Modelling Scheduling Systems with Exhaustive Priority Service

Muhammad Faisal Hayat, Tauqir Ahmad, Muhammad Afzaal, Khadim Asif, H. M. Shahzad Asif, Yasir Saleem

Department of Computer Science & Engineering
University of Engineering and Technology, Lahore (Pakistan)
fsl.hayat@gmail.com

Abstract: Scheduling systems are an integral part of edge nodes in modern communication systems. These nodes aggregate the incoming traffic flows and groomed traffic is sent to core network. The aggregation is done on the basis of scheduling criteria. In this work, we have analyzed a scheduling system where high priority traffic is exhaustively served. The modelling approach we have used is Markov chain modelling. It is assumed that buffers available are of finite nature. Important performance measures such as blocking and waiting probabilities and mean flow time have been analyzed to give an insight into the system behavior. To prove the accuracy of analytical modelling, all results have been simulated.
[Muhammad Faisal Hayat. **Modelling Scheduling Systems with Exhaustive Priority Service.** *Life Sci J* 2012;9(4):74-80] (ISSN:1097-8135). http://www.lifesciencesite.com. 12

Keywords: Optical burst switching, Joint edge-core node

1. Introduction:

In modern communication networks, edge node's major functionality is to provide a fair service to all incoming flows which are using edge nodes as access points. This fairness is achieved using some scheduling mechanism. A lot of scheduling policies have been proposed and it depends upon the particular scenario which policy suits the system requirements. There are several mathematical treatments possible to model a scheduling system. In general, such a system is many queue single server queueing system. Many results from queueing literature can be extended to analyze such a system. It is very helpful to categorize the research in this context. The recent survey has been given by [35]. This categorization is based on type of service, switch over rates, and the system size. There are other categorizations, presented by [25]. The most common types of service are exhaustive, non-exhaustive or gated service. In exhaustive service there are also many flavors. In general, in exhaustive service, the server provides service to a particular queue till the queue is empty. In a gated service system, the server switches to a queue and serves exhaustively only those customers which were present in the queue at switch-over time. In non-exhaustive systems, the server serves only one customer in a queue but queues can be polled in many different fashion like in a cyclic manner or some pre-defined order [13]. The switchover time is also an important parameter. It is the time taken by server to switch to another queue after service completion. This time is usually very small as compared to service time and is ignored in many studies; however this can affect the performance of a scheduling system where it cannot be ignored in comparison with service time. The capacity of system is also an important parameter to take into account. Most of studies have assumed infinite queue capacities for modelling such scheduling systems. However, real systems always have finite capacities, therefore queues with finite capacities should be considered for true performance evaluations. Blocking probability, which is an important performance measure in real networks is only applicable when system has limited capacity. In literature, first study of multi-queue single server was done by [19]. The first study on communication networks was done in early 70s in order to model the time-sharing systems. Leibowitz in [16] first studied a cyclic queueing system with constant switch over times. Exhaustive and gated service with null switch-over times have been analyzed by Cooper and Murray [5], [6]. Bux and Truong [3] gave a very general analysis for exhaustive service discipline with any number of queues. Models with asymmetric service were also presented by many authors like Lee [15]. He presented an analysis for two-queue system. Lee analyzed one queue for exhaustive service and limited discipline was used for the other queue. Kuehn [13] analyzed the round-robin queueing system. He derived results for GI/G/1 queues based on cylic service time and general switch-over rates. He also extended the results for batch service and re-transmission with constant bit-error rate [14]. All studies discussed above used infinite queue capacity. The models with exhaustive and non-exhaustive service discipline with finite queue capacity are not so common in the

literature. Single buffer systems were first discussed by Chung [4] and Takine [29], [30], [31]. Magalhaes in [20] used M/M/1/1 queues for a multi-queue system. Titenko [32] calculated moments of the waiting time for single buffer multi-queue system. Takagi [26] used M/G/1/n for finding Laplace transforms of cycle times with exhaustive service. Tran-Gia have done several studies in this regard [33], [34]. He used imbedded Markov chain for analysis of a non-exhaustive queueing system with finite buffers. In [1] authors presented a markov chain analysis of cyclic finite queue system, with non-zero switch-over times. A very good extension of early work was presented by Takagi [27]. In addition [35] covers a good overview of available models. Closed-form solutions have been given by some authors but mostly these results are available for single-buffer systems. There are several ways to give priority to involved queues in a scheduling system [25], which includes schemes using more visits to higher priority queues or exhaustive/gated service of higher priority queues. These schemes are known as queue priority schemes. Manfield [21], [22] studied one exhaustive queue and n - 1 limited service queues. Message priority schemes are another form of priority schemes, in which priority is done within one queue [25]. A priority scheme in which the highest priority is served while visit to a queue was analyzed by [37], Karvelas and Leon-Garcia [12]. Regarding application areas, Nagle in [24] proposed and analyzed fairqueueing system which is mapable to many scheduling algorithms without any priority considerations. Ibe and Trivedi [11] purposed stochastic Petri Net models for exhaustive, gated and limited service queue scheduling discipline. Takagi [28] proposed three main areas in communication networks where polling models can be used for modelling scheduling policies. Bruneel and Kim [2], Grillo [9] and Levy [17] analyzed several examples of communication networks including ATM. In late 90s, people also studied many recent systems using multiqueue models like [38] and [36]. They analyzed polling systems involved in communications over IEEE 802.11 WLANs. Bluetooth systems were studied using multi-queue polling models by [23]. In this paper an exhaustive priority scheduling system with non-negligible switchover time and queues with finite capacity is considered. The paper is originally presented in [10] and it is extended here with more results. and discussions. Additionally, background of the work is properly extended to have reference for future research. The system is a multi-queue single server system and is evaluated with very high switch-over rate as compared to the service rate; therefore its effect is ignored when switch-over happens with a service in the current queue. However, when the server switches without any service the switch-over time has to be taken into account. The main quantities of interest are the mean number of customers, the mean waiting time, and the mean blocking in a queue.

2. Materials and Methods:

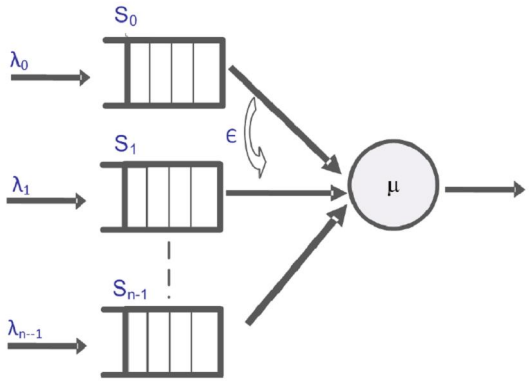

Fig. 1. Model for queueing system

An n-queue single server system is used model scheduling problem under consideration. The size of queues is represented by s_i where (i = 0, 1, …, n - 1). Single server provides the service with exponentially distributed service time which is represented by μ_i and it is the same for all types of customers. The arrival rate in different queues is shown by λi where index i serves for the same purpose as above. The server switches from one queue to the next and switch-over time is only taken into account when the server jumps to the next queue from an empty queue. It means that we are ignoring switchover time when a service takes place and server jumps to the next queue with a service in the current queue. Switchover rate is represented by ε. and it is also considered a single parameters for all queues. The queues are numbered ; 0; 1; 2 … :n - 1. And descending order is used to show priority. It means a queue with number 0 has the highest priority. The working of system is explained below. The server starts with the queue 0 and service all available customers exhaustively if queue is not empty, after the highest queue is empty it jumps over to the next queue in priority and serves one customer from it. After serving one customer it again polls the queue 0 for any available customer. If a queue is empty, the server switches to the next queue with delay of switch over time. The server keeps on going down to priority if all high priority queues are empty. As a thumb rule the

server jumps to the queue 0 after serving one in any queue and it goes down to lower priority queue if higher priority queues are empty. The switch over time is only used when the server switches over without serving any from the current queue.

2.1 Analysis of Model

To analyze the model consisting of a single server and multiple priority queues as described in the last section above, we take a vector T = [Q_1;Q_2; … ;Q_n; i], where i is used as queue index and is also used in diagram to show the queue being served. The process involved here is modelled then with the help of a Markov Chain. For an n-queue system n + 1 parameters are used to represent a state of the Markov process in the steady state. To explain better using Fig. 2, we can draw the state diagram for a two queue system. Three parameters Q_1,Q_2,L are used, to represent the state as shown in Fig. First parameter describes the number of customers in queue in the first queue, second parameter represents the number of customers in the second queue and the last one describes the state of server. L = 1 shows first queue is being served which is having capacity s_1 and L = 2 shows the server is busy with queue 1. First two parameters range upto s_i +1, where si is the queue size and 1 shows an extra customer in with server. The state space is having three-dimensions and it is visualized as shown in Fig. The state space shown actually helps to understand the flow of process, however, it cannot be extended for more queues. Nevertheless, the same rules can be used to formulate the problem in computer programs to extend the Markov chain for large number of queues. The symmetry of the system also helps to generalize the rule of evolution of Markov process involved. Using state space, the state probabilities p can be calculated by solving a system of linear equations given by:

$$PQ = 0 \tag{1}$$

Where P = [p_1, p_2, …], is probability vector and Q is an infinitesimal generator matrix given by

$$Q = \begin{bmatrix} -2\lambda & 0 & 0 & \epsilon & 0 & \cdots \\ 0 & -2\lambda & \epsilon & 0 & 0 & \cdots \\ 0 & 0 & -(2\lambda+2\epsilon) & \epsilon & 0 & \cdots \\ 0 & 0 & \epsilon & -(2\lambda+2\epsilon) & 0 & \cdots \\ \lambda & 0 & \lambda & 0 & -(2\lambda+\epsilon) & \cdots \\ 0 & \lambda & 0 & \lambda & \epsilon & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

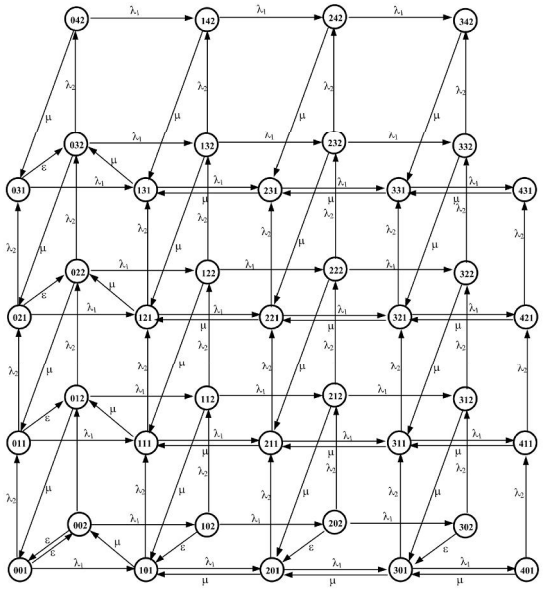

Fig. 2. State space of a system with two queues

Using state probabilities, the mean number of customers in a queue and the mean number of customers in the system can be found using equations below. Mean number of customers in the system:

$$E[X] = \sum_{i=0}^{s+1} i p_i \tag{2}$$

Mean number of customers in the queue:

$$E[Q] = \sum_{i=0}^{s+1} (i-1) p_i \tag{3}$$

From above equations using Little's law [18], it can be found Mean flow time (time spent in system)

$$E[T_w] = \frac{E[Q]}{\lambda} \tag{4}$$

Mean waiting time (time spent in queue):

$$E[T_f] = \frac{E[X]}{\lambda} \tag{5}$$

State space is easy to extend to form a generalized system for n-queues. It requires now n + 1 parameters to represent a state. It is not at all easy to draw state diagram for large number of queues, however, the same construction rules evaluated for two dimensional state diagram may be extended to formulate a system of linear equations for higher dimensions. The various performance measures can also be given in a general form as shown by equation given below. The mean number in the system and the

mean number in queue are given by

$$E[X_1] = \sum_{i_n=0}^{s_n} \cdots \sum_{i_2=0}^{s_2} \sum_{i_1=0}^{s_1+1} i_1 P(i_1, i_2, \cdots, i_n, 1) + \sum_{i_n=0}^{s_n} \cdots \sum_{i_2=0}^{s_2+1} \sum_{i_1=0}^{s_1} i_1 P(i_1, i_2, \cdots, i_n, 2) + \cdots + \sum_{i_n=0}^{s_n+1} \cdots \sum_{i_2=0}^{s_2} \sum_{i_1=0}^{s_1} i_1 P(i_1, i_2, \cdots, i_n, n)$$

$$E[Q_1] = \sum_{i_n=0}^{s_n} \cdots \sum_{i_2=0}^{s_2} \sum_{i_1=2}^{s_1+1} (i_1 - 1) P(i_1, i_2, \cdots, i_n, 1) + \sum_{i_n=0}^{s_n} \cdots \sum_{i_2=0}^{s_2+1} \sum_{i_1=1}^{s_1} i_1 P(i_1, i_2, \cdots, i_n, 2) + \cdots + \sum_{i_n=0}^{s_n+1} \cdots \sum_{i_2=0}^{s_2} \sum_{i_1=1}^{s_1} i_1 P(i_1, i_2, \cdots, i_n, n) \quad (7)$$

The waiting probability can be calculated using equation 8, which sums up all state probabilities, where a queue is neither empty nor full and blocking probability can be calculated by summing up all the state probabilities where a queue is full as given in equation 9.

$$P_{w_1} = \sum_{i_n=0}^{s_n} \cdots \sum_{i_2=0}^{s_2} \sum_{i_1=1}^{s_1} P(i_1, i_2, \cdots, i_n, 1) + \sum_{i_n=0}^{s_n} \cdots \sum_{i_2=0}^{s_2+1} \sum_{i_1=0}^{s_1-1} P(i_1, i_2, \cdots, i_n, 2) + \cdots + \sum_{i_n=0}^{s_n+1} \cdots \sum_{i_2=0}^{s_2} \sum_{i_1=0}^{s_1-1} P(i_1, i_2, \cdots, i_n, n) \quad (8)$$

$$P_{b_1} = \sum_{i_n=0}^{s_n} \cdots \sum_{i_2=0}^{s_2} P(s_1 + 1, i_2, \cdots, i_n, 1) + \sum_{i_n=0}^{s_n} \cdots \sum_{i_2=0}^{s_2+1} P(s_1, i_2, \cdots, i_n, 2) + \cdots + \sum_{i_n=0}^{s_n+1} \cdots \sum_{i_2=0}^{s_2} P(s_1, i_2, \cdots, i_n, n) \quad (9)$$

3. Results and Discussion:

Various characteristics measures have been used to analyze the behaviour of presented system. Mean queue sizes, flow time, waiting and blocking probabilities have been studied using different arrival rates or total load to the system in this section. Additionally, the effect of arrival rates of lower priority queues on high priority queue has been demonstrated to show the validity of exhaustive priority scheduling. The model assumes ignorable switch over times as compared to service time. Maximum queue sizes for all the plots are same and equal to 10. For clarity and ease of understanding, all figures except Fig. 3 and Fig. 4 are plotted using three priority classes. All results have been plotted with simulations points on analytical curves. Fig. 3 and Fig. 4 are using a scheduling system with four priority classes. Fig. 3 shows the mean number of customer in all queues for varying arrival rates. It is clear that the highest priority queue has minimum number for all arrival rates. The mean number of customers in all queues grows slowly for low arrival rate or total load, but increases rapidly after a certain load for low priority queues which is different for different queues. However, the high priority queues continues to show the same steady behaviour. The low priority queues reaches to the saturation point, which is approximately equal to 1. The behaviour of exhaustive priority discipline is quite clear from the plot. The same conclusions can be drawn from Fig. 4. Here, all high priorities experience comparative blocking probabilities and low priorities approach to 1 much earlier for all different arrival rates. In Fig. 5. the mean flow time is plotted against varying arrival rates in all queues of the priority system. It is quite clear that low priorities face comparatively high flow time. The customer which are blocked are not included in the calculations of flow time. Fig. 6 depicts the same kind of behaviour where the difference is only that now customers blocked have been included in calculations. In Fig. 7, the mean flow time has been plotted against varying arrival rates. The arrival rate of the highest priority queue is fixed and we can easily observe that increase of arrival rate in lower priority queues does not much influence the flow time of highest priority queue. Next in Fig. 8 the mean waiting probability has been plotted against varying arrival rates in all queues. The same behaviour is observed as the mean flow time plot discussed earlier. The waiting probability increases and after a certain saturation point it tends to decrease for all queues. The decrease in lowest priority queue is most rapid, and the reason of that is the sharp increase in blocking probability which resultantly reduces the mean waiting time of waiting customers. Fig. 9 shows that only increasing the arrival rate in lower priority classes has negligible effect on highest priority which validates the exhaustive service scheduling again. In the last Fig. 10 switch over time has been plotted against blocking probabilities for a system of three queues with fixed arrival rate in all queues equal to 0.3. It can be seen that increasing the switch over time has very little effect on the performance measures. Definitely this can only be assumed if the chosen switch over rate is much higher than service rate.

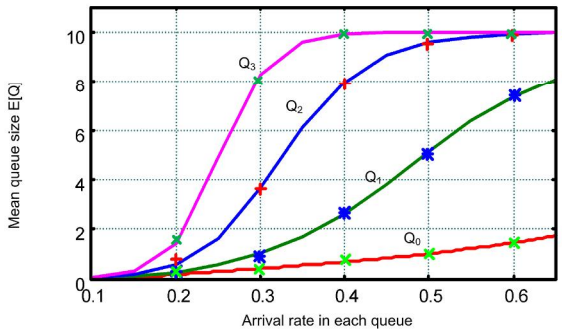

Fig. 3. Effect of varying all arrival rates on queue sizes with maximum queue capacity = 10 for a four-queue system.

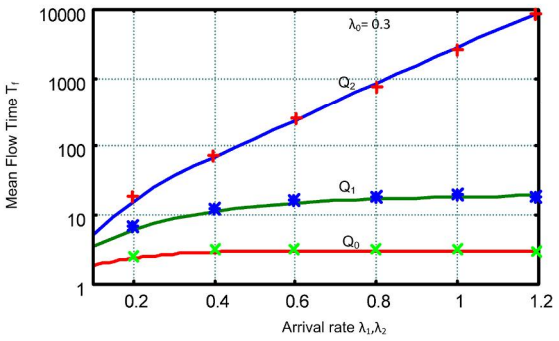

Fig. 7. Varying only λ_1; λ_2, mean flow time $E[T_f]$ of all queues with maximum queue capacity = 10 for a three-queue system. Blocked customers are not taken.

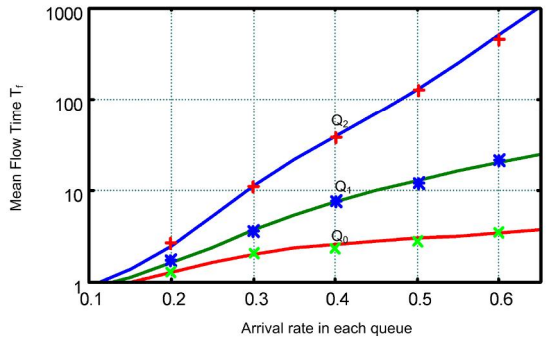

Fig. 5. Mean flow time E[Tf] with maximum queue capacity = 10 for a three-queue system vs arrival rate. all rates are varied simultaneously and blocked customers are not taken for mean flow time calculations.

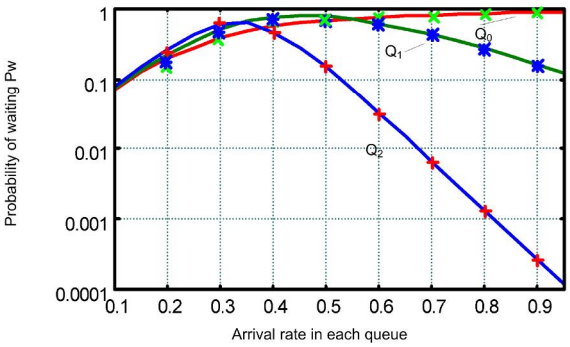

Fig. 8. Waiting probability Pw vs arrival rates where all rates are varied simultaneously with s = 10. Blocked customers are included in calculations.

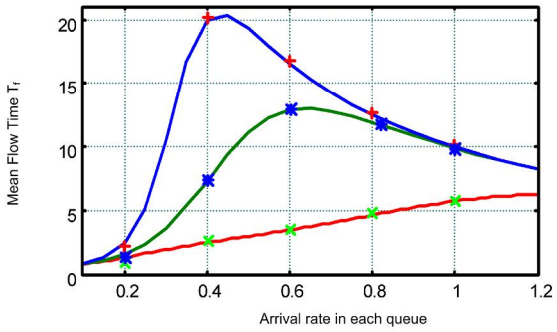

Fig. 6. Mean flow time E[Tf] vs arrival rates, where s = 10 for a three-queue system. Blocked customers are considered for mean flow
time.

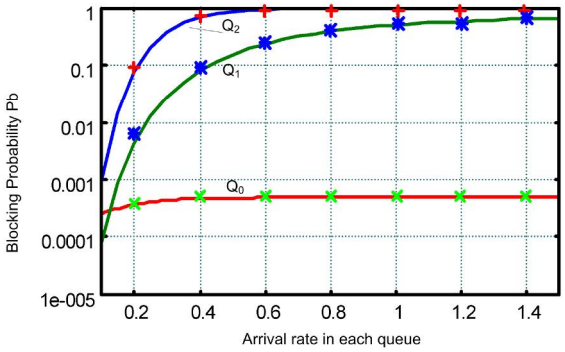

Fig. 9. Varying λ_1; λ_2, for Arrival rates vs blocking probability P_b s = 10 and system is a three-queue system.

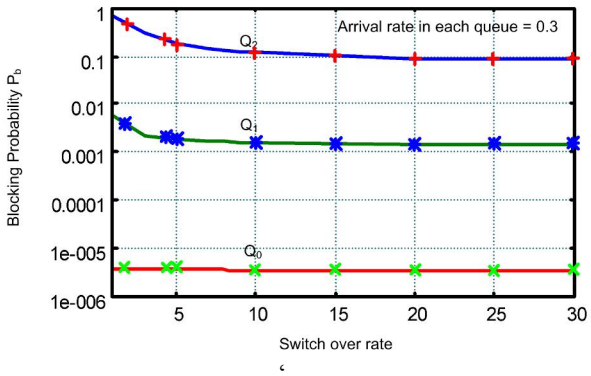

Fig. 10. Switch over time vs blocking probability P_b with maximum queue capacity = 10 and $\lambda_i = 0:3$; i = 0; 1; 2

4. Conclusion:

In this paper, a Markov chain model is introduced to model an exhaustive priority scheduling system where the involved buffers are limited and switch over time is only considerable when it happens with empty queues. The Markov chain is extended to a system with few queues but it is computationally not feasible to extend the analysis for very large number of queues. However, the analysis presented a real insight into the system behavior under consideration.

5. References

[1] K. Aziz, van As Harmen R., and S. Sarwar. Performance of Non-exhaustive Cyclic Service Systems with Finite Queues and Non-zero Switchover Times. In INMIC 2007. IEEE International Multitopic Conference, Pakistan, 2007.

[2] H. Bruneel and B. G. Kim. Discrete-time Models for Communication Systems Including ATM. Boston: Kluwer, 1993.

[3] W. Bux and H. L. Truong. Mean-delay Approximation for Cyclic-Service Queueing Systems. Performance Evaluation, 3(3):187–196, 1983.

[4] H. U. Chung and W. Jung. Performance Analysis of Markovian Polling Systems with Single Buffers. Performance Evaluation, 19(4):303–315, 1994.

[5] R. B. Coooper and G. Murray. Queues Served in Cyclic Order. The Bell Systems *Technical Journal, 48:675–689, 1969.*

[6] R. B. Cooper. Queues Served in Cyclic Order : Waiting Times. The Bell Systems Technical Journal, 49:399–413, 1970.

[7] M. Eisenberg. Two Queues with Changeover Times. Operations Research, 19:386–401, 1971. [8] M. Eisenberg. Queues with Periodic Service and Changeover Times. Operations Research, 20:440/451, 1972.

[9] D. Grillo. Polling Mechanism Models in Communication Systems - Some Application Examples. Stochastic Analysis of Computer and Communication Systems, Amsterdam: North- Holland, pages 659–698, 1990.

[10] M. F. Hayat, F. Z. Khan, and A. Lezanska. Performance modelling of a priority scheduling system with exhaustive service, finite capacity and switchover. IEEE Symposium on Computers and Communications (ISCC), pages 964 – 966, June 2011.

[11] O. C. Ibe and K. S. Trivedi. Stochastic Petri Net Models of Polling Systems. IEEE Journal for Selected Areas in Communications, 8(9):1649–1657, 1990.

[12] D. Karvelas and A. P. Leon Garcia. Performance of integrated packet voice/data token-passing rings. IEEE Journal for Selected Areas in Communications, SAC-4, 6 (Sept.):823–832, 1986.

[13] P. J. Kuehn. Multiqueue Systems with Nonexhaustive Cyclic Service. The Bell Systems Technical Journal, 58(3):671–699, 1979.

[14] P. J. Kuehn. Performance of arq-protocols for hdx transmission in hierarchical polling systems. Perform. Eval. Journal, 1, 1 (Jan.):19–30, 1981.

[15] D.-S. Lee. A two-queue model with exhaustive and limited service disciplines. Stochastic Models, 12(2):285–305, 1996.

[16] M. A. Leibowitz. An Approximate Method for Treating a Class of Multiqueue Problems. IBM J. Res. Develop., 5:204–209, 1961.

[17] H. Levy and M. Sidi. Polling Systems: Applications, Modeling and Optimization. IEEE Transactions on Communications, 38(10):1750–1760, 1990.

[18] J. D. C. Little. A proof for the queuing formula: L= ë w. Operations Research, 9(3):383–387, 1961.

[19] C. Mack, T. Murphy, and N. Webb. The efficiency of N machines unidirectionally patrolled by one operativewhen walking times and repair times are constants. J. Roy. Statist. Soc., B 19:166–172, 1957.

[20] M. N. Magalhaes, D. C. McNickle, and M. C. B. Salles. Outputs from a Loss System with Two Stations and a Smart (Cyclic) Server. Investigacion Oper., 16(1-3):111–126, 1998. [21] D. R. Manfield. Analysis of a polling

system with priorities. In IEEE GlobeCom, San Diego, Calif., 1983.
[22] D. R. Manfield. Analysis of a priority polling system for twoway traffic. IEEE Trans. Commun., COM-33, 9 (Sept.):1001– 1006, 1985.
[23] D. Miorandi, A. Zanella, and G. Pierobon. Performance Evaluation of Bluetooth Polling Schemes: An Analytical Approach. ACM Mobile Networks Applications, 9(2):63–72, 2004.
[24] J. B. Nagle. On Packet Switches with Infinite Storage. IEEE Transactions on Communications, 35(4):435–438, 1987.
[25] H. Takagi. Queuing analysis of polling models. ACM Computing Surveys, 20(1):5–28, 1988.
[26] H. Takagi. Application of polling models to computer networks. Computer Networks and ISDN Systems, 22(3):193–211, 1991.
[27] H. Takagi. Queueing analysis of polling models: progress in 1990-1994. Frontiers in Queueing: Models and Applications in Science and Technology, CRC Press, Boca Raton, Florida, (Chapter 5):119–146, 1997.
[28] H. Takagi. Analysis and Applications of Polling Models. Performance Evaluation, LNCS-1769:423–442, 2000.
[29] T. Takine, Y. Takahashi, and T. Hasegawa. Performance Analysis of a Polling System with Single Buffers and its Application to Interconnected Networks. IEEE Journal on Selected Areas in Communication, SAC-4(6):802–812, 1986.
[30] T. Takine, Y. Takahashi, and T. Hasegawa. Analysis of a Buffer Relaxation Polling System with Single Buffers. Proceedings of the Seminar on Queuing Theory and its Applications, May 11- 13, Kyoto Univ., Kyoto, Japan, pages 117–132, 1987.
[31] T. Takine, Y. Takahashi, and T. Hasegawa. Modelling and Analysis of a Single-buffer Polling System Interconnected with External Networks. INFOR., 28(3):166–177, 1990.
[32] I. M. Titenko. On Cyclically Served Multi-Channel Systems with Losses. Avtom. Telemekh., 10(10):88–95, 1984.
[33] P. Tran-Gia and T. Raith. Approximation for Finite Capacity Multiqueue Systems. Conf. on Measurement, Modelling and Evaluation of Computer Systems, Dortmund, Germany, 1985. [34] P. Tran-Gia and T. Raith. Multiqueue Systems with Finite Capacity and Nonexhaustive Cyclic Service. International Seminar on Computer Networking and Performance Evaluation, Tokyo, Japan, 1985. [35] V. M. Vishnevskii and O. V. Semenova. Mathematical models to study the polling systems. Automation and Remote Control, 67(2):173–220, 2006.
[36] V. M. Vishnevsky, A. I. Lyakhov, and N. N. Guzakov. An adaptive polling strategy for ieee 802.11 pcf. Proceedings of 7th International Symposium on Wireless Personal Multimedia Communications, Abano Terme, Italy, 1:87–91, 2004.
[37] F. Y., T. Yanaraj, and S. Yoshida. An approximate analysis for a multiqueue with a non-preemptive priority and cyclic service. Trans. Inst. Electron. Znf. Commun. Eng., J70-A, 9 (Sept.), (in Japanese).:1351–1354, 1987. [38] E. Ziouva and T. Antonakopoulos. Improved IEEE 802.11 PCF Performance Using Silence Detection and Cyclic Shift on Stations Polling. IEE Proceedings Communications, 150(1):45– 51, 2003.

8/11/2012