Int j simul model 21 (2022) 2, 273-283
Original scientific paper

LOAD BALANCING IN POLLING SYSTEMS UNDER DIFFERENT POLICIES VIA SIMULATION OPTIMIZATION

Uncu, N.
Department of Industrial Engineering, Adana Alparslan Turkes Science and Technology University, Adana, Turkey
E-Mail: nuncu@atu.edu.tr

Abstract
A polling system is a type of queueing system in which a single server serves many queues. In contrast to classical queueing systems, the server switches between queues and serves the customers under different routing and service policies. The problem addressed in this study is to define the load balancing problem in polling systems, where the balancing behaviour is dependent on server shifts, not the distribution of customers among queues. Although the load balancing problem in multi-server queuing systems is very common, modelling the load balancing problem in polling systems is the novelty of this study. Furthermore, not only the performance of the system has been analysed as in the previous simulation studies but also the balanced queues with optimal routing probabilities for multi-class queues under different routing and service policies are achieved by using Arena-OptQuest.
(Received in February 2022, accepted in April 2022. This paper was with the author 2 weeks for 2 revisions.)

Keywords: Multi-Class Queues, Polling Systems, Routing, Simulation

1. INTRODUCTION

There are numerous types of queueing systems, which are classed based on the number of queues and servers, customer behaviour, and service policies, among other factors. Polling system is one of the queueing systems which have multiple queues with one server switching between the queues under different routing and service policies. The customers in the system are in competition for accessing to a common single server. Computer communication, transportation, manufacturing, robotics, and road traffic networks are among the most common uses of polling systems. The load balancing problem arises in such systems that aim to distribute traffic between multiple servers in the queueing system. This problem is similar to the line balancing problem in manufacturing systems. However, it is modelled differently in polling systems due to the single server and the server's ability to shift between queues.

In today's world, where the population and data volume are increasing rapidly, the use of scarce resources in both physical and virtual systems will be a problem to be tackled. Therefore, the importance of well-structured polling systems that serve many customers with a single server will increase. The policies used in polling systems have a significant impact on the system's performance. The routing policies describes the order of queues that server visits one from another and are named by static (cyclic/periodic), probabilistic (random) and dynamic. Static routing policies are predetermined and do not change throughout the process. In probabilistic models, the server routes randomly, while in dynamic routing policies algorithms are applied to route the server based on the current state of the system. Another important point in polling systems is the behaviour of the server that is determined by the service policies. In general, service policies are divided into two categories: server-driven policies and customer-driven policies. The difference is that customer-related rules influence service decisions in some systems, whereas server time or other factors can influence service policy decisions in others. These can be categorized mainly into three types: 1) limited, 2) gated, 3) exhaustive. These service policies are extended to more than these main types in [1] and [2]. In limited policy sometimes called *l*-limited or *k*-limited policy, the queue is served for a fixed number of entity or a specified time. In a gated service policy, the number of entities in the queue at the instant when the server arrived is served, but next arrivals during the service must wait until the next visit of the server. In an exhaustive service, the queue is served until it becomes empty. The time of visit is called as the vacation time, while the time to switch between queues is switchover time.

Although the concept of polling systems dates back to a work by [3], a particularly noteworthy study on the fundamental concepts of polling systems and analytical inferences with some modelling challenges can be found in [4]. Since then, some analytical and theoretical studies have been taken place in the literature. A number of numerical examples are given in [5] for determining the effect of customer priorities on the performance of two-queue polling systems under mixed gated/exhaustive service policies. In [6], the discrete time analysis of gated time-limited service policy for MAP/PH/1 queue is given. The key performances in the polling systems are mean waiting times in each queue, overall waiting time or sometimes the fairness between the queues are other performance outputs as emphasized in [7]. In terms of analytical models of polling systems, it has also been emphasized that some necessary conditions should be considered regarding the behaviour modelling of *k*-limited polling systems under heavy-traffic [8].

A mean waiting time approximation method is proposed for cyclic polling models with three [9] and four queues in [10]. They compared the time-limited service policy with dynamic routing rule that routes the server to the most loaded queue. The method allows explicit minimization of the weighted sum of the mean waiting times with respect to visiting frequencies of the queues in the routing table. In [11], the Markov chain of nonhomogeneous finite-source queueing model was proposed to describe the performance of a multi-terminal system subject to random breakdowns under different polling service policies.

The rest of the studies are mainly on the applications of polling systems in different areas. An ATM network system with priorities for two types of customers are studied and analytical solutions are given in [12]. Two queueing models with state dependent setups are suggested and applied in logistics [13]. They considered one queue with high priority and exhaustively served, whereas the other as low priority queue is served according to *k*-limited strategy. Two echelon supply chain system with simulation is applied in [14]. They compared the results based on minimum lead times of cyclic routing under different service policies. An attractive application area of polling systems is fluid systems in [15]. They analysed the performances of the polling system with cyclic and probabilistic policies. An application of polling scheme in the computer systems provides an efficient approach that acts as time-limited policy to minimize the CPU time and maximize the I/O performance for ultra-low latency storage devices in [16]. An adaptive scheme of polling for cyclic and gated service systems are examined by generating function method in [17]. A continuous system case that a fluid switched over two queues for random time-limited policy is studied in [18]. A recent study of polling system belongs to [19] analysed the performance of a tandem queue modelled as a polling system in an aluminium manufacturing process. The policies for service and routing of the server have been applied in different systems. Applications in appointment driven systems and landside operations of container ports can be found in [20] and [21].

The optimization of polling systems with exact methods has been widely studied in the literature. Although simulation modelling of such systems is capable of giving the approximate performance evaluations, there is a few works on simulation models. One of the salient studies belongs to [22]. They applied the exact methods and simulation to obtain performance measures of polling system operating under exhaustive, gated and mixed service policies without considering routing policies. The simulation of polling systems with time-limited service with random and cyclic routing policies are given in [23]. Most of the previous studies have attended on the performance of the polling systems rather than the optimal policies. For more on applications of polling systems, [24] can be investigated. An actual review of polling systems can be found in [25].

Most of the previous studies focus on the performance analysis of the polling systems in terms of mean waiting time minimization, in this study load balancing between queues in terms of average waiting times are considered to satisfy all the customers waiting for the service in different queues. The polling systems with three and four queues studied in [10] are extended to gated and exhaustive service policy. Optimal routes are obtained by using *OptQuest* which is an optimization tool embedded in Arena. The contribution of this study is to provide the simulation of polling systems under different service and routing policies while balancing average waiting times in the queues. The results of this study are evaluated using the routing tables of polling models in [9] and [10].

2. PROBLEM DEFINITION

Polling systems have an N-queue structure, with only one server serving all of them based on varied routing and service policies. The arrival rate of queue i (i = 1, 2, ..., N) is λ_i and the service rate of each queue is μ_i (i = 1, 2, ..., N). The switch over time between the queues is d_{ij} (i = 1, 2, ..., N, j = 1, 2, ..., N) the time spent by the server moving to the queue i from the queue j. An illustration of polling systems is represented in Fig. 1, for three-queues and four-queues, respectively.

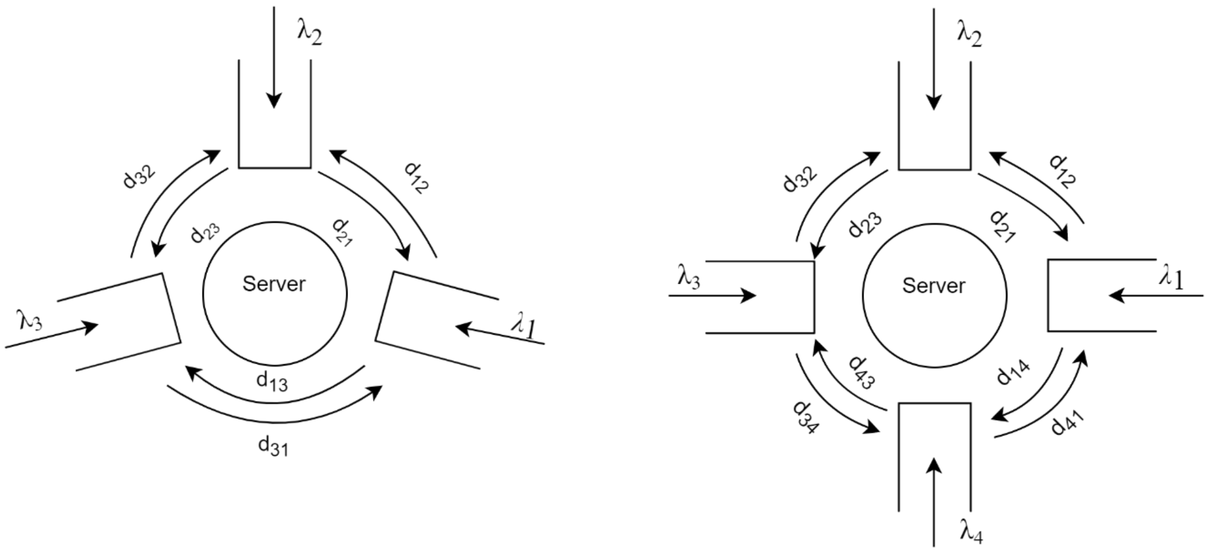

Figure 1: Illustration of polling systems.

Customers and queues are expected to be identical in this study since the polling mechanism is assumed to be homogeneous. Customers, on the other hand, arrive at various rates in each queue. The standard deviation is used to minimize the difference between the expected average waiting time in each line and the overall expected average waiting time in order to balance the average waiting time of queues. The objective function is given as:

$$\text{Min } z=\sqrt{\left(\sum_{i=1}^{N}\left(E[\overline{W}_{qi}]-E[\overline{W}_{q}]\right)^2\right)/N} \tag{1}$$

where the expected average waiting time at queue i and the overall expected average waiting time are given respectively, $E[\overline{W}_{qi}]$ and $E[\overline{W}_q]$.

The server is assumed to move between queues under different routing and service policies within a constant switch-over time. The related policies and parameters for three and four queues polling systems are given in Table I and Table II, respectively.

Table I: The simulation parameters of models with three queues.

Models	Routing policy	Service policy	Common parameters	
Model 3Q-1 [a, b]	Cyclic	Time-Limited	Service Rate	Arrival Rates
Model 3Q-2	Probabilistic	Time-Limited	$\mu = 1$	$\lambda_1 = 0.54$
Model 3Q-3 [a, b]	Cyclic	Gated	Switch-over Time	$\lambda_2 = 0.24$
Model 3Q-4	Probabilistic	Gated	$d = 1$ min	$\lambda_3 = 0.06$
Model 3Q-5 [a, b]	Cyclic	Exhaustive		
Model 3Q-6	Probabilistic	Exhaustive		

[a, b] Routes from studies in [9] and [10], respectively.

Table II: The simulation parameters of models with four queues.

Models	Routing policy	Service policy	Common parameters	
Model 4Q-1 [c]	Cyclic	Time-Limited	Service Rate	Arrival Rates
Model 4Q-2	Probabilistic	Time-Limited	$\mu = 1$	$\lambda_1, \lambda_3 = 0.02$
Model 4Q-3 [c]	Cyclic	Gated	Switch-over Time	$\lambda_2, \lambda_4 = 0.005$
Model 4Q-4	Probabilistic	Gated	$d = 1$ min	
Model 4Q-5 [c]	Cyclic	Exhaustive		
Model 4Q-6	Probabilistic	Exhaustive		

[c] Route from study in [10].

The simulation models of polling systems are created by Arena software. It is challenging due to the logic of Arena depends on the entity movements. However, in polling models the server moves among the queues. Hence, some extra blocks are used to overcome this drawback. The block models are only given for the three-queue models as an instance for simulation models.

2.1 Time-limited and gated polling models

The time-limited and gated policy can be modelled in a similar manner in terms of service policy. In the time-limited, it is assumed that the server visits each queue for 5000 minutes, while in the gated policy the server ends its visit after processing the last part that is the one when the server enters the queue. The Arena blocks for both policies are illustrated in Fig. 2. However, additional HOLD blocks are used in the gated system after the *RouteforTx* to prevent the server processing the parts that are not in the system when the server enters that queue.

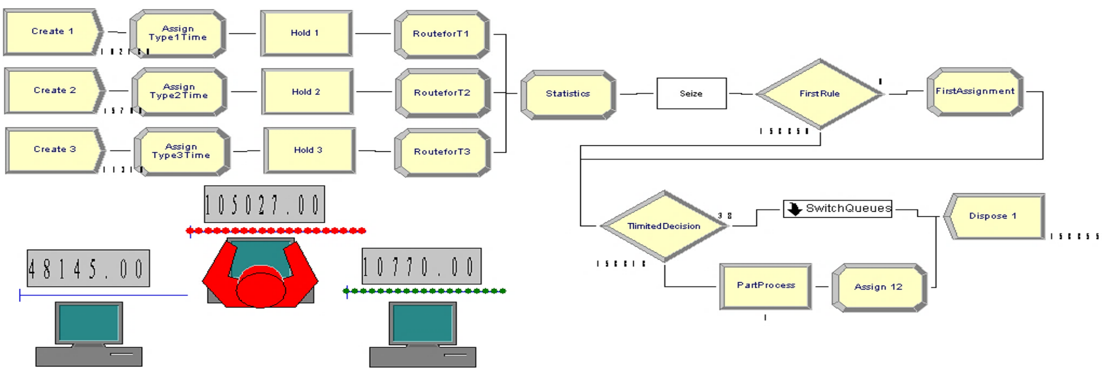

Figure 2: Model for time-limited and gated policy.

2.2 Exhaustive polling model

The first group of modules that creates the parts includes CREATE-ASSIGN-Signal-HOLD blocks and is shown in Fig. 3. This group creates and holds the customers in each queue waiting the idle signal to start the next visit of the server.

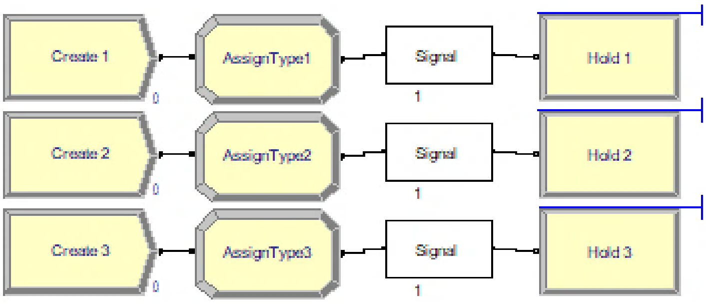

Figure 3: The group of blocks for entity generation and service policy design.

In Fig. 4, the SEQUENCE module with STATION-ROUTE blocks is used to provide the route for the server. PICKUP-SEARCH- DROPOFF blocks enable the entities to be served by server by sending them to PROCESS block when the server switches to the corresponding queue. For only optimization model, DECIDE blocks are added for routing the server based on a probability of choosing the next queue to visit. Throughout the simulation and optimization, the optimal values of probabilities are obtained as a result of minimum deviation from the expected average waiting times at each queue.

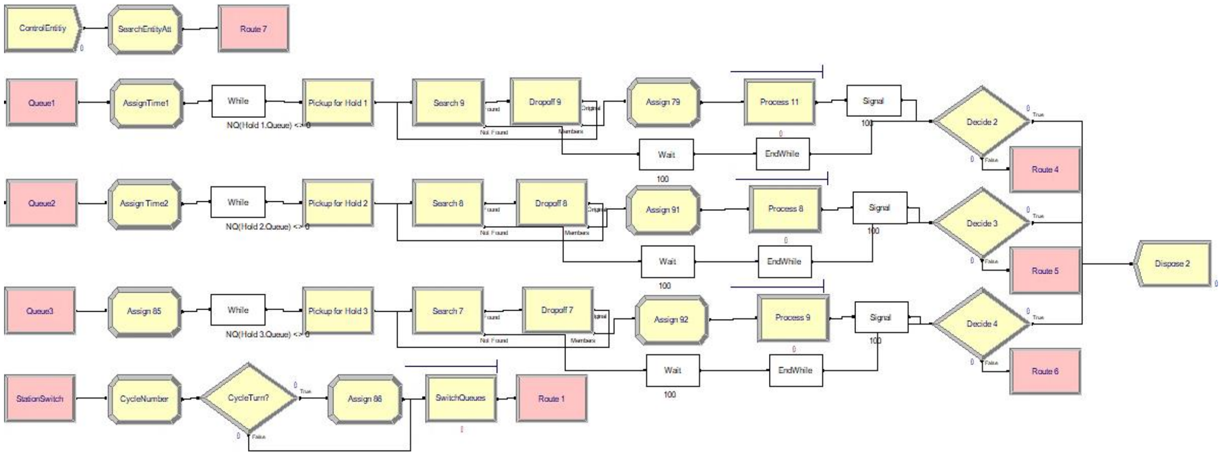

Figure 4: Stations and blocks in the exhaustive model.

3. RESULTS

In order to achieve 95 % confidence level, the simulation is run for 10 replications and one year period with a warm up period of one month. The static routings are taken from the study of [9] and [10]. The sequence of visits of server to each queue are set up by random probabilities changing through optimization process. In the time limited, the server switches to another queue based on the probabilities assigned in each optimization run. In the model for the gated policy, the server visits randomly to each queue by starting with initial probabilities of visiting each queue and serves the queue down to the last customer who is the one when the server arrives that queue.

The results for three-queue polling models with different policies are given in Table III. The coefficient of variation (*CV)* that allows to rank the dispersion level is calculated for each model. For each of the service policy, the probabilistic routing suggests the best balance between the queues visited by the server while compared with other given cyclic/static routes based on the *CV*. Although, the average waiting times in time-limited model are higher than the other models, the lowest *CV* indicates that the best balance is obtained in time-limited with the probabilistic routing policy. Additionally, the expected waiting times and the overall expected waiting time in the system are the lowest in exhaustive model among all the routing policies.

Table III: Simulation results of three-queue models for different policies.

Models	**Service policy**	**Route**	**Average waiting times at queues** (min)				**Std. dev.**	***CV***	**Rank**
			$\overline{W}_{q1}$	$\overline{W}_{q2}$	$\overline{W}_{q3}$	$\overline{W}_{q}$			
3Q-1	Time-limited	121213	1245.67	4062.76	10385.76	5231.40	3821.83	0.73	8
		12123121231213	2115.21	3318.13	7375.13	4269.49	2250.26	0.53	6
3Q-2		Probabilistic	3595.38	3477.81	3591.94	3555.04	261.70	0.07	1
3Q-3	Gated	121213	10.37	13.10	81.74	35.07	33.02	0.94	9
		12123121231213	15.25	15.22	22.87	17.78	3.60	0.20	3
3Q-4		Probabilistic	16.29	23.28	20.34	19.97	2.87	0.14	2
3Q-5	Exhaustive	121213	6.75	16.12	34.65	19.17	11.59	0.60	7
		12123121231213	7.24	13.87	25.89	15.67	7.72	0.49	5
3Q-6		Probabilistic	9.43	15.28	17.12	13.94	4.02	0.29	4

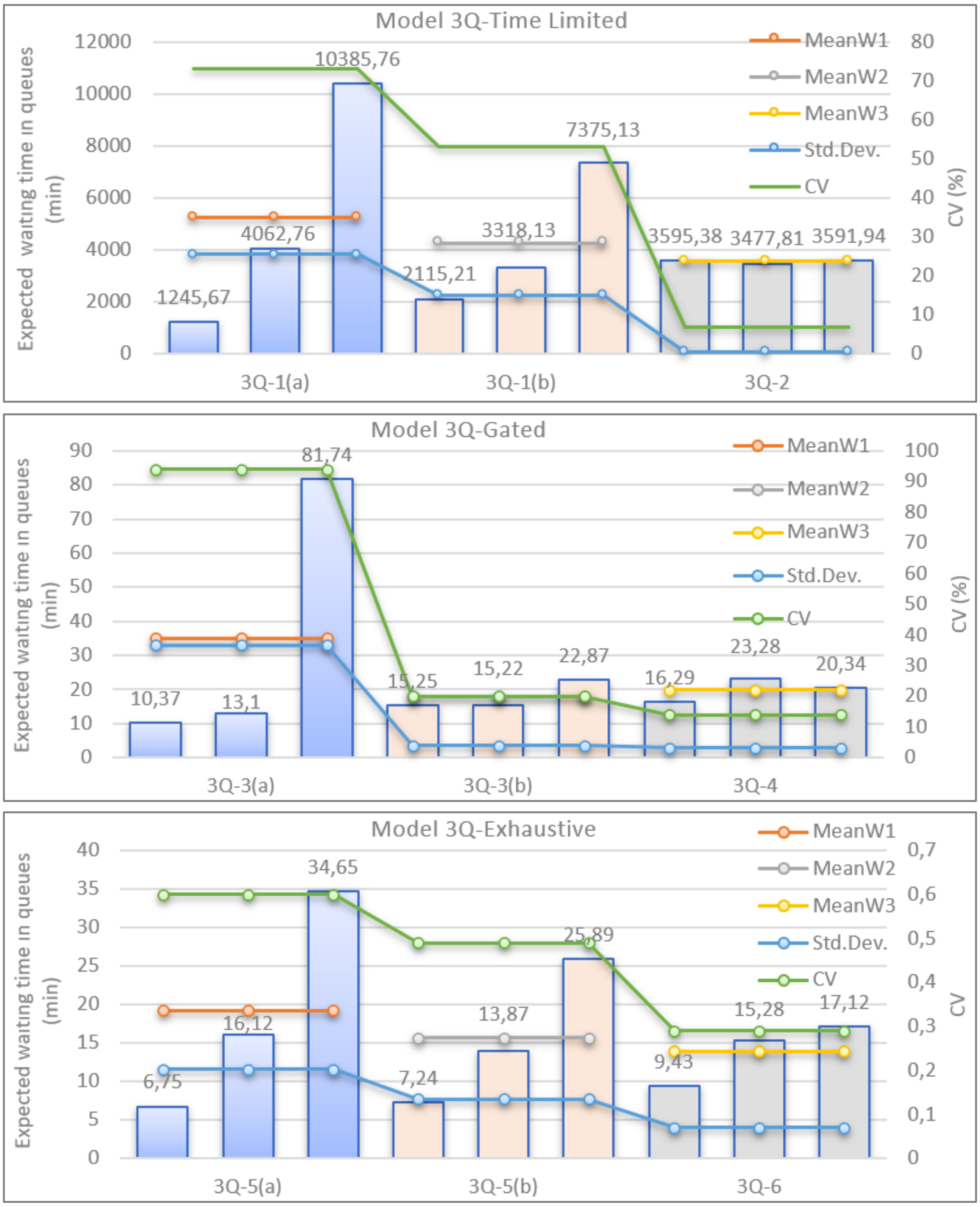

Figure 5: Comparison of the expected waiting times for three-queue polling models.

The comparison of the expected waiting times based on the routing policies for each model is illustrated in Fig. 5. It is obvious that there is a high dispersion between the expected waiting times of queues in the refereed studies in which the routing tables were obtained by approximation methods.

Similar in three-queue model, the best balance is achieved in the probabilistic model for each service policy. The first one is the time-limited policy and it is followed by the gated policy based on their *CV* ranking. However, the lowest average waiting times and overall mean are achieved by the exhaustive policy. The results are illustrated and compared in Fig. 6.

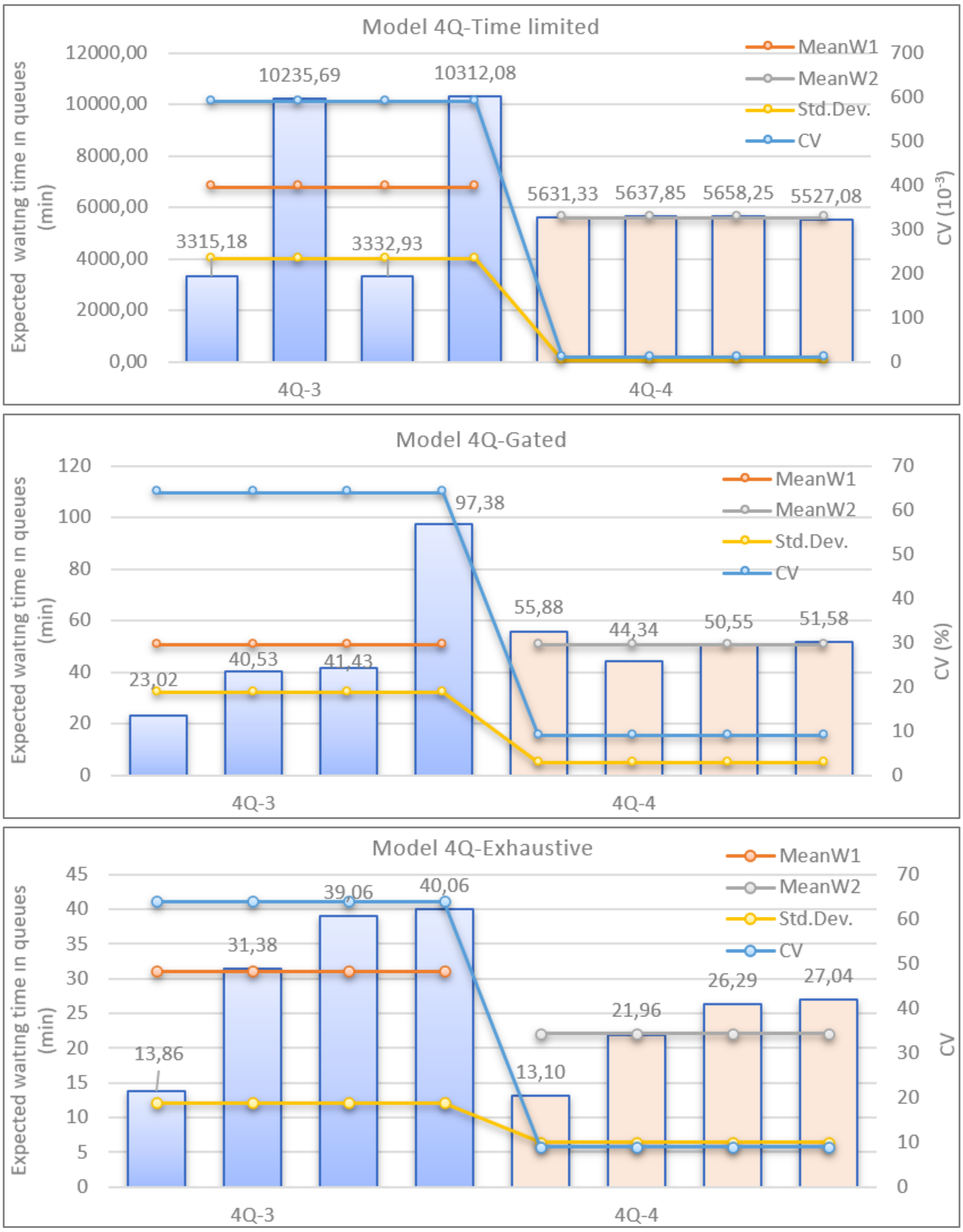

Figure 6: Comparison of the expected waiting times for four-queue polling models.

Table IV: Simulation results of four-queue models for different policies.

Models	Service policy	Route	Average waiting times at queues (min)					Std. dev.	CV	Rank
			$\overline{W}_{q1}$	$\overline{W}_{q2}$	$\overline{W}_{q3}$	$\overline{W}_{q4}$	$\overline{W}_{q}$			
4Q-1	Time-limited	123143	3315.18	10235.69	3332.93	10312.08	6798.97	4012.61	0.59	5
4Q-2		Prob.	5631.33	5637.85	5658.25	5527.08	5613.63	50.95	0.01	1
4Q-3	Gated	123143	23.02	40.53	41.43	97.38	50.59	27.99	0.55	6
4Q-4		Prob.	55.88	44.34	50.55	51.58	50.59	4.12	0.08	2
4Q-5	Exhaustive	123143	13.86	31.38	39.06	40.06	31.09	12.12	0.39	4
4Q-6		Prob.	13.10	21.96	26.29	27.04	22.10	6.40	0.29	3

In optimization phase, the probabilities are randomly assigned to the pairs of switches initially. As a result of the simulation optimization, the optimal routing probabilities of each queue visited by the server are obtained and given in Table V and Table VI for three and four queue models, respectively. The total probability of each pair of switches of the server from one of all the queues is hundred percent. The probabilities are the percent of service time that the server switches between corresponding pair of queues. For instance, 100 % percent between queue 2 and queue 1 in Model 3Q-2 means that the server switches from queue 2 to queue 1 without switching from queue 1 to queue 2 in any time of its service. It visits queue 1 only from queue 3 with 100 % of its time. In terms of optimal routing probabilities, the explanation is similar for all other pairs of queues.

Table V: Routing probabilities for model 3Q-2/4/6.

From queue	To queue	Routing probabilities		
		Model 3Q-2 Time-limited	Model 3Q-4 Gated	Model 3Q-6 Exhaustive
1	2	63.29	87.44	1.50
1	3	36.71	12.26	99.5
2	1	100	35.13	76.43
2	3	0	64.87	23.57
3	1	100	100	2
3	2	0	0	98

Table VI: Routing probabilities for Model 4Q-2/4/6.

From queue	To queue	Routing probabilities		
		Model 4Q-2 Time-limited	Model 4Q-4 Gated	Model 4Q-6 Exhaustive
1	2	0	60.63	0
1	3	0	39.37	100
1	4	100	0	0
2	1	100	0	96.01
2	3	0	54.31	3.96
2	4	0	45.69	0.03
3	1	0	53.04	0
3	2	100	0	100
3	4	0	46.96	0
4	1	0	52.56	0
4	2	0	47.44	100
4	3	100	0	0

Based on the optimal probabilities, the trade-off between the number of switches and expected average waiting times for three-queue polling model simulation results is illustrated in Fig. 7 as an instance. The results for the time-limited policy are not depicted in the figure because a fixed number of switches as 105 times occur in the defined simulation period. The graph indicates no significant relationship between average waiting times and the number of switches for different policies. However, it can be concluded that in the time-limited policy the number of switches is constant and less than other policies. If there is a noticeable effect of the number of switches to the cost in the system, then the time-limited policy can be preferable to others.

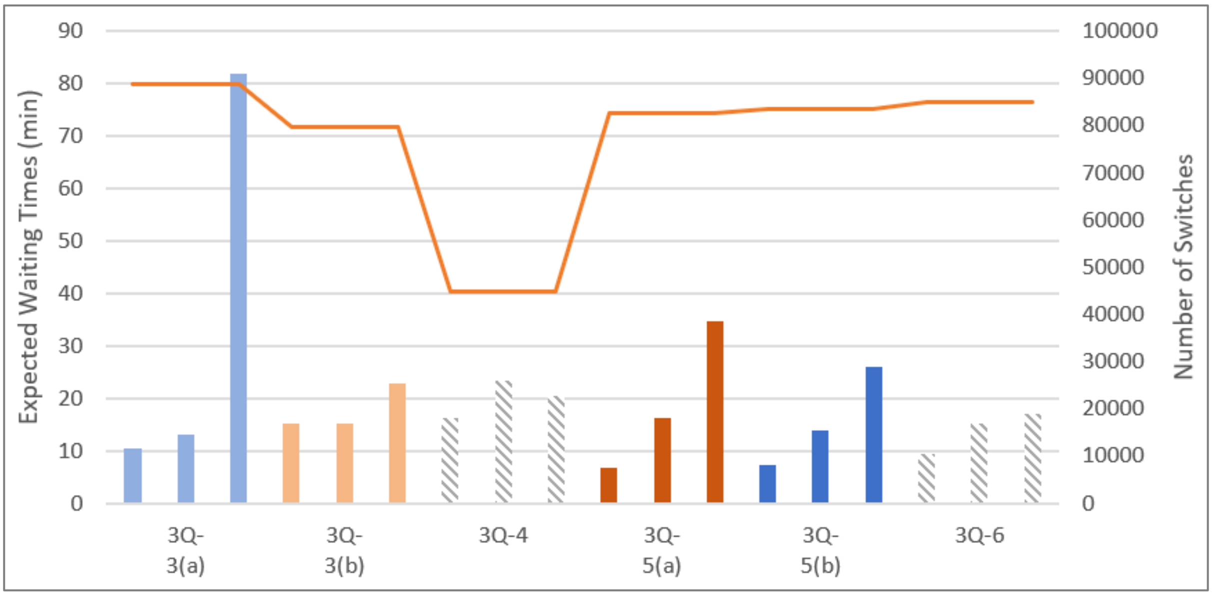

Figure 7: Relation between the expected waiting times and the number of switches.

4. CONCLUSION

Despite the fact that polling systems have been studied since the 1950s, there is still a large study and application area that needs to be investigated and analysed through simulation due to the applicability of various routing and service policies that result in varying performance levels. The decision of using such policies should be evaluated before design of the system or new algorithms and online searching algorithms may be embedded in the real-world systems. In most of the previous studies, static or predetermined routing or service policies are widely analysed by exact and approximation methods. They mainly focus on performance analysis or mean waiting time minimization. In this study, the load balancing problem in queueing systems are modelled for polling systems. Time-limited, gated and exhaustive service policies with predefined routing tables and the probabilistic routing policy for three-queue and four-queue polling systems are simulated, and optimized to balance the waiting times between the queues while improving the system responsiveness. The probabilistic routing policy for each service policy resulted as the best among other proposed models considering minimum dispersion of average waiting times between queues. Another analysis is conducted for determining the relationship between the number of switches and average waiting times for three-queue models only as an instance. However, there is no noticeable relationship. But, if the number of switches is not negligible, the time-limited policy is better than the gated or exhaustive policies which result in higher waiting times in the system. If the gated or exhaustive policy is preferred due to the low waiting times, there should be negligible amount of cost for switchover time.

For further research, the load balancing problem in polling systems can also be applied to appointment-driven queueing systems, fluid systems, wireless systems, and many forms of server routing problems. Other service policies, such as dynamic, priority-based, and decrementing or mixed policies, can be analysed using the simulation models for polling systems.

REFERENCES

[1] Bishoff, W. (2001). Analysis of M/G/1-queues with setup times and vacations under six different service disciplines, *Queueing Systems*, Vol. 39, No. 4, 265-301, doi:10.1023/A:1013992708103

[2] Vishnevskii, V. M.; Semenova, O. V. (2006). Mathematical methods to study the polling systems, *Automation and Remote Control*, Vol. 67, No. 2, 173-220, doi:10.1134/S0005117906020019

[3] Mack, C. (1957). The efficiency of *N* machines uni-directionally patrolled by one operative when walking time is constant and repair times are variable, *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 19, No. 1, 173-178, doi:10.1111/j.2517-6161.1957.tb00254.x

[4] Takagi, H. (1986). *Analysis of Polling Systems*, MIT Press, Cambridge

[5] Boon, M. A. A.; Adan, I. J. B. F. (2009). Mixed gated/exhaustive service in a polling model with priorities, *Queueing Systems*, Vol. 63, No. 1, 383-399, doi:10.1007/s11134-009-9115-z

[6] Alfa, A. S. (1998). Discrete time analysis of MAP/PH/1 vacation queue with gated time-limited service, *Queueing Systems*, Vol. 29, No. 1, 35-54, doi:10.1023/A:1019123828374

[7] Boxma, O. J.; Levy, H.; Yechiali, U. (1992). Cyclic reservation schemes for efficient operation of multiple-queue single-server systems, *Annals of Operations Research*, Vol. 35, No. 3, 187-208, doi:10.1007/BF02188704

[8] Suman, R.; Krishnamurthy, A. (2022). Analysis of two-station polling queues with setups using continuous time Markov chain, *arXiv*, Paper 2202.10045, 27 pages, doi:10.48550/arXiv.2202.10045

[9] Boxma, O. J.; Levy, H.; Weststrate, J. A. (1990). Optimization of polling systems, *Proceedings of the 14th IFIP WG7.3 International Symposium on Computer Performance Modelling, Measurement and Evaluation*, 341-361

[10] Bertsimas, D.; Xu, H. (1993). *Optimization of Polling Systems and Dynamic Vehicle Routing Problems on Networks*, Massachusetts Institute of Technology, Cambridge

[11] Almási, B. (1999). A queuing model for a non-homogeneous polling system subject to breakdowns, *Annales Universitatis Scientiarum Budapestinensis de Rolando Eotvos Nominatae: Sectio Computatorica*, Vol. 18, 11-23

[12] Avram, F.; Gómez-Corral, A. (2006). On the optimal control of a two-queue polling model, *Operation Research Letters*, Vol. 34, No. 3, 339-348, doi:10.1016/j.orl.2005.05.005

[13] Winands, E. M. M.; Adan, I. J. B. F.; van Houtum, G. J.; Down, D. G. (2009). A state-dependent polling model with *k*-limited service, *Probability in the Engineering and Informational Sciences*, Vol. 23, No. 2, 385-408, doi:10.1017/S0269964809000217

[14] Kamath, B. N.; Bhattacharya, S. (2007). Lead time minimization of a multi-product, single-processor system: a comparison of cyclic policies, *International Journal of Production Economics*, Vol. 106, No. 1, 28-40, doi:10.1016/j.ijpe.2006.04.002

[15] Czerniak, O.; Yechiali, U. (2009). Fluid polling systems, *Queueing Systems*, Vol. 63, No. 1, 401-435, doi:10.1007/s11134-009-9129-6

[16] Lee, G.; Shin, S.; Jeong, J. (2022). Efficient hybrid polling for ultra-low latency storage devices, *Journal of Systems Architecture*, Vol. 122, Paper 102338, 12 pages, doi:10.1016/j.sysarc.2021.102338

[17] Semenova, O. V.; Bui, D. T. (2018). Method of generating functions for performance characteristic analysis of the polling systems with adaptive polling and gated service, Dudin, A.; Nazarov, A.; Moiseev, A. (Eds.), *Information Technologies and Mathematical Modelling, Queueing Theory and Applications*, Springer, Cham, 348-359, doi:10.1007/978-3-319-97595-5_27

[18] Kapodistria, S.; Saxena, M.; Boxma, O. J.; Kella, O. (2021). Workload analysis of a two-queue fluid polling model, *arXiv*, Paper 2112.04819, 27 pages, doi:10.48550/arXiv.2112.04819

[19] Boon, M. A. A.; Winands, E. M. M. (2022). Open problems for critically loaded *k*-limited polling systems, *Queueing Systems*, Published online, 3 pages, doi:10.1007/s11134-022-09770-x
[20] Creemers, S.; Lambrecht, M. (2010). Queueing models for appointment-driven systems, *Annals of Operations Research*, Vol. 178, No. 1, 155-172, doi:10.1007/s10479-009-0646-9
[21] Deniz, E.; Tuncel, G.; Yalçınkaya, Ö.; Esmer, S. (2021). Simulation of multi-crane single and dual cycling strategies in a container terminal, *International Journal of Simulation Modelling*, Vol. 20, No. 3, 465-476, doi:10.2507/IJSIMM20-3-559
[22] Azadeh, A.; Sheikhalishahi, M.; Yousefi, N. (2014). An efficient computer simulation-based approach for optimization of complex polling systems with general arrival distributions, *Simulation*, Vol. 90, No. 12, 1346-1359, doi:10.1177/0037549714556018
[23] Uncu, N. (2016). Simulation modelling and analysis of polling systems, *Proceedings of the 1st International Mediterranean Science and Engineering Congress*, 355-359
[24] Boon, M. A. A.; van der Mei, R. D.; Winands, E. M. M. (2011). Application of polling systems, *Surveys in Operations Research and Management Science*, Vol. 16, No. 2, 67-82, doi:10.1016/j.sorms.2011.01.001
[25] Vishnevsky, V.; Semenova, O. (2021). Polling systems and their application to telecommunication networks, *Mathematics*, Vol. 9, No. 2, Paper 117, 30 pages, doi:10.3390/math9020117