Performance Analysis of IoT-based Healthcare Heterogeneous Delay-sensitive Multi-Server Priority Queuing System

Barbara Kabwiga Asingwire[1], Alexander Ngenzi[2], Louis Sibomana[3], Charles Kabiri[4]
African Centre of Excellence in Internet of Things, College of Science and Technology, University of Rwanda[1]
Department of Computer Engineering, Busitema University, Uganda[1]
African Centre of Excellence in Internet of Things, College of Science and Technology, University of Rwanda[2, 4]
African Centre of Excellence in Internet of Things, College of Science and Technology[3]
University of Rwanda and National Council for Science and Technology, Rwanda[3]

***Abstract*—Previous studies have considered scheduling schemes for Internet of Things (IoT)-based healthcare systems like First Come First Served (FCFS), and Shortest Job First (SJF). However, these scheduling schemes have limitations that range from large requests starving short requests, process starvation that results in long time to complete if short processes are continuously added, and performing poorly under overloaded conditions. To address the mentioned challenges, this paper proposes an analytical model of a prioritized scheme that provides service differentiation in terms of delay sensitive packets receiving service before delay tolerant packets and also in terms of packet size with the short packets being serviced before large packets. The numerical results obtained from the derived models show that the prioritized scheme offers better performance than FCFS and SJF scheduling schemes for both short and large packets, except the shortest short packets that perform better under SJF than the prioritized scheme in terms of mean slowdown metric. It is also observed that the prioritized scheme performs better than FCFS and SJF for all considered large packets and the difference in performance is more pronounced for the shortest large packets. It is further observed that reduction in packet thresholds leads to decrease in mean slowdown and the decrease is more pronounced for the short packets with larger sizes and large packets with shorter sizes.**

***Keywords*—Delay tolerant; delay sensitive; internet of things; mean slowdown; prioritized scheme**

I. Introduction

The recent advances in technologies have led to the emergence of Internet of Things (IoT) [1], [2] that interconnects everything around us, including sensors, devices and systems and also supports a range of applications. IoT has been applied in several domains including but not limited to remote health monitoring [3]. IoT-enabled remote health monitoring systems have huge advantages over traditional health monitoring systems and are likely to improve the future of healthcare monitoring and emergency management.

In remote health monitoring, the IoT-based physical monitoring devices need to transmit collected data in real time, with low latency and in a highly reliable way so as to ensure accurate monitoring of patients. This is because healthcare systems are highly time-sensitive and require minimal delay.

Specifically, it is required that medical emergencies are given precedence in reporting over other regular services [4]. Further to this, transmission services for medical signals should be classified based on the different signal requirements. Besides, low latency is important for healthcare environments such that in cases of emergencies timely notification allows the medical personnel responsible to respond accordingly [5],[6].

The traditional computing server scheduling schemes are not ripe enough to provide services to IoT based healthcare services due to the heterogeneity of IoT applications and traffic which require different levels of service guarantees [7].

Healthcare IoTs may tolerate delays ranging from milliseconds to microseconds [8], [9]. Increase in the data size leads to increase in delay for the healthcare IoT applications and for time-sensitive applications the delay may vary from milliseconds to minutes [8], [10], and this worsens the performance of real time healthcare IoTs [11], [12].

While, scheduling traffic in healthcare systems, the following issues need to be addressed [13]:

1) Emergent medical situations should be given precedence in reporting than those with regular importance. This is because excessive delays in the transmission of emergent medical situations may deteriorate health services to patients. To address this issue, this study prioritizes delay sensitive packets over delay tolerant packets.

2) Transmission services for non-emergent medical situations should be differentiated by their heterogeneous delay sensitivities with regards to different application purposes. Applying absolute priority rule can maintain the transmission priorities among different medical levels, but may lead to tremendously large waiting delays for "less important" packets and yet the "less important" medical packets are also critical components of patients' health profiles. To address this issue, service differentiation is implemented, in this study, to differentiate the traffic based on the delay sensitivity of the traffic and also based on the size of each packet, with the short packets being serviced before large packets in order to improve on the number of requests served per unit time.

3) Healthcare IoT devices generate huge volumes of healthcare data which results in high data traffic that causes network congestion and high latency [18]. By servicing short packets before the large packets, the number of packets served will increase hence reducing the congestion.

Recent developments within the research community provide numerous scheduling schemes for IoT-based healthcare systems namely: First Come First Served (FCFS) [19], Shortest Job First (SJF) [24], [25], preemptive resume service priority [20]. Unfortunately, these schemes have limitations that range from large requests starving short requests [19], process starvation that results in a long time to complete if short processes are continuously added [24], to high priority requests starving lower priority requests [6].

To address the above limitations, this study formulates an analytical framework for the performance evaluation of IoT-based healthcare heterogeneous delay-sensitive multi-server priority queuing system based on the formulated packet transmission scheduling.

The contribution of this paper is two-fold. Firstly, the study developed models of mean slowdown for the prioritized scheduling scheme for IoT-based healthcare monitoring systems. Secondly, the performance of the proposed models is evaluated against the FCFS and SJF scheduling schemes. The rest of the paper is organized as follows: Section II is related work. The analytical models are presented in Section III, while Section IV presents the performance evaluation, discussions are presented in Section V, conclusion in Section VI and future work is presented in Section VII.

II. Related Work

First-Come-First-Served (FCFS) scheduling scheme applied in [19] is the simplest scheduling policy where requests are served according to their order of arrival. As a non-pre-emptive scheduling discipline, once a request has a server, it runs to completion. One of the major drawbacks of FCFS scheme is that the emergent healthcare packets are completely starved of service and this increases the average waiting time of emergent healthcare packets which may result into serious issues in healthcare including death.

Therefore, scheduling techniques that provide fairness to all competing packets is required in the allocation of resources to prevent starvation of some packets.

Preemptive resume service priority introduced in [20] is a scheduling scheme where incoming traffic are prioritized into normal and emergency traffic, where normal traffic has low priority and emergency traffic has high priority. This scheduling is based on preemptive priority mechanism where a higher priority traffic is serviced before a low priority traffic but each category of traffic is served in a FCFS order. A lower priority traffic is preempted on arrival of emergent traffic and the lower priority traffic could be dropped if the buffer is full so as not to cause data loss or delay of sensitive traffic. However, the weakness of preemptive resume service priority is that when high priority rate exhibits high arrival rate, the low priority traffic is starved. Hence, there is need to place a threshold on the amount of high priority traffic to be serviced during high arrival rate of high priority traffic so as not to starve the low priority traffic.

A priority-aware truthful mechanism for scheduling delay constrained medical packet transmissions in IoT-based healthcare networks is proposed in [13]. The study considered multiclass health packets from the biosensors arriving randomly at each gateway and their delay-constrained transmission requests are immediately reported to the base station. The base station schedules the transmissions by including the priority and the delay constraints of medical packet transmissions. However, the limitation of this scheme is that; the absolute prioritized transmission used naturally results in a non-preemptive priority queueing, where under high arrival rates of higher priority medical packets, the lower priority medical packets are starved. In addition, the servers (channels) are taken to be homogeneous implying same characteristics, which in reality is not the case being that different channels have different characteristic and can be modeled as heterogeneous servers.

In [22], a dynamic scheduling of beyond-WBAN medical packet transmissions is modeled by *M/G/K* queues with a Poisson packet arrival, generally distributed service (transmission) time and priority disciplines. The system consists of a gateway, a number of heterogeneous biosensors worn on different parts of the human body and the Base Station (BS). The BS serves the packets in a priority order with emergent medical packets being given a higher priority over those with regular importance. In scheduling, some channels are completely reserved for emergent medical packets and the balance of the channels are reserved for non-emergent channels. However, when channels are completely partitioned for each packet class, the use of the un-utilized channels of one class of packets cannot be used by other classes of users and therefore the capacity is wasted.

T. Aladwani [23] proposed to use fog computing between sensors and cloud computing to reduce the amount of data that is transported between the cloud and the sensors. In addition, the authors improved task scheduling algorithm by making the main factor in giving priority to tasks their importance regardless of their length. The authors proposed a new method of scheduling called Tasks Classification and Virtual Machines Categorization (TCVC) based on tasks importance. Tasks that are received by IoT are classified based on their importance into three classes: high importance, medium importance, and low importance tasks based on the patient's health status. In scheduling, critical tasks take high importance, important tasks take medium importance, and general tasks take low importance. The limitation of this scheme is matching the virtual machine's capability to the important of tasks, and also under high arrival rate of higher priority tasks, the lower priority tasks are starved of service.

SJF scheduling policy has been used in scheduling tasks in healthcare systems, for example, an innovative IoT based remote healthcare monitoring system by using Free RTOS with priority scheduling based on SJF is proposed in [24]. The proposed system provides vital health information and live video of a patient who is located in a rural area. A framework that utilizes the 5G network's low-latency, high bandwidth

functionality to detect COVID-19 using chest X-ray or CT scan images, and to develop a mass surveillance system to monitor social distancing, mask wearing, and body temperature using the SJF policy is proposed in [25]. The weakness of the SJF scheduling algorithm is that it gives priority to tasks based only on their length. This leads to unfairness, as the large tasks must be waiting in the tasks list until the smallest tasks finish execution even if it is important.

In summary, the limitations of the existing studies include; lack of a fair scheduling scheme that prioritizes traffic in the system without penalizing other classes of traffic, lack of scheme that caters for the dynamic changes in the periods, starvation of emergent healthcare packets, and lack of optimized frameworks and algorithms in allocation of system resources.

In contrast to the existing work reported in the literature, this study proposes an analytical model that will aid in studying and analyzing performance of healthcare monitoring systems considering different packet sizes and packet thresholds.

III. System Model

The healthcare monitoring system consists of heterogeneous healthcare monitoring data packets from different independent sensors mounted on the body to monitor different health situations. In the considered system model as shown in Fig. 1, the heterogeneous data packets generated by the different sensors arrive randomly to the network gateway following a Poisson process, the Poisson distribution has been found to approximate well the arrival patterns of healthcare data packets [26], [27].

Fig. 1 shows the queue system model with healthcare data packets generated from different sensor nodes mounted on the body.

The gateway is required to immediately declare a transmission packet along with the corresponding packet priority based on the time sensitivity of the packets, this is done at classifier 1, where requests are classified into delay sensitive and delay tolerant based on their delay requirements, for example EEG/ECG/EMG has delay requirement of less than 250ms, Glucose monitoring less than 20ms, Blood pressure less than 750ms, Endoscope imaging less than 500ms [13]. Examples of delay tolerant traffic include access to a patient's Electronic Health Records; home tele-monitoring, medication dispenser data, etc. [14].

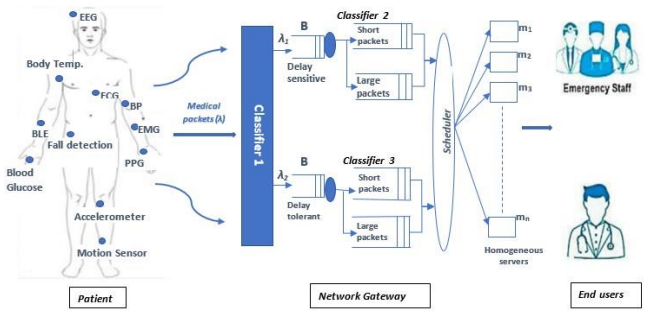

Fig. 1. Queue System Model.

A major requirement in scheduling transmissions of multiclass healthcare packets with different criticality is the priority awareness [28]. For each queue of the delay sensitive or delay tolerant classes, packets are queued in the buffers assumed to be infinite. For each of the delay sensitive and delay tolerant classes, the data packets are further classified as short or large based a threshold. Short packets are chosen for the next execution before the large packets, the idea being to reduce the average waiting time for other packets awaiting execution. After classifying the packets by their sizes, the packets are forwarded to the scheduler which allocates the packets to the different servers. This scheduling scheme considers shared servers for each priority class. Considering the diversities in terms of packet sizes, the transmission time of healthcare packets can be represented by a generic random variable, that is, follows the general service distribution [13]. In particular, the service rate of packets will follow the exponential distribution [21]. The probability density function of an exponential distribution is given as [13]:

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

where μ is the service rate and x is the size of the packet. The proposed policy is a delay sensitive non-preemptive size-based scheduling policy where packets are classified as delay sensitive or delay tolerant at the first priority level and also on their sizes, namely (x_s) and large (x_l). For each delay sensitive or delay tolerant classes, short packets are served before large packets. Within each class, packets are served in a FCFS order using multiple servers. The system model can be represented as a multi-server queue. For each queue of the delay sensitive or delay tolerant classes, packets are queued in the buffers assumed to be infinite. The queue model can be formulated under the following assumptions:

The arrival rate follows the Poisson process with parameter λ_i, i = 1,2, where λ_1 is the arrival rate of delay sensitive packets and λ_2 is the arrival rate of delay tolerant packets.

The service times of each server is independent and identically distributed exponential random variable with.

parameter μ_i, i = 1,2, where μ_1 is the service rate of delay sensitive packets and μ_2 is the service rate of delay tolerant packets.

There are m servers through which the service is provided.

The capacity of each server is finite, N.

The above system can be represented as an *M/M/m/N* queue system, where the first *M* represents random arrivals of packets following the Poisson process, the second *M* represents exponentially distributed service time, with *m* servers each of finite capacity *N*.

A. Mathematical Background

Denote the probability density function of a packet of size x as $f(x)$ defined in equation 1. The cumulative distribution function is then given as: $F(x) = \int_o^x f(t)dx$.

Using a naive definition of packet size based on threshold x_t which may be dynamic, all packets that have sizes less than or equal to x_t are said to be short, whereas packets that are larger than x_t are said to be large.

The load due to packets with sizes less than or equal to x_t is given as $\rho_{x_t} = \lambda \int_0^{x_t} tf(t)dt = \frac{\lambda}{\mu}(1 - e^{-\mu x_t}) - x_t e^{-\mu x_t}$ [15], where μ is the service rate of packets, while the load due to packets with sizes greater than x_t is given as.

$$\rho_{xl} = \lambda \int_{x_t}^{\infty} tf(t)dt = \lambda e^{-\mu x_t}\left(x_t + \frac{1}{\mu}\right)$$

The steady state equations of the *M/M/m/N* queue model are derived as follows:

The probability that there are packets in the system is given as [17]:

$$P_n = \begin{cases} \frac{\rho^n}{n!}P_o, 1 \le n \le m \\ \frac{\rho^n}{m!}\left(\frac{\rho}{m}\right)^{n-m} P_o, m < n \le N-1 \end{cases} \quad (2)$$

where P_o is the probability that the system is empty and is given by;

$$P_o = \left[\sum_{n=0}^{m} \frac{\rho^n}{n!} + \sum_{n=m+1}^{N} \frac{\rho^n}{m!}\left(\frac{\rho}{m}\right)^{n-m}\right]^{-1} \quad (3)$$

The expected waiting time in the queue can be deduced as

$$W_q = \frac{1}{\lambda}\sum_{n=m}^{N}(n-m)\,P_n \quad (4)$$

Hence,

$$W_q = \sum_{n=m}^{N}(n-m)\frac{\rho^n}{\lambda m!}\left(\frac{\rho}{m}\right)^{n-m} P_o \quad (5)$$

We next define the expressions for the mean response time under FCFS and SJF, which will be used to compare with the prioritized scheduling scheme. An arriving packet to the FCFS queue has to wait for all packets it finds in the queue upon arrival. The mean response time of a packet of size x_s in an M/G/m/FCFS system is given as [15].

$$T^{FCFS}(x_s) = x_s + W^{FCFS}(x_s) \quad (6)$$

where $W^{FCFS}(x_s) = \frac{\overline{\lambda x_s^2}}{2(1-\rho_{x_s})}$ and $\rho_{x_s} = \frac{\lambda}{m\mu}$

Under SJF, the shortest packet in the queue is given priority. Therefore, at every instant, the next packet to be serviced is the smallest one in the queue. A packet of size x_s is then delayed by packets in the system that is less or equal than its size. The mean response time of the packet of size x_s under SJF is given as [15].

$$T^{SJF}(x_s) = x_s + W^{SJF}(x_s) \quad (7)$$

where $W^{SJF}(x_s) = \frac{\overline{\lambda x_s^2}}{2(1-\rho_{x_s})^2}$ and $\rho_{x_s} = \frac{\lambda}{m\mu}$, m being the number of servers.

B. Model for Delay Sensitive Packets

Consider a tagged packet arriving to a delay sensitive queue, two scenarios arise, the first scenario is when the tagged packet finds in the queue short delay sensitive packets being serviced, including at least one delay sensitive large packet, the second scenario includes the tagged packet arriving to a delay sensitive queue with only short packets. We consider scenario one where at least one delay sensitive large packet is found in service.

Assuming the tagged delay sensitive short packet, its service will be delayed by all delay sensitive short packets it finds in the queue and the remaining service of the large packets it finds in the servers when it arrived. The mean response time for the delay sensitive short packet of size x_s is given as [15]:

$$T(x_{ts}) = x_{ts} + W(x_{ts}) + W_r(x_{ls}) \quad (8)$$

where

$$W(x_{ts}) = \sum_{n=m}^{N}(n-m)\frac{\rho_{x_{ts}}^n}{\lambda_1 m!}\left(\frac{\rho_{x_{ts}}}{m}\right)^{n-m} P_0^{x_{ts}} \quad (9)$$

and

$$P_o^{x_{ts}} = \left[\sum_{n=0}^{m}\frac{\rho_{x_{ts}}^n}{n!} + \sum_{n=m+1}^{N}\frac{\rho_{x_{ts}}^n}{m!}\left(\frac{\rho_{x_{ts}}}{m}\right)^{n-m}\right]^{-1} \quad (10)$$

$$\rho_{x_{ts}} = \lambda_1 \int_0^{x_{ts}} tf(t)dt$$

$$W_r(x_{ls}) = \sum_{n=m}^{N}(n-m)\frac{\rho_{x_{ls}}^n}{\lambda_1 n!}P_o^{x_{ls}} \quad (11)$$

Where

$$P_0^{x_{ls}} = \left[\sum_{n=o}^{m}\frac{\rho_{x_{ls}}^n}{n!} + \sum_{n=m+1}^{N}\frac{\rho_{x_{ls}}^n}{m!}\left(\frac{\rho_{x_{ls}}}{m}\right)^{n-m}\right]^{-1} \quad (12)$$

and $\rho_{x_{ls}} = \lambda_1 \int_{x_{ts}}^{\infty} tf(t)dt$

On the other hand, the delay sensitive large packet is delayed by all delay sensitive short packets found in the queue plus all delay sensitive large packets found in the queue, and the mean service time of the large packets the tagged large packet finds in the servers when it arrived. In addition, all delay sensitive short packets that arrive after the tagged large packet is in the queue will be served before the tagged large packet. The mean response time for the delay sensitive large packet of size x_l is given as:

$$T(x_{ls}) = x_{ls} + 2W(x_{ts}) + W(x_{ls}) + W_r(x_{ls}) \quad (13)$$

The term $2W(x_{ts})$ is the contribution from delay sensitive short packets found in the queue and the delay due to the delay sensitive short packets that arrive after the tagged large packet is in the queue, $W(x_{ts})$ and $W_r(x_{ls})$ are as given in equations 9 and 11 respectively and.

$$W(x_{ls}) = \sum_{n=m}^{N}(n-m)\frac{\rho_{x_{ls}}^n}{\lambda_1 m!}\left(\frac{\rho_{x_{ls}}}{m}\right)^{n-m} P_o^{x_{ls}} \quad (14)$$

where,

$$P_o^{x_{ls}} = \left[\sum_{n=0}^{m}\frac{\rho_{x_{ls}}^n}{n!} + \sum_{n=m+1}^{N}\frac{\rho_{x_{ls}}^n}{m!}\left(\frac{\rho_{x_{ls}}}{m}\right)^{n-m}\right]^{-1} \quad (15)$$

C. Model for Delay Tolerant Packets

Consider a tagged packet arriving to a delay tolerant queue. In case the tagged packet is a short delay tolerant packet its service will be delayed by all delay sensitive short packets, all delay sensitive large packets and all delay tolerant short packets found in the queue. In addition, the short delay tolerant packet will be delayed by all delay sensitive short and large packets that arrive after the tagged delay sensitive short packet is in the queue will be served before the tagged delay tolerant short packet is serviced. The mean response time for the delay tolerant short packet of size x_{sd} is given as:

$$T(x_{sd}) = x_{sd} + 2W(x_{ts}) + 2W(x_{ls}) + W(x_{td}) \quad (16)$$

where,

$$W(x_{td}) = \sum_{n=m}^{N}(n-m)\frac{\rho_{x_{td}}^{n}}{\lambda_2 m!}\left(\frac{\rho_{x_{td}}}{m}\right)^{n-m} P_o^{x_{td}} \quad (17)$$

and

$$P_o^{x_{td}} = \left[\sum_{n=0}^{m}\frac{\rho_{x_{td}}^{n}}{n!} + \sum_{n=m+1}^{N}\frac{\rho_{x_{td}}^{n}}{m!}\left(\frac{\rho_{x_{td}}}{m}\right)^{n-m}\right]^{-1} \quad (18)$$

$$\rho_{x_{td}} = \lambda_2 \int_0^{x_{td}} t f(t) dt$$

The term $2W(x_{ts})$ is as explained for equation 13.

For the case of the tagged large delay tolerant packet its service will be delayed by all delay sensitive short packets, all delay sensitive large packets, all delay tolerant short packets and all delay tolerant large packets found in the queue. In addition, the tagged large delay tolerant packet will be delayed by short and large delay sensitive packets that arrive after the tagged delay tolerant large packet is in the queue will be served before the tagged delay tolerant large packet. The mean response time for the delay tolerant large packet of size x_{ld} is given as:

$$T(x_{ld}) = x_{ld} + 2W(x_{ts}) + 2W(x_{ls}) + W(x_{ld}) \quad (19)$$

where

$$W(x_{ld}) = \sum_{n=m}^{N}(n-m)\frac{\rho_{x_{ld}}^{n}}{\lambda_2 m!}\left(\frac{\rho_{x_{ld}}}{m}\right)^{n-m} P_o^{x_{ld}} \quad (20)$$

and

$$P_o^{x_{ld}} = \left[\sum_{n=0}^{m}\frac{\rho_{x_{ld}}^{n}}{n!} + \sum_{n=m+1}^{N}\frac{\rho_{x_{ld}}^{n}}{m!}\left(\frac{\rho_{x_{ld}}}{m}\right)^{n-m}\right]^{-1} \quad (21)$$

and $\rho_{x_{ld}} = \lambda_2 \int_{x_{td}}^{\infty} t f(t)\, dt$

The term $2W(x_{ts})$ is the contribution from delay sensitive short packets found in the queue and the delay due to the delay sensitive short packets that arrive after the tagged large packet is in the queue, $2W(x_{ls})$ is the contribution from delay sensitive large packets and delay sensitive large packets that arrive after the tagged delay tolerant large packet is in the queue.

In the next section, we present the performance evaluation of the derived models in terms of mean slowdown.

IV. Performance Evaluation

In order to evaluate the performance of the proposed IoT-based healthcare monitoring system, the derived models are used to plot graphs using MATLAB and in particular Simulink package was used [16]. Simulink provides a graphical editor, customizable block libraries, and solvers for modeling and simulating dynamic systems. It is integrated with MATLAB, enabling one to incorporate MATLAB algorithms into models and exporting simulation results to MATLAB for further analysis.

The performance of the proposed system is evaluated using mean slowdown as the performance metrics. Mean slowdown is the normalized response time, i.e., the ratio of the response time of a packet to the size of that packet. Unlike mean response time which tends to be representative of the performance of just a few big packets since they count the most in the mean because their response times tend to be highest [32], slowdown is a useful metric to analyze fairness of a scheduling scheme.

The paper investigates how the prioritized scheduling (PS) scheme performs compared to the FCFS and SJF scheduling schemes for short and large packets. The effect of key parameters such as packet sizes on mean slowdown is investigated.

A. Model Parameters

Table I shows the hypothetical parameters used in the analysis which is consistent with parameters used in literature [29], [30]. The packet arrival rate and service rate follow Poisson distribution [15].

TABLE I. Implementation Parameters

Parameter	Value
Number of servers, *m*	10 [29]
The maximum number of health data packets in the queue *N*	150 [30]
Packets arrival rate λ	6.549 packets/second [31]
Packets service rate μ	8.8 packets/second [31]
The average packet size x_t	100 Kb [13]
Threshold size of the packet size *xts*	75 Kb [13]

B. Evaluation of the mean Slowdown with Packet Sizes for Delay Sensitive Packets

This section presents the performance of the packets in terms of mean slowdown while varying packet sizes for delay sensitive packets.

Fig. 2 shows the mean slowdown of delay sensitive short packets under FCFS, SJF, and PS schemes where short packets are packets with sizes less or equal to $x_s = 75Kb$. It is also observed that some shorter packets experience lower mean slowdown under SJF than under the PS scheme. The situation is however very different as the sizes of packets increase, the PS scheme performs better than FCFS and SJF by offering lower mean slowdown. It is shown that the difference in performance between the PS scheme and SJF and FCFS is more pronounced as the packet sizes increase for short packets. It can be observed that in all cases, the FCFS scheme performs

worse than SJF and PS scheme for all packet sizes for short packets. We can also see from Fig. 2 that the PS scheme performs much more closely with FCFS and SJF for small packet sizes for short packets.

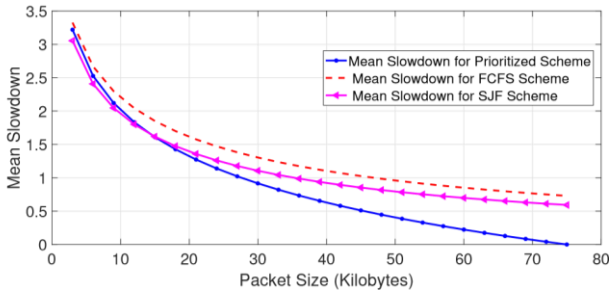

Fig. 2. Mean Slowdown for Delay Sensitive Short Packets under PS, SJF and FCFS Schemes.

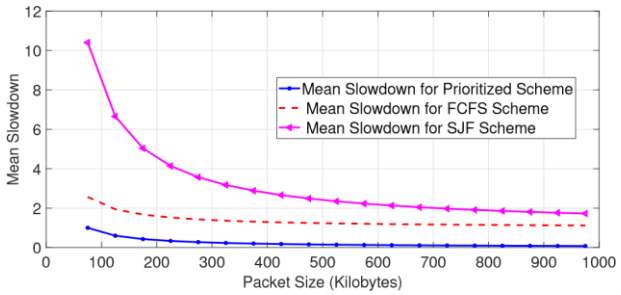

Fig. 3. Mean Slowdown for Delay Sensitive Large Packets under PS, SJF and FCFS Schemes.

Fig. 3 shows the mean slowdown of delay sensitive large packets under FCFS, SJF, and PS schemes where large packets are packets with sizes greater than x_s = 75 bytes. It can be observed from the figure that the PS scheme performs better than FCFS and SJF scheduling policies regardless of the packet size for large packets. In turn, FCFS also performs better than the SJF scheme for all large packet sizes considered. This is because under FCFS, there is a mix of short and large packets resulting into lower mean slowdown, whereas under SJF, large packets are serviced last and will always experience higher mean slowdown. The difference in performance is much more pronounced for shorter packet sizes, however as the packet sizes increase, the performance becomes closer as the mean slowdown values are closer.

C. Evaluation of the Mean Slowdown with Packet Sizes for Delay Tolerant Packets

This section presents the performance of the packets in terms of mean slowdown for the PS scheme in comparison with the FCFS and SJF scheduling schemes for delay tolerant packets.

Fig. 4 shows results of PS scheme in comparison with FCFS and SJF scheduling schemes for delay tolerant short packets. It can see that the SJF scheme performs better than the PS scheme for shorter packet sizes, this is because delay tolerant short packets are delayed by delay sensitive large packets which is not the case under SJF where there are only short packets, however as the packet sizes increase, the PS scheme performs better than SJF by offering lower mean slowdown. Similar to Fig. 2, it can be observed that in all cases, the FCFS scheme performs worse than SJF and PS scheme for all packet sizes for short packets. It is observed that the difference in performance between the PS scheme, SJF and FCFS is more pronounced as the packet sizes increase for short packets. In general, the PS scheme performs better than SJF and FCFS as the packet sizes increase for short packets.

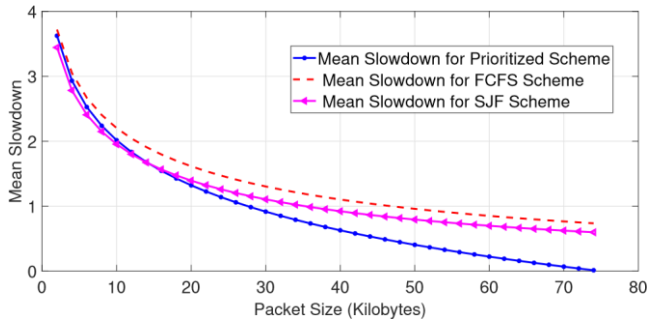

Fig. 4. Mean Slowdown for Delay Tolerant Short Packets under Prioritized, SJF and FCFS Schemes.

Fig. 5 shows results of PS scheme in comparison with FCFS and SJF scheduling schemes for delay tolerant large packets. It is observed that for the considered packet sizes, the PS scheme performs better than FCFS and SJF schemes by offering lower mean slowdown; the FCFS in turn is observed to offer lower mean slowdown than SJF scheme. It is further observed that the difference in mean slowdown is higher for shorter packet sizes and closer when the packet sizes increase. The performance between PS, FCFS and SJF schemes differ specifically for shorter packets where SJF performs worse than FCFS which in turn performs worse than the PS scheme.

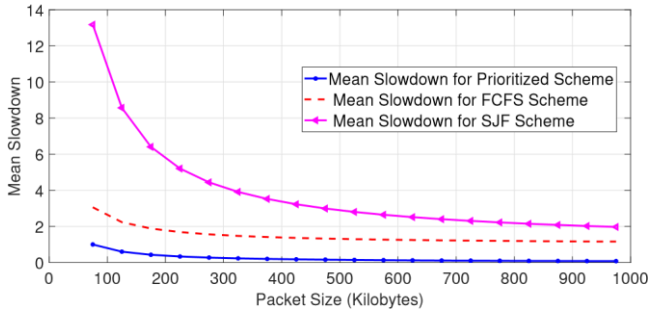

Fig. 5. Mean Slowdown for Delay Tolerant Large Packets under Prioritized, SJF and FCFS Schemes.

D. Evaluation of the Effect of Packet Threshold on Mean Slowdown for the PS Scheme for Delay Sensitive Packets

This section presents the performance of the packets in terms of mean slowdown for the PS scheduling scheme for different thresholds for delay sensitive packets. In doing this, the effect of the variation of the packet threshold in terms of size is investigated.

The results of the effect of varying the packet threshold on the mean slowdown for delay sensitive short packets are shown in Fig. 6. It can be observed that the decrease in the packet threshold leads to a reduction in the mean slowdown of delay sensitive short packets. The reduction in mean slowdown is observed to be more pronounced as the packet sizes increase, however for smaller packet sizes, the packet threshold has very

little effect. When the packet thresholds are reduced, it means the number of shorter packets are reduced hence the reduction in the mean slowdown.

Fig. 7 shows the variation of mean slowdown for delay sensitive large packets under the PS scheme for different packet thresholds. It can be observed that the decrease in the packet threshold reduces the mean slowdown of delay sensitive large packets. The reduction in mean slowdown is observed to be more pronounced for large packets with smaller sizes, however as the sizes of the delay sensitive packets increase, the packet threshold has very little effect. When the packet thresholds are reduced, the large packets with shorter sizes experience a more reduced mean slowdown due to the reason presented in Fig. 6.

E. Evaluation of the Effect of Packet Threshold on Mean Slowdown for the PS Scheme for Delay Tolerant Packets

This section presents the performance of the packets in terms of mean slowdown for PS scheduling scheme for different packet thresholds for delay tolerant packets as shown in Fig. 8 and 9.

Fig. 8 shows the variation of mean slowdown for delay tolerant short packets under the PS scheme for different packet thresholds. It can be observed that when the packet thresholds are reduced, the mean slowdown of delay tolerant short packets is reduced. The reduction in mean slowdown is observed to be more pronounced as the packet sizes increase, however for smaller packet sizes, the packet threshold has minimal effect and this is similar to the observation noted for delay sensitive short packets in Fig. 6. When the packet thresholds are reduced, the number of shorter packets is reduced hence the reduction in the mean slowdown.

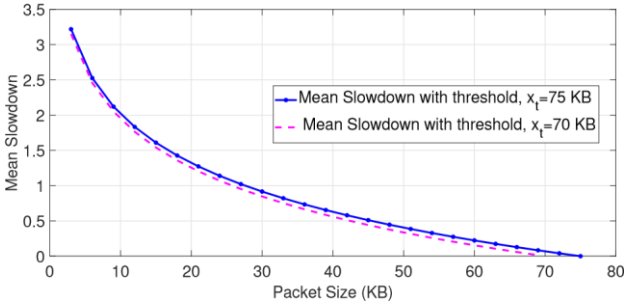

Fig. 6. Mean Slowdown for Delay Sensitive Short Packets under Prioritized Scheme for different Thresholds.

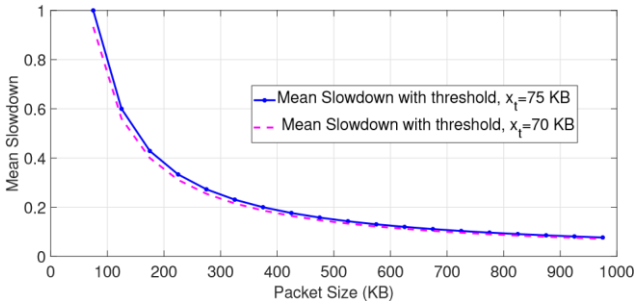

Fig. 7. Mean Slowdown for Delay Sensitive Large Packets under PS Scheme for different Thresholds.

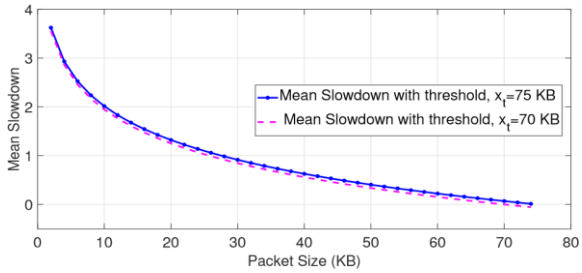

Fig. 8. Mean Slowdown for Delay Tolerant Short Packets under PS Scheme for different Thresholds.

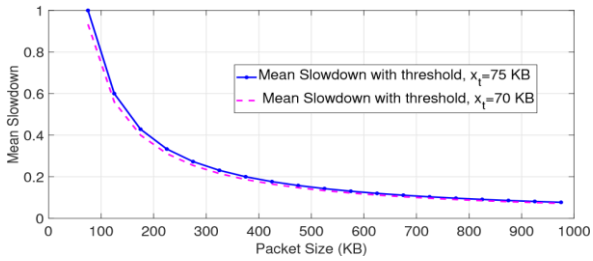

Fig. 9. Mean Slowdown for Delay Tolerant Large Packets under PS Scheme for different Thresholds.

Fig. 9 shows the variation of mean slowdown for delay tolerant large packets under the PS scheme for different packet thresholds. It can be observed that the decrease in the packet threshold reduces the mean slowdown of delay tolerant large packets. The reduction in mean slowdown is noted to be more pronounced for large packets with smaller sizes, however as the sizes of the delay tolerant packets increase, and the packet threshold has minimal effect on the mean slowdown. When the packet thresholds are reduced, the large packets with shorter sizes experience a more reduced mean slowdown due to increased number of large packets with shorter packet sizes.

V. Discussion

This study developed analytical models of mean slowdown for the PS scheme where incoming packets are prioritized based on the delay requirement and size of the packets and serviced using multiple servers. The effect of varying packet sizes on the mean slowdown under the PS is investigated in comparison with the FCFS and SJF scheduling policies. Results from the derived models show that the largest short packets perform better under the PS scheme than under the SJF and FCFS schemes. Similar observation has been noted by SWAP policy which also favors short packets to the expense of delaying large ones within the queue [15]. On the other hand, all large packets perform better under the PS scheme compared to the FCFS and SJF schemes. By giving priority to short packets under the PS scheme, more packets are served and hence large packets do not have to wait for so long for service. Large packets perform worse under FCFS scheme because their services are interrupted by large packets whose sizes may be larger. Similar explanations hold for the SJF scheme where large packets remain in the queue for a long time and may even lead to starvation.

VI. Conclusion

The PS scheduling scheme has been modeled and evaluated for varying packet sizes and thresholds. The numerical results

obtained from the derived models show the PS scheme generally reduces the mean slow down for most of the packet sizes considered. The comparison of the PS scheme with FCFS and SJF show that the PS scheme is superior in reducing the mean slowdown except for the few shortest short packets under SJF. The performance difference is more pronounced for the large packets with shorter sizes. It is also observed that short packets which are much shorter perform better under SJF than under the Prioritized scheme, however as the packet sizes increase, the PS scheme offers better performance than FCFS and SJF. It is further observed that when the packet threshold is reduced, the mean slowdown packets are reduced and the reduction is more pronounced for the short packets with larger sizes and large packets with shorter sizes.

VII. Future Work

In this paper, numerical results for the PS scheduling scheme using multiple homogeneous servers are presented. In the future, it will be interesting to investigate the effect of using heterogeneous servers on the performance, and also the effect of varying arrival and service rates.

References

[1] Ala, M. Guizani, M. Mohammad and M. Aledhari, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," IEEE Communications Surveys & Tutorials, vol. 17, no. 4, pp. 2347-2376, 2015.

[2] 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.

[3] H. Bhatia, S. N. Panda and D. Nagpa, "Internet of Things and its Applications in Healthcare-A Survey," in 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), Noida, India, 2020.

[4] C. Yi and Jun Cai, "Transmission Management of Delay-Sensitive Medical Packets in Beyond Wireless Body Area Networks: A Queueing Game Approach," IEEE Transactions on Mobile Computing, vol. 17, no. 9, pp. 2209 - 2222, 15 January 2018.

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

[6] 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. Mobile Computing., vol. 16, no. 9, pp. 2422-2435, September 2017.

[7] N. Nasser, L. Karim and T. Taleb, "Dynamic multilevel priority packet scheduling scheme for wireless sensor network," IEEE Transaction on Wireless Communication, vol. 12, no. 4, p. 1448–1459, 2013.

[8] 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.

[9] S.C. Hung, D. Liau, S-Y. Lien and K-C. Chen, "Low latency communication for Internet of Things," in IEEE/CIC International Conference on Communications in China, 2015.

[10] T.N Gia, M. Jiang, A-M Rahmani and T. Westerlund, "Fog computing in healthcare internet of things: A case study on ecg feature extraction," in IEEE International Conference on Computer and Information Technology, 2015.

[11] G. Lee, W. Saad W and M. Bennis, "An Online Optimization Framework for Distributed Fog Network Formation With Minimal Latency," IEEE Transactions on Wireless Communications, vol. 18, no. 4, pp. 2244 2258, 2019.

[12] H. Gupta, D. A. Vahid Dastjerdi , S.K. Ghosh and R. Buyya, "iFogSim: A toolkit for modeling and simulation of resource management techniques in the Internet of Things, Edge and Fog computing environments," Journal of Software: Practice and Experience, vol. 47, no. 9, pp. 1275-1296, 2017.

[13] Y. Changyan and J. Cai, "A Truthful Mechanism for Scheduling DelayConstrained Wireless Transmissions in IoT-Based Healthcare Networks," IEEE Transactions on Wireless Communications, vol. 18, no. 2, pp. 912 - 925, December 2018.

[14] K. Park, J. Park and J. Lee, "An IoT System for Remote Monitoring of Patients at Home", Journal of Applied Sciences, March 2017.

[15] I. A. Rai and M. Okopa, "Modeling and evaluation of swap scheduling policy under varying job size distributions", The Tenth International Conference on Networks, IARIA, pp. 115-120, 2011.

[16] D. K. Chaturvedi, "Modeling and Simulation of Systems Using Matlab and Simulink," CRC Press, 2010.

[17] P. J. Smith, A. Firag, P. A. Dmochowski, and Mansoor Shafi, "Analysis of the M/M/N/N Queue with Two Types of Arrival Process: Applications to Future Mobile Radio Systems", Journal of Applied Mathematics, 2012.

[18] S. Shukl, M. F. F. Hassan, M. K. Khan, L. T. Jung and A. Awang, "Ananalyticalmodel to minimizethe latency in healthcare internet-ofthings in fog computing environment," PLoS ONE, vol. 14, no. 11, pp. 1-31, 2019.

[19] S. El Kafhali and K. Salah, "Performance Modeling and Analysis of IoTenabled Healthcare Monitoring Systems," The Institute of Engineering and Technology (IET) Journals, pp. 1-12, 18 September 2018.

[20] I. Awan, M. Younas and W. Naveed, "Modelling QoS in IoT Applications," in International Conference on Network-Based Information Systems, 2014.

[21] C. Yi and J. Cai, "A Truthful Mechanism for Scheduling DelayConstrained Wireless Transmissions in IoT-Based Healthcare Networks," IEEE, pp. 1-14, 2018.

[22] Y. Changyan and J. Cai, "Transmission Management of Delay-Sensitive Medical Packets in Beyond Wireless Body Area Networ ks: A Queueing Game Approach," IEEE Transactions on Mobile Computing, vol. 17, no. 9, pp. 2209 - 2222, January 2018.

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

[24] M. A. Deepika.N, 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, 2019.

[25] M. Shamim Hossain; Ghulam Muhammad; Nadra Guizani, "Explainable AI and Mass Surveillance System-Based Healthcare Framework to Combat COVID-I9 Like Pandemics," IEEE Network, vol. 34, no. 4, July/August, 2020.

[26] D. Niyato, E. Hossain and S. Camorlinga, "Remote patient monitoring service using heterogeneous wireless access networks: architecture and optimization," IEEE Journal on Selected Areas in Communications, vol. 27, no. 4, pp. 412-423, 2009.

[27] H. Su and X. Zhang, "Battery-dynamics driven TDMA MAC protocols for wireless body-area monitoring networks in healthcare applications," IEEE Journal on Selected Areas in Communications, vol. 27, no. 4, p. 424–434, 2009.

[28] S. Rashwand and J. Misic, "Two-tier WBAN/WLAN Healthcare Networks: Priority Considerations," in IEEE/GLOBECOM, 2012.

[29] K. Salah and S. El Kafhali, "Performance Modeling and Analysis of IoT-enabled Healthcare Monitoring Systems,"

[30] K. Salah and S. El Kafhali, "Performance Modeling and Analysis of Hypoexponential Network Servers", Journal of Telecommunications System, vol. 65, no. 4, pp. 717-728, 2017.

[31] 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," in Proc. 6th IEEE Annual Consumer Communications & Networking Conference, Jan.2019, pp.1-7.

[32] 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.