Performance Analysis of Delay and Size-Dependent Scheduling for IoT-based Healthcare Traffic using Heterogeneous Multi-Server Priority Queueing System

1st Barbara Kabwiga Asingwire
African Centre of Excellence in Internet of Things (ACEIoT), University of Rwanda
Department of Computer Engineering, Busitema University, Kampala, Uganda
kezabarbara@gmail.com

2nd Louis Sibomana
African Centre of Excellence in Internet of Things (ACEIoT), University of Rwanda
National Council for Science and Technology, Rwanda
Kigali, Rwanda
lewis.sis@gmail.com

3rd Alexander Ngenzi
ACEIoT, University of Rwanda
Kigali, Rwanda
yngenzi37@gmail.com

4th Charles Kabiri
ACEIoT, University of Rwanda
Kigali, Rwanda
chakabiri@googlemail.com

***Abstract*—Time is very crucial in Internet of Things (IoT)-based healthcare applications and any delay can lead to dangerous scenarios, including patient deaths. The Earliest Deadline First (EDF) scheduling mechanism has been recommended for use in IoT-based applications in healthcare systems. However, the EDF performs poorly under overloaded conditions since priority is given to high-priority packets with close deadlines. In order to overcome the limitation of EDF, models for the prioritized scheduling (PS) scheme was proposed. The PS scheme is an improvement of the EDF scheme for IoT-based healthcare applications. The PS scheme uses a heterogeneous multi-server priority queuing system to provide service differentiation by prioritizing short packets over large packets and delay sensitive packets are serviced before delay tolerant packets. In this paper, the analytical models of the PS scheduling scheme are validated using the Simulink MATLAB R2023b tool. The performance measures are obtained for the mean slowdown for delay sensitive and delay tolerant packets as a function of the sizes of packets. In addition, the performance measures in terms of throughput as a function of packet sizes for short and large delay sensitive packets are obtained. We compare the results obtained from simulations with the analytical results for mean slowdown and throughput. Simulation results.**

***Index Terms*—EDF, Heterogeneous, Mean slowdown, Prioritized Scheme, Size, Throughput**

I. Introduction

The Internet of Things (IoT) has been empowered by recent technological advancements to create a unified network of interlinked devices, sensors, and systems [1]. This capability has extended the scope of IoT's application in multiple domains, including remote healthcare monitoring, rendering it a dynamic and formidable technology [2].

The implementation of IoT technology in remote healthcare monitoring has several benefits compared to conventional healthcare monitoring approaches [3]. It is anticipated that IoT will advance emergency management and healthcare monitoring in the near future. In IoT-based healthcare monitoring, instantaneous and dependable delivery of collected data is crucial to ensure precise patient monitoring, given that healthcare applications rely on real-time data with minimum latency.

Medical emergencies necessitate priority treatment over regular services [3]. Moreover, the services offered for medical data packets should be distinguished based on signal demands. In emergency scenarios, low latency is crucial for healthcare traffic to enable prompt response by healthcare professionals [4], [5]. However, traditional server scheduling schemes used in computing servers are unsuitable for delivering services to IoT-based healthcare applications due to the heterogeneity of servers and varying service requirements [6]. Therefore, it is essential to enhance standard server scheduling algorithms by considering the servers' heterogeneity and different service requirements to meet users' expectations efficiently.

According to [7], there is an increase in delay for healthcare IoT packet transmission as the data size increases, with delays ranging from milliseconds to minutes for time-sensitive applications.

Several scheduling techniques for IoT-based healthcare monitoring systems have been proposed in recent studies, including the Earliest Deadline First method (EDF) [4], Rate-Monotonic [4], preemptive resume service priority [10], and Dynamic Transmission Mechanism-L priority (DTM-L) [5]. Despite their effectiveness, these techniques have certain limi-

tations, such as process starvation. This can lead to prolonged delays for lengthy processes to complete their service if shorter processes are repeatedly introduced [4], poor performance when operating under overloaded conditions, these techniques may not be optimal for multiprocessor systems, low throughput [11]. Moreover, in scenarios with high arrival rates of higher-priority applications, these techniques may cause lower-priority applications to starve [5].

Previous literature has commonly assumed IoT servers to be homogeneous, with similar devices and equal service rates [12]. In our earlier work [6], we proposed a two-level priority system based on the size and delay of healthcare packets, assuming server homogeneity. However, the IoT ecosystem comprises of a variety of heterogeneous devices that operate at different service rates [9]. In addition, in a multi-server system, replacing outdated or misbehaving servers with newer or more powerful ones leads to server heterogeneity [13]. As a result, when developing scheduling algorithms for IoT healthcare monitoring, it is crucial to take into account the heterogeneity of servers and their capabilities.

This study addresses the aforementioned challenges by introducing analytical models that evaluate delay and size-dependent priority-aware scheduling for IoT-based healthcare packets using heterogeneous multi-server priority queuing systems. The model's performance is evaluated based on mean slowdown and throughput. Mean slowdown is the normalized response time, which is the ratio of the packet's response time to its size [14]. Throughput refers to the amount of data that can be transferred within a given time frame [15].

The contribution of this study is that the performance of the analytically developed PS scheduling models are validated using simulations with mean slowdown and throughput as performance metrics. The remainder of the paper is structured as follows: review of previous work with a focus on the server characteristics and scheduling strategies used in healthcare monitoring systems is discussed in Section II. Section III presents the analysis of the expressions. Model validation and comments are discussed in Section IV, while Section V presents the conclusion.

II. RELATED WORK

The EDF scheduling approach, which prioritizes requests according to their absolute deadlines, was proposed in [4]. Data packets with close deadlines receive higher priority compared to those with distant deadlines, which are given lower priority. However, EDF has a significant disadvantage in situations of high load since it focuses on packets that are near their deadlines, causing delays for other packets that have sufficient time to meet their deadlines. Hence, it is essential to develop scheduling techniques that give preference to packets with tight deadlines while not considerably extending the deadline for packets with more extended deadlines.

The Rate-Monotonic (RM) algorithm considered in [4], prioritizes the tasks with the shortest duration. The method is extremely predictable since scheduling choices are determined a priori under this technique. However, precomputation is required whenever changes occur in task parameters. A task's duty cycle is another factor that the algorithm uses to determine priority, with lower duty cycle activities receiving greater priority.

In their work, Sharif et al. [16] presented a mechanism for scheduling tasks and allocating resources based on priorities. The mechanism, known as priority-based task-scheduling and resource-allocation (PTS-RA), can assign different priorities to tasks by considering their emergency levels, which are calculated based on data from a patient's smart wearable devices. The mechanism can determine whether a task should be processed locally at the hospital workstations or in the cloud. This is aimed at reducing the total task processing time and the bandwidth cost as much as possible. One potential drawback of PTS-RA task scheduling and resource allocation in edge computing for health monitoring systems is the increased complexity of the system. This approach requires sophisticated algorithms and decision-making processes to determine the appropriate priorities for tasks and allocate resources efficiently. Additionally, there may be challenges related to data privacy and security when collecting and processing data from patients' wearable devices.

Iqbal et al. [17] presented a smart patient health monitoring system (PHMS) based on an optimized scheduling mechanism using IoT-tasks orchestration architecture to monitor vital signs of remote patients. The proposed smart PHMS consists of two core modules, a healthcare task scheduling module and optimization of healthcare services using a real-time IoT-based task orchestration architecture. The experimental results reveal that an optimized scheduling mechanism reduces the tasks starvation compared to a conventional fair emergency first (FEF) scheduling mechanism. However, the proposed work did not integrate predictive analytics with IoT to forecast vital signs to improve the performance of IoT-based healthcare services.

In order to improve the performance of static scheduling algorithms, a new method called Tasks Classification and Virtual Machines Categorization (TCVC) based on tasks importance was proposed by T. Aladwani [18]. Tasks are classified based on the importance of the patient's health status that is, high, medium and low importance. The method was applied with the Max-Min scheduling algorithm and the performance was found to outperform the First Come First Serve (FCFS), Shortest Job First (SJF) and Max-Min scheduling algorithms in terms of total execution time, total waiting time, and total finish time. However, under high arrival rate of high important tasks, low important tasks are starved of service.

A new cloud scheduling architecture called IADA was presented in [19]. The architecture was aimed at improving previous methods by using a dynamic classification scheme for workload variations instead of a segmented classification. The approach utilizes resources more efficiently and ensures compliance with Quality of Service requirements through the application of machine learning techniques, heuristics, and a Bayesian changepoint detection algorithm for real-time analysis. However, the study did not address how to physically

place virtual machines to minimize performance degradation and comply with QoS requirements.

In this paper, we deal with two traffic streams submitted to heterogenous shared servers and ordered by a non-preemptive priority scheduling discipline. We propose an analysis of the mean slowdown for each stream of traffic assuming that arrivals are Poisson and the server is work-conserving and non-preemptive. The intention of the study is to validate the theoretical results from our previous works in [6] using simulations.

III. Analysis

A. System Model

The proposed system model consists of various healthcare packets that originate from several distinct sensors mounted on a patient's body to track various health conditions, as illustrated in Figure 1. The healthcare packets produced by the sensors arrive at the network gateway randomly and have been shown to be well approximated by the Poisson process, as reported in [21]. Examples of delay-sensitive packets include EEG/ECG/EMG with a delay limit of not more than 250 ms, glucose monitoring with a delay limit of not more than 20 ms, blood pressure monitoring with a delay requirement of not more than 750 ms, and endoscope imaging with a delay requirement of not more than 500 ms [8].

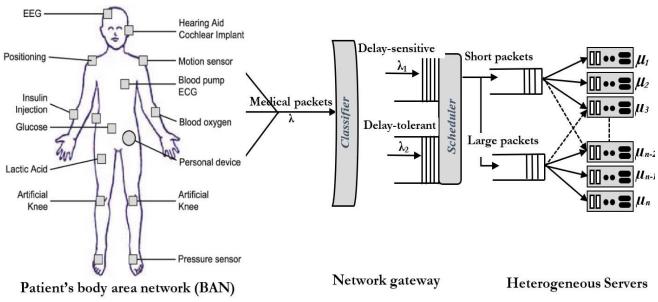

Fig. 1. System model [6]

When a packet enters the network gateway, the classifier immediately assigns the packet a priority based on its level of delay sensitivity and predetermined requirements such as the maximum tolerable delay.

On the other hand, medication dispenser data, home telemonitoring, access to a patient's electronic health records, etc. are some examples of delay tolerant packets [22]. Delay-sensitive packets are given priority over delay-tolerant packets. The scheduler then receives the packets and classifies them into short and large packets depending on the set threshold size. Large packets are serviced after short packets.

Assumptions

The system model is a heterogeneous multi-server with an infinite capacity queue, developed under the following assumptions:

- Packet arrival rate follows a Poisson distribution function with parameter $\lambda_i; i = 1, 2,$, in which case λ_1 represents arrival rate of delay sensitive packets while λ_2 represents arrival rate of delay tolerant packets [21].
- Each server's service times follow an exponential distribution with parameter $\mu_i; i = 1, 2, ..., c$, in which case μ_i is the service rate of server M_i [21].
- The service is offered via a variety of c heterogeneous servers.
- Each server has infinite capacity [23].

The system model is represented as an $M/M_i/c$ queue, where M denotes random packet arrival following a Poisson distribution, M_i denotes the exponentially distributed service time of server i, and c represents the number of heterogeneous servers with infinite capacity.

B. Prioritized Scheduling Scheme

Priority awareness is the most crucial criterion when scheduling the service of multiclass healthcare packets that possess various levels of urgency [**?**]. In this PS system, packets are divided into two priority levels: in terms of delay requirements, the first priority level classifies packets into delay-sensitive and delay-tolerant, and short or large packets at the second priority level, depending on a predetermined threshold. The working of the PS scheme is shown by the flow diagram in Fig. 2. To increase the number of packets serviced in a given amount of time, short packets are given preference in service over large packets. Buffers are considered to have infinite capacity for each queue of packets that are delay-sensitive or delay-tolerant. Similar assumptions were made in the performance evaluation of IoT-enabled healthcare monitoring systems [23].

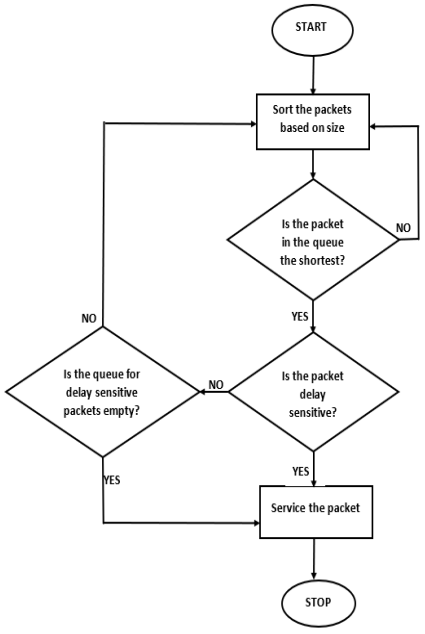

Fig. 2. Flow diagram showing the working of the PS scheme [6]

The packets are then sent to the scheduler, which distributes them to other shared heterogeneous servers. Concerning server

allocation, three popular allocation strategies have been used in literature [9]: the fastest server first (FSF) allocation, which sends the packet to the fastest free server first; the slowest server first (SSF) allocation, which sends the packet to the slowest free server first; and the randomly chosen server (RCS) allocation, which sends the subsequent packet in the queue to any idle server at random. In this study, we consider the FSF allocation policy since it has been proven to be better than the others [9].

Given the differences in term of sizes of packets, the service rate of healthcare packets can be modeled using the exponential distribution [8], [12].

The exponential probability density function is given in [8] as:

$$f(x) = \mu e^{-\mu x}, x \geq 0, \mu \geq 0. \tag{1}$$

where the service rate is given as μ.

The proposed PS policy is a non-preemptive, delay-aware, size-based scheduling policy. At the first priority level, the PS policy classifies packets into delay-sensitive or delay-tolerant and on packet sizes, namely short (x_s) and large (x_l) at the second priority level. Short packets are prioritized over large packets for each class of delay-sensitive or delay-tolerant packets. Utilizing heterogeneous multiple servers, packets belonging to the same class are served in first-come, first-served (FCFS) order.

C. Mathematical background

This study assumes that the servers are ordered in decreasing service rate, that is, $\mu_1 > \mu_2 > ... > \mu_c$. The implication of this, is that, μ_1 is faster than μ_2, and μ_2 is faster than μ_3, etc. The service rate of the servers can be defined by [9].

$$M_i = \begin{cases} \sum_{j=1}^{i} \mu_j & i < c \\ \sum_{j=1}^{c} \mu_j & i \geq c \end{cases} \tag{2}$$

Eq. 2 shows that M_i is a variable and may be expressed in two different ways depending on whether the system has less than c servers or packets (in which case one server serves one packet at a time) and when the system contains at least c packets.

The expressions for the mean response time under the EDF policy are then defined, and the EDF policy is used to compare with the prioritized scheduling scheme. Under the EDF scheme, the server processes packets having the smallest deadline among all of the waiting packets. For a two priority class, the waiting time of packets under the EDF scheme is given in [23].

$$W_s = W_o + \rho_s W_s + \rho_d max(0, W_d - D_{d,s}) \tag{3}$$

$$W_d = W_o + \rho_s W_s + \rho_d W_d + \rho_s min(W_d, D_{d,s}) \tag{4}$$

where W_o is the mean waiting time required to finish the service of the packet being served when the tagged packet arrives. In this case, $W_o = \frac{\sum_{t=1}^{2} \lambda_t E(x_t^2)}{2}$.

W_s is the average waiting time for delay sensitive packets, W_d is the average waiting time for delay tolerant packets, ρ_s is the load resulting from delay sensitive packets, ρ_d is the load resulting from delay tolerant packets.
$D_{d,s} = d_d - d_s$, where d_d is the deadline offset of delay tolerant packets and d_s is the deadline offset for delay sensitive packets.

D. The PS scheme: Delay sensitive packets

A tagged delay-sensitive packet that arrives to a short packet-only delay-sensitive queue under the PS scheme will be delayed all delay-sensitive short packets found in the queue and the average waiting time for the tagged short delay-sensitive packet of size x_s is given in [6].

$$W(x_{ss}) = \frac{P_{oss} m_c^c \rho_{x_{ss}}^{c+1}}{\lambda_1 (\pi_{i=1}^{c} m_i)(1 - \rho_{xss})^2} \tag{5}$$

where
$\rho_{xss} = \lambda_1 \int_0^{x_{ss}} t f(t) dt = \frac{\lambda_1}{m_c}(1 - e^{-m_c x_{ss}}) - x_t e^{-m_c x_{ss}}$
and

$$P_{o_{ss}}^{-1} = \sum_{n=0}^{c-1} \frac{\lambda_1^n}{\pi_{i=1}^{n} (\sum_{j=1}^{i} \mu_j)} + \left(\frac{m_c^c}{\pi_{i=1}^{c} m_i} \right) (1 - \rho_{xss})^{-1} \rho_{x_{ss}}^{c}$$

In the same way, a large delay-sensitive packet that has been tagged will experience a delay not only from other large delay-sensitive packets in the queue but also from short delay-sensitive packets in the queue. Furthermore, the large delay-sensitive packet that has been tagged will be served only after all the short delay-sensitive packets that arrived after it in the queue have been serviced. The average waiting time for a large delay-sensitive packet of size x_l is provided in [6].

$$\overline{W(x_{ls})} = 2W(x_{ss}) + W(x_{ls}) \tag{6}$$

where $W(x_{ss})$ is as given in (5) and

$$W(x_{ls}) = \frac{P_{oss} (m_c)^c \rho_{x_{ls}}^{c+1}}{\lambda_1 (\pi_{i=1}^{c} m_i)(1 - \rho_{xls})^2} \tag{7}$$

Also,

$$P_{o_{ls}}^{-1} = \sum_{n=0}^{c-1} \frac{\lambda_1^n}{\pi_{i=1}^{n} (\sum_{j=1}^{i} \mu_j)} + \left(\frac{m_c^c}{\pi_{i=1}^{c} m_i} \right) (1 - \rho_{xls})^{-1} \rho_{x_{ls}}^{c}$$

and $\rho_{xls} = \lambda_1 \int_{x_t}^{\infty} t f(t) dt = \lambda_1 e^{-m_c x_t} (x_t + \frac{1}{m_c})$

E. The PS scheme: Delay tolerant packets

The PS scheme chooses the delay tolerant packets after servicing all delay tolerant packets.

In case the tagged packet happens to be a short delay-tolerant one, the service of the tagged packet will be delayed by all short and large delay-sensitive packets, as well as short delay-tolerant packets already present in the queue. Furthermore, all short and large delay-sensitive packets that come after the tagged short delay-tolerant packet in the queue will cause a delay for the short delay-tolerant packet. The service of short and large delay-sensitive packets that arrive

after the tagged short delay-tolerant packet is added to the queue will take place before the tagged short delay-tolerant packet is serviced. The average waiting time for the delay-tolerant short packet of size x_{sd} is given in [6] as

$$\overline{W(x_{sd})} = 2W(x_{ss}) + 2W(x_{ls}) + W(x_{sd}) \tag{8}$$

where $W(x_{ss})$ and $W(x_{ls})$ are as given in (5) and (7) respectively. Here,

$$W(x_{sd}) = \frac{P_{osd}(m_c)^c \rho_{x_{sd}}^{c+1}}{\lambda_2(\pi_{i=1}^c m_i)(1-\rho_{xsd})^2} \tag{9}$$

and

$$P_{o_{sd}}^{-1} = \sum_{n=0}^{c-1} \frac{\lambda_1^n}{\pi_{i=1}^n(\sum_{j=1}^i \mu_j)} + \left(\frac{m_c^c}{\pi_{i=1}^c m_i}\right)(1-\rho_{xsd})^{-1}\rho_{x_{sd}}^c$$

where $\rho_{xsd} = \lambda_2 \int_0^{x_t} tf(t)dt$

The same analysis can be applied to a tagged large delay-tolerant packet. In this case, this packet will be served after all short delay sensitive, large delay sensitive, short delay tolerant, and large delay tolerant packets in the queue have been served. Moreover, any short or large delay-sensitive packets that arrive after the tagged large delay-tolerant packet has that been added to the queue will be serviced before the tagged packet. The average waiting time for the delay tolerant large packet of size x_{ld} is given in [6] as:

$$\overline{W(x_{ld})} = 2W(x_{ss}) + 2W(x_{ls}) + W(x_{sd}) + W(x_{ld}) \tag{10}$$

where $W(x_{ss})$, $W(x_{ls})$ and $W(x_{sd})$ are as given in (5), (7) and (9) respectively. Here

$$W(x_{ld}) = \frac{P_{osd}(m_c)^c \rho_{x_{ld}}^{c+1}}{\lambda_2(\pi_{i=1}^c m_i)(1-\rho_{xld})^2}$$

and

$$P_{o_{ld}}^{-1} = \sum_{n=0}^{c-1} \frac{\lambda_2^n}{\pi_{i=1}^n(\sum_{j=1}^i \mu_j)} + \left(\frac{m_c^c}{\pi_{i=1}^c m_i}\right)(1-\rho_{xld})^{-1}\rho_{x_{ld}}^c$$

where $\rho_{xld} = \lambda_2 \int_{x_{td}}^{\infty} tf(t)dt$.

In the next section, simulations are run in order to assess the accuracy of the theoretical results with results obtained using simulations.

IV. Model Validation and Comments

To validate the analytical models, we implement the PS scheduling policy in the Simulink MATLAB R2023b tool [20]. Simulink is a graphical programming environment for modeling, simulating and analyzing dynamic systems. It can be used to build and simulate the queuing system in a block diagram form. We considered the Poisson distribution [21] to generate the arrival rates of both delay sensitive packets, λ_1 and arrival rates of delay tolerant packets, λ_2. To enusre a high load value, the arrival rate is varied from 0 to 6.549 packets per second. The server's service times are generated using the exponential distribution with parameters, μ_i, where i is the service state of the packets for server i. The packet sizes are distributed by the exponential distribution that has the coefficient of variability equal to one, since the standard deviation of an exponential distribution is equal to its mean.

Each server is considered to have infinite capacity [23]. The server type is selected based on the Fastest Server First (FSF) allocation policy since it has been proven to be better than the others [9]. The particular combination of workload parameters used to generate the packets is shown in Table I. This set of parameters results in a total load equal to 0.9, which is considered high load. These parameters are consistent with those used in the literature. [8], [23], [24], [25].

Using the Simulink MATLAB tool, we obtained performance measures for the mean slowdown for delay sensitive short and large packets as a function of the sizes of packets and mean slowdown for delay tolerant short and large packets as a function of packet sizes. In addition, we obtained performance measures for throughput as a function of packet sizes for short and large delay sensitive packets. We compare the results obtained from simulations with the analytical results for mean slowdown as a function of packet sizes. Since the results are in agreement, we consider the analytical model (and the simulation results) validated.

A. Simulation Parameters

In this section, the simulation parameters are presented.

TABLE I
SIMULATION PARAMETERS

Parameter	Value
Number of servers, m	5 [24]
Packet arrival rate, λ	0 to 6.549 packets/second [25]
Service rate for the multi-servers, μ	1,2,3,4,5 packets/second [24]
High system load, ρ	0.9 [23]
Average packet size, x_t	100 Kb[8]
Threshold of the packet size, x_{ts}	75 Kb [8]

B. Evaluation of the mean slowdown of packet sizes for delay sensitive packets

In this section, the performance of the PS scheduling scheme is compared for analytical and simulation results for delay sensitive packets in terms of mean slowdown, the performance of the EDF policy is included for purposes of comparison. Fig. 3 presents the variation of mean slowdown of delay-sensitive short packets with packet size for the EDF and PS schemes for the analytical models and for the corresponding detailed simulation, where short packets have sizes less than or equal to $x_s = 75$ Kb. The results demonstrate that the analytic model has very good overall agreement with the MATLAB simulation estimates. The mean slowdown is slightly underestimated for short packets, shorter than about 30 Kb. The overall agreement between the analytic and simulation estimates is excellent, and the small discrepancy for short packets can easily be taken into account due to time taken in prioritization of the different packets.

Similar observations are noted in Fig. 4 for the mean slowdown against packet size for delay sensitive large packets,

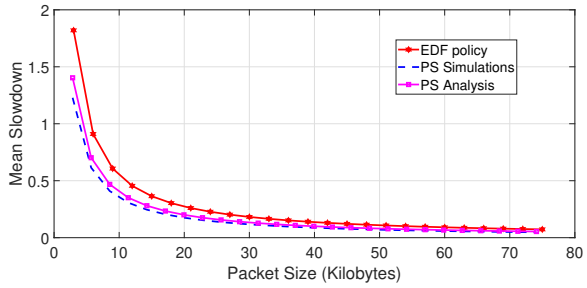

Fig. 3. Mean slowdown vs packet size for for delay sensitive short packets

where the analytical model and the corresponding simulations are in close agreement. Again, large packets are packets with sizes greater than $x_s = 75$ Kb. The mean slowdown is slightly underestimated for shorter packets, shorter than about 600 kb. The agreement between the analytic and simulation results is in good agreement, and the small discrepancy for shorter packets can easily be taken into account due to time taken in prioritization of the different packets.

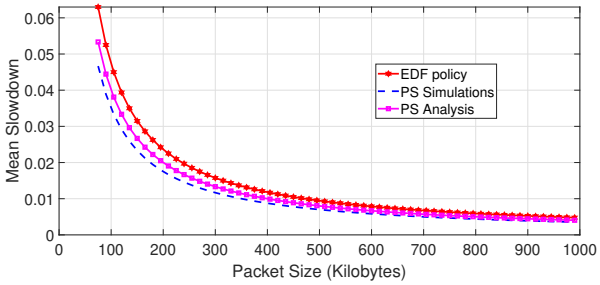

Fig. 4. Mean slowdown vs packet size for delay sensitive large packets

C. Evaluation of the mean slowdown of packet sizes for delay tolerant packets

In this section, the performance of the PS scheduling scheme is compared for analytical and simulation results for delay-tolerant packets in terms of mean slowdown. The performance of the EDF scheme is included for purposes of comparison. The study validates the analytical models for the PS scheduling policy, we varied the arrival rate and fixed the service rate to achieve high load and ascertain the effect of varying the packet sizes on the mean slowdown for delay tolerant packets. In Figure 5 we plot the mean slowdown versus packet size obtained from the analytic model as well as from the simulations. It is noted that the PS scheme is better than the EDF scheme, especially for shorter packet sizes. We note the representative validation results being near in perfect agreement with the analytic models. Similar to Figure 3, the mean slowdown is slightly underestimated for short packets, shorter than about 30 Kb. The overall agreement between the analytic and simulation estimates is excellent, and the small discrepancy for short packets can easily be taken into account due to time taken in prioritization of the different packets.

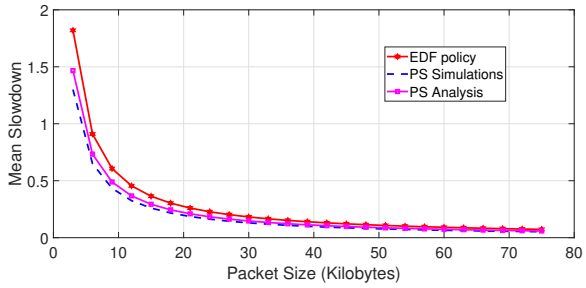

Fig. 5. Mean slowdown vs packet size for delay tolerant short packets

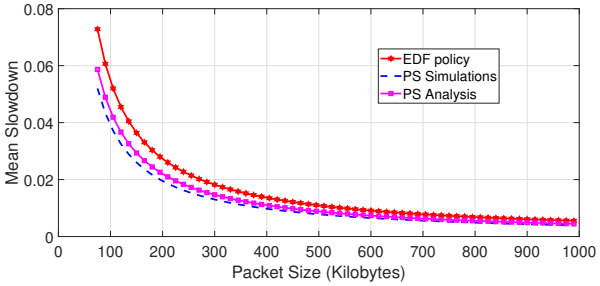

Fig. 6. Mean slowdown vs packet size for delay tolerant large packets

Fig. 6 presents the mean slowdown against packet size for delay tolerant large packets. The arrival rate is varied while the service rate is fixed to ensure high load. In this case, large packets are packets with sizes greater than $x_s = 75$ Kb. The results demonstrate that the analytic model is in very good agreement with the MATLAB simulation results. It is also observed that the mean slowdown is slightly underestimated for shorter packets, shorter than about 600 Kb.

D. Evaluation of throughput of packet sizes for delay sensitive packets

In this section, the validated results of the PS scheduling scheme is presented in terms of throughput for delay-sensitive packets. The performance of the EDF scheme is included for purposes of camparison. Figure 7 and 8 shows the validation results for PS scheduling schemes for delay-sensitive short and large packets in terms of throughput. We observe that the simulation validates the model. Similarly, we observe that the analytic mean slowdown as a function of packet size for delay tolerant short and large packets is in excellent agreement with the simulation results. Note also that the mean slowdown for siumulations results underestimates mean slowdown for shorter packets in both cases, this observation was noted in the previous sections. In all cases, it can be observed that as packet sizes increase, the throughput also increases.

In the next section, the conclusion and future work are presented.

V. Conclusion and Future Work

We have validated the analytical models of the PS scheduling scheme with exponentiallly distributed packet sizes. In

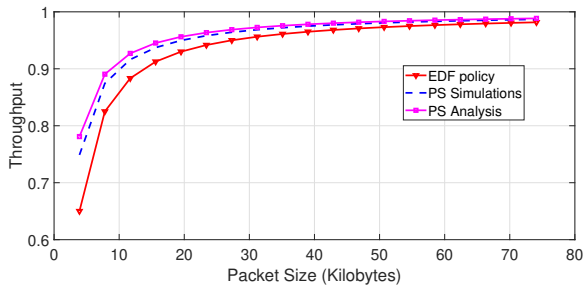

Fig. 7. Throughput vs packet size for delay sensitive short packets

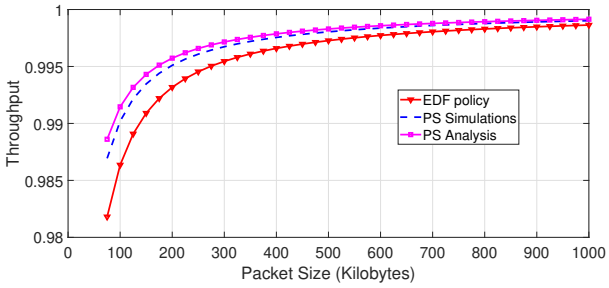

Fig. 8. Throughput vs packet size for delay sensitive large packets

addition, the performance of the EDF scheme is included for purposes of comparison. We presented results for delay sensitive and delay tolerant packets for short and large packets in each case. In addition, we evaluated the performance of the system using mean slowdown and throughput as performance measures. Both theoretical and simulation results show that the proposed mechanism can meet all design requirements for both short and large packets. Given these validations, the analytical models can be used to evaluate the performance of PS scheme. In future research, we will investigate the performance of the PS scheme under other packet size distributions like the Bounded Pareto Distribution where there is a mix of a large fraction of short packets and a small fraction of large packets but the large packets can cause congestion in the system, in addition to implementing a threshold on packet sizes.

REFERENCES

[1] H. Zhang, J. Li, B. Wen, Y. Xun and J. Liu, "Connecting intelligent," IEEE Internet of Things, vol. 5, no. 4, p. 1550–1560, June 2018.

[2] H. Bhatia, S. N. Panda and D. Nagpa, "Internet of Things and its Applications in Healthcare-A Survey," Proc. International Conference on Reliability, Infocom Technologies and Optimization, Noida, India, 2020.

[3] C. Yi and Jun Cai, "Transmission Management of Delay-Sensitive Medical Packets in Beyond Wireless Body Area Networks: A Queueing Game Approach," IEEE Trans. Mob. Comput., vol. 17, no. 9, pp. 2209-2222, January, 2018.

[4] E. Gomes, M.A.R Dantas and P. Plentz, "A Real-Time Fog Computing Approach for Healthcare Environment," Springer, pp. 85-95, 2019.

[5] C. Yi and J. Cai, "A priority-aware truthful mechanism for supporting multi-class delay-sensitive medical packet transmissions in e-health networks," IEEE Trans. Mob Comput., vol. 16, no. 9, pp. 2422-2435, September, 2017.

[6] B. K. Asingwire, A. Ngenzi, L. Sibomana and C. Kabiri, "Performance Analysis of IoT-based Healthcare Heterogeneous Delay-sensitive Multi-Server Priority Queuing System," International Journal of Advanced Computer Science and Applications, Vol. 12, No. 10, 2021.

[7] A. M. Rahmani, T. N. Gia, B. Negash, A. Anzanpour, I. Azimi, M. Jiang and P. Liljeberg, "Exploiting smart e-Health gateways at the edge of healthcare Internet-of-Things: A fog computing approach," Future Generation Computer Systems, vol. 78, no. 2, pp. 641-658, 2018.

[8] Y. Changyan and J. Cai, "A Truthful Mechanism for Scheduling Delay-Constrained Wireless Transmissions in IoT-Based Healthcare Networks," IEEE Trans Mob Comput., vol. 18, no. 2, pp. 912-925, December, 2018.

[9] H. S. Narman, M. Hossain, M. Atiquzzaman, and H. Shen, "Scheduling Internet of Things Applications in Cloud Computing," Annals of Telecommunications, vol. 72, pp. 79–93, 2017.

[10] N. Deepika, M. Anand, K. Sudhaman, "Internet Connected e-Healthcare System with Live Video Monitoring using LWIP Stack and SJF Priority Scheduling," International Journal of Recent Technology and Engineering, Vol. 8, No. 4, November, 2019.

[11] X. Ma, Z. Wang, S. Zhou, H. Wen and Y. Zhang, "Intelligent Healthcare Systems Assisted by Data Analytics and Mobile Computing," Wireless Communications and Mobile Computing, 2018.

[12] C. Yi and J. Cai, "A Truthful Mechanism for Scheduling Delay-Constrained Wireless Transmissions in IoT-Based Healthcare Networks," IEEE Trans. on Wireless Communications, vol. 17, no. 9, pp. 912-925, February, 2019.

[13] D. Efrosinin, N. Stepanova and J. Sztrik 4, "Algorithmic Analysis of Finite-Source Multi-Server Heterogeneous Queueing Systems," MDPI Journal of Mathematics, vol. 9, no. 20, pp. 2-24, 2021.

[14] M. Okopa, D. Turatsinze, T. Bulega and J. Wampande, "Revenue Maximization Based on Slowdown in Cloud Computing Environments". Australasian Journal of Computer Science, vol 4, pp. 1-16, 2017.

[15] P. D. Mankar, Z. Chen, M. A. Abd-Elmagid, N. Pappas, and H. S. Dhillon, "Throughput and Age of Information in a Cellular-Based IoT Network," IEEE Transactions on Wireless Communications, vol 20, no. 12, pp. 8248-8263, December, 2021.

[16] Z. Sharif, L. T. Jung, M. Ayaz, M. Yahya and S. Pitafi, "Priority-based task scheduling and resource allocation in edge computing for health monitoring system," Journal of King Saud University of Computer and Information Sciences, vol. 35, no. 2, pp. 544-559, February 2023.

[17] N. Iqbal, Imran, S. Ahmad, R. Ahmad, and D. Kim, "A Scheduling Mechanism Based on Optimization Using IoT-Tasks Orchestration for Efficient Patient Health Monitoring," Journal of Sensors, vol. 21, no. 16, August, 2021.

[18] T. Aladwani, "Scheduling IoT Healthcare Tasks in Fog Computing Based on their Importance," Procedia Computer Science vol. 163, pp. 560-569, 2019.

[19] V. Meyer, M. L. da Silva, D, F. Kirchoff, C. A.F. De Rose, "IADA: A dynamic interference-aware cloud scheduling architecture for latency-sensitive workloads," Journal of Systems and Software, vol. 194, no. C, December, 2022.

[20] D. K. Chaturvedi, "Modeling and Simulation of Systems Using MATLAB® and Simulink®; CRC Press: Boca Raton, FL, USA, 2017.

[21] A. Salh, L. Audah, M. Alhartomi. K. Soon, S. H. Alsamhi, F. A. Almaki, Q. Abdullahi. A, Saif and H. Algethami, "Smart Packet Transmission Scheduling in Cognitive IoT Systems: DDQN Based Approach," IEEE Access, vol. 10, no. 4, pp.50023-50035, 2022.

[22] K. Park, J. Park and J. Lee, "An IoT System for Remote Monitoring of Patients at Home", Journal of Applied Sciences, vol. 7, no. 3, pp. 1-23, March, 2017.

[23] A. M. Muwumba, G. N. Justo, L. V. Massawe and J. Ngubiri, "Priority EDF Scheduling Scheme for MANETs," Proc. International Conference on Communications and Networking in China, pp. 66-76, 2020

[24] B. Nansamba, M. Okopa, B. K Asingwire, and K. S. Kaawaase, "Pricing Scheme for Heterogeneous Multi-server Cloud Computing System," Australasian Journal of Computer Science, vol 4, pp. 32-43, 2017.

[25] C. Majumdar, M. Lopez-Benitez, and S. N. Merchant, "Experimental Evaluation of the Poisson Process of Real Sensor Data Traffic in the Internet of Things," Proc. IEEE Annual Consumer Communications & Networking Conference, pp. 1-7, January, 2019.