996.
4. C. Blondia and O. Casals. Statical Multiplexing of VBR Source: A Matrix-Analytic Approach. *Performance Evaluation*, 16:5–20, 1992.
5. Ram Chakka. *Performance and Reliability Modelling of Computing Systems Using Spectral Expansion*. PhD thesis, University of Newcastle upon Tyne (Newcastle upon Tyne), 1995.
6. Ram Chakka. Spectral Expansion Solution for some Finite Capacity Queues. *Annals of Operations Research*, 79:27–44, 1998.
7. Ram Chakka and Tien Van Do. The MM $\sum_{k=1}^K CPP_k/GE/c/L$ G-Queue and Its Application to the Analysis of the Load Balancing in MPLS Networks. In *27th Annual IEEE Conference on Local Computer Networks (LCN 2002)*, 6-8 November 2002, Tampa, FL, USA, *Proceedings*, pages 735–736, 2002.
8. Ram Chakka and Tien Van Do. The MM $\sum_{k=1}^K CPP_k/GE/c/L$ G-Queue with Heterogeneous Servers: Steady state solution and an application to performance evaluation. *Performance Evaluation*, 64:191–209, March 2007.

9. Ram Chakka, Tien Van Do, and Zsolt Pandi. Generalized Markovian queues and applications in performance analysis in telecommunication networks. In D. D. Kouvatsos, editor, *the First International Working Conference on Performance Modelling and Evaluation of Heterogeneous Networks (HET-NETs 03)*, pages 60/1–10, July 2003.
10. Ram Chakka, Tien Van Do, and Zsolt Pandi. A Generalized Markovian Queue and Its Applications to Performance Analysis in Telecommunications Networks. In D. Kouvatsos, editor, *Performance Modelling and Analysis of Heterogeneous Networks*, pages 371–387. River Publisher, 2009.
11. Ram Chakka, Enver Ever, and Orhan Gemikonakli. Joint-state modeling for open queuing networks with breakdowns, repairs and finite buffers. In *15th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*, pages 260–266. IEEE Computer Society, 2007.
12. Ram Chakka and Peter G. Harrison. Analysis of MMPP/M/c/L queues. In *Proceedings of the Twelfth UK Computer and Telecommunications Performance Engineering Workshop*, pages 117–128, Edinburgh, 1996.
13. Ram Chakka and Peter G. Harrison. The Markov modulated CPP/GE/c/L queue with positive and negative customers. In *Proceedings of the 7th IFIP ATM Workshop*, Antwerp, Belgium, 1999.
14. Ram Chakka and Peter G. Harrison. A Markov modulated multi-server queue with negative customers - the MM CPP/GE/c/L G-queue. *Acta Informatica*, 37:881–919, 2001.
15. Ram Chakka and Peter G Harrison. The MMCPP/GE/c queue. *Queueing Systems: Theory and Applications*, 38:307–326, 2001.
16. Ram Chakka and Isi Mitrani. Multiprocessor systems with general breakdowns and repairs. In *SIGMETRICS*, pages 245–246, 1992.
17. Ram Chakka and Isi Mitrani. Heterogeneous multiprocessor systems with breakdowns: Performance and optimal repair strategies. *Theor. Comput. Sci.*, 125(1):91–109, 1994.
18. M. Chien and Y. Oruc. High performance concentrators and superconcentrators using multiplexing schemes. *IEEE Transactions on Communications*, 42:3045–3050, 1994.
19. D. Kouvatsos. Entropy Maximisation and Queueing Network Models. *Annals of Operations Research*, 48:63–126, 1994.
20. Tien Van Do. Comments on multi-server system with single working vacation. *Applied Mathematical Modelling*, 33(12):4435–4437, 2009.
21. Tien Van Do, Ram Chakka, and Peter G. Harrison. An integrated analytical model for computation and comparison of the throughputs of the UMTS/HSDPA user equipment categories. In *MSWiM '07: Proceedings of the 10th ACM Symposium on Modeling, analysis, and simulation of wireless and mobile systems*, pages 45–51, New York, NY, USA, 2007. ACM.
22. Tien Van Do, Nam H. Do, and Ram Chakka. Performance evaluation of the high speed downlink packet access in communications networks based on high altitude platforms. In Khalid Al-Begain, Armin Heindl, and Miklós Telek, editors, *ASMTA*, volume 5055 of *Lecture Notes in Computer Science*, pages 310–322. Springer, 2008.
23. Tien Van Do, Udo R. Krieger, and R. Chakka. Performance modeling of an apache web server with a dynamic pool of service processes. *Telecommunication Systems*, 39(2):117–129, 2008.
24. Tien Van Do, Denes Papp, Ram Chakka, and Mai X T Truong. A Performance Model of MPLS Multipath Routing with Failures and Repairs of the LSPs. In

- D. Kouvatso, editor, *Performance Modelling and Analysis of Heterogeneous Networks*, pages 27–43. River Publisher, 2009.
25. Steve Drekic and Winfried K. Grassmann. An eigenvalue approach to analyzing a finite source priority queueing model. *Annals OR*, 112(1-4):139–152, 2002.
 26. R. V. Evans. Geometric Distribution in some Two-dimensional Queueing Systems. *Operations Research*, 15:830–846, 1967.
 27. Enver Ever, Orhan Gemikonakli, and Ram Chakka. A mathematical model for performability of beowulf clusters. In *Annual Simulation Symposium*, pages 118–126. IEEE Computer Society, 2006.
 28. Enver Ever, Orhan Gemikonakli, and Ram Chakka. Analytical modelling and simulation of small scale, typical and highly available beowulf clusters with breakdowns and repairs. *Simulation Modelling Practice and Theory*, 17(2):327–347, 2009.
 29. R.J. Fretwell and D.D. Kouvatso. ATM traffic burst lengths are geometrically bounded. In *Proceedings of the 7th IFIP Workshop on Performance Modelling and Evaluation of ATM & IP Networks*, Antwerp, Belgium, 1999. Chapman and Hall.
 30. H. R. Gail, S. L. Hantler, and B. A. Taylor. Spectral analysis of M/G/1 type Markov chains. Technical Report RC17765, IBM Research Division, 1992.
 31. Winfried K. Grassmann. The use of eigenvalues for finding equilibrium probabilities of certain markovian two-dimensional queueing problems. *INFORMS Journal on Computing*, 15(4):412–421, 2003.
 32. Winfried K. Grassmann and Steve Drekic. An analytical solution for a tandem queue with blocking. *Queueing System*, (1-3):221–235, 2000.
 33. B. Haverkort and A. Ost. Steady State Analysis of Infinite Stochastic Petri Nets: A Comparing between the Spectral Expansion and the Matrix Geometric Method. In *Proceedings of the 7th International Workshop on Petri Nets and Performance Models*, pages 335–346, 1997.
 34. ITU-T Recommendation Y.2001. *General overview of NGN*, 2005. Geneve, Switzerland.
 35. D. D. Kouvatso. A maximum entropy analysis of the G/G/1 Queue at Equilibrium. *Journal of Operations Research Society*, 39:183–200, 1998.
 36. U. R. Krieger, V. Naoumov, and D. Wagner. Analysis of a Finite FIFO Buffer in an Advanced Packet-Switched Network. *IEICE Trans. Commun.*, E81-B:937–947, 1998.
 37. G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. ASA-SIAM Series on Statistics and Applied Probability, 1999.
 38. Isi Mitrani. Approximate solutions for heavily loaded markov-modulated queues. *Perform. Eval.*, 62(1-4):117–131, 2005.
 39. Isi Mitrani and Ram Chakka. Spectral expansion solution for a class of Markov models: Application and comparison with the matrix-geometric method. *Performance Evaluation*, 23:241–260, 1995.
 40. V. Naoumov, U. R. Krieger, and D. Warner. Analysis of a Multi-Server Delay-Loss System With a General Markovian Arrival Process. In S. R. Chakravorthy and A. S. Alfa, editors, *Matrix-analytic methods in stochastic models*, volume 183 of *Lecture Notes in Pure and Applied Mathematics*. Marcel Dekker, September 1996.
 41. M. F. Neuts. *Matrix Geometric Solutions in Stochastic Model*. Johns Hopkins University Press, Baltimore, 1981.
 42. Z. Niu and Y. Takahashi. An Extended Queueing Model for SVC-based IP-over-ATM Networks and its Analysis. In *Proceedings of IEEE GLOBECOM*, pages 490–495, 1998.

43. I. Norros. A Storage Model with Self-similar Input. *Queueing Systems and their Applications*, 16:387–396, 1994.
44. V. Paxman and S. Floyd. Wide-area traffic: The failure of Poisson modelling. *IEEE/ACM Transactions on Networking*, 3(3):226–244, 1995.
45. Emilia Rosti, Evgenia Smirni, and Kenneth C. Sevcik. On processor saving scheduling policies for multiprocessor systems. *IEEE Trans. Comp*, 47:47–2, 1998.
46. L. P. Seelen. An Algorithm for Ph/Ph/c queues. *European Journal of Operational Research*, 23:118–127, 1986.
47. Charalambos Skianis and Demetres Kouvatsos. An Information Theoretic Approach for the Performance Evaluation of Multihop Wireless Ad Hoc Networks. In D. D. Kouvatsos, editor, *Proceedings of the Second International Working Conference on Performance Modelling and Evaluation of Heterogeneous Networks (HET-NETs 04)*, pages P81/1–13, Ikley, UK, July 2004.
48. W. J. Stewart. *Introduction to Numerical Solution of Markov Chains*. Princeton University Press, 1994.
49. H. T. Tran and T. V. Do. An iterative method for queueing systems with batch arrivals and batch departures. In *Proceedings of the 8th IFIP Workshop on Performance Modelling and Evaluation of ATM&IP Networks*, pages 80/1–13, Ikley, UK, 2000.
50. H. T. Tran and T. V. Do. A new iterative method for systems with batch arrivals and batch departures. In *Proceedings of CNDS'00*, pages 131–137, San Diego, USA, 2000.
51. H. T. Tran, T. V. Do, and T. Ziegler. Analysis of MPLS-compliant Nodes Deploying Multiple LSPs Routing. In *Proceedings of the First International Working Conference on Performance Modelling and Evaluation of Heterogeneous Networks, HET-NETs*, pages 26/1–26/10, July, 2003.
52. Hung T. Tran and Tien Van Do. Computational Aspects for Steady State Analysis of QBD Processes. *Periodica Polytechnica, Ser. El. Eng*, pages 179–200, 2000.
53. Hung T. Tran and Tien Van Do. Generalised invariant Subspace based Method for Steady State Analysis of QBD-M Processes. *Periodica Polytechnica, Ser. El. Eng.*, 44:159–178, 2000.
54. Hung T. Tran and Tien Van Do. Comparison of some Numerical Methods for QBD-M Processes via Analysis of an ATM Concentrator. In *Proceedings of 20th IEEE International Performance, Computing and Communications Conference, IPCCC 2001*, Pheonix, USA, 2001.
55. V. L. Wallace. *The Solution of Quasi Birth and Death Processes Arising from multiple Access Computer Systems*. PhD thesis, University of Michigan, 1969.
56. Adam Wierman, Takayuki Osogami, Mor Harchol-Balter, and Alan Scheller-Wolf. How many servers are best in a dual-priority M/PH/k system? *Perform. Eval.*, 63(12):1253–1272, 2006.
57. Patrick Wüchner, János Sztrik, and Hermann de Meer. Finite-source M/M/S retrial queue with search for balking and impatient customers from the orbit. *Computer Networks*, 53(8):1264–1273, 2009.
58. Y. Zhao and Winfried K. Grassmann. A numerically stable algorithm for two server queue models. *Queueing Syst.*, 8(1):59–79, 1991.