Simulation Studies of Waiting Time Approximation for the 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:- In this paper we extend our work on dual queues with multi-class non-pre-emptive prioritised customers with finite waiting room to investigate waiting times. We consider different rates of arrival for the single and dual queueing systems, and examine the way waiting times are effected. Distribution fits are given for high class customers, and the waiting time for high class for various low class arrival rates are investigated.

Key-Words:- dual queue, simulation, finite queue, waiting times, distribution fits.

1 Introduction

In our work on the performance characteristics of the MPDQ, we established the usefulness of the dual queue to providers of communications services [1]. The demands required for the services may vary from a telephone call through a mobile phone network, to the need for a document to be printed in an office network or a web page across the Internet. We take our prior work on the MPDQ further and investigate the waiting time behaviour of customers under various single and dual queue models with different queueing regimes. This will provide further insight to these providers as to the viability and quality of service expectations a multi priority queue can deliver. We showed in terms of loss that a priority scheme was beneficial for the dual queue with more than two classes. Indeed, there have been other designs aimed and proven to reduce congestion. Previous work on the dual queue included simulations based on actual MPEG files [2]. More recently, the scheme has been adapted to wireless local area networks [3].The analysis showed that dual queue improved performance characteristics over the FIFO discipline. Some analysis on waiting times for specific traffic intensity was undertaken. Here we investigate average waiting times, and also provide distribution fits of the waiting times for specific customers.

As in our previous work on the MPDQ, we analyse a fixed buffer size with various queueing disciplines for customers of different classes. In the case of a single queue with finite waiting room and customers of two classes, complicated solutions for waiting times were obtained by using matrix-analytic methods [4],[5]. The complicated nature of obtaining an explicit solution for multi-priority queues illustrates the need for simulation.

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. 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.

2 Model

The queueing system is illustrated below

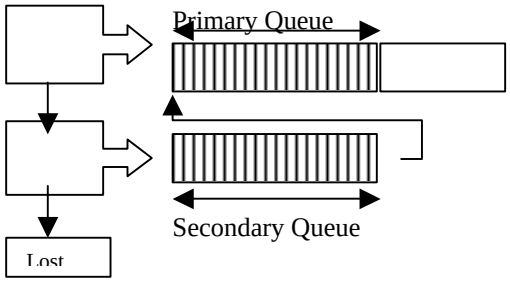


Fig. 1 *Dual Queue model*

Figure 1 illustrates the dual queue, where if an 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. 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. The simulation models here contain two, three, four and five class customers.

2.1 Queueing disciplines

Again, 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, with 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.

The four queueing disciplines analysed here for use in both the single and dual queue simulations are First In First Out (FIFO), Last In First Out (LIFO), Lowest Class First (LCF) and Highest Class First (HCF). These regimes govern how customers are moved through the queues. HCF and LCF re-organise the customers after each arrival. 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 [1].

2.2 Simulation set up

As in our prior work of the MPDQ, Arena was used for the simulations here[6]. A total of 10 simulation runs with simulation time of 15,000 units per run for each queueing model was evaluated. 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. Arena has the ability to store a wide variety of performance characteristics. For the simulations, the final analyses were simplified by using the Batch/Truncate option. This feature 'lumps' replications together and from this we obtain an overlaid picture of a typical system at any point through the simulation period. All maximum values refer to the maximum of all simulation runs, not just a single run, for each respective model.

To be consistent for comparison with our prior work, the buffer size/s (waiting space) for arriving customers was again fixed at size 10 for the single, 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.

Arrival and Service Rates

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

3 Waiting time statistics

We now provide the statistics for each of the four models. The waiting times give us further clues as to the best model for the number of classes. In this section we present both single and dual queue statistics under the same arrival/service rates. The statistics as given in the following subsections tables are as follows: W_q^i = Average waiting time in queue *i* for any customer, W_s^i =Average waiting time in the system for class *i* customers, M_s^i =The maximum waiting time in the system for class *i* customers.

3.1 Model I - 2 Classes

When considering the average waiting time for any customer, as seen in Table 2, the single FIFO is clearly the best. This statistic offers a rough guide to the efficiency of the queue. By comparing the single and dual queue waits for any customer, for all disciplines the sum of waiting times in the dual queue exceeds the single queue times.

The average waiting time in the system by class statistics are close for the single and dual queues for each regime. If we wish to ensure that the highest class or lowest class spends the shortest time possible in the system/queue, the priority disciplines show distinct advantages over the non-priority schemes.

The HCF queues is the best for Class 1 customers, yet the dual queue method offers no advantage in terms of waiting time. Class 2 traffic for HCF is the worst off of all schemes. The LCF scheme is very poor toward Class 1 customers in waiting times. LIFO offers the poorest guarantees for maximum waits. In general, the dual queue shows little improvement to justify a case for the 2 class models.

Queue regime	W_q^1	$W_q^{1,2}$	$W_s^{1,2}$	$W_s^{1,2}$	$M_s^{1,2}$	$M_s^{1,2}$
	single	**dual**	**single**	**dual**	**single**	**dual**
HCF	6.81	5.06 4.02	3.28 10.2	3.83 10.2	13.8 34.6	12.5 37.3
FIFO	5.1	5.15 3.71	8.99 8.14	8.99 8.14	25.7 24.7	25.7 24.7
LCF	6.41	5.11 3.37	19.9 3.04	19.7 3.49	70.2 12.3	65.8 13.2
LIFO	6.87	5.25 3.79	8.77 8.11	8.72 8.45	386 493	330 410

Table 2 Waiting time statistics Model I

Whilst in our previous study LCF was clearly the best of the four disciplines when we consider loss, this is not the case here[1]. The time spent in the primary queue by any customer is fairly close, whereas for the secondary queue there are varying average waits. Whilst HCF has the fastest average time in the first queue, it has the poorest in the second – this may lie with the fact that primary queue will contain more Class 1 customers as time continues than any of the other models. It is for similar reasons that the LCF has the least waiting time in the secondary queue, as all Class 2 customers are in the primary queue. HCF has the highest average and maximum loss in both Class 1 and Class 2, with LCF boasting the lowest loss for Class 1 and FIFO for Class 2

The HCF model delivers lowest time in the system to Class 1 customers. Interestingly, LCF gives marginally better times for Class 2, however significantly poorer times for Class 1 in the reverse model. Of the non-priority model, LIFO is the 'fairer' of the two, bringing the average time of the two classes to close levels.

3.2 Model II - 3 Classes

Now with a third class, the dual queueing scheme improves in terms of waiting time statistics. All combined time in the queue, time in the system and maximum time's statistics in Table 3 show the value of the dual queue. Furthermore, the priority models exhibit major improvements in waiting over their non-priority counterparts. The LIFO again performs poorly. However if we wish to have evenness in terms of waiting times, it certainly achieves this - at the risk of waiting on some occasions around 60 times longer than the other models. When we consider the importance of first class traffic, the HCF delivers. The waiting times are over 3 times less than the next best model. With this swift service of class 1 customers, a follow-on effect occurs for the 2nd class, with this too having the best waiting time statistic of all the models. However the 3rd class is the poorest in HCF.

Queue regime	W_q^1	$W_q^{1,2}$	$W_s^{1,2,3}$	$W_s^{1,2,3}$	$M_s^{1,2,3}$	$M_s^{1,2,3}$
	single	**dual**	**single**	**dual**	**single**	**dual**
HCF	20.4	9.74 7.4	6.97 8.08 37.2	6.73 7.12 23.6	25.5 45.8 111	24.7 38.6 76.4
FIFO	21.9	9.56 6.85	26.1 25.3 24.5	16.4 15.4 15.1	56.2 55.1 51.6	44.6 40.9 42.4
LCF	23.3	8.81 6.61	105 12.2 5.53	41.5 11.1 6.1	224 55.7 25	179 47.6 22.8
LIFO	20.6	9.64 7.09	21.7 21.6 25.4	14.4 14.3 16.6	1480 1190 1320	463 460 1100

Table 3 Waiting time statistics Model II

The dual queue now outperforms the single queue in terms of both average waiting time for any customer in the queue, and class wise in the system. The LIFO has serious problems with extreme maximum waiting times. The problem lies with a system that sees customers never being served. The FIFO shows evenness for all classes, with improvement in the dual queue. For service providers, the decision of FIFO or one of the priority regimes could be governed by either the maximum threshold or average waits. The priorities here over better average service, whereas the FIFO offer better maximum thresholds. Our prior work showed the priority regimes to be superior, and combined with the above results, this seems the best here[1]

3.3 Model III - 4 Classes

The introduction of another class strengthened the case for the HCF model. Whilst the 1st, 2nd and 3rd class customers received marginally better waiting times under the dual queue, the 4th class benefited from significant improvement. This may be an important factor if considering the value of the 'common' class. The HCF stands alone for waiting times, showing at least 50% improvement in waiting times over its rivals for 1st and 2nd classes.

Queue regime	W_q^1	$W_q^{1,2}$	W_s^{1-4}	W_s^{1-4}	M_s^{1-4}	M_s^{1-4}
	single	**dual**	**single**	**dual**	**single**	**dual**
HCF	22.3	10 8.26	9.28 7.6 8.15 36.2	9.07 7.13 7.95 21.1	39.4 49.2 51.9 103	32.4 40.3 56 99.9
FIFO	22.9	10.1 8.02	29.2 27.4 26.8 25.9	20.7 16.6 16.7 15.5	88.5 74.8 80.2 89.2	54.8 55.1 53.1 52.2
LCF	23.4	9.68 6.24	258 44.8 17.8 7.69	75.2 32.6 15.3 6.95	738 201 87.9 40.3	330 153 104 31.1
LIFO	20.1	9.67 7.15	32.5 25.4 26.4 21.5	16.8 18.7 15.7 14.4	1170 1100 1150 1490	403 468 567 669

Table 4 Waiting time statistics Model III

The dual queue again offers superior waiting times for all regimes, making it the best choice for four classes. FIFO consistently gives an evenness across classes, which may be preferential to service providers wanting this quality of service criteria.

3.5 Model IV - 5 Classes

The 5-class model was included to further investigate a trend appearing for waiting time (and indeed loss[1]). This is that the middle classes are suffering high levels of loss with respect to the low/high classes, yet are receiving shorter waiting times. In the 5 class models, this trend continued. The 2nd and 3rd classes in the HCF model were the fastest through the system. 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. With more classes, the overall system is slowed down, hampering waits for the high class customer.

Queue regime	W_q^1	$W_q^{1,2}$	W_s^{1-5}	W_s^{1-5}	M_s^{1-5}	M_s^{1-5}
	single	**dual**	**single**	**dual**	**single**	**dual**
HCF	6.01	5.33 10	12.5 8.54 8.48 10.5 13.6	14.1 8.79 8.45 10.2 12.6	53.8 60.4 65 112 198	63.6 74.7 86.9 105 195
FIFO	7.45	6.61 12.3	19.5 13.6 11.8 11 9.18	19.5 13.6 11.8 11 9.18	84.3 79.8 82.1 81.6 82.3	84.3 79.8 82.1 81.6 82.3
LCF	5	5.63 7.79	23 15.8 9.82 6.74 4.48	27.3 17 10.7 7.25 4.75	164 172 112 99 87.9	134 118 85.6 59.5 38.1
LIFO	6.31	5.15 9.32	17.2 11.7 10.4 9.91 9.96	16.1 12.2 9.74 8.93 8.08	120 176 192 325 318	164 177 137 236 191

Table 5 Waiting time statistics Model IV

An interesting result is the marginal 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 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 arrives here, especially under the dual queue.

3.6 Summary

From our prior work on the performance characteristics in terms of loss, we concluded the dual queue scheme for 3 or more gave better results. Furthermore, the HCF regime was seen as the best[1]. Here, the results further solidify these findings, with the waiting times superior for the dual queue. If we have 3 or more classes, there is a strong case for the HCF regime, with maximum waits the lowest, and class waits excellent for high-class customers.

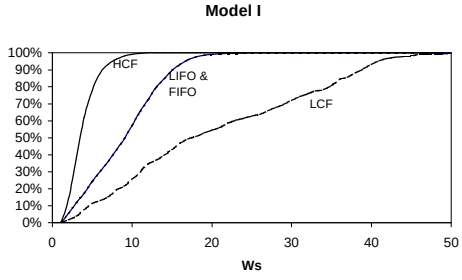
4 Waiting time distributions for HCF using the dual queue regime

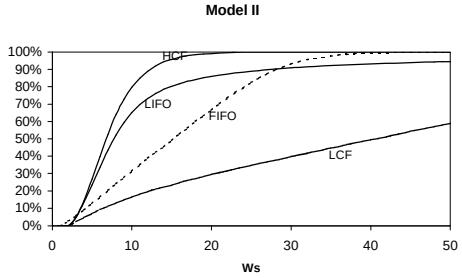
This section is arranged into three areas of analysis. Firstly, for each of the four models for the dual queueing scheme, the cumulative distribution function was calculated using a distribution fit for the highest class only. Each figure represents $P(W_s^1 < x)$.

Secondly, we complete a probability distribution fit for the dual HCF customers across models, and compare the findings. Finally, for Models I-IV, the arrival rate of the lowest class was varied. In this way, the effect of a small mean to large mean arrival of the least important customer can be compared.

4.1 CDF of first class customers in the dual queue

The distribution functions below are constructed using the results from the simulations described in section 2. For models I-IV, the first-class customers CDF is given for each of the queueing disciplines.





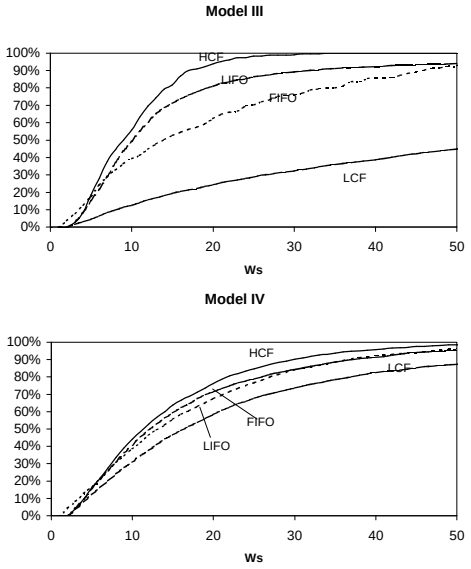


Fig. 2 CDF of waiting times for Class 1, Models I-IV

As the arrival and service rates are different for the models, we cannot make direct model comparisons. However, from Figure 2, the CDF are closest to each other in model IV, and move away as the models decrease in number of classes. The LCF discipline is poorest, especially in models II and III. The HCF shows superior waiting time probabilities for all models, however the margin closes between disciplines as the classes increase.

When comparing the LIFO and FIFO queues, the LIFO performs well through the middle of the CDF. However the presence of customers at the front of the queue (ie the first-in customers), sees the presence of extremely large waiting times in times of congestion, and the LIFO CDF appears asymptotic to its upper tail. This has been found to be the case in other queueing models, such as M/G/1 LIFO, where approximations were used[7]. In the previous section, we saw that the LIFO had the lowest levels of loss in many of the four Models. For an increase in quality of service, the introduction of a time-out discard limit would increase the loss, but should also have the effect of reducing these 'extreme values' for LIFO waiting times.

4.2 Distribution fit for DQ HCF Class 1

To model the CDF, a distribution fit for each of the HCF curves was undertaken. The curve fitting is undertaken using maximum likelihood estimators. The choice of distribution was made by choosing the distribution with the smallest squared error. The results of Chi-square and Kolmogorov-Smirnov goodness-of-fit tests are also shown in Table 6. These results are presented in the form of p-values; the p-value is the largest value of the type-I error probability that allows the distribution to fit the data.

Model	Distribution	Square error	c^2 p-value	K-S Test p-value
HCF (IV)	Γ (10.1,1.25)	0.000345	0.145	> 0.15
HCF (III)	Γ (4.27, 1.89)	0.004952	0.136	> 0.15
HCF (II)	Γ (2.61, 2.11)	0.000597	< 0.005	-
HCF (I)	Γ (1.37, 2.07)	0.000671	0.0086	0.0711

Table 6 Distribution fits for HCF Class 1 by Model

From the findings, the best fit for each model was the gamma distribution. In Figure 3, the plots for the distributions given in Table 6 are displayed. In the Table, $\Gamma(a, b)$ refers to the gamma distribution with location parameter α and scale b. The technique of determining the final values of alpha and beta in the function follow a numerical scheme for α < 1.5[6],[8]. For α >= 1.5, three inverted approximations to the gamma distributions based on the Burr family of distributions are used, depending upon the values of α[9],[10]. In addition, to calculate the natural logarithm of the Γ(Y) function, a polynomial approximation is used[11].

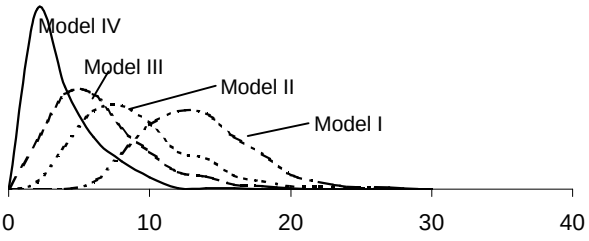


Fig. 3 Waiting time distribution for HCF Class 1, Models I-IV

4.3 Waiting times behaviour with respect to low class customer arrival rates

To analyse the effect of varying low class arrival times, the waiting times of all classes in the four models were examined. For all the disciplines, the waiting times follow similar patterns for each model for the same type of disciplines. We look specifically at the behaviour of the high-class customer when the low class customer's arrival rate is varied. For the FIFO scheme, all four models had the lowest class customer waiting the shortest amount of time, and followed the patterns seen in Figure 4.

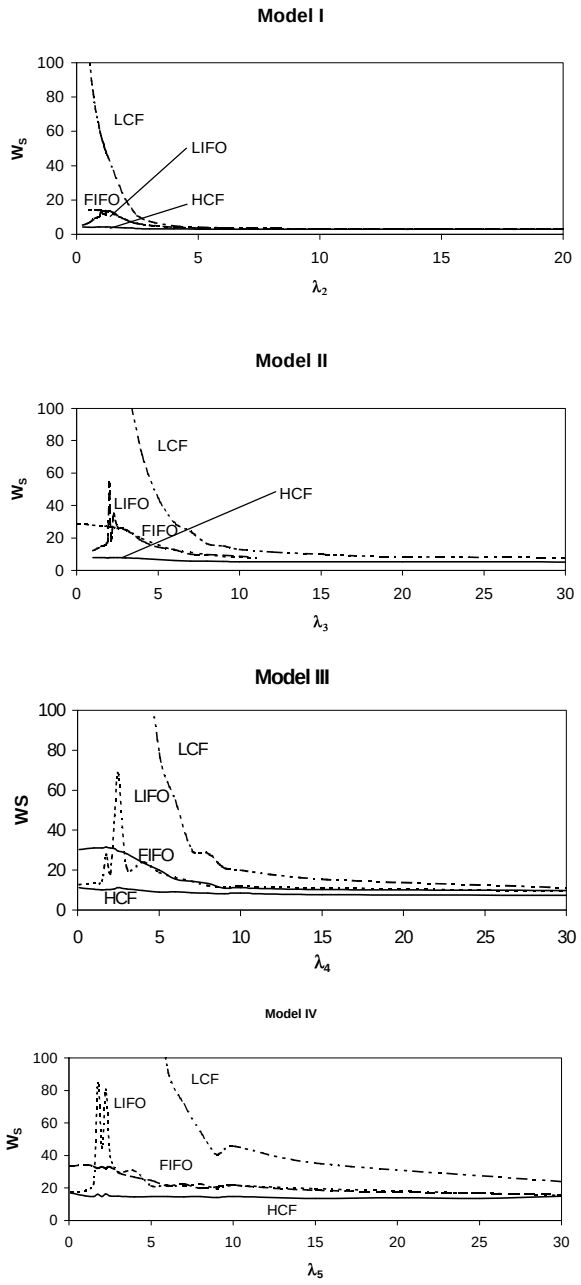


Fig. 4 Class 1 Waiting Times for Models I-IV by $l_{\mathrm{mod}\,el}$

From Figure 4, the LIFO again displays its asymptotic behaviour, ie. as $l_{i+1} \to 0, W_q^{i+1} \to \infty$, where *i* is the model number. The LIFO has early periods of instability. This instability increases as the number of classes increase, and hints of this can be seen in Figure 4. Once again, the HCF discipline is superior to all the others. However, it is best when the arrival rates are low.

5 Concluding Remarks

We have presented a new combination of schemes called the MPDQ and explored some of its waiting time characteristics under various disciplines. As a scheme combining priorities with a dual queue, MPDQ The HCF discipline for 3 and 4 classes performed well, whereas the LCF and LIFO showed volatility. For service providers, the introduction the HCF is worth investigation, with the final decision governed by quality of service constraints.

References:

[1] A. Bedford, P. Zeephongsekul, Simulation Studies on the Performance Characteristics of Multi Priority Dual Queue (MPDQ) with Finite Waiting Room and Non-Preemptive Scheduling, *PSOR'01*

[2]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.

[3] Ravindra S. Ranasinghe, Lachlan L. H. Andrew, David A. Hayes, David Everitt, Scheduling disciplines for multimedia WLANs: Embedded round robin and wireless dual queue, *Proc. IEEE Int. Conf. Commun.*, 2001, pp. 1243-1248.

[4] 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.

[5] Dietmar Wagner, Waiting times of a finite-capacity multi-server model with non-preemptive priorities, *European Journal of Operational Research 102*, 1997, pp. 227-241.

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

[7] Joseph Abate, Ward Whitt, Limits and approximations for the M/G/1 LIFO waiting-time distribution, *European Journal of Operational Research 20*, 1997, pp. 199-206.

[8] D.T. Philips, C.S. Beightler, Procedures for Generating Gamma Variates with Non-Integer Parameters Sets, Jour. Stat. Comp. Simul., Vol. 1, 1972, pp. 197-208.

[9] D.J. Wheeler, An Approximation for Simulation of Gamma Distributions, Jour. Stat. Comp. Simul., Vol. 3, 1975, pp. 225-232.

[10] P.R. Tadikamalla, Computer Generation of Gamma Random Variables, C.A.C.M., Vol. 28, No.5, 1978

[11] P.C Pike, I.D. Hill, Algorithm 291. Logarithm of the Gamma Function, C.A.C.M., Vol. 9, No. 9, 1966, pp. 684.