Simulation Studies on the Performance Characteristics of Multi Priority Dual Queue (MPDQ) with Finite Waiting Room and Non-Preemptive Scheduling

ANTHONY BEDFORD
PANLOP ZEEPHONGSEKUL
Department of Statistics and Operations Research
RMIT University
Plenty Road, Bundoora East, Victoria
AUSTRALIA

Abstract:- Due to the complexity of solving single and dual queues with multi-class non-pre-emptive prioritised customers with finite waiting room, a simulation approach is used to investigate their viability. Investigation of loss probabilities and utilisation levels for various arrival rates are analysed. Models with up to 5 classes are simulated and comparisons are made across queueing disciplines.

Key-Words:- dual queue, simulation, finite queue, non-preemptive priority under different service regimes, loss, utilisation.

1 Introduction

Communication systems are under continuing strain in the face of increasing demand for more information and faster service. There are many various schemes in place for the transfer of information in communication networks. These schemes are used for a wide variety of communications network demands. With telecommunications networks becoming larger each day, efficient and effective queueing methods are needed to analyse new schemes aimed at faster communication through networks.

There have been a wide variety of methods studied that reduce congestion of communication systems. Many differentiate customers through marking and dropping processes [1],[2]. Others use time-marking and derivatives of this to allocate a degree of fairness in service, such as self-clocked fair queueing (SCFQ) and credit based fair queueing (CBFQ) [3],[4]. In recent times, the idea of prioritising traffic has received interest. In many cases these schemes have proven effective in providing quality of service for users of the networks. When we mix a priority scheme with a series or network of queues, the solution to such systems becomes complex. As far as loss of customers is concerned, it has been shown to decrease logarithmically as buffer size increases [6]. What we analyse here is a fixed buffer si e with various queueing disciplines for customers of different classes.

The need for differentiating customers has arisen recently due to concerns with loss for real-time mediums. In communications systems, this may be a mobile phone call, a video transmission, or a streaming audio application.

Whilst this problem was solved for service centres with infinite waiting room, only recently has it been investigated for finite waiting room in the case of a single queue [7]. The analysis has involved looking at two classes of customers, one high and one low class, where the high class has precedence over the lower class whilst waiting in the queue. This scenario sets up interesting results for communications service centres.

We wish to take this analysis further and beyond the capabilities of obtaining an analytical solution. As the problem has been solved theoretically for a single queue (with limited useability due to priority and queue type), this research is targeted at combining this key result to a new scheme called dual queueing [8]. This scheme has two queues of finite space where a customer, upon arrival, if finding the first queue full, waits in the second queue if there is room. When a space becomes vacant in the first queue, a customer at the front of the second queue enters the back of the first, which is the queue that has the service centre at the front of it. Previous work on the dual queue included simulations based on actual MPEG files [8]. The analysis showed that dual queue improved performance characteristics over the FIFO discipline. We aim to combine the dual queue idea

with that of a priority scheme, with the anticipation that prioritised traffic coupled with the dual queue will enhance quality of service for customers. Furthermore, there are a variety of queueing disciplines that will be investigated, such as First In First Out (FIFO), Last In First Out (LIFO), High Class First (HCF), and Low Class First (LCF). These queueing disciplines will be investigated via computer simulations.

2 Complexities of solving analytically

The problem of solving the dual queueing problem analytically is mathematically complex. Consider the lowest level of a single and dual queue with prioritised traffic of only two classes of customers. If we are to investigate the state generation of a single queue with finite waiting space with two classes of customers $M_1 + M_2 / M /1/ N$, the dimensions of the irreducible infinitesimal generator of the system is given by

$$A_{c_1} \in \Re^{\left(c_1^2+c_1+1\right)\times\left(c_1^2+c_1+1\right)} \quad (1)$$

where c_1 is the queue length. This matrix forms part of the linear system generating all transitional states of the queueing model that is given by

$$\bar{p}^T A_{c_1} = 0 \quad (2)$$

where $\bar{p}^T$ is the vector of the steady-state distribution of the continuous-time Markov chain containing the unique normalised non-negative solution once solved. The solution of such a system requires the use of a recursive algorithm that is an exhaustive process for values of $c_1 > 3$. Furthermore, for a dual queueing system with the same number of classes $M_1 + M_2 / M /1/ N_{c_1} + N_{c_2}$ with waiting space c_1 for the primary queue and c_2 for the secondary queue, the dimensions of the irreducible generator matrix of the system is given by

$$A_{c_1,c_2} \in \Re^{\left(\frac{1}{2}(c_1+1)\left(c_2^2+3c_2+c_1+2\right)\right)\times\left(\frac{1}{2}(c_1+1)\left(c_2^2+3c_2+c_1+2\right)\right)} \quad (3)$$

As for a single queue, the stationary state distribution is obtained by solving

$$\bar{p}^T A_{c_1,c_2} = 0 \quad (4)$$

The dual queue requires more exhaustive demands on computational resources than for a single queue. This is due to the rapidly increasing size of A_{c_1,c_2} as c_1 and c_2 increase. Through our investigation, it becomes quickly apparent that it is impractical to solve systems with a total queueing capacity beyond five. For the single queueing model with c_1 = 5, the size of A_5 is 31x 31, and for a dual queue with c_1 = 2 and c_2 =3, it is 33 x 33. (Double the size of these respective queues and we have generator matrices of size 111 x 111 and 150 x 150). The Matrix-Analytic method is used to solve such systems. For the single queue, this method is practical for smaller systems, however the rapidly increasing size of the system tends to make large queue sizes difficult to solve.

To gain some insight into the behaviour of single and dual queues with various queueing disciplines and priorities, we have undertaken computer simulations. Furthermore, we extend the application of these schemes to situations with more than two priorities. This has not been solved either theoretically or through simulation. It is seen to be far too complex at this stage to be solved theoretically for more than two classes of customers. Furthermore, using simulation, comparisons can be undertaken for the single and dual queue schemes for more than two priorities. The simulation can allow for differing queueing disciplines other than the solved priority disciplines discussed so far, which assume HCF.

The paper is organised as follows. In Section 3 we define the model for the dual queue, the queueing disciplines and the simulation design. In Section 4 we discuss the results in terms of loss and utilisation. We summarise the findings and suggest further research in Section 5. We begin with a simple two-class scheme. This is extended up to a five-class scheme.

3 Model

The queueing set-up is illustrated below

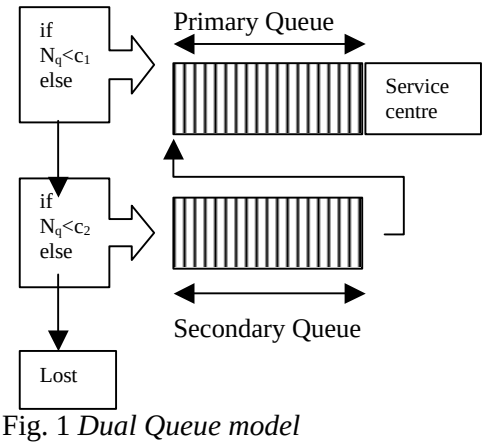

Fig. 1 *Dual Queue model*

Figure 1 illustrates the dual queue. N_q = number in queue and c_i = capacity of queue *i*. If arriving customers meet a full primary queue, then it waits in the secondary queue. If the secondary queue is also full then the arriving data is lost. The dual queue has the ability to have independent queueing disciplines if

need be.If we are to consider only the primary queue in Figure 1, this is the single queue with losses. Upon arrival to the system, if the service centre is busy then the arriving customer waits in the primary queue given there is sufficient space. If there is no space in the queue (or buffer) then the customer is lost.

3.1 Queueing disciplines

As there are multiple classes of customers simulated here, we have assigned arrivals based on decreasing demand as class increases. The first class packet is of utmost importance when considering time and loss constraints. If the scheme is to be adopted without cost but rather importance in terms of applications, then this customer may be data that is for a live event, or a streaming video/audio type. It typically is chosen to represent customers that if not given fast service, will have an impact on applications susceptible to loss.

We use the exponential distribution for both the arrival rate and the service times of customers. For the arrival of the data, we are assuming that the arrival process is independent for the classes. The customers here are considered of uniform length, that is, there are no differing sizes of customers with respect to their occupancy within the queue (ie-uniform batch sizes of 1). By considering two to five classes, we can compare how the introduction of more classes changes the behaviour of the queue.

Now a brief summary of the four queueing disciplines analysed here for use in both the single and dual queue simulations. First we consider the first in first out (FIFO) queueing discipline. This is the simplest, and is also known as first come first serve. Because of its easy to implement sequential handling of customers in waiting, it is analysed first and is common to many communications networks. Next is the last in first out (LIFO) (or last come first serve) queueing discipline. This discipline is not unusual to communications networks. Lowest class first (LCF) is one of two priority disciplines at the heart of the model. In this discipline, lowest class customers jump to the head of the queue behind any already present customers of the same class. Highest Class First (HCF) is the most important model here. In this discipline, highest-class customers jump to the head of the queue behind any already present customers of the higher or the same class.

The dual queue model is combined here with prioritised traffic so as to investigate a splitting of what may be viewed as unfairness. By having a dual queue in place, the strong bias towards HCF and LCF models to their respective prioritised customers allows for some traffic of a lower class to move through the queues, unlike a FIFO or LIFO single queue model.

3.2 Simulation set up

Arena was used for the simulations here[11]. Due to need for a diverse range of analysis, Arena was chosen because of its flexibility. Figure 1 contains a sample of one of the simulation screens in Arena. A total of 10 simulation runs with simulation time of 15,000 units per run for each queueing model was evaluated. The multiple runs were used to initialise the models with different random number seeds. All arrival, service and statistical values are given in the same time scale so comparisons between queueing models and model types could be made. This is uniform only for each of the classes. All maximum values refer to the maximum of all simulation runs, not just a single run, for each respective model.

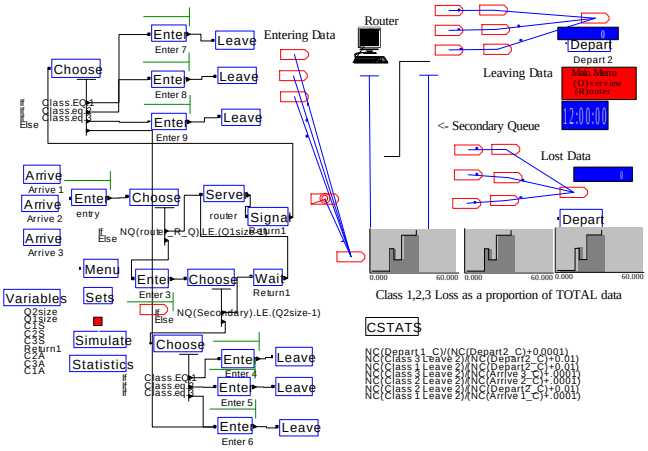

Figure 1 Arena Model - Overview

For all classes we have allocated the first class customer as the rarest arrival and longest in service. The rationale behind this is that the ‘rarer’ traffic will in many cases be the more demanding on system resources and can be seen as either the most or least valuable, depending upon the type of queueing discipline. Examples of high class high demand traffic could include videoconference links, streaming audio or streaming video. As more classes are introduced, the performance between them may not be as efficient as a few classes. For the simulations, the buffer size/s (waiting space) for arriving customers was fixed. For a single queue, the size was 10, and for a dual queue the size was 5 for each queue. Table 1 contains the arrival and service rates for the four models used here. Model I contains 2 classes, Model II, 3 classes and so on.

Model	$\lambda_1;\mu_1$	$\lambda_2;\mu_2$	$\lambda_3;\mu_3$	$\lambda_4;\mu_4$	$\lambda_5;\mu_5$
I	5 ; 1	2 ; 0.2			
II	15 ; 2.5	10 ; 1.5	5 ; 0.5		
III	60 ; 5	30 ; 2.5	15 ; 1.5	5 ; 0.5	
IV	120 ;10	60 ; 5	30 ; 2.5	15 ; 1.5	5 ; 0.5

Table 1 Arrival and service rates for the models

When considering the differing queueing disciplines for each system, the FIFO is considered the baseline. As it requires no re-organisation, it is the simplest. It would be a poor choice to employ a prioritised method of a more complicated nature if the enhancements were marginal over a FIFO/LIFO discipline.

4 Performance Characteristics

Many criteria could be considered depending upon the nature of the communications system. Consequently, we present all statistics without any policies such as time-out thresholds, loss thresholds and throughput constraints. We have chosen a simple loss of approximately 5% for the models. A pilot simulation was undertaken to determine the loss levels for various arrival/service rate combinations. As seen in Table 1, the idea was to double the arrival and service rates as the number of classes increased. In this way, by the time one reaches Model IV, the high-class customer was rare and demanding on system resources.

Next, we define the performance characteristics used in the tables: L_s = Average loss for all classes; L_s^i = Average loss for class *i*; $M\ell$ = Maximum of all loss; $M\ell^i$ = Maximum loss for class *i; U* = Utilisation. For each of the models, we will investigate the probabilities on the basis of single versus dual queue, then class wise, and finally summarise overall.

4.1 Model I - 2 Classes

From Table 2, we have the performance characteristics for the 2-class model. First, we compare the single and dual queue designs. When looking at the average loss by class probabilities and overall average loss, the single queue design is superior for all regimes with the exception of LIFO. The maximum loss by class and overall maximum loss probabilities again show that the single queue design is superior for all regimes. When considering utilisation, the probabilities are close for both single and dual queue designs. If a low utilisation is desirable, the HCF model is the best, being marginally superior to the LIFO regime. For this Model, the single queue outperforms the dual queue.

When considering the performance of traffic class wise, the results are varied. On average loss by class, Class 1 performs best under a LCF regime, for both single and dual queues. The loss levels are vastly superior under LCF than any other regime, with an average loss of only 0.026, half that of the next best loss level. The same conclusion can be drawn when considering maximum loss. For Class 2, LCF is also the best choice.

Queue regime	$L_s^{1,2}$	$L_s^{1,2}$	$M\ell^{1,2}$	$M\ell^{1,2}$
	single	**dual**	**single**	**dual**
HCF	0.052 0.053	0.062 0.064	0.0762 0.0603	0.098 0.137
FIFO	0.053 0.057	0.053 0.057	0.0722 0.0785	0.072 0.079
LCF	0.026 0.031	0.039 0.042	0.0384 0.0397	0.058 0.089
LIFO	0.062 0.065	0.048 0.047	0.0987 0.111	0.091 0.130
	$\frac{L_s}{M\ell}$	$\frac{L_s}{M\ell}$	U	U
	single	**dual**	**single**	**dual**
HCF	0.056 0.0637	0.068 0.15	0.937	0.938
FIFO	0.0589 0.0831	0.059 0.083	0.941	0.941
LCF	0.0305 0.0409	0.043 0.086	0.954	0.952
LIFO	0.069 0.122	0.050 0.174	0.938	0.944

Table 2 Loss probabilities, Overall loss, and Utilisation Model I

For this Model, it is clear that the dual queue offers no advantages over the single queue designs for the arrival and service rates tested. The LCF design offers the lowest class wise loss levels and overall loss levels under a single queue design. The LCF single queue design appears the best choice for this Model. With only 2 classes, preliminary analysis suggests the dual queue offers no advantages to traffic.

4.2 Model II - 3 Classes

Now with a third class, the dual queueing designs exhibit major improvements over the single queueing designs. Table 3 now depicts a reversal of the losses seen in the 2-class models. The loss probabilities improve dramatically in favour of the dual queueing models. The levels of loss are at their lowest for the priority schemes in comparison to the single models. The dual queueing designs are superior over the single scheme for almost all performance characteristics. Of the dual queueing designs, the best queueing regime in terms of average loss by class probabilities and overall average loss is LCF. For maximum loss by class and overall maximum loss probabilities, the dual queue is clearly superior over the single queue designs for all queueing regimes. Noticeably, the priority regimes show the largest reduction in loss from single to dual queue designs.

On a class basis, again the LCF is superior, but only for the dual queue design. If we were confined to using a single queue design, the LCF is one of the poorest designs in terms of class wise loss. When considering maximum loss by class, the HCF is best for Class 1, LIFO for Class 2, and LCF for Class 3. The FIFO

regime is the best for single queue designs, yet poorest in the dual queue designs. The overall loss levels are poor for the single queue schemes.

Queue regime	$L_s^{1,2,3}$	$L_s^{1,2,3}$	$M\ell^{1,2,3}$	$M\ell^{1,2,3}$
	single	**dual**	**single**	**dual**
HCF	0.132 0.126 0.145	0.0503 0.0618 0.0618	0.181 0.175 0.207	0.064 0.088 0.081
FIFO	0.116 0.135 0.122	0.0632 0.0653 0.0529	0.145 0.16 0.144	0.14 0.214 0.134
LCF	0.148 0.142 0.156	0.0388 0.041 0.0473	0.174 0.16 0.21	0.097 0.08 0.069
LIFO	0.165 0.115 0.122	0.0516 0.0532 0.059	0.214 0.138 0.156	0.094 0.076 0.070
Queue regime	L_s $M\ell$	L_s $M\ell$	U	U
	single	**dual**	**single**	**dual**
HCF	0.162 0.227	0.064 0.083	0.984	0.927
FIFO	0.144 0.174	0.063 0.187	0.988	0.935
LCF	0.18 0.227	0.047 0.082	0.993	0.940
LIFO	0.15 0.185	0.06 0.077	0.988	0.937

Table 3 Loss probabilities, Overall loss / Utilisation Model II

Most of the dual queue designs, with the exception of maximum loss under FIFO, show considerable improvement over single queue designs. Due to the decrease in loss levels, utilisation is also lower in the dual schemes. This may be an important factor when considering the systems ability to run near capacity. The average loss of class points to the LCF model as the best. However it is the HCF which has the lowest maximal loss. It may be difficult to decide which of LCF and HCF is best for 3 classes. In overall terms, it seems that the HCF is best, with superior maximum loss rates and utilisation and comparable overall maximum loss.

4.3 Model III - 4 Classes

The introduction of another class strengthened the case for the HCF and LCF models. The loss probabilities are best for the priority disciplines when considering maximum loss. The LCF is particularly good. In both the HCF and LCF models, the middle classes suffer the highest lost. The nature of arrivals for the other schemes does not see such results.
When considering overall loss of the system, the priority schemes are the best. Notice the high levels of loss for the single class schemes, which are far above acceptable levels especially considering the maximum loss levels.

Queue regime	$L_s^{1,2,3,4}$	$L_s^{1,2,3,4}$	$M\ell^{1,2,3}$	$M\ell^{1,2,3,4}$
	single	**dual**	**single**	**dual**
HCF	0.092 0.166 0.13 0.144	0.0203 0.0431 0.0361 0.0283	0.125 0.333 0.227 0.324	0.0392 0.0632 0.0567 0.0625
FIFO	0.111 0.133 0.137 0.137	0.0456 0.0814 0.047 0.0601	0.6 0.5 0.162 0.183	0.0769 0.152 0.121 0.127
LCF	0.108 0.087 0.115 0.091	0.0114 0.0124 0.0194 0.0187	0.2 0.117 0.15 0.122	0.0307 0.0271 0.04 0.0302
LIFO	0.085 0.078 0.064 0.083	0.0669 0.0428 0.0239 0.0342	0.125 0.106 0.0953 0.107	0.148 0.154 0.0462 0.0879
Queue regime	L_s $M\ell$	L_s $M\ell$	U	U
	single	**dual**	**single**	**dual**
HCF	0.169 0.75	0.033 0.063	0.97367	0.90678
FIFO	0.16 0.437	0.064 0.143	0.98135	0.91044
LCF	0.109 0.142	0.018 0.031	0.98834	0.92806
LIFO	0.0866 0.113	0.037 0.1	0.97091	0.91443

Table 4 Loss probabilities Model III

4.4 Model IV - 5 Classes

The 5-class model was included to further investigate a trend appearing for loss. This is that the middle classes are suffering high levels of loss with respect to the other classes. In the 5-class models, this trend continued on from the 4-class model. The dual queueing scheme this time improved only for classes 3, 4, and 5 over the single queue. It may seem that this scheme may have the system too full of middle class customers to allow high-class customers the chance of arrival. The dual scheme may disadvantage the high class in its two-time wait. It is becoming a rare event and the single queue benefits high class by letting it jump to the front immediately.

All loss probabilities are small, with the LCF operating at the lowest loss levels. An interesting result is the difference between single and dual queue for the HCF. The dual queue shows the improvement for the 2nd, 3rd and 4th classes in the dual model. As discussed, it would seem the increase in classes sees the decrease in quality for the first class of customer. The LIFO model is clearly the best. This model is beneficial to rare arrivals as a rare arrival will usually find waiting customers in front of them. The LIFO gives the last customer the advantage of jumping the queue, something that benefits the rare arrivals.
The dual queue model slows down the severity of LIFO model, with it showing no loss for all but 4th class customers.

Queue regime	$L_s^{1,2,3,4,5}$	$L_s^{1,2,3,4,5}$	$M\ell^{1,2,3,4,5}$	$M\ell^{1,2,3,4,5}$
	single	**dual**	**single**	**dual**
HCF	0 0.00491 0.00308 0.00151 0.00197	0.00546 0 0 0.00037 0.00155	0 0.0278 0.0175 0.00309 0.00518	0.0159 0 0 0.00095 0.00417
FIFO	0.00128 0.0071 0.00995 0.00852 0.00449	0.00128 0.0071 0.00995 0.00852 0.00449	0.00441 0.0149 0.0588 0.0128 0.0112	0.00441 0.0149 0.0588 0.0128 0.0112
LCF	0 0 0.00020 0.00020 0	0.00526 0.00101 0 0.00125 0.00316	0 0 0.0011 0.00103 0	0.0147 0.00303 0 0.00235 0.00926
LIFO	0 0 0.00035 0.0032 0.0137	0 0 0 0.00098 0	0 0 0.00128 0.0147 0.0426	0 0 0 0.00073 0
Queueing Discipline	L_s $M\ell$	L_s $M\ell$	U	U
	single	**dual**	**single**	**dual**
HCF	0.00225 0.00806	0.00070 0.00134	0.6027	0.60065
FIFO	0.00772 0.0182	0.00772 0.0182	0.6064	0.6064
LCF	0.00017 0.00075	0.00143 0.00305	0.60082	0.60724
LIFO	0.00322 0.0117	0.00051 0.00124	0.60257	0.59982

Table 5 Loss probabilities by Model IV

The LCF model has the worst overall loss when changing from single to dual queue. The others equal or better improvement in loss levels. The LIFO performs the best.

5 Concluding Remarks

We have presented a new combination of schemes called the MPDQ and explored some of its probabilistic characteristics under various queueing disciplines. The need for using a simulation approach was discussed, and evidence given of the difficulty in obtaining an exact result analytically. As a scheme combining priorities with a dual queue, the HCF discipline for 3 and 4 classes performed well, whereas the LCF and LIFO showed volatility in certain situations. For service providers, the introduction of the MPDQ under a HCF or FIFO discipline is worth further investigation, with the final decision governed by quality of service constraints. Further follow up issues worth investigating include

- Analysing waiting times and the influence on the models
- The arrival of batches of non-uniform length
- Modifying the arrival distribution to allow for 'bursty' traffic
- A network analysis involving multiple simultaneous simulation, extending the model described in Figure 1.
- Investigating the effect of changing buffer size

References:

[1] Miguel A. Labrador and Sujata Banerjee, Enhancing Application Throughput by Selective Packet Dropping, *Proceedings of IEEE International Conference on Communications (ICC)*, Vancouver, Canada, 1999, pp.1217-1222.

[2] Wu-chang Feng, Dilip D. Kandlur, Debanjan Saha, Kang G. Shin, Adaptive Packet marking for Maintaining End-to-End Throughput in a Differentiated Services Internet, *IEEE – ACM Transactions on Networking*, Vol. 7, No. 5, 1999, pp. 685-697.

[3] S.J. Golestani, A self-clocked fair queueing scheme for broadband applications, *Proceedings of the IEEE Infocom*, 1994, pp 636-646.

[4] K.T. Chan, B. Bensaou and D.H.K. Tsang, Credit-Based Fair Queueing (CBFQ), *IEEE Electronics Letters*, Vol.33, No.7, March 1997, pp. 584-585.

[5] Jeffery Fritz, Caught Up On Video, *Data Communications*, Oct 21, 1999, pp 51-55.

[6] Cynthia Bagwell Tipper, ATM Cell Delay and Loss for Best-Effort TCP in the Presence of Isochronous Traffic, *IEEE Journal on Selected Areas in Communications*, Vol. 13, No. 8, Oct 1995, pp. 1457-1463.

[7] D Wagner, U Krieger, Analysis of a finite buffer with non-preemptive priority scheduling, *Communications in Statistics-Stochastic Models*, Vol 15, No. 2, 1999, pp. 345-365.

[8]David A. Hayes, Michael Rumsewicz, Lachlan L. H. Andrew, Quality of Service Driven Packet Scheduling Disciplines for Real-Time Applications: Looking Beyond Fairness, *Infocom 1999, IEEE*, pp. 405-412.

[9] Andrew Odlyzko, Paris Metro Pricing for the Internet, *Proceedings of ACM Conference on Electronic Commerce*, 1999, pp. 140-147.

[10] Andrew Odlyzko, "The Current State and Likely Evolution of the Internet", *Proceedings Globecom '99*, IEEE, pp 1869-1875.

[11] W. David Kelton, Randall P. Sadowski, Deborah A. Sadowski , *Simulation with Arena*, McGraw-Hill, 1998

[12] Masaklatsu Ogawa, Takashi Sueoka, Takeshi Hattori, Priority Based Wireless Packet Communication with Admission and Throughput Control, *Proceedings of the 51st IEEE Conference on Vehicular Technology*, 2000, pp. 370-374.