Perfomance Modelling of Differentiated Services in 3G Mobile Networks

Irfan Awan and Khalid Al-Begain
Mobile Computing and Communications Research Group
Department of Computing, University of Bradford
BD7 1DP, Bradford, UK
{I.Awan, K.Begain}@bradford.ac.uk

Abstract

One of the main features of the third Generation (3G) mobile networks is their capability to provide different classes of services; especially multimedia and real-time services in addition to the traditional telephony and data services. These new services, however, will require higher Quality of Service (QoS) constraints on the network mainly regarding delay, delay variation and packet loss. Additionally, the overall traffic profile in both the air interface and inside the network will be rather different than used to be in today's mobile networks. Therefore, providing QoS for the new services will require more than what a call admission control algorithm can achieve at the border of the network, but also continuous buffer control in both the wireless and the fixed part of the network to ensure that higher priority traffic is treated in proper way. This paper proposes and analytically evaluates a buffer management scheme that is based on multi-level priority and Complete Buffer Sharing (CBS) policy for all buffers at the border and inside the wireless network. The analytical model is based on the G/G/1/N censored queue with single server and R ($R \geq 2$) priority classes under the Head of Line (HOL) service rule for the CBS scheme. The traffic is modelled using the Generalised Exponential distribution. The paper presents an analytical solution based on the approximation using the Maximum Entropy (ME) principle. The numerical results show the capability of the buffer management scheme to provide higher QoS for the higher priority service classes.

Keywords
3G mobile networks, performance evaluation, maximum entropy (ME) principle, queueing network model (QNM), generalised exponential (GE) distribution, head-of-line (HOL) discipline, complete buffer sharing (CBS) rule.

1 Introduction

The Third Generation (3G) mobile networks are under deployment in many regions in the world. In Europe, the Universal Mobile Telecommunications System (UMTS) has been implemented almost by every major mobile network operator offering new mostly multmedia based services. At the moment, the volume of data traffic in mobile networks is still moderate but it is obvious that as more and more customers join the new opportunities offered by 3G, the volume and the nature of the traffic at both the border and inside the mobile network will be very different. Furthermore, many of the new services will require more strict quality of service (QoS) guarantees than the simple data transmission, mainly regarding packet loss, delay and delay variation. Therefore, it is important to implement suitable algorithms to provide prioritization between services and to guarantee preferential treatment to the packets belonging to the higher priority services. This leads to a traffic with Differetiated Services (DS).

Considering the UMTS architecture [1], two levels of QoS assuring algorthms can be identified. First, at the air interface (UTRAN), efficient call admission control algorithms (CACA) can be implemented [2] to limit the volume of the traffic entering the network. However, CACA are not enough to prevent congestions and delay inside the mobile or the fixed parts of the network. Additionally, due to the hierarchical architecture of UMTS, multiple traffic stream from different connections belonging to different classes of services will aggregate at the Gateway Servicing GPRS Node (GSGN) [1] which serves as a gateway towards the public data network (the Internet). In this node, there is a need to implement a buffer management scheme that provides differentiated service.

Finite buffer queues with service and space priorities

 ISBN 3-936150-25-7

are of great importance towards effective congestion control and quality of service (QoS) protection in high speed telecommunication networks.
Many queueing systems with priorities have been explored by Cohen [3] and various applications of the analytical results of priority queues to data communication systems are surveyed by de Moraes [4]. A stable infinite capacity G/G/1 queue with a single server and priority classes under either Preemptive-Resume (PR) or Head-of-the-Line (HOL) scheduling disciplines has been analysed in [5], by applying the method of entropy maximisation (MEM). MEM has also been used in [6] to study a stable single class G/G/1/N censored queue with a single server and First-Come-First-Served (FCFS) scheduling discipline.
A stable G/G/1/N censored queue with a single server, finite capacity and priority classes under complete buffer sharing (CBS) scheme is an important building block in the performance of communication networks. The analysis of such queue is very difficult to tackle using the classical queueing theory. To the authors' knowledge, no exact or approximate closed-form solutions for a stable G/G/1/N censored queue with service priorities have appeared in the literature.
This paper presents further advances of maximum entropy (ME) towards the approximate analysis of a stable G/G/1/N censored queue with a single server, and, $R\,(R \geq 2)$ priority classes under HOL service rule for CBS scheme.
The paper is organised as follows: The ME solution for a stable G/G/1/N censored queue with service priorities is characterised in Section 2. Marginal and aggregate performance distributions are presented in Section 3. Numerical validation results against simulation, involving Generalised Exponential (GE) interarrival and service time distributions, are included in Section 4. Section 5 includes conclusions and remarks for future work.
Remarks

- *The GE Distribution*

The GE distribution is an interevent-time distribution of the form

$$F(t) = P(W \leq t) = 1 - \tau e^{-\sigma t},\ t \geq 0, \quad (1)$$
$$\tau = 2/(C^2+1), \quad (2)$$
$$\sigma = \tau\nu, \quad (3)$$

where W is a mixed-time random variable (rv) of the interevent-time, whilst $(1/\nu, C^2)$ are the mean and squared coefficient of variation (SCV) of rv W. The GE distribution is versatile, possessing pseudo-memoryless properties which makes the solution of many GE-type queueing systems and networks analytically tractable [7].

2 ME Analysis of GE/GE/1/N Priority Queue

Consider a single server finite capacity GE/GE/1/N queue at equilibrium with $R\,(R > 1)$ distinct priority classes of jobs (indexed from 1 to R in descending order of priority) such that

- the total buffer capacity is N for a CBS scheme.

- the interarrival and service times per class are distributed according to a GE distribution under HOL service rule in conjunction with CBS buffer management scheme.

Notation

For each class $i\,(i = 1, 2, \ldots, R)$, let λ_i be the mean arrival rate, C_{ai}^2 be the interarrival time SCV, μ_i be the mean service rate and C_{si}^2 be the service time SCV.
Focusing on a stable GE/GE/1/N queue, let at any given time

n_i $(0 \leq n_i \leq N)$ be the number of class $i\,(i = 1, 2, \ldots, R)$ jobs in the queue (waiting and/or receiving service)

$\mathbf{S} = (n_1, n_2, \ldots, n_R, \omega)$ be a joint queue state, where $\omega\,(1 \leq \omega \leq R)$ denotes the class of the current job in service and $\sum_{i=1}^{R} n_i \leq N$ (n.b., for an idle queue $\mathbf{S} \equiv \mathbf{0}$ with $\omega = 0$)

$\mathbf{Q}$ be the set of all feasible states $\mathbf{S}$

$\mathbf{n} = (n_1, n_2, \ldots, n_R)$ be an aggregate joint queue state (n.b., $\mathbf{0} = (0, \ldots, 0)$)

$\mathbf{\Omega}$ be the set of all feasible states $\mathbf{n}$

Remarks

- The arrival process for each class $i\,(i = 1, 2, \ldots, R)$ is assumed to be censored i.e., a job of class $i\,(i = 1, 2, \ldots, R)$ will be lost if on arrival finds N jobs at the queue.

2.1 Prior Information

For each state $\mathbf{S}$, $\mathbf{S} \in \mathbf{Q}$, and class i $(i = 1, 2, \ldots, R)$ the following auxiliary functions are defined:

$$
\begin{aligned}
n_i(\mathbf{S}) &= \text{the number of class } i \text{ jobs present in state } \mathbf{S}, \\
s_i(\mathbf{S}) &= \begin{cases} 1, & \text{if } \omega = i, \\ 0, & \text{otherwise.} \end{cases} \\
h_i(\mathbf{S}) &= \begin{cases} 1, & \text{if } n_i(\mathbf{S}) > 0, \\ 0, & \text{otherwise.} \end{cases} \\
f_i(\mathbf{S}) &= \begin{cases} 1, & \text{if } \sum_{i=1}^{R} n_i(\mathbf{S}) = N \ \& \ \omega = i, \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}
$$

Suppose all that is known about the state probabilities $\{P(\mathbf{S})\}$ is that they satisfy the

- Normalisation constraint

$$\sum_{\mathbf{S} \in \mathbf{Q}} P(\mathbf{S}) = 1, \tag{4}$$

and that the following marginal mean value constraints per class i exist:

- Server utilisation, U_i $(0 < U_i < 1)$,

$$\sum_{\mathbf{S} \in \mathbf{Q}} s_i(\mathbf{S}) P(\mathbf{S}) = U_i, \ i = 1, 2, \ldots, R; \tag{5}$$

- Busy state probability per class, θ_i $(0 < \theta_i < 1)$,

$$\sum_{\mathbf{S} \in \mathbf{Q}} h_i(\mathbf{S}) P(\mathbf{S}) = \theta_i, \ i = 1, 2, \ldots, R; \tag{6}$$

- Mean queue length, L_i $(U_i \leq L_i < N)$,

$$\sum_{\mathbf{S} \in \mathbf{Q}} n_i(\mathbf{S}) P(\mathbf{S}) = L_i, \ i = 1, 2, \ldots, R; \tag{7}$$

- Full buffer state probability, ϕ_i $(0 < \phi_i < 1)$,

$$\sum_{\mathbf{S} \in \mathbf{Q}} f_i(\mathbf{S}) P(\mathbf{S}) = \phi_i, \ i = 1, 2, \ldots, R; \tag{8}$$

satisfying the flow balance equations, namely

$$\lambda_i (1 - \pi_i) = \mu_i U_i, \ i = 1, 2, \ldots, R; \tag{9}$$

where π_i is the blocking probability that an arriving job of class i finds the queue full.

The choice of mean value constraints (4) - (8) is based on the type of constraints used for the ME analysis of a stable multiple class queues with or without priorities (c.f., [7, 8, 9]). Note that if additional constraints are used, it is no longer feasible to capture a computationally efficient ME solution in closed-form. As a consequence, this will have adverse implications towards the creation of a cost-effective queue-by-queue decomposition algorithm for arbitrary queueing network models (QNMs). Conversely, the removal of one or more constraints from the set (4) - (8) will result into a ME solution of reduced accuracy.

2.2 A Universal Maximum Entropy Solution

A universal form of the state probability distribution $\{P(\mathbf{S}), \mathbf{S} \in \mathbf{Q}\}$ can be characterised by maximising the entropy functional

$$H(P) = -\sum_{s} P(\mathbf{S}) \log P(\mathbf{S}), \tag{10}$$

subject to constraints (4) - (8). By employing Lagrange's method of undetermined multipliers, the ME solution is expressed by

$$P(\mathbf{S}) = \frac{1}{Z} \prod_{i=1}^{R} g_i^{s_i(\mathbf{S})} \xi_i^{h_i(\mathbf{S})} x_i^{n_i(\mathbf{S})} y_i^{f_i(\mathbf{S})}, \ \forall \, \mathbf{S} \in \mathbf{Q}; \tag{11}$$

where Z, the normalising constant, is clearly given by

$$Z = \sum_{\mathbf{S} \in \mathbf{Q}} \left(\prod_{i=1}^{R} g_i^{s_i(\mathbf{S})} \xi_i^{h_i(\mathbf{S})} x_i^{n_i(\mathbf{S})} y_i^{f_i(\mathbf{S})} \right), \tag{12}$$

and $\{g_i, \xi_i, x_i, y_i, \ i = 1, 2, \ldots, R\}$ are the Lagrangian coefficients corresponding to constraints (5) - (8), respectively.

Remarks

Although constraints (5) - (8) are not known priori, nevertheless it is assumed that these constraints exist. This information, therefore, has been incorporated into the ME formalism (4) - (10) in order to characterise the form of the joint state probability (11). An efficient computational implementation of the ME solution (11), however, requires the prior estimation of the Lagrangian coefficients. This can be achieved by making GE-type buffer size invariance assumptions with regard to Lagrangian coefficients $\{g_i, \xi_i, x_i, \ i = 1, 2, \ldots, R\}$ together with asymptotic connections to an infinite capacity GE/GE/1 queue (c.f., [7]).

Aggregating (11) over all feasible states $\mathbf{S} \in \mathbf{Q}$, and after some manipulation, the joint aggregate ME queue length distribution $\{\mathrm{P}(\mathbf{n}), \ \mathbf{n} \in \Omega\}$ is given by:

$$P(\mathbf{0}) = \frac{1}{Z}. \tag{13}$$

$$P(\mathbf{n}) = \frac{1}{Z} \left(\prod_{i=1}^{R} x_i^{n_i} \xi_i^{h_i(\mathbf{n})} \right) \left(\sum_{j=1 \wedge n_j > 0}^{R} g_j y_i^{f_i(\mathbf{n})} \right), \quad \forall \, \mathbf{n} \in \mathbf{\Omega} - \{\mathbf{0}\}; \tag{14}$$

where $h_i(\mathbf{n}) = 1$, if $n_i > 0$, or 0 otherwise and $f_i(\mathbf{n}) = 1$, if $\sum_{j=1}^{R} n_j = N$, or 0, otherwise, for $i = 1, 2, \ldots, R$.

2.3 Recursive Relationships

Taking advantage of the ME product-form solution (13)-(14) and applying the generating function approach [10], recursive expressions for marginal utilisations $\{U_i, i = 1, \ldots, R\}$, aggregate state probabilities $\{P(n),\ n = 0, \ldots, N\}$, marginal state probabilities $\{P_i(n_i), n_i = 0, 1, \ldots, N\}$ and marginal mean queue lengths $\{L_i, i = 1, \ldots, R\}$ can be obtained.

2.3.1 Marginal Utilisations

It can be observed that the marginal utilisations $\{U_i, i = 1, \ldots, R\}$ are clearly defined by $U_i = \sum_{\mathbf{S}\in\mathbf{Q}} s_i(\mathbf{S})P(\mathbf{S})$ and after some manipulation, they take the following universal form, namely

$$U_i = \frac{1}{Z} g_i \xi_i x_i \left(\sum_{v=1}^{N} y_i^{\gamma(v)} C^{(i)}(v-1) \right),\ i = 1, 2, \ldots, R; \tag{15}$$

where Z is the normalising constant and can be derived from the above eqaution (15) as follows:

$$Z = 1 + \sum_{i=1}^{R} g_i \xi_i x_i \left(\sum_{v=1}^{N} y_i^{\gamma(v)} C^{(i)}(v-1) \right). \tag{16}$$

where $\gamma(v) = 1$, if $v = N$ or 0, otherwise. $C^{(i)}(v)$ can be calculated recursively using the following expressions:

$$\begin{aligned} C^{(i)}(v) = (1-\xi_i)x_i C^{(i)}(v-1) + C(v) - x_i^N \\ C(v-N) + x_i^{N+1}\xi_i C^{(i)}(v-N-1), \end{aligned} \tag{17}$$

for $v = 1, 2, \ldots, N-1;\ i = 1, 2, \ldots, R$ with initial conditions

$$C^{(i)}(v) = \begin{cases} 0, & v < 0, \\ 1, & v = 0, \end{cases}$$

where $C(v) = C_R(v)$ and

$$\begin{aligned} C_r(v) = x_r C_r(v-1) + C_{r-1}(v) - (1-\xi_r)x_r \\ C_{r-1}(v-1) - \xi_r x_r^{N+1} C_{r-1}(v-N-1), \end{aligned} \tag{18}$$

$r = 1, 2, \ldots, R$ with initial conditions

$$C_r(v) = \begin{cases} 0, & v < 0, \\ 1, & v = 0, \\ \xi_1 x_1^v, & v > 0. \end{cases}$$

2.3.2 Marginal State Probabilities

Aggregating ME solution $\{P(\mathbf{S}), \mathbf{S} \in \mathbf{Q}\}$ and defining an appropriate z-transform [10], after some manipulation, the following recursive expressions for the marginal probabilities can be obtained:

$$P_i(0) = \frac{1}{Z}\left(1 + \sum_{j=1\wedge j\neq i}^{R} g_j \xi_j x_j \sum_{v=0}^{N-1} C_i^{(j)}(v) y_j^{\delta(v)}\right), \tag{19}$$

$$P_i(n) = \frac{1}{Z}\xi_i x_i^n \left(\sum_{j=1}^{R} g_j E_j \sum_{k=1\wedge k\neq i}^{R} \sum_{v=0}^{N-n-F} C_i^{(j)}(v) y_j^{\delta(v)} \right), \tag{20}$$

where $E_j = \xi_j x_j$ if $j \neq i$ or 1 ow, $F = 1$ if $j \neq i$ or 0 ow.

The coefficients $\{C_i^{(j)}(v), v = 0, 1, \ldots, N-1, (i,j) \in [1, R]\}$ can be determined by the following recursive formulae:

$$C_i^{(j)}(v) = \begin{cases} C^{(j)}(v) - x_i C^{(j)}(v-1) + (1-\xi_i)x_i \\ C_i^{(j)}(v-1) + \xi_i x_i^{N+1} C_i^{(j)}(v-N-1), i \neq j \\ C^{(j)}(v) - x_i C^{(j)}(v-1) + x_i^N C_i^{(j)}(v-N), \\ i = j \end{cases} \tag{21}$$

with initial condition $C_i^{(j)}(v) = 0$ if $v < 0$, or 1, if $v = 0$, where $C^{(j)}(v)$ is determined by (17).

2.4 The Blocking Probability

A universal form for the marginal blocking probabilities $\{\pi_i,\ i = 1, 2, \ldots, R\}$ of a stable multiple class GE/GE/1/N queue can be approximately established, based on GE-type probabilistic arguments, by the following expression:

$$\pi_i \quad = \quad \frac{1}{Z}\left(\sum_{v=0}^{N} \delta_i(v)(1-\sigma_i)^{N-v} P(v) \right), \tag{22}$$

where $\delta_i(v) = \frac{r_i}{r_i(1-\sigma_i)+\sigma_i}$,$\sigma_i = 2/(1 + C_{a\,i}^2)$ and $r_i = 2/(1 + C_{s\,i}^2)$ and $P(v)$ are the aggregate probabilities.

2.5 The Lagrangian Coefficients

It is assumed, as in earlier works (c.f., [7, 8, 9]), that the Lagrangian coefficients $\{g_i, \xi_i, x_i,\ i = 1, \ldots, R\}$ of the ME solution (11) for GE-type queues and networks are largely invariant to the buffer threshold size $N_i\,(i = 1, 2, \ldots, R)$. These coefficients can be, therefore, approximated via closed form asymptotic queueing theoretic expressions based on the ME solution of the corresponding infinite capacity GE/GE/1 queue at equilibrium (c.f., [5]). Using the flow balance condition (9) and the closed-form expressions for the normalising constant, Z, the aggregate probabilities $\{P(n), n = 0, 1, \ldots, N_1\}$ and the blocking probabilities $\{\pi_i, i = 1, \ldots, R\}$, the Lagrangian coefficients $\{y_i, i = 1, 2, \ldots, R\}$ can be recursively determined (c.f., [11]):

3 Numerical Results

This section presents typical numerical experiments in order to illustrate the credibility of the proposed ME solution against simulation. Moreover, it demonstrates the applicability of ME results as simple but cost-effective performance evaluation tools for assessing the effect of external multiple class GPRS traffic at the GE/GE/1/N/HOL queue.

The numerical study focuses on two data packet classes representing typical Internet applications, namely, 12.5 KBytes (class 1, e.g., email) and 62.5 KBytes (class 2, e.g., web browsing) and , respectively. The parameterization also involves mean arrival rates and SCV of inter-arrival and service times. It is assumed that the GPRS partition consists of one frequency providing total capacity of 171.2 Kbps. Without loss of generality, the evaluation study focuses on marginal performance metrics of utilisation, mean response time and mean queue length per class. Numerical tests are carried out to verify the relative accuracy of the ME algorithm against simulation at 95 % confidence intervals based on the Queueing Network Analysis Package (QNAP-2) [12], using the same assumptions and input parameterization as the ones used for the analytic ME solution (c.f., Figs. 1-3). It can be observed that the ME results are very comparable to those obtained via simulation. Moreover, it can be seen that the interarrival-time SCV has an inimical effect, as expected, on the mean response time per class (c.f., Fig. 3). Results shows that high priority calls face less mean response times as compared to the low priority calls.

Moreover, relative comparisons to assess the impact at varying degrees of interarrival time SCVs and buffer size, N, at the GE/GE/1/N/HOL queue upon ME generated mean queue lengths are presented in Figs. 4-5, respectively. It can be seen that the analytically established mean queue lengths deteriorate rapidly with increasing external interarrival-time SCVs (or, equivalently, average batch sizes) beyond a specific critical value of the buffer size which corresponds to the same mean queue length for two different SCV values. It is interesting to note, however, that for smaller buffer sizes in relation to the critical buffer size and increasing mean batch sizes, the mean queue length steadily improves with increasing values of the corresponding SCVs. This 'buffer size anomaly' can be attributed to the fact that, for a given arrival rate, the mean batch size of arriving bulks increases whilst the interarrival time between batches increases as the interarrival time SCV increases, resulting in a greater proportion of arrivals being blocked (lost) and, thus, a lower mean effective arrival rate; this influence has much greater impact on smaller buffer sizes.

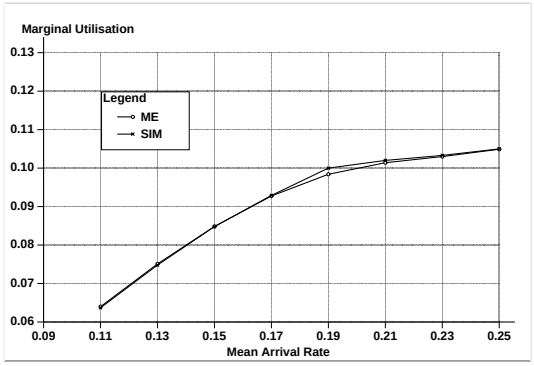

Figure 1: Marginal Utilisations for Class 1 Calls

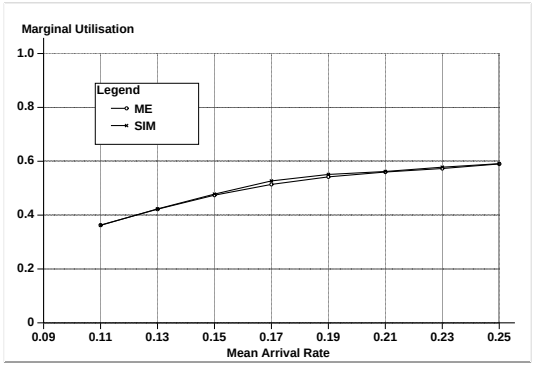

Figure 2: Marginal Utilisations for Class 2 Calls

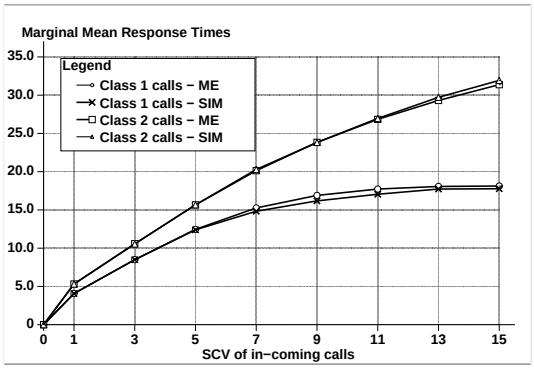

Figure 3: Effect of varying degrees of SCV on Mean Response Time

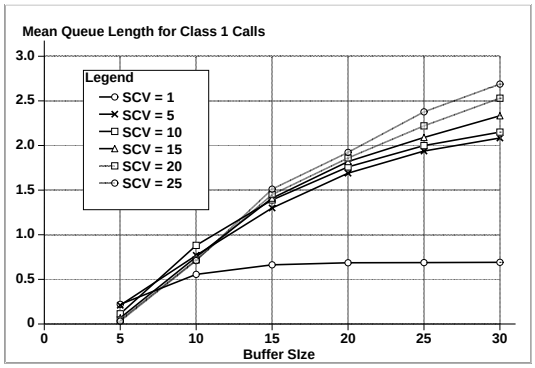

Figure 4: Effect of varying degrees of SCV on MQLs of Class 1 at different buffer sizes

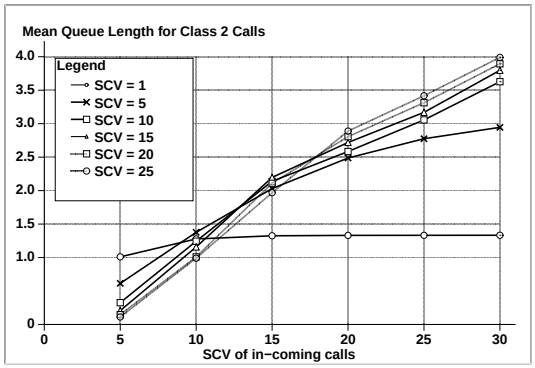

Figure 5: Effect of varying degrees of SCV on MQLs of Class 2 at different buffer sizes

4 Conclusions

This paper presents an analytical model to evaluate performance of the wireless network where different applications are given preferential treatment in order to provide quality of service. In this context, a G/G/1/N censored queue with single server and R ($R \geq 2$) priority classes under the HOL service rule for the CBS scheme has been analysed using ME principle. Closed form expressions for marginal utilsations, state probabilities and blocking probabilities are presesnted. Typical numerical experiments show the capability of the buffer management scheme to provide higher QoS for the higher priority service classes.

References

[1] Bernhard H. Walke *Mobile Radio Networks, Networking and Protocols,* . WILEY.

[2] K. Al-Begain and A. Zreikat ,"Interference Based CAC for Up-link Traffic in UMT S Networks", *in Proceedings of World Wireless Congress, 2002*, 28-31 June 2002, pp. 298 - 303, San Francisco, USA.

[3] Cohen, J. W. The Single Server Queue. Revised edition, *North-Holland Publishing Company*, Amsterdam (First edition: 1969).

[4] de Moraes, L. F. M. Priority Scheduling in Multiaccess Communication, Stochastic Analysis of Computer and Communication Systems, H. Takagi (ed.), *Elsevier Science Publishers* (North-Holland), Amsterdam, pp. 699-732, 1990.

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

[6] D.D. Kouvatsos, Maximum Entropy and the G/G/1/N Queue, *Acta Informatica*, Vol. 23, (1986), pp. 545-565.

[7] D.D.Kouvatsos, Entropy Maximisation and Queueing Network Models, *Annals of Operation Research* 48, (1994), pp. 63-126.

[8] D.D.Kouvatsos and I.U.Awan, MEM for Arbitrary Closed Queueing Networks with RS-Blocking and Multiple Job Classes, *Annals of Operations Research 79*, (1998), pp. 231-269.

[9] D.D.Kouvatsos and Xenios N.P., MEM for Arbitrary Queueing Networks with Multiple General servers and repetitive-Service Blocking, *Performance Evaluation*, Vol. 10, (1989), pp. 169-195.

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

[11] I.U. Awan and D.D.Kouvatos, Maximum Entropy Analysis of Arbitrary Queueing Network Models with Service priorities, *Research Report* RS-07-01, Performance Modelling and Engineering Research Group, Department of Computing, University of Bradford, August, (2001).

[12] M. Veran and D. Potier, QNAP-2: A Portable Environment for Queueing Network Modelling Techniques and Tools for Performance Analysis, D.Potier (ed.), North Holland, pp. 25-63, 1985.