

IET Communications

Special issue Call for Papers

**Be Seen. Be Cited.
Submit your work to a new
IET special issue**

Connect with researchers and experts in your field and share knowledge.

Be part of the latest research trends, faster.


[Read more](#)



The Institution of
Engineering and Technology

ORIGINAL RESEARCH

Delay and size-dependent priority-aware scheduling for IoT-based healthcare traffic using heterogeneous multi-server priority queueing system

Barbara Kabwiga Asingwire^{1,2}  | Louis Sibomana^{1,3} | Alexander Ngenzi¹ | Charles Kabiri¹

¹African Centre of Excellence in Internet of Things (ACEIoT), University of Rwanda, Kigali, Rwanda

²Department of Computer Engineering, Busitema University, Tororo, Uganda

³National Council for Science and Technology, Kigali, Rwanda

Correspondence

Barbara Kabwiga Asingwire, ACEIoT, Univer.
Email: kezabarbara@gmail.com

Abstract

Internet of Things (IoT) based healthcare applications are time-sensitive and any delay can cause alarming situations, including death of patients. The Earliest Deadline First (EDF) scheduling scheme has been proposed for use in IoT-based healthcare applications. However, the EDF scheme performs poorly under overloaded conditions due to giving highest priority to packets that are close to missing their deadlines. Some studies have proposed the use of Priority EDF to overcome the challenges of EDF; however, Priority EDF still favours higher priority queues which increases the waiting times of lower priority queues. In order to overcome the limitation of EDF and its variants, this paper proposes a system model for a prioritized scheduling (PS) scheme. The PS scheme is an improvement of the Earliest Deadline First (EDF) scheme and its variants for IoT-based healthcare applications. The PS scheme uses a heterogeneous multi-server priority queueing system to provide service differentiation by prioritizing short packets over large packets and delay sensitive packets are serviced before delay tolerant packets. Numerical results demonstrate that the PS scheme minimizes the mean slowdown for both delay sensitive short and large packets at low and high load values. Additionally, the PS scheme performs better than the EDF and Priority EDF schemes in terms of reducing mean slowdown of packets and the PS scheme performs better than the EDF in terms of throughput for all packet sizes at both low and high load values. The performance improvement in terms of throughput is more pronounced at high load values. This addresses the challenge of the EDF scheme which performs poorly under overloaded conditions and the challenge of the Priority EDF scheme which favours higher priority queues at the expense of low priority queues.

1 | INTRODUCTION

Recent technological advancements have given the Internet of Things (IoT) the ability to seamlessly connect devices, sensors, and systems, creating a cohesive network of interconnected technology [1]. It has potential applications in a variety of fields, such as remote healthcare monitoring, making it a versatile and powerful technology [2].

The application of IoT in remote healthcare monitoring offers an advantage over traditional healthcare monitoring methods, and is expected to improve emergency management and healthcare monitoring in the future. To ensure accurate patient monitoring, IoT-based healthcare monitoring requires

that the collected data be delivered instantly, with minimal delay, and in a highly reliable manner. This is because healthcare applications require real-time data with minimal delay.

Medical emergencies must typically be reported before other regular services [3]. Additionally, the services for medical packets must be differentiated based on the demands of the signals. In emergency situations, it is important for healthcare traffic to have low latency to allow the healthcare personnel to respond on time [4, 5]. However, due to the heterogeneity of IoT servers and applications with varying degrees of service requirements, traditional computing server scheduling schemes cannot deliver services to IoT-based healthcare services [6]. Therefore, the standard server scheduling algorithms should be enhanced to

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *IET Communications* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

efficiently schedule packets by taking into account the heterogeneity of servers and various service requirements in order to satisfy user expectations.

Furthermore, as the data size increases, there is an increase in delay for healthcare IoT packets [7]. These delays can range from milliseconds to minutes for time-sensitive applications.

In order to properly schedule healthcare traffic, the following factors need to be considered [8]:

- 1) Critical medical information transmissions should be given a guarantee. This is because prolonged transmission delays for urgent medical information deliveries can cause patients' health to deteriorate and potentially result in their death. To address this challenge, delay-sensitive packets are given precedence over delay-tolerant packets in this study.
- 2) Delivery of medical traffic should be differentiated by its heterogeneous nature in terms of delay requirements. Maintaining an absolute priority rule over all medical applications may result in abnormally huge delays for "delay-tolerant" applications. Additionally, "delay-tolerant" medical applications are important elements of patients' health characteristics. To overcome this challenge, this study considers priority service differentiation. To increase the number of packets serviced per unit time, service differentiation of medical packets is implemented based on the size of each packet, with short packets given priority over large packets, in addition to delay sensitivity. In this case, delay-sensitive packets are given priority over delay-tolerant packets.
- 3) Healthcare IoT devices are analyzed based on the assumption that the servers are homogeneous, which is unrealistic because IoT consists of various heterogeneous devices and applications [9]. To address this issue, the IoT servers are assumed to be heterogeneous in order to account for differences in capacity, processing speed etc.

Recent studies have provided a number of scheduling techniques for IoT-based healthcare monitoring systems, including Earliest Deadline First (EDF) [4], Priority EDF [10], Rate-Monotonic [4], preemptive resume service priority [11], and Dynamic Transmission Mechanism-L priority (DTM-L) [5]. However, these techniques have drawbacks, such as process starvation, which can result in long delays for long processes to finish service if short processes are introduced repeatedly [4], poor performance under overloaded conditions, not being optimal for multiprocessors, low throughput [12], and higher priority applications starving lower priority applications under high arrival rates of higher priority applications [5, 10].

In literature, IoT servers are widely assumed to be homogeneous (having equal service rates and consisting of similar devices) [13, 14]. Furthermore, our previous work [6] introduced the idea of applying priority at two levels based on the delay and size of healthcare packets, based on the assumption that the servers are homogeneous. However, IoT consists of various heterogeneous devices that operate at different service rates [9]. In addition, a multi-server system becomes heterogeneous when outdated or misbehaving servers are replaced with newer or more powerful ones [15]. Therefore, when

designing server scheduling algorithms for IoT healthcare monitoring, heterogeneous servers and their capabilities should be well-considered.

To overcome the above challenges, this study develops analytical models for the evaluation of delay and size-dependent priority-aware scheduling for IoT-based healthcare packets using heterogeneous multi-server priority queuing systems. The model performance is assessed in terms of mean slowdown and throughput. The normalized response time, or the ratio of the packet's response time to its size, is referred to as mean slowdown [16], while throughput refers to the actual data a network can transfer within a given time frame [17].

This study makes two contributions. The first contribution is that the study develops a system model for the evaluation of delay and size-dependent priority-aware scheduling for IoT-based healthcare packets using a heterogeneous multi-server priority queuing system. The second contribution is that the performance of the developed model is assessed against the EDF and Priority EDF scheduling schemes using mean slowdown and throughput as performance metrics. The remainder of the paper is structured as follows: related work is discussed in Section 2. Section 3 presents the system model and analytical expressions. System performance evaluation is discussed in Section 4, while Section 5 presents the conclusion.

2 | RELATED WORK

This section presents a review of existing work, focusing on scheduling policies used for allocating services in healthcare monitoring systems, as well as the characteristics of servers used.

The EDF scheduling scheme, which assigns priorities to requests based on their absolute deadlines, was proposed in [4]. Data packets with short deadlines are assigned higher priority, while those with long deadlines are assigned lower priority. However, one of the major drawbacks of EDF is that it performs poorly under overloaded conditions, as it prioritizes packets close to their deadlines, resulting in delays for other packets that still have time to meet their deadlines. Therefore, there is a need to develop scheduling techniques that prioritize packets with short deadlines without significantly increasing the mean waiting time for packets with longer deadlines.

Analytical EDF Priority schedulers which favour higher priority queues thereby reducing their waiting times have been proposed in [10]; however, favoring higher priority queues end up increasing the waiting times of lower priority queues.

The Rate-Monotonic (RM) algorithm, a static scheduling algorithm, is considered in [4]. In the RM algorithm, the task with the shortest period has the highest priority. Under this mechanism, scheduling decisions are made a priori, making the algorithm highly predictable. However, changes in task parameters require precomputation. The algorithm also assigns priority based on a task's duty cycle, with lower duty cycle tasks having higher priority. Due to the static nature of the algorithm, the priorities of tasks are fixed, which may not be suitable for dynamic changes in task periods. Therefore, there is a need to

develop scheduling policies that can adapt to dynamic changes in task periods.

A data scheduling approach for monitoring inter-Wireless Body Area Network (WBAN) systems was developed in [18] to meet the Quality of Service (QoS) requirements in Wireless Body Area Network (WBAN) networks. The authors proposed using a critical delay parameter to prioritize and group packets into an aggregated frame for transmission to the medical server. This approach aims to ensure QoS in terms of delay, throughput, and packet loss for applications running on sensor nodes. However, using an aggregate approach can significantly increase latency.

Iqbal et al. [19] proposed a real-time IoT-based task orchestration system to produce autonomous healthcare tasks and control the deployment of mission-critical healthcare tasks. The system uses an optimized time-sensitive task allocation approach. The results show that the optimized scheduling approach reduces task starvation by 14% and task failure by 17% compared to the fair emergency first (FEF) task allocation approach. However, to provide reliable medical treatment, the scheduling mechanism requires additional vital sign data.

A task scheduling technique called HealthEdge was proposed in [20], based on data gathered about human health status. HealthEdge evaluates whether a task should be carried out locally or remotely in the cloud, and assigns separate processing priorities for different tasks. Its aim is to reduce the total processing time when dealing with health emergencies in smart homes for healthcare. The authors developed an optimization problem for balancing task allocation between edge workstations and remote cloud datacenters. However, the study did not provide information about the communication protocols used.

A priority mechanism called the Dynamic Transmission Mechanism-L (DTM-L) was proposed in [5] for multi-class delay-sensitive medical packet transmissions. DTM-L integrates a delay control approach for classifying beyond-WBAN traffic into distinct packet priorities. However, the delay control approach can introduce more delays for lower priority medical packets, resulting in poor overall performance of DTM-L. There is a need to modify the DTM-L scheme to create a balance between complexity and performance, and to enable a mechanism that differentiates packets based on priorities.

A new cloud scheduling architecture called IADA was presented in [21]. It aimed to improve upon previous methods by using a dynamic classification scheme for workload variations instead of a segmented classification. This approach utilizes resources more efficiently and ensures compliance with Quality of Service requirements through the use of machine learning techniques, heuristics, and a Bayesian changepoint detection algorithm for real-time analysis. However, the paper did not address how to physically place virtual machines to minimize performance degradation and meet QoS requirements.

An incentive-compatible scheduling method for electronic healthcare networks with delay-sensitive medical packets was presented by Yi et al. in [22]. The proposed system sends transmission requests to the base station with their delay requirements, and the transmission requests arrive arbitrarily at each gateway. The base station then employs a priority queue to

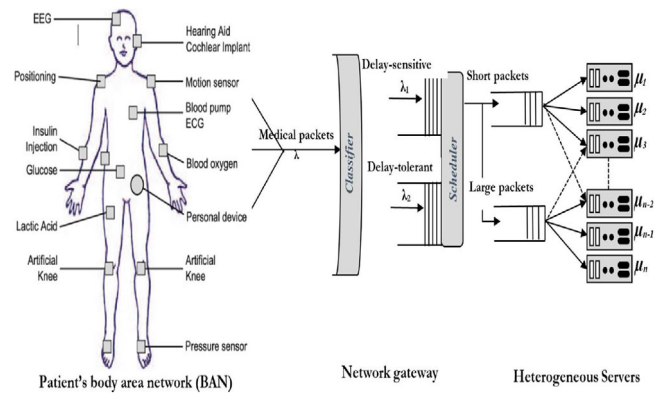


FIGURE 1 System model.

determine the transmission order, with requests having a higher delay sensitivity being given a higher queuing priority. The transmission of existing, lower-priority medical packets in the system can be interrupted by newly incoming medical packets with a higher priority. If no higher priority packets are present, the preempted packets resume service. To schedule the service, a $M/D/K$ queue with medical-grade priority was used, and the mathematical analysis of the packet waiting time was performed. However, the lower priority packets may be starved due to the high arrival rate of higher priority packets.

The existing studies on scheduling policies for allocating services in healthcare monitoring systems have several shortcomings such as poor performance under overloaded conditions, starvation of low priority traffic under high arrival rate of high priority traffic, and the algorithm being static (not catering for changes in priority of tasks). This paper suggests an analytical model for evaluating the effectiveness of healthcare monitoring systems while taking into account various latency requirements, packet sizes, and the heterogeneity of the servers. This approach is a contrast to the previous studies described in the literature.

3 | SYSTEM MODEL AND ANALYSIS

This section presents the system model and performance analysis of the proposed PS scheme against the EDF scheme.

3.1 | 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 [23].

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.

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].

On the other hand, medication dispenser data, home tele-monitoring, access to a patient's electronic health records etc. are some examples of delay tolerant packets [24]. 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.

Assumption. 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 λ_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 [23].
- Each server's service times follow an exponential distribution with parameter μ_i ; $i = 1, 2, \dots, c$, in which case μ_i is the service rate of server M_i [23].
- The service is offered via a variety of c heterogeneous servers.
- Each server has infinite capacity [26].

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.

3.2 | Prioritized scheduling scheme

Priority awareness is the most crucial criterion when scheduling the service of multiclass healthcare packets that possess various levels of urgency [25]. 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 Figure 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 [26].

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

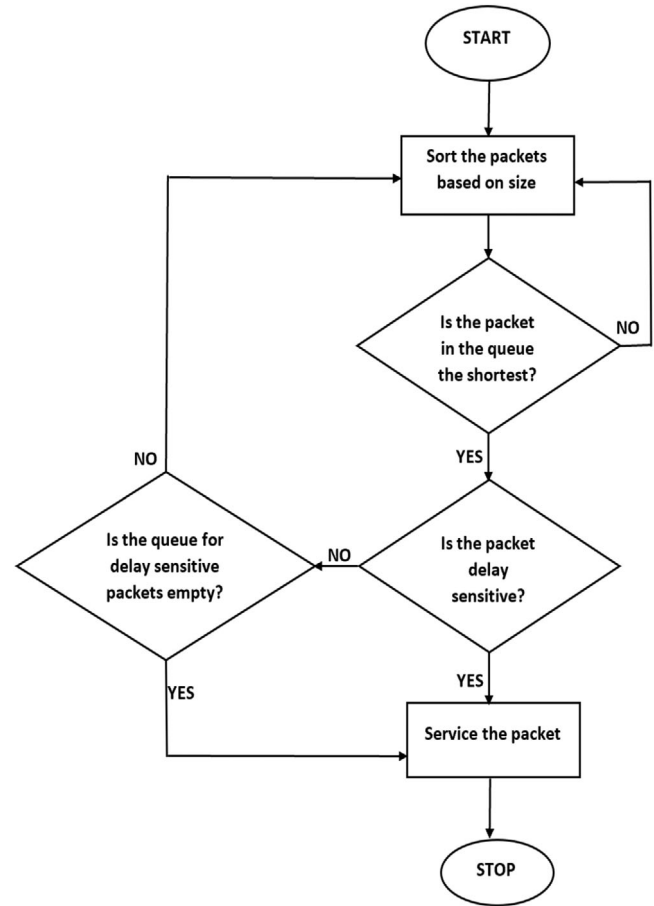


FIGURE 2 Flow diagram showing the working of the PS scheme.

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, 13].

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

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

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.

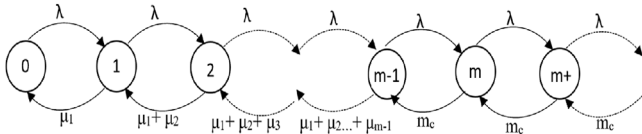


FIGURE 3 State transition diagram.

3.3 | Mathematical background

This study assumes that the servers are ordered in decreasing service rate, that is, $\mu_1 > \mu_2 > \dots > \mu_c$. 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].

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

Equation (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.

For $i = 1$; the system is in state 1, only one packet is available, and the fastest server is used for processing. For $i = 2$; the system is in state 2 and has two packets present, and the system uses two servers to provide service. The service rate for the above scenario is provided in [9] as.

$$M_i = \begin{cases} 0 & i = 0 \\ \mu_1 & i = 1 \\ \mu_1 + \mu_2 & i = 2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \mu_1 + \mu_2 + \dots + \mu_c & i \geq c \end{cases} \quad (3)$$

The $M/M_i/c$ state transition diagram is shown in Figure 3.

Using the state transition diagram in Figure 3, the probability that a newly incoming packet will be delayed because every server is busy may be generalized as [9]:

$$P_c = \frac{\lambda^c}{\mu_1(\mu_1 + \mu_2)(\mu_1 + \mu_2 + \mu_3) \dots + m_c}, \quad (4)$$

where $m_c = \mu_1 + \mu_2 + \dots + \mu_c$. From 4, it can be noted that μ_1 occurs c times and has the highest effect on the generalized probability. In the same way, μ_2 has the second highest effect on the probability since it occurs $(c - 1)$ times, and μ_c occurs once and has the least effect on the probability. The probability of having n packets present may be stated as

$$P_n = \begin{cases} \frac{\lambda^n}{\pi_{i=1}^n (\sum_{j=1}^i \mu_j)} P_0 & n < c \\ \frac{\lambda^n}{(\pi_{i=1}^c m_i) (m_c)^{n-c}} P_0 & n \geq c \end{cases}, \quad (5)$$

where $m_i = \sum_{k=1}^i \mu_k$ and $m_c = \sum_{k=1}^c \mu_k$. Using the fact that the sum of the probabilities P_n is 1,

$$\left[\sum_{n=0}^{c-1} \frac{\lambda^n}{\pi_{i=1}^n (\sum_{j=1}^i \mu_j)} + \sum_{n=c}^{\infty} \frac{\lambda^n}{(\pi_{i=1}^c m_i) (m_c)^{n-c}} \right] P_0 = 1,$$

making P_0 the subject,

$$P_0^{-1} = \sum_{n=0}^{c-1} \frac{\lambda^n}{\pi_{i=1}^n (\sum_{j=1}^i \mu_j)} + \left(\frac{m_c}{\pi_{i=1}^c m_i} \right) \sum_{n=c}^{\infty} \left(\frac{\lambda}{m_c} \right)^n, \quad (6)$$

but $\sum_{n=c}^{\infty} \left(\frac{\lambda}{m_c} \right)^n = (1 - \rho)^{-1} \rho^c$

$$P_0^{-1} = \sum_{n=0}^{c-1} \frac{\lambda^n}{\pi_{i=1}^n (\sum_{j=1}^i \mu_j)} + \left(\frac{m_c}{\pi_{i=1}^c m_i} \right) (1 - \rho)^{-1} \rho^c. \quad (7)$$

The average number of packets in waiting or in execution can be determined as follows:

$$\begin{aligned} N_q &= \sum_{n=c}^{\infty} (n - c) P_n \\ &= \sum_{n=c}^{\infty} (n - c) P_0 \frac{\lambda^n}{(\pi_{i=1}^c m_i) (m_c)^{n-c}} \\ &= \frac{P_0 (m_c)^c \rho^c}{(\pi_{i=1}^c m_i)} \sum_{n=c}^{\infty} (n - c) \rho^{n-c}, \end{aligned}$$

but $r = n - c$

$$\begin{aligned} N_q &= \frac{P_0 (m_c)^c \rho^c}{(\pi_{i=1}^c m_i)} \sum_{r=0}^{\infty} r \rho^r \\ &= \frac{P_0 (m_c)^c \rho^{c+1}}{(\pi_{i=1}^c m_i)} \sum_{r=0}^{\infty} r \rho^{r-1}, \end{aligned}$$

since $\sum_{r=0}^{\infty} r \rho^{r-1} = \frac{1}{(1 - \rho)^2}$

$$N_q = \frac{P_0 (m_c)^c \rho^{c+1}}{(\pi_{i=1}^c m_i) (1 - \rho)^2}. \quad (8)$$

The mean waiting time experienced by a packet can be deduced using Little's Law [27] as

$$W_x = \frac{P_0 (m_c)^c \rho_x^{c+1}}{\lambda (\pi_{i=1}^c m_i) (1 - \rho_x)^2}, \quad (9)$$

where ρ_x is the load due to packets of size x . Using 7, the throughput Th can be formulated as

$$Ib = \sum_{i=1}^c (1 - P_o),$$

where P_o is given as

$$P_o = \left[\sum_{n=0}^{c-1} \frac{\lambda^n}{\pi_{i=1}^n (\sum_{j=1}^i \mu_j)} + \left(\frac{m_c^c}{\pi_{i=1}^c m_i} \right) \frac{\rho^c}{(1 - \rho)} \right]^{-1}. \quad (10)$$

Using a simplistic definition of packet size based on the potentially dynamic threshold x_t , all packets with size less than or equal to x_t are referred to as short, while those with size more than x_t are referred to as large. The load caused by packets with size less than or equal to x_t is given as $\rho_{x_t} = \lambda \int_0^{x_t} tf(t)dt = \frac{\lambda}{\mu}(1 - e^{-\mu x_t}) - x_t e^{-\mu x_t}$ [27], where the load due to packets having size greater than x_t is given as $\rho_{x_t} = \lambda \int_{x_t}^{\infty} tf(t)dt = \lambda e^{-\mu x_t} (x_t + \frac{1}{\mu})$.

The expressions for the mean response time under the EDF and Priority EDF policies are then defined, and the EDF and Priority EDF policies are 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 [26].

$$W_s = W_o + \rho_s W_s + \rho_d \max(0, W_d - D_{d,s}), \quad (11)$$

$$W_d = W_o + \rho_s W_s + \rho_d W_d + \rho_s \min(W_d, D_{d,s}), \quad (12)$$

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, $W_o = \frac{\sum_{i=1}^2 \lambda_i E(x_i^2)}{2}$.

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.

On the other hand, under the Priority EDF scheme, the server processes packets having the smallest deadline among all of the waiting packets, however, for non-preemptive priority queue, the server gives service to the packet with the smallest deadline among all of the waiting packets. For a two priority class, class 2 packets have no chance to be processed before any packet of class 1. The mean waiting time for class 1 under the Priority EDF scheme is given in [10].

$$W_1 = \frac{W_o}{(1 - \rho_1)}, \quad (13)$$

where ρ_1 is the load due to packets of class 1. $W_o = \frac{\sum_{i=1}^N \lambda_i E(x_i^2)}{2}$, where N is the number of classes of packets considered.

The mean waiting time for class 2 under the Priority EDF scheme is given as.

$$W_2 = \frac{W_o}{(1 - \rho_1)(1 - \rho_1 - \rho_2)}, \quad (14)$$

where ρ_2 is the load due to packets of class 2.

3.3.1 | Models for delay sensitive packets under the PS scheme

Consider a short packet that is marked as delay-sensitive and arrives to a short packet-only delay-sensitive queue. All delay-sensitive short packets found in the queue will delay the short packet that is tagged. The average waiting time for the tagged short delay-sensitive packet of size x_s is given as

$$W(x_{ss}) = \frac{P_{o_{ss}} m_c^c \rho_{x_{ss}}^{c+1}}{\lambda_1 (\pi_{i=1}^c m_i) (1 - \rho_{x_{ss}})^2}, \quad (15)$$

where

$$\rho_{x_{ss}} = \lambda_1 \int_0^{x_{ss}} tf(t)dt = \frac{\lambda_1}{m_c} (1 - e^{-m_c x_{ss}}) - x_t e^{-m_c x_{ss}},$$

and

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

Similarly, a tagged large delay-sensitive packet is delayed by delay-sensitive short packets found in the queue, in addition to large delay-sensitive packets found in the queue. Additionally, the tagged large delay-sensitive packet is serviced after all delay-sensitive short packets that arrive after it in the queue. The mean waiting time for the delay-sensitive large packet of size x_l is given as

$$\overline{W(x_{ls})} = 2W(x_{ss}) + W(x_{ls}), \quad (16)$$

where $W(x_{ss})$ is as given in (15) and

$$W(x_{ls}) = \frac{P_{o_{ls}} (m_c)^c \rho_{x_{ls}}^{c+1}}{\lambda_1 (\pi_{i=1}^c m_i) (1 - \rho_{x_{ls}})^2}. \quad (17)$$

Also,

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

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

3.3.2 | Model for delay tolerant packets under the PS scheme

Considering a newly arrived tagged delay tolerant packet that joins a queue for delay tolerant packets.

If the tagged packet is a short delay-tolerant packet, all delay-sensitive short and large packets, and all short delay-tolerant packets found in the queue would delay the service of the tagged short delay-tolerant packet. Additionally, the short delay-tolerant packet will be delayed by all short and large delay-sensitive packets that follow the tagged short delay-sensitive packet in the queue. Before servicing the tagged short delay-tolerant packet, the short and large delay-sensitive packets that arrive after the tagged delay-sensitive short packet is in the queue will be serviced. The average waiting time for the delay-tolerant short packet of size x_{sd} is then expressed as:

$$\overline{W}(x_{sd}) = 2W(x_{ss}) + 2W(x_l) + W(x_{sd}), \quad (18)$$

where $W(x_{ss})$ and $W(x_l)$ are as given in (15) and (17), respectively. Here,

$$W(x_{sd}) = \frac{P_{osd}(m_c)^c \rho_{x_{sd}}^{c+1}}{\lambda_2 (\pi_{i=1}^c m_i) (1 - \rho_{x_{sd}})^2}, \quad (19)$$

and

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

where $\rho_{x_{sd}} = \lambda_2 \int_0^{x_{sd}} tf(t) dt$.

Consider a recently arrived, tagged, large delay-tolerant packet. 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. Additionally, short and large delay sensitive packets that arrive after the tagged large delay tolerant packet in the queue will be serviced before the tagged large delay tolerant packet. The average waiting time for the delay tolerant large packet of size x_{ld} can be expressed as:

$$\overline{W}(x_{ld}) = 2W(x_{ss}) + 2W(x_l) + W(x_{sd}) + W(x_{ld}), \quad (20)$$

where $W(x_{ss})$, $W(x_l)$ and $W(x_{sd})$ are as given in (15), (17) and (19), respectively. Here

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

and

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

where $\rho_{x_{ld}} = \lambda_2 \int_0^{x_{ld}} tf(t) dt$.

The developed models are evaluated in terms of mean slowdown as the performance metric in the next section.

TABLE 1 Implementation parameters.

Parameter	Value
Number of servers, m	5 [27]
Packet arrival rate, λ	0 to 6.549 packets/second [28]
Service rate for the multi-servers, μ	1,2,3,4,5 packets/second [27]
Low system load, ρ_l	0.5 [26]
High system load, ρ_h	0.9 [26]
$D_{2,1}$ is the difference between deadlines for delay sensitive and delay tolerant packets	3[26]
d_1 is the constant deadline offset for delay sensitive packets	2[26]
Average packet size, x_s	100 Kb[8]
Threshold of the packet size, x_{ts}	75 Kb [8]

4 | PERFORMANCE EVALUATION

The performance of the developed IoT-based healthcare monitoring model is evaluated in this section using Matlab. The primary metrics of interest in this study are mean slowdown and throughput. Mean slowdown is often used as a measure of system performance instead of mean response time because it takes into account both the packet's response time and its processing requirements [16]. Furthermore, as larger packets contribute more to the mean and often have higher response times, the mean response time is more representative of the performance of a few packets. In contrast, mean slowdown can only be significantly improved if the slowdown of a larger portion of all packets is affected. On the other hand, throughput measures the proportion of the total service rate when the system is busy.

This study compares the performance of the PS, EDF and Priority EDF scheduling schemes for both short and large packets. By comparing the performance of packets under the PS, EDF and Priority EDF schemes, the study considers the use of heterogenous servers. The study conducts investigations to determine the impact of key parameters, such as packet size and load, on the mean slowdown and throughput.

4.1 | Implementation parameters

In this section, the implementation parameters are presented.

The hypothetical parameters used in the study are shown in Table 1. These parameters are consistent with those used in the literature [8, 26–28].

4.2 | Evaluation of the mean slowdown of packet sizes for delay sensitive packets

Figure 4 shows the variation of mean slowdown of delay-sensitive short packets with packet size under the EDF, Priority EDF and PS schemes, where short packets have sizes less than or equal to $x_s = 75$ Kilobytes. Expressions (11), (13) and (15)

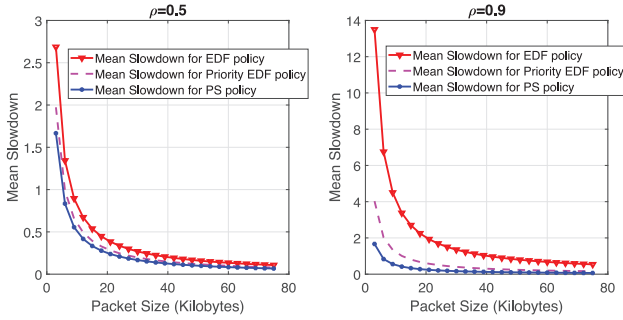


FIGURE 4 Mean slowdown vs packet size for delay sensitive short packets.

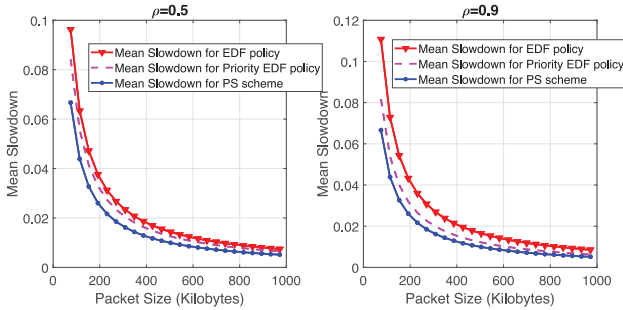


FIGURE 5 Mean slowdown vs packet size for delay sensitive large packets.

are used to obtain Figure 4. It can be observed that, at both low system load values of $\rho = 0.5$ and high system load of $\rho = 0.9$, delay-sensitive short packets experience lower mean slowdown under the PS scheme in comparison to the EDF and Priority EDF schemes. In all cases, it is clear that as packet size increases, the performance of delay-sensitive short packets under PS, Priority EDF and EDF schemes become closer. Additionally, it can be observed that the performance under the PS scheme of delay-sensitive short packets is much better compared to EDF and Priority EDF at high system loads. By favoring high-priority packets under the EDF and Priority EDF schemes, the mean slowdown of lower-priority packets is increased.

Similarly, Figure 5 shows the variation of mean slowdown with packet size for large, delay-sensitive packets for the EDF, Priority EDF and PS schemes, where large packets are packets with sizes greater than $x_s = 75$ Kilobytes. Expressions (12), (11) and (16) are used to obtain Figure 5. It can be observed that delay-sensitive large packets exhibit better performance under the PS scheme compared to the EDF and Priority EDF schemes for all the considered load values. In all cases, it can be further observed that the performance of delay-sensitive large packets for PS, EDF and Priority EDF is closer at low load values of $\rho = 0.5$, but there is a noticeable difference in performance at high load value of $\rho = 0.9$.

4.3 | Evaluation of the mean slowdown of packet sizes for delay tolerant packets

In this section, the performance of the PS scheduling scheme is compared to that of the EDF and Priority EDF scheduling

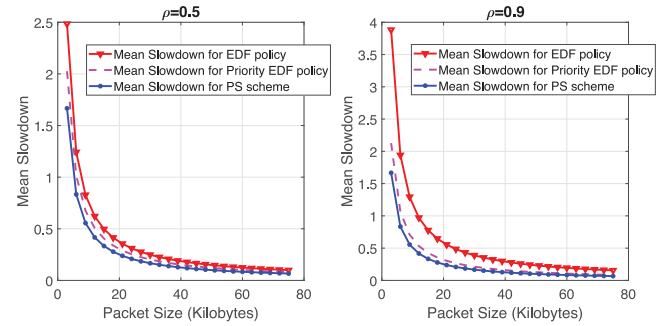


FIGURE 6 Mean slowdown vs packet size for delay tolerant short packets.

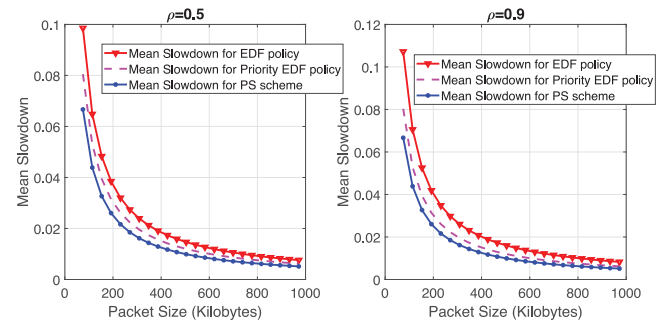


FIGURE 7 Mean slowdown vs packet size for delay tolerant large packets.

schemes for delay-tolerant packets in terms of mean slowdown. The study examines the effect of packet size on the mean slowdown of delay-tolerant packets under PS, EDF and Priority EDF schemes.

Figure 6 illustrates the mean slowdown of delay-tolerant short packets under the EDF, Priority EDF and PS scheduling schemes, where short packets have sizes of less than or equal to $x_s = 75$ Kilobytes. To obtain this figure, expressions (12), (14) and (18) were used. It can be observed that, at both low system load values of $\rho = 0.5$ and high system loads of $\rho = 0.9$, the PS scheme outperforms the EDF and Priority EDF schemes for delay-tolerant short packets. It can also be seen that as packet sizes increase, the performance of delay-tolerant short packets under PS, Priority EDF and EDF schemes become more similar. Additionally, it can be observed that the performance of delay-sensitive short packets under the PS scheme is much better compared to the EDF and Priority EDF schemes at high system load $\rho = 0.9$.

Figure 7 illustrates the mean slowdown of the delay-tolerant large packets under the EDF, Priority EDF and PS schemes, where large packets are packets with sizes greater than $x_s = 75$ Kilobytes. The results for Figure 7 were obtained by using expressions (12), (14) and (20). At low system load values of $\rho = 0.5$, it can be observed that delay-tolerant large packets perform slightly better under the PS scheme compared to the EDF scheme. However, the performance of delay-tolerant large packets under the PS scheme is significantly better than the EDF and Priority EDF schemes for high system load values of $\rho = 0.9$. It can also be seen that the performance of the PS scheme is even higher for smaller delay-tolerant large packet sizes.

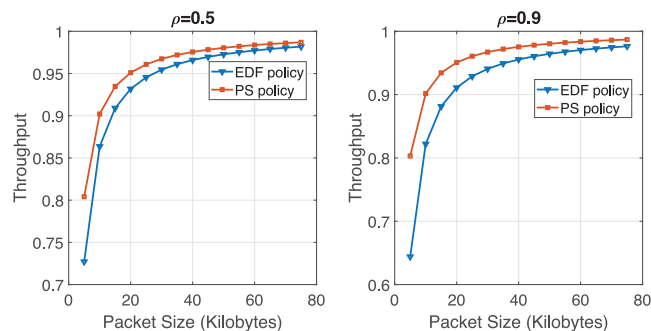


FIGURE 8 Throughput vs packet size for delay sensitive short packets.

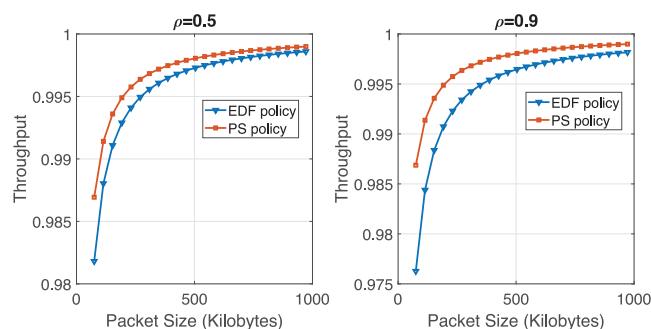


FIGURE 9 Throughput vs packet size for delay sensitive large packets.

4.4 | Evaluation of throughput of packet sizes for delay sensitive packets

In this section, the performance of the PS scheduling scheme in terms of throughput for delay-sensitive packets is compared to that of the EDF scheduling scheme in order to investigate the impact of varying packet sizes on throughput.

Results for throughput of PS in comparison to EDF for delay-sensitive short packets are shown in Figure 8. It can be shown that increasing packet sizes results in higher throughput for both low and high load values of $\rho = 0.5$ and high load, $\rho = 0.9$, respectively. Additionally, it can be seen that the PS scheme outperforms the EDF scheme for delay-sensitive packets.

Figure 9 compares the performance of the PS and EDF schemes in terms of throughput for large, delay-sensitive packets. It can be observed that as packet sizes increase, the throughput also increases. Furthermore, it is clear that the PS scheme outperforms the EDF scheme in terms of throughput, both at low load (with a value of $\rho = 0.5$) and high load (with a value of $\rho = 0.9$). Additionally, it is demonstrated that under high load, the PS scheme performs even better than the EDF scheme. This is due to the more frequent interruptions of the service of delay-sensitive large packets that occur at high load values under the EDF scheme, which is not the case under the PS scheme.

In the next section, the discussion of results is presented.

5 | DISCUSSIONS

In this paper, analytical models are developed to examine the mean slowdown of the PS scheme, where incoming packets are served by multiple heterogeneous servers and are assigned priorities based on their size and delay. The performance of these models is evaluated using mean slowdown and throughput as the performance metrics, and is compared to that of the EDF and Priority EDF scheduling scheme.

This involves examining the effect of packet size variation and load on the mean slowdown. Numerical results from the derived models demonstrate that for both low and high load values, delay sensitive packets (both short and big) perform better under the PS scheme compared to the EDF and Priority EDF schemes. It is further noted that PS scheme performs much better than EDF and Priority EDF schemes at high load values compared to low load values. The worse performance under the EDF scheme especially at high load is because EDF gives highest priority to packets close to their deadlines, leading to delays for other packets that still have the capacity to meet their deadlines.

Similar observations have been noted by a number of variants of EDF priority scheduling policies proposed in previous studies where new constraints were added to existing EDF scheme to improve performance [10, 26].

It is also observed that the throughput value increases according to packet size until it reaches the saturated value. Throughput increases as a result of increase in packet sizes due to increased amount of data sent. Similar observations were noted by the study that investigated the influence of packet size on network throughput [29].

It is further observed that delay sensitive packets perform worse under the EDF scheme than under the PS scheme in terms of throughput. This is due to the fact that, unlike in the PS scheme where delay sensitive packets are always processed prior to delay tolerant packets if any, the EDF scheme allows some delay tolerant packets to be processed before some delay sensitive packets. Similar observations have been noted by authors in [26]. The performance of PS over EDF and Priority EDF schemes is observed to be even much better at high load compared to low load values. The observation is due to the fact that at high load values, the number of packets increase causing more frequent interruptions to delay sensitive short packets from delay tolerant packets under the EDF and Priority EDF schemes, which is not the case under the PS scheme. Overall, the results of this study provide valuable insights into the impact of packet size and load on the performance of PS scheme, and demonstrate the potential benefits of using the PS scheme as an alternatives to the EDF and Priority EDF schemes for packet scheduling in multi-server systems.

6 | CONCLUSION AND FUTURE WORK

With varying packet sizes and load, the PS system has been modeled and evaluated. The results show that the PS scheme generally reduces the mean slowdown for both delay-sensitive

and delay-tolerant packets at both low and high load values. As a result, the PS scheme performs better than the EDF and Priority EDF systems for both short and large packets at all load values. It has also been observed that delay-sensitive packets perform better under the PS scheme compared to the EDF scheme in terms of throughput. Additionally, the throughput is observed to increase as the packet size increases until it reaches a saturated value. In future research, we will investigate the effect of using a threshold on the number of delay-sensitive packets served before serving delay-tolerant packets at high arrival rates of delay-sensitive packets, in addition to implementing a threshold on packet sizes. Furthermore, we intend to validate the proposed analytical model using simulations.

AUTHOR CONTRIBUTIONS

Barbara Asingwire: Conceptualization, formal analysis, methodology, project administration, resources, validation, visualization, writing - original draft, writing - review and editing. Louis Sibomana: Conceptualization, methodology, supervision, writing - review and editing. Alexander Ngenzi: Conceptualization, methodology, supervision, writing - review and editing. Charles Kabiri: Conceptualization, methodology, supervision, writing - review and editing.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

FUNDING INFORMATION

This research was not supported by funding.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Barbara Kabwiga Asingwire  <https://orcid.org/0000-0003-3668-3854>

REFERENCES

- Zhang, H., Li, J., Wen, B., Xun, Y., Liu, J.: Connecting intelligent things in smart hospitals using NB-IoT. *IEEE Internet Things* 5(3), 1550–1560 (2018)
- Bhatia, H., Panda, S.N., Nagpa, D.: Internet of Things and its applications in healthcare: A survey. In: *Proceedings of International Conference on Reliability, Infocom Technologies and Optimization*. IEEE, Piscataway (2020)
- Yi, C., Cai, J.: Transmission management of delay-sensitive medical packets in beyond wireless body area networks: A queuing game approach. *IEEE Trans. Mob. Comput.* 17(9), 2209–2222 (2018)
- Gomes, E., Dantas, M.A.R., Plentz, P.: A real-time fog computing approach for healthcare environment. In: *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, pp. 85–95. Springer, Cham (2019)
- Yi, C., Cai, J.: A priority-aware truthful mechanism for supporting multi-class delay-sensitive medical packet transmissions in e-health networks. *IEEE Trans. Mob. Comput.* 16(9), 2422–2435 (2017)
- Asingwire, B.K., Ngenzi, A., Sibomana, L., Kabiri, C.: Performance analysis of IoT-based healthcare heterogeneous delay-sensitive multi-server priority queuing system. *Int. J. Adv. Comp. Sci. Appl.* 12(10), 666–673 (2021)
- Rahmani, A.M., Gia, T.N., Negash, B., Anzanpour, A., Azimi, I., Jiang, M., Liljeberg, P.: Exploiting smart e-Health gateways at the edge of healthcare Internet-of-Things: A fog computing approach. *Future Gener. Comput. Syst.* 78(2), 641–658 (2018)
- Changyan, Y., Cai, J.: A truthful mechanism for scheduling delay-constrained wireless transmissions in IoT-based healthcare Networks. *IEEE Trans Mob Comput.* 18(2), 912–925 (2018)
- Narman, H.S., Hossain, M., Atiquzzaman, M., Shen, H.: Scheduling Internet of Things applications in cloud computing. *Ann. Telecommun.* 72, 79–93 (2017)
- Mukakanya Muwumba, A., Justo, G.N., Massawe, L.V., Ngubiri, J.: Priority EDF scheduling scheme for MANETs. *Communications and Networking*. ChinaCom 2019. *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, vol. 312, pp. 66–76. Springer, Cham (2020)
- Deepika, N., Anand, M., Sudhaman, K.: Internet connected e-healthcare system with live video monitoring using LWIP stack and SJF priority scheduling. *Int. J. Recent Technol. Eng.* 8(4), 3362–3368 (2019)
- Ma, X., Wang, Z., Zhou, S., Wen, H., Zhang, Y.: Intelligent healthcare systems assisted by data analytics and mobile computing. *Wireless Commun. Mobile Comput.* 2018, 3928080 (2018)
- Yi, C., Cai, J.: A truthful mechanism for scheduling delay-constrained wireless transmissions in IoT-based healthcare networks. *IEEE Trans. on Wireless Communications* 17(9), 912–925 (2019)
- Shukl, S., Hassan, M.F.F., Khan, M.K., Jung, L.T., Awang, A.: An analytical model to minimize the latency in healthcare Internet of things in fog computing environment. *PLoS ONE* 14(11), 1–31 (2019)
- Efrosinin, D., Stepanova, N., Sztrik, J.: Algorithmic analysis of finite-source multi-server heterogeneous queueing systems. *MDPI J. Math.* 9(20), 2–24 (2021)
- Okopa, M., Turatsinze, D., Bulega, T., Wampande, J.: Revenue maximization based on slowdown in cloud computing environments. *Australasian J. Comp. Sci.* 4, 1–16 (2017)
- Mankar, P.D., Chen, Z., Abd-Elmagid, M.A., Pappas, N., Dhillon, H.S.: Throughput and age of information in a cellular-based IoT network. *IEEE Trans. Wireless Commun.* 20(12), 8248–8263 (2021)
- Zhong, L., He, S., Lin, J., Wu, J., Li, X., Pang, Y., Li, Z.: Technological requirements and challenges in wireless body area networks for health monitoring: A comprehensive survey. *Sensors* 22(9), 2–22 (2022)
- Iqbal, N., Imran, Ahmad, S., Ahmad, R., Kim, D.: A scheduling mechanism based on optimization using IoT-tasks orchestration for efficient patient health monitoring. *J. Sens.* 21(16), 5430 (2021)
- Wang, H., Gong, J., Zhuang, Y., Shen, H., Lach, J.: Healthedge: Task scheduling for edge computing with health emergency and human behavior consideration in smart homes. In: *Proceedings of IEEE International Conference on Big Data*, pp. 1213–1222. IEEE, Piscataway (2017)
- Meyer, V., da Silva, M.L., Kirchoff, D.F., De Rose, C.A.F.: IADA: A dynamic interference-aware cloud scheduling architecture for latency-sensitive workloads. *J. Syst. Softw.* 194(C), 111491 (2022)
- Yi, C., Alfa, A.S., Cai, J.: An incentive-compatible mechanism for transmission scheduling of delay-sensitive medical packets in E-health networks. *IEEE Trans. Mob. Comput.* 15(10), 2424–2436 (2016)
- Salh, A., Audah, L., Alhartomi, M., Soon, K., Alsamhi, S.H., Almaki, F.A., Abdullahi, Q., Saif, A., Algethami, H.: Smart packet transmission scheduling in cognitive IoT systems: DDQN based approach. *IEEE Access* 10(4), 50023–50035 (2022)
- Park, K., Park, J., Lee, J.: An IoT system for remote monitoring of patients at home. *J. Appl. Sci.* 7(3), 1–23 (2017)
- Ala, A., Chen, F.: Appointment scheduling problem in complexity systems of the healthcare services: A comprehensive review. *J. Healthcare Eng.* 2022, 5819813 (2022)
- Muwumba, A.M., Justo, G.N., Massawe, L.V., Ngubiri, J.: Priority EDF scheduling scheme for MANETs. In: *Proceedings of International Conference on Communications and Networking in China*, pp. 66–76. IEEE, Piscataway (2020)
- Nansamba, B., Okopa, M., Asingwire, B.K., Kaawaase, K.S.: Pricing scheme for heterogeneous multi-server cloud computing system. *Australasian J. Comp. Sci.* 4, 32–43 (2017)

28. Majumdar, C., Lopez-Benitez, M., Merchant, S.N.: Experimental evaluation of the Poisson process of real sensor data traffic in the Internet of Things. In: Proceedings of IEEE Annual Consumer Communications & Networking Conference, pp. 1–7. IEEE, Piscataway (2019)
29. Xiong, Y., Chang, Y., Hu, M., Li, J.: Packet-size based overlapping user grouping in MU-MIMO systems. In: IEEE Wireless Communications and Networking Conference. IEEE, Piscataway (2019)

How to cite this article: Asingwire, B.K., Sibomana, L., Ngenzi, A., Kabiri, C.: Delay and size-dependent priority-aware scheduling for IoT-based healthcare traffic using heterogeneous multi-server priority queuing system. *IET Commun.* 17, 1877–1887 (2023). <https://doi.org/10.1049/cmu2.12675>