Delay Control in Multi-Class Multi-Server RST-based Communication Networks Via M/K Setting

Amin Babiker A/Nabi Mustafa†, Muawia Mohammed Ahmed Mahmoud††,

Ismail El-Azhary†††, and Huda Adibah Mohd Ramli††††

†Department of Communications Engineering, Faculty of Engineering, Al-Neelain University Khartoum, Sudan
††Department of Control Engineering, Faculty of Engineering, Al-Neelain University, Khartoum, Sudan
†††Department of Computer Engineering, Faculty of Engineering, Al-Neelain University, Khartoum, Sudan
††††Department of Electrical and Computer Engineering, international Islamic university Malaysia

Summary
In this paper, analysis of non-preemptive priority queues with multiple servers and multiple priority classes, which is based on Residual Service Time (RST) is utilized to calculate the total waiting time for system customers. Definitely, this waiting time is determined by several parameters such as the number of servers, the order of the priority class to which the customer belongs and the total number of customers in the system. Analysis results show that the same waiting time can be attained by different Priority class order/ Number of Servers combinations. Accordingly, these parameters could be used in determining the desired level of performance expressed in terms of waiting time. The obtained results of waiting times and their relations with priority class orders and Number of Servers could help a lot in justifying and supporting this proposed RST-Based Analysis.

Key words:
non-preemptive priority queues, residual service time, total waiting time, RST-based analysis.

1. Introduction

One of the most powerful mathematical tools for making quantitative analysis of computer networks and communication systems is the queuing theory [1]. Analytical techniques based on queuing theory provide a reasonably good fit to reality. They may play a very important role in studying the effect of load changes, forming a good base for design purposes and for making necessary performance projections [2], [3]. To characterize computer communication networks performance the average delay required to deliver a packet (a message) from origin to destination is measured or calculated. Delay considerations have a strong influence on the choice and performance of network routing, flow control and congestion control algorithms.

In computer networks, there are several models describing the behavior of both preemptive and non-preemptive queuing systems. In the non-preemptive queuing systems, it is assumed that always the highest priority job is selected by the server with no interruptions allowed until the job is completed. On the other hand, in the preemptive queuing systems, models allow job interruption if a higher priority job is submitted. In this paper we will focus our discussion on the non-preemptive priority queuing systems. Several researchers have treated delays encountered by jobs on non-preemptive priority queuing systems where only limited number of priority classes is considered. D. Lee [4] and G. Horvath [5] have considered non-preemptive queuing systems with two priority classes namingly high and low-priority. Moreover, Landry and Stavrakakis have developed a three-priority queuing policy that can be applied to the distributed queue dual bus (DQDB)[6]. Multiple priority classes are rarely discussed in literature. Developing a generalized model for waiting time for multi-class multi-server systems would be critically needed to design newer networks where multiple priority classes can be implemented. In this paper, multiple priority classes are considered during the calculations of delays encountered by jobs using multiple servers, non-preemptive systems. The use of queuing theory often requires making simplifying assumptions to perform meaningful yet close to reality analysis. In general more realistic assumptions result in highly complex analytical expressions which tender an extremely difficult analysis. It is sometimes impossible to obtain accurate quantitative delay predictions on the basis of queuing models that make use of very realistic assumptions.[7].

The paper is organized as follows. In Section II, we provide a background for priority queuing systems where the wait time for each priority class with one server is derived. The derived relation for the wait time is then expanded to multiple servers' case as will be shown in Section III. A numerical examples and results discussion are given in Section IV. In Section V the conclusions are given.

2. Background for Priority Queuing Systems

The analysis of Priority Queuing is based on the analysis

Manuscript received December 5, 2013
Manuscript revised December 20, 2013

of M/G/1 system in which customers arrival rate follows a Poisson Process with rate λ and the customers service times have a general distribution (M stands for memory less systems.[8].

In priority queuing systems the arriving customers are divided into n priority classes such that for class k, the priority of class k where 0<k< n is higher than priority of class k+1.

The arrival rate and the first two moments of service times of each priority class are denoted as:

$$\lambda_k , \overline{x_k} = 1/\mu_k \ \& \ \overline{x_k^2}$$

Arrivals of all classes are assumed to be independent, Poisson and independent of the service times.

Non-preemptive priority rule dictates that a customer undergoing service is allowed to complete service without being interrupted.

To determine the average delay for each priority class, the following parameters are defined according to the standard notation in [7]:

$N_Q^k \equiv$ Average number in queue for priority class k

$W_k \equiv$ Average queuing time for priority class k

$\rho_k = \lambda_k/\mu_k \equiv$ System utilization for priority class k

$R \equiv$ Mean residual service time

The overall system utilization is less than unity. Then

$$\rho_1 + \rho_2 + \rho_3 + \cdots + \rho_n < 1 \quad (1)$$

The customer waiting time w, is composed of two components:

The mean residual service time R which is the time required to complete the service of the undergoing service customer.

The time required for the service of all queued customers.

The system service rate is μ then average service time of a given customer is 1/μ Assuming that there are NQ queued customers in the system, then the total service time for all customers is

$$N_Q/\mu \quad (2)$$

Then the total wait time can be given by:

$$W = R + N_Q/\mu \quad (3)$$

Applying (3) for the highest priority class

$$W_1 = R + N_Q^1/\mu_1 \quad (4)$$

From Little's Theorem, it is known that

$$N = \lambda W \quad (5)$$

Where λ is the average customers' arrival rate. Considering the highest priority class, expression (5) becomes:

$$N_Q^1 = \lambda W_1 \quad (6)$$

Using expression (6) in equation (4), the first priority waiting time can be described as

$$W_1 = \frac{R}{(1-\rho_1)} \quad (7)$$

Where ρ is the utilization factor, which is defined as the ratio of the average customers' arrival rate to the average service rate $\rho = \lambda/\mu$

There is a similar expression for the second priority class except that, there is additional delay due to high priority customers that arrive while this second priority class customer is waiting in a queue. This additional delay should be taken into account

Then W2 is given by

$$W_2 = R + N_Q^1/\mu_1 + N_Q^2/\mu_2 + \lambda_1 W_2/\mu_1 \quad (8)$$

Rearranging and using Little's Theorem, the waiting time for the second priority class becomes:

$$W_2 = \frac{R + \rho_1 W_1}{1-\rho_1-\rho_2} = \frac{R}{(1-\rho_1-\rho_2)(1-\rho_1)} \quad (9)$$

Intuitively, for any priority class k, Wk, can be given by

$$W_k = \frac{R}{(1-\rho_1 \cdots - \rho_k)(1-\rho_1 - \cdots - \rho_{k-1})} \quad (10)$$

The average delay per customer of class k is composed of two components, the service time plus the waiting time (Queuing time). Then the average delay Tk is given by:

$$T_k = \frac{1}{\mu} + W_k \quad (11)$$

It can be shown that, the residual service time in single server systems, is given by:

$$R = \frac{1}{2}\sum_{i=1}^{n} \lambda_i \overline{x_i^2} \quad (12)$$

3. Extension to Multiple Servers Case

The above formula cannot be extended to multiple servers' case (multiple communication channels from the communication systems point of view) due to the fact that, the residual service time is complex to formulate mathematically in a fashion simple enough to enable calculating the average customer waiting time.To overcome this problem, the proposed solution is to assume that the service times for all priority classes are identically and exponentially distributed.

Consider the M/M/m system in which customers arrive according to a Poisson process while service times are exponentially distributed, it can be shown that, using

Markov Chains, the probability of n customers in the system is given by:

$$p_n = p_0 \frac{(m\rho)^n}{n!}, \quad n \le m \tag{13}$$

$$p_n = p_0 \frac{m^m \rho^n}{m!}, \quad n > m \tag{14}$$

Where ρ is the utilization factor, m is the number of servers (Communication Channels), ρ0 is the probability of 0 customers in the system.

Since

$$\sum_{n=0}^{\infty} p_n = 1$$

Then using (13) and (14), one can write ρ0 as follows:

$$p_0 = \left[1 + \sum_{n=1}^{m-1} \frac{(m\rho)^n}{n!} + \sum_{n=m}^{\infty} \left(\frac{(m\rho)^n}{m!} \times \frac{1}{m^{n-m}} \right) \right]^{-1} \tag{15}$$

The first term on the left side of (15) can be simplified to

$$1 + \sum_{n=1}^{m-1} \frac{(m\rho)^n}{n!} = \sum_{n=0}^{m-1} \frac{(m\rho)^n}{n!}$$

And the second term on the left side of (15) can be simplified to

$$\sum_{n=m}^{\infty} \left(\frac{(m\rho)^n}{m!} \times \frac{1}{m^{n-m}} \right) = \frac{m^m}{m!} \frac{\rho^m}{(1-\rho)}$$

Then (15) becomes:

$$p_0 = \left[\sum_{n=0}^{m-1} \frac{(m\rho)^n}{n!} + \frac{m^m}{m!} \frac{\rho^m}{(1-\rho)} \right]^{-1} \tag{16}$$

The queuing probability is the probability that an arrival will find all servers busy and hence it will be forced to wait in a queue. This probability gives a powerful measure for the evaluation of the performance of different communication systems.

Equation (15) shows that, the queuing probability PQ is given by:

$$p_Q = \sum_{n=m}^{\infty} p_n = p_0 \frac{m^m}{m!} \frac{\rho^m}{(1-\rho)} \tag{17}$$

Where P0 is given by Equation (16).

The expected number of customers waiting in queue (not in service) is given by:

$$N_Q = \sum_{n=0}^{\infty} n \times p_{m+n} \tag{18}$$

Since

$$E(x) = \sum_{\forall i} x_i \times f(x_i)$$

Equation (14) states:

$$p_n = p_0 \frac{m^m \rho^n}{m!}$$

Then

$$p_{m+n} = \frac{m^m \rho^{m+n}}{m!}$$

Then after few mathematical manipulations, NQ can be shown to be:

$$N_Q = \sum_{n=0}^{\infty} n p_0 \frac{m^m \rho^{m+n}}{m!} = p_0 \frac{(m\rho)^m}{m!} \rho \times \frac{1}{(1-\rho)^2} \tag{19}$$

From the expression of PQ given in (17), P0 can be written as

$$p_0 = \frac{p_Q \times m!(1-\rho)}{(m \times \rho)^m}$$

Substituting for P0 in (19) and simplifying, NQ can be written as

$$N_Q = p_Q \frac{\rho}{(1-\rho)}$$

By using Little's Theorem in (5), then the average time W the customer has to wait in queue can then be given by:

$$W = \frac{\rho \times p_Q}{\lambda(1-\rho)} \tag{20}$$

The utilization factor ρ for a given priority class i. is given by

$$\rho_i = \frac{\lambda_i}{m\mu} \tag{21}$$

Then

$$\rho = \sum_{i=1}^{n} \rho_i \tag{22}$$

From equation (7), (20), (21), the residual service time R can be written as

$$R = \frac{p_Q}{m\mu} \tag{23}$$

Equation (23) can be used in the calculation of the customer waiting time, in multiple servers' non-preemptive queuing systems as follows:

Substituting (23) in (10) gives:

$$W_k = \frac{p_Q \big/ m\mu}{(1-\rho_1-\ldots-\rho_{k-1})(1-\rho_1\ldots-\rho_k)} \quad (24)$$

Where P0 and PQ are given by (16) and (17) respectively.

4. Numerical Demonstration and Discussion

The above detailed equations that describe the customer waiting times for different priority classes in multiple servers (multiple communication channels) non-preemptive priority queuing systems, were used in writing a simple computer simulation program. Specifying the required parameters and inputs, the simulation program was used in obtaining waiting times corresponding to different priority classes as described by Equation (24).

The first run of the simulation program assumes the following set of values for different parameters:

Number of servers - communication channels: m=8

Number of priority classes: k=10

Utilization factors for all priority classes: $\rho i = 0.085$ $(1 \leq i \leq 10)$

System service rate per server -communication channel–: $\mu = 16$

Accordingly, Table1 contains results representing waiting times for different priority classes. These results are plotted as shown in Figure 1.

Table 1:Waiting times for different priority classes and servers with U=16.

K	M									
	1	2	3	4	5	6	7	8	9	10
1	0.0515	0.0239	0.015	0.0107	0.0081	0.0065	0.0053	0.0045	0.0039	0.0034
2	0.062	0.0288	0.0181	0.0129	0.0098	0.0078	0.0064	0.0054	0.0046	0.004
3	0.0762	0.0353	0.0222	0.0158	0.012	0.0096	0.0079	0.0067	0.0057	0.005
4	0.0958	0.0444	0.0279	0.0199	0.0152	0.0121	0.0099	0.0084	0.0072	0.0062
5	0.1242	0.0576	0.0361	0.0257	0.0196	0.0157	0.0129	0.0109	0.0093	0.0081
6	0.1672	0.0775	0.0487	0.0346	0.0264	0.0211	0.0174	0.0146	0.0125	0.0109
7	0.2374	0.1101	0.0691	0.0492	0.0375	0.03	0.0247	0.0208	0.0178	0.0155
8	0.3636	0.1686	0.1058	0.0753	0.0575	0.0459	0.0377	0.0318	0.0272	0.0237
9	0.6266	0.2905	0.1824	0.1298	0.0991	0.079	0.0651	0.0548	0.0469	0.0408
10	1.3367	0.6198	0.389	0.2769	0.2113	0.1686	0.1388	0.1169	0.1001	0.087
U=16										

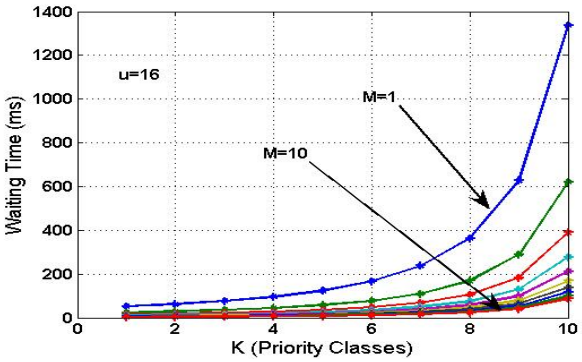

Figure 1: Waiting times Vs Priority classes for multi-servers for

U=16

The second run of the simulation program assumes the same above used set of values for different parameters except that, system service rate per server (communication channel) is given by: $\mu = 4$.

Again, Table 2 contains results representing waiting times for different priority classes. These results are plotted as shown in Figure 2.

Table 2:Waiting times for different priority classes and servers with U=4.

K	M									
	1	2	3	4	5	6	7	8	9	10
1	0.206	0.0955	0.06	0.0427	0.0326	0.026	0.0214	0.018	0.0154	0.0134
2	0.2482	0.1151	0.0722	0.0514	0.0392	0.0313	0.0258	0.0217	0.0186	0.0162
3	0.3048	0.1413	0.0887	0.0631	0.0482	0.0385	0.0316	0.0266	0.0228	0.0198
4	0.3833	0.1777	0.1116	0.0794	0.0606	0.0484	0.0398	0.0335	0.0287	0.025
5	0.4966	0.2303	0.1445	0.1029	0.0785	0.0627	0.0516	0.0434	0.0372	0.0323
6	0.6689	0.3102	0.1947	0.1386	0.1058	0.0844	0.0695	0.0585	0.0501	0.0436
7	0.9497	0.4404	0.2764	0.1968	0.1502	0.1198	0.0986	0.083	0.0712	0.0618
8	1.4542	0.6743	0.4233	0.3013	0.2299	0.1835	0.151	0.1271	0.109	0.0947
9	2.5063	1.1622	0.7295	0.5192	0.3963	0.3162	0.2602	0.2191	0.1878	0.1632
10	5.3467	2.4793	1.5562	1.1077	0.8453	0.6745	0.5551	0.4674	0.4006	0.3481
U=4										

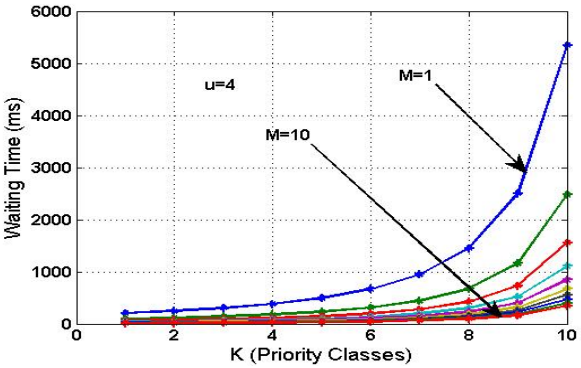

Figure 2: Waiting times Vs Priority classes for multi-servers for

U=4

5. Discussion of the Results

The tables and plots speak for themselves and they say that delay times for all classes improves significantly with increased number of servers and delay times worsens for same servers number but with increased number of priority classes.

As shown by Table 3 which describes the results of the first run, the waiting times for the first priority class decrease as the number of servers increases. For example the waiting time for the first priority class in the case of single server system is found to be 0.0515 ms. However, the value decrease to 0.0107 and 0.0045 ms in the cases of 4 and 8 servers respectively.

Table 3: Similar Waiting Times for different M/K pairs (first run)

Pair NO.	Number of servers	Priority class	Waiting time (ms)
1	4	2	0.0129
1	7	5	0.0129
2	5	3	0.012
2	6	4	0.0121
3	5	2	0.0098
3	7	4	0.0099
4	9	1	0.0039
4	10	2	0.004
5	7	1	0.0053
5	8	2	0.0054
6	8	1	0.0045
6	9	2	0.0046
7	6	1	0.0065
7	7	2	0.0064
8	4	3	0.0158
8	6	5	0.0157
9	2	5	0.0576
9	5	8	0.0575

Again, dealing with the fourth priority class, the waiting time in the case of single server system is found to be 0.0199 and 0.0084 ms in the cases of 4 and 8 servers respectively.

Dealing with the same number of servers but with different priority classes, it is clear that waiting times increase with the decreased number of priority classes. For example, the waiting time in the case of single server first priority class is found to be 0.0515 ms. This value increase as we deal with second and third priority classes to 0.0620 and 0.0762 ms respectively till it reaches 1.3367 ms in the case of tenth priority class.

Again, the waiting time in the case of 4 servers first priority class is found to be 0.0107 ms. This value increases considerably when we deal with second and third priority classes to 0.0129 and 0.0158 ms respectively till it reaches 0.2769 ms in the case of tenth priority class.

Table 4 describes the results of the second run. Again, waiting times decrease for the same priority class as the number of servers increases. For example, the waiting time for the first priority class in the case of single server is found to be 0.206 ms. However, this value decreases to 0.0427 and 0.0180 ms in the cases of 4 and 8 servers respectively.

Additional worth mentioning point is that, waiting times obtained in the second run are greater than their counterparts obtained in the first run. This is expected since the value of the service rate per server μ used in the second run is less than that used in the first one.

Again, dealing with fourth priority class, the waiting time in the case of single server is found to be 0.3833 ms. This value decreases to 0.0794 and 0.0335 in the cases of 4 and 8 servers respectively.

Table 4: Similar Waiting Times for different M/K pairs (second run)

Pair NO.	Number of servers	Priority class	Waiting time (ms)
1	4	2	0.0514
1	7	5	0.0516
2	5	3	0.0482
2	6	4	0.0484
3	6	1	0.026
3	7	2	0.0258
4	8	5	0.0434
4	10	6	0.0436
5	7	1	0.0214
5	8	2	0.0217
6	6	2	0.0313
6	7	3	0.0316

Again, dealing with the same number of servers but with different priority classes, it is clear that, waiting times increase with the increased number of priority classes. For example, the waiting in the case of single server first priority class is found to be 0.206 ms. This value increases when we deal with second and priority classes to 0.2482 and 0.3048 ms respectively till it reaches 5.3467 ms in the case of tenth priority class.

Again, the waiting time in the case of 4 servers first priority class is found to be 0.0427 ms. This value increases considerably when we deal with second and third priority classes to 0.0514 and 0.0631 ms respectively till it reaches 1.1077 ms in the case of tenth priority class.

Additional worth mentioning point shown by Fig3.1 is that, above a certain no. of servers the expected effect on waiting time is hardly noticeable. The reason is that, when the number of arriving customers is comparable with the number of servers the system performance – in terms of waiting time – tends to approach a steady state. Accordingly, any further increase in the number of severs will hardly affect the system performance.

Another worth mentioning point is that, referring to Table 3.1, the waiting time within the second priority class in the case of four servers is the same as that obtained in the case of seven servers within the fifth priority class. This case is not unique, several pairs of cells, as shown in Table 3.1 and Table 3.2, hold approximately the same value for waiting time. Some of these cases are extracted from tables 3.1 and 3.2 and are explicitly shown in Tables 3.3 and 3.4 respectively.

6. Conclusions

The assumption that the service times for all priority classes are identically and exponentially distributed led to

the possibility of extending the analysis of non-preemptive priority queuing systems to the multiple servers case (multiple communication channels). The extension is based on the developed formula for the residual service time R.

Discussion presented by the previous, section V. dictates that, if it is possible to change the priority class and/or the number of servers, then these parameters could be used in determining the desired level of performance expressed in terms of waiting time.

The possibility of changing the above mentioned parameters leads to the possibility of deciding the desired service cost barrier. This is true provided that most service costing systems rely on different performance parameters particularly, priority levels and average delays.

References

[1] Tanen, A. 1996 "Computer Networks ", 3rd ed. Prentice Hall of India, New Delhi.

[2] Stalling, W. 1998 "High Speed Networks", Prentice Hall, Upper Saddle River. New Jersey.

[3] Enns, S. T. and Sangjin Choi, "Use of GI/G/1 Queuing Approximation to test tactical parameters for the simulation of MRP systems", Simulation Conference, 2002. Proceedings of the Winter Volume 2, 8-11, Dec. 2002, Page(s): 1123 - 1129 vol.2

[4] Duan-Shin Lee, "A generalized non-preemptive priority queue", INFOCOM '95. Fourteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Bringing Information to People. Proceedings. April 1995 Page(s):354 - 360 vol.1

[5] Gabor Horvath, "A Fast Matrix-Analytic Approximation for the Two Class GI/G/1Non-Preemptive Priority Queue", 12th International conference on analytical and stochastic modelling Techniques and Applications ASMTA 2005 in Conjunction with 19th European Conference on Modelling and Simulation, June 2005.

[6] Randall Landry and Ioannis Stavrakakis, "Queuing study of 3-priority policy with distinct service strategies", IEEE/ACM Trans. on Networking, vol.1, no. 5, October 1993, Page(s):576-589.

[7] Bertsekas, D. & Gallger, R. 1992, "Data Networks", 2nd ed. Prentice Hall Engle wood cliffs, New Jersey.

[8] Silva, F. & Serra, D. 2003, "Locating Emergency Services with priority rules: The priority Queuing Location Problem", 27th Conference of National Statistics and Investigational Operations.

[9] Jens Schmitt, "On average and Worst Case Behavior in Non-Preemptive Priority Queuing" Darmstadt University of Technology, 2001.

Amin B. A/N Mustafa obtained his B.Sc. and M.Sc. from the University of Khartoum in 1990 and 2001, respectively. He obtained his Ph.D. from Al Neelain University in 2007. . He was the Head of the Dept. of Computer Engineering from 2001 to 2004. Then, he became the Vice Dean. He has been the Dean since 2009. His research areas include QoS in telecommunications, traffic engineering, and service costing. Dr Amin is a member of the Sudan Engineering Council.

Muawia Mohamed Ahmed Mahmoud obtained his B.Sc. from the University of Bagdad in 1987. He obtained the M.Sc. and the PhD from Sudan university of Science and Technogy in 2000 and 2007 respectively. He started his career in networks installation and Maintenance in many establishments in Sudan. Later, he switched to the academic field as Lecturer and Assistant professor in Electronic Engineering. He is now the Head of the Dept of Control Engineering in the Faculty of Engineering, AL Neelain University, Khartoum, Sudan.

Ismail El-Azhary received his BSc (Hons) degree from the University of Khartoum (Sudan) in1979. In 1989 he obtained his PhD degree from the University of Bradford (UK). He joined Omdurman Islamic University in 1992 as an assistant professor. Then, he moved to the Sudan University of Science and Technology as an associate professor in 1994. He became the Dean of the Faculty of Engineering, Al Neelain University from 2000 to 2005. His research interests include: QoS of networks, technical communication, digital typography, e-learning, antennas and propagation, and embedded systems. Dr El-Azhary is a Chartered Engineer and member of the IET.