Diffusion Analysis Improves Scalability of IoT Networks to Mitigate the Massive Access Problem

Erol Gelenbe[1] and Mert Nakıp
Inst. Theoretical & Applied Informatics
Polish Academy of Sciences
Gliwice, Poland 44–100
Email: {seg, mnakip}@iitis.pl
& Yaşar University
35100 Bornova/İzmir, Turkey
& [1]Lab. I3S Université
Côte d'Azur, 06200 Nice, France

Dariusz Marek
Silesian Uiversity of Technology
Fac. Autom. Control, Electr. & CS
Gliwice, Poland 44–100
Email: dariusz.marek@polsl.pl

Tadeusz Czachorski
Inst. Theoretical & Applied Informatics
Polish Academy of Sciences
Gliwice, Poland 44–100
Email: tadek@iitis.pl

***Abstract*—A significant challenge of IoT networks is to offer Quality of Service (QoS) and meet deadline requirements when packets from a massive number of IoT devices are forwarded to an IoT gateway. Many IoT devices tend to report their data to their wired or wireless network gateways at closely correlated instants of time, leading to congestion known as the Massive Access Problem (MAP), which increases the probability that the IoT data will not meet its required deadlines. Since IoT data loses much of its value if it arrives to destination beyond a required deadline, MAP has been extensively studied in the literature. Thus we first take a queueing theoretic view of the problem, and also use a Diffusion Approximation to gain insight into the IoT traffic statistics that affect MAP. Then we introduce the Quasi-Deterministic Transmission Policy (QDTP) which significantly alleviates MAP when the average traffic rate grows beyond a given level and substantially reduces the probability that IoT data deadlines are missed. The results are validated using real IoT data which has been placed in IP packets for transmission.**

***Index Terms*—Internet of Things (IoT), Scheduling, Massive Access Problem, Queueing Theory, Quasi-Deterministic Transmission Policy, Diffusion Approximations**

I. Introduction

Sensors for health applications, monitoring of areas which are of difficult access such as remote areas or large civil engineering structures, and geophysical characteristics [1]–[4] are among the numerous motivations for the development of sensor networks, giving rise to the Internet of Things (IoT) where the number of connected devices is rapidly increasing with the needs of autonomous systems, smart cities and smart vehicles [5], [6]. This increase in connectivity results in high traffic rates, causing congestion at the Physical Random Access Channels (PRACH) that service these systems [7], [8] which is known as the Massive Access Problem (MAP). When IoT data is used to control a complex distributed system, the needs for synchronization of the data to present a coherent view of the system can lead to further constraints regarding packet delays [9].

This research has been supported by the European Commission H2020 Program through the IoTAC Research and Innovation Action, under Grant Agreement No. 952684.

MAP was addressed in early research [10]–[12] through adaptive routing to reduce congestion in networks with multiple paths and gateways, and via information theoretic techniques to reduce the amount of traffic that is sent [13]. More recent work [7], [8], [14]–[29] commonly assumes that IoT traffic arrives at random, leading to solutions to MAP that include Access Class Barring (ACB) [8], [14], [16], [23]–[26], Cognitive Machine-to-Machine (M2M) communication [17], [19], game theory [18], clustering of devices [20], [27], data rate adaptation [21], Spread-spectrum, Non-Orthogonal Multiple Access (NOMA) [22], Interference Cancellation (SIC) [7], the use of CSMA/CA or slotted-ALOHA [15], [28] and collision awareness [29]. Other work [30]–[35] has suggested proactive network solutions for MAP, which can include techniques such as adaptive traffic ofloading to storage areas or less congested gateways [36].

On the other hand it was empirically shown that the IoT traffic at the MAC-layer is predictable via specific machine-learning techniques that can help to identify distinct IoT traffic classes in [37]. Thus in [38], the predictability of IoT traffic is used to address the MAP in a "predictive network" framework with a Joint Forecasting-Scheduling (JFS) system that allocates the time slots of a single frequency channel for IoT devices based on a forecast of traffic generation. In [39] a Multi-Scale Algorithm is suggested to improve the performance of the JFS system while [40] has suggested the extension of JFS for multi-frequency channels. The results of this work has shown that predictive networks, in particular JFS, are promising for the solution to the MAP and that the lightweight scheduling heuristics are crucial since optimal schedules are difficult to achieve in practice due to their high computational requirements.

In recent work [41] we show that, for any given and fixed IoT packet traffic rate, the statistics of the actual instants at which IoT packets are sent has a significant impact on the number of packets which are received before the packets' deadlines expire. We showed that "smoothing" the traffic by setting the individual packet transmission times in a uniformly distributed manner will considerably improve the system's performance. We used queueing analysis to select the randomization instants, and to predict which packets would not meet their deadlines and should be eliminated before transmission to improve the chances that other packets meet their deadlines.

In the present paper, we seek the optimum instants of packet transmission which will maximize the probability that deadlines are met. Since exact analytical solutions for queueing systems with general arrival and service processes are not readily available, we first use Diffusion Approximations [42], [43] to compute the probability that an IoT packet will meet its deadline. Based on this analysis we show that the probability that the packet does not meet its deadline is an increasing function of the Squared Coefficient of Variation (SCV) of the interarrival times. Thus we observe that for any given fixed average arrival rate, and any service time distribution of the IoT data, the best possible schedule (in the sense of minimizing the IoT date that misses its deadline) is deterministic, with data being sent at fixed time intervals whose value is identical to the inverse of the arrival rate.

However such a deterministic schedule cannot be applied exactly: indeed, at low traffic rates deterministic schedules may unnecessarily delay some of the data, while not providing any significant improvement in the amount of data that misses its deadline. Thus in the sequel we develop a Quasi-Deterministic Transmission Policy (QDTP) which is used when traffic rates are moderate or high, and is turned off for low traffic rates.

DAs also provide us with mathematically based insight with regard to the randomization result obtained earlier in [41], since the uniformly distributed randomization reduces the SCV of the interarrival times for the publicly available empirical data set which is in the range 1.6 to 2.18, down to approximately 0.33. Furthermore the QDTP results in an SCV close to zero, hence it also minimizes the probability of missing deadlines.

The rest of this paper is organized as follows. Section II describes the problem we wish to solve, while Section III introduces the diffusion process based analysis and introduces our main result concerning the Quasi-Deterministic Transmission Policy which aims to reduce the probability that deadlines are not respected. Section IV details the QDTP and evaluates it on the experimental data that we use from [44]. Finally section V summarizes our main conclusions and suggests further research.

II. The Probability of Meeting Deadlines

We model the channel that offers access to the IoT gateway as a single server queue with a service time that is related the length of each individual IoT packet that is being sent, with queuing theory techniques that are widely used [45]–[49].

Each IoT device that uses this gateway forwards packets structured from data in the form of "bursts" of multiple bits seen in the data set [44]. These bits would normally be packetized in some form, for instance basrd on the relatively recent LoRa-WAN standard which is used for IoT devices [50]. However we simplify matters and assume that the packetizations is based on IP packets with $21-Byte$ headers, which is not too different from LoRa-WAN, followed by payload bytes that contain the IoT device bit Bursts. The data transmission times are then selected in a manner that is compatible with the measurements reported in [50] were and effective data transmission rate in free space of 5400 $bits/sec$ was measured; beyond that rate, it was found that bit error rates became significant. Based on the measured average packet length of 22.47 $Bytes$, we see that the average packet transmission time would be 33.33 $msec$, which will be used in the sequel.

The IoT traffic data set [44] that we have used contains the traffic patterns of $10,000$ IoT devices, whose traffic generation patterns belong to one of the following classes: Fixed Bit Aperiodic, Fixed Bit Periodic, Variable Bit Aperiodic and Variable Bit Periodic. For a device j which sends a burst of Bits at time $a_{j,n}$, the deadline beyond which the burst is of "no value" is denoted by $\Delta_{j,n} > 0$; thus the bits must arrive at the gateway and be served and transferred out of the gateway by time $a_{j,n}+\Delta_{j,n}$. The evaluations we will conduct are based on different fixed values of the deadline.

In turn, the gateway will then process the successive packets to extract the required information, or to pass on the packet to some other system such as a Cloud server.

Let $a_1 \geq 0, a_2 \geq a_1,\ ..\ a_{n+1} \geq a_n,\ ...$ be the successive transmission times to the gateway of data packets by the various IoT devices that are connected, which is the First-In-First-Out ordered set of all data transmission instants from all IoT devices $\{a_{j,n}\}$.

On the other hand $S_1, S_2,\ ..\ S_n,\ ...$ are the successive durations of occupancy of the channel to the gateway by these successive packets which we call the "service times", so that in the numerical examples considered later in the paper we will take $E[S_n] = 0.33\ msec$. Each packet will also have a deadline that is denoted $\Delta_n \geq 0$ for the $n-th$ packet and which we discuss below.

The waiting W_n experienced by the $n-th$ packet is the delay that separates its arrival instant from its departure instant, and it is given by the well known Lindley's recursive equation [45], [48]:

$$W_{n+1} = [W_n + S_n - a_{n+1} + a_n]^+\ ,\ n = 0, 1, 2,\ ... \quad (1)$$

where we take $a_0 = 0$ and $[X]^+ = 0$ when $X < 0$, and $[X]^+ = X$ if $X \geq 0$. The response time R_n experienced by the $n-th$ packet is then defined as:

$$R_n = W_n + S_n, \quad (2)$$

i.e. the waiting time plus its service time. Note that $A(t)$ the number of arrivals by time t and the arrival rate of packets (when the arrival rate does not vary with time and the corresponding limit exists) are defined as:

$$A(t) = \sum_{n=1}^{\infty} 1[a_n \leq t],\ \lambda = \lim_{t\to\infty} \frac{A(t)}{t}. \quad (3)$$

When the arrival instants $\{a_n\}_{n\geq 0}$ constitute a random process, and the $\{S_n\}_{n\geq 0}$ are random variables, we define the probability Π_n that the $n-th$ packet misses its deadline, i.e.:

$$\Pi_n = Prob[R_n > \Delta_n],\ and\ \Pi = \lim_{n\to\infty} \Pi_n, \quad (4)$$

which is the probability that by the time a packet exits the gateway, its deadline has expired. Note that the "clock instant" when the $n-th$ packet's deadline expires is $a_n + \Delta_n$.

In the sequel we will assume that the transmission channel characteristics are fixed, and that all packet transmission times are proportional to their sizes and are drawn from the same distribution, so that the service times are samples of the same random variable S with mean value $E[S] = 0.33\ ms$. On the other hand, we will assume that the arrival rate of IoT packets λ depends on the number of active IoT devices in the system, which we denote by M.

The problem we now address is how to take appropriate scheduling decisions for each value of the workload λ, and in particular we will consider how to **optimize the instants at which the packets are actually transmitted, which we call t_n, so as to minimize Π.** Note that $t_n \geq a_n$, while a_n is the instant at which the IoT device data is available for transmission.

Thus in effect, the simple scheduling algorithm we describe, which we call Quasi-Deterministic Traffic Policy (QDTP), will transform each a_n into a new value t_n, and W_n will be transformed into W_n^* where:

$$W_{n+1}^* = [W_n^* + S_n - (t_{n+1} - t_n)]^+, \quad (5)$$

so that the new response time becomes:

$$R_n^* = [t_n - a_n] + W_n^* + S_n\ , \quad (6)$$

since we need to include the delay introduced by the QDTP algorithm itself into the response time of the $n-th$ packet that is being sent. Note that the new probability that the deadline is not met, also becomes:

$$\Pi_n^* = Prob[R_n^* > \Delta_n],\ and\ \Pi^* = \lim_{n\to\infty} \Pi_n^*, \quad (7)$$

A. Interarrival and Service Time Statistics

In the sequel we will assume that the channel characteristics between the IoT devices and the gateway are fixed, and that all packets are drawn from the same population having given packet length characteristics. Thus we assume that all S_n are independent random variables with the same probability distribution (i.e. they are i.i.d. or independent and identically distributed), with given mean $E[S]$ and SCV:

$$C_B^2 = \frac{E[S^2]}{(E[S])^2} - 1. \quad (8)$$

We also assume that we will operate under variable load λ (the arrival rate) but under stable conditions, i.e. $\lambda E[S] < 1$.

For a given $\lambda = (E[a_{n+1} - a_n])^{-1}$, we will assume that we restrict ourselves to the i.i.d. case but we have the freedom of choosing the probability distribution function that best fits our needs to minimize Π, and we define:

$$C_A^2 = \frac{E[(a_{n+1} - a_n)^2]}{(E[a_{n+1} - a_n])^2} - 1. \quad (9)$$

In fact, our analysis will show that by setting C_A^2 as small as possible, we can minimize Π.

1) An Example of IoT Data Statistics: To illustrate some of the statistics obtained from real IoT date available from the Open Source Repository [44], in Figure 1 we show the distribution of the amount of data in Bits from data Bursts emanating from the IoT devices in the data set. We observe that the average number of Bits per Burst is 6.8167 and the SCV of the number of Bits per Burst is 1.9506. There are only 2000 Bursts with more than 50 Bits per Burst. For those Bursts with less than 50 Bits, the average number of Bits per Burst is 6.778 and the SCV of the number of Bits per Burst is 1.7633.

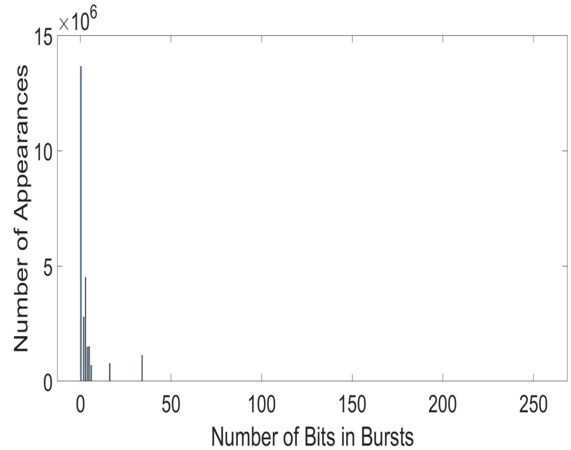

Fig. 1. The histogram of the number of Bits transmitted per Burst, by the set of IoT devices, obtained from the Open Source Data Set [44].

In Figure 2, we have used the same data set, but assumed that each burst from an IoT device is sent in the form of an IP packet with a $21-Byte$ header, and that the Bits belonging to the burst are stored in $8-Bit$ Bytes inside each packet, so that we exhibit the histogram of the length of the packets in Bytes.

In this data set, only 2000 packets are longer than $30-Bytes$. We note that the overall average packet length is $22.4744-Bytes$ with an SCV of 0.0031. If we only consider *just* those packets whose length is less than $30-Bytes$ the statistics are hardly different, since the average packet length is $22.4695-Bytes$ with an SCV of 0.0028 and the amount of payload data transmitted per burst is on average around $12-Bits$.

In the numerical examples of the next Sections III and IV, we normalize the average arrival rate λ based on the LoRaWAN bit-rate mentioned at the beginning of Section II, so that the average packet length of $22.4744\ Bytes$ is transmitted in $33.33\ msec$, and the maximum traffic rate $\lambda = 1$ corresponds to 30 packets/sec

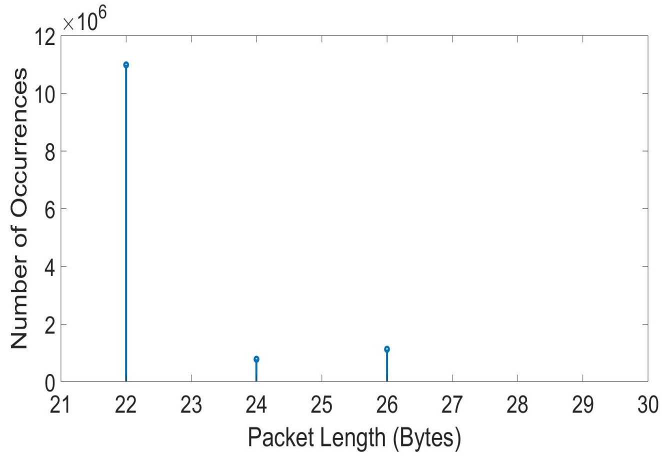

Fig. 2. The histogram of the number of Bytes transmitted per packet by the set of IoT devices, obtained from the Open Source Data Set [44], where we have assumed that the Bursts have been forwarded in the form of IP packets.

III. Using the Diffusion Approximation

While the probability density of the response time in steady-state:

$$F_R(t) = \lim_{n\to\infty} Prob[R_n \leq t], \tag{10}$$

is known in the case when at least one of the interarrival or service times are exponential (G/M/1, M/G/1 systems) [48], there is no easy way to obtain it exactly for both arbitrary interarrival and service time distributions. Therefore, we will use a diffusion approximation [51], [52] to determine the probability $F_R(\Delta)$ that the response time of the station is shorter than the deadline Δ, where we assume that the deadline is identical for all the IoT packets:

$$\Pi = \lim_{n\to\infty} \Pi = 1 - F_R(\Delta)\ . \tag{11}$$

In the diffusion approximation, the number of customers $N(t)$ in a single server queue is modeled by the diffusion process $X(t)$ on the interval $[0, +\infty)$ with probability density function $f(x, t; x_0)$ which approximates the queue length probability $p(n, t; x_0)$ for the initial condition x_0 at $t = 0$. In steady state, when $f(x) = \lim_{t\to\infty} f(x, t; x_0)$ the diffusion model yields [51]:

$$f(x) = \begin{cases} \frac{\lambda p_0}{-\beta}(1 - e^{zx}), \text{for} \quad 0 < x \leq 1 \\ \frac{\lambda p_0}{-\beta}(e^{-z} - 1)e^{zx}, \text{for} \quad x \geq 1, \quad z = \frac{2\beta}{\alpha}, \\ when\ \beta < 0,\ so\ that\ p_0 = 1 - \frac{\lambda}{\mu}. \end{cases} \tag{12}$$

where:

$$\lambda = \frac{1}{E[a_{n+1} - a_n]}, \quad \mu = \frac{1}{E[S_n]}, \tag{13}$$

$$\beta = \lambda - \mu, \quad \alpha = \lambda C_A^2 + \mu C_B^2. \tag{14}$$

Note that p_0 is obtained by the relation $p_0 + \int_0^\infty f(x)dx = 1$, and its value is also known from queueing theory. On the basis of (12) we can also calculate the mean number of customers in the system:

$$\begin{aligned} E[N] &\approx \int_0^\infty x f(x) dx \\ &\approx \frac{\lambda p_0}{-\beta}[\int_0^1 x(1 - e^{zx})dx + \int_1^\infty x(e^{-z} - 1)e^{zx}dx], \\ &\approx \frac{\lambda p_0}{-\beta}[0.5 - \frac{1}{z}] = \left[0.5 + \frac{C_A^2\varrho + C_B^2}{2(1-\varrho)}\right]\varrho. \end{aligned} \tag{15}$$

Finally, using Little's law, the mean response time in steady-state can also be obtained:

$$E[R] = \frac{E[N]}{\lambda}. \tag{16}$$

We note that $E[R]$ is monotone increasing in C_A^2 and C_B^2 but we need the probability distribution function of R to determine the probability of respecting the deadline. Therefore, we compute the response time with the use of the diffusion process [43]. It is can be obtained as a first passage time, since the time $X(t)$ needs to travel from the point $x = x_0$, corresponding to the queue length at the moment when a new packet joins it, to the first following instant when $x = 0$ when the packet has left the queue.

The probability density function $\gamma_{x_0,0}(t)$ of the distribution of first passage time from $x = x_0$ to $x = 0$, i.e. probability density that the process hits $x = 0$ the barrier at time t after starting at $x = x_0$ for time $t = 0$, is given by [46]:

$$\gamma_{x_0,0}(t) = \frac{x_0}{\sqrt{2\pi\alpha t^3}} e^{-\frac{(x_0+\beta t)^2}{2\alpha t}}.$$

For an arrival that finds the diffusion at level x_0 on arrival, this is simply the probabiity density function of the response time, because when this customer's service is complete it will leave the queue in the empty state, and the diffusion at level $x = 0$. Considering that the customer arrives when the system is at steady-state, the density of x_0 is simply $f(x_0)$, i.e. the stationary distribution of the diffusion process given in the expression (12).

We therefore obtain the probability density function of the response time as:

$$f_R(t) = \int_0^\infty \frac{x}{\sqrt{2\pi\alpha t^3}} e^{-\frac{(x+\beta t)^2}{2\alpha t}} f(x) dx, \tag{17}$$

and **the probability that the deadline is missed** is simply:

$$\Pi = 1 - \int_0^\Delta f_R(\tau) d\tau. \tag{18}$$

In order to illustrate these results, we provide two figures that show how Π varies with C_A^2, λ and Δ. Both Figures 3 and 4 also show, for comparison purposes, the actual values of the SCV of inter-arrival times C_A^2 of the real data set [44] with vertical bars, marked as ranging from approximately 1.6 to 2.18. The vertical bar for 0.33 which is close to the the value for the randomization policy of arrival instants developed in a recent paper [41], which reduces the value of C_A^2 by selecting the instants at which data is sent from individual devices using a uniform distribution over a deterministic interval of length $\Delta - E[R]$, when $E[R] \leq \Delta$.

In Figure 3, we show the probability of missing the deadline (y-axis) in logarithmic scale (to the base ten) for a given value of Δ, different values of λ, and with C_A^2 varying over a wide range. Here the average service rate and the SCV of service time are both fixed to 1. We see that as C_A^2 and λ increase, Π increases significantly.

Similar results are shown in Figure 4 for different values of Δ and a fixed value of λ, and C_A^2 varying over a wide range, showing that as Δ decreases and C_A^2 increases, Π increases significantly.

In Figure 5, we detail the probability of missing the deadline for the real data set of [44], plotted against the number of IoT devices M being used. For each value of M, we also give the arrival rate λ normalized against the measured average bit rate from the M sources and the corresponding real values of the SCV of interarrival times C_A^2. We see that the value of M or of the corresponding λ has the principal effect in determining the measured fraction Π of the data transfers which do not meet the deadline.

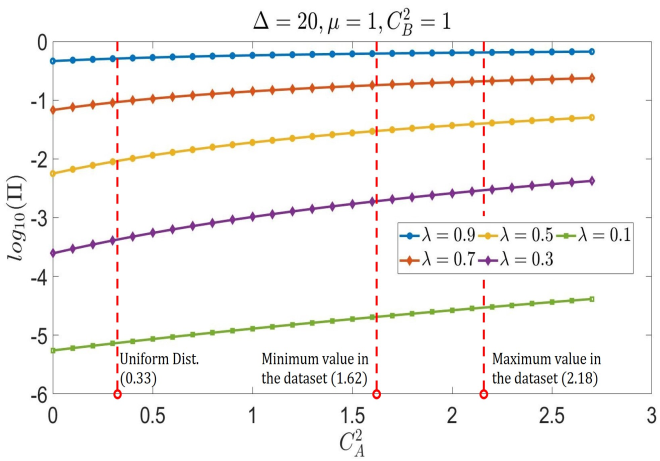

Fig. 3. The probability of missing the deadline (y-axis) in logarithmic scale (to the base ten), estimated using the diffusion approximation, increases significantly as C_A^2 and λ increase, for a fixed value of Δ. The average service rate μ and the SCV of service time C_B^2 are both fixed to 1.

IV. Quasi-Deterministic Transmission Policy (QDTP)

For $0 \leq a_1 \leq a_2 \leq \ \ldots\ a_n \leq a_{n+1} \leq \ \ldots$, the list of successive Burst dates of all the IoT devices, and λ the overall average arrival rate of the bursts, let $t_n \geq a_n,\ n = 1,\ 2,\ \ldots$ be the instants at which the Bursts are actually sent from the IoT devices, and define the deterministic quantity $D = \frac{1}{\lambda}$, which is identical to the average inter-arrival time. The "Quasi-Deterministic Transmission Policy" (QDTP) which is meant to reduce the value of C_A^2 and hence reduce Π the probability of missing deadlines, is defined as follows:

1) Set $n = 1$,
2) Send Packet 1 at $t_1 = a_1$,
3) Set $n \leftarrow n + 1$,

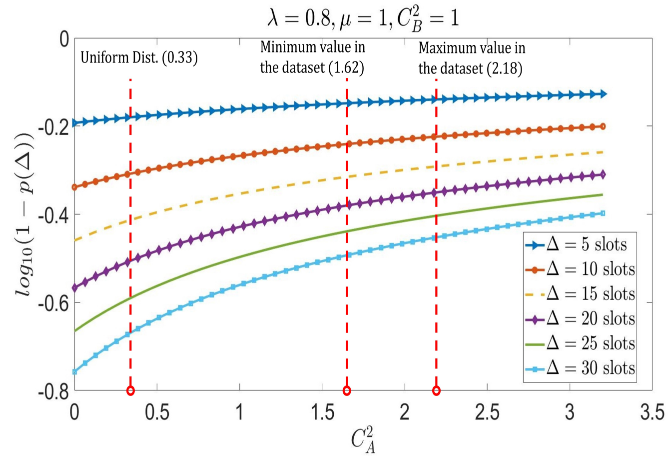

Fig. 4. The probability of missing the deadline (y-axis) in logarithmic scale (to the base ten), estimated with the diffusion approximation, increases significantly as C_A^2 increases and the the deadline Δ measured in slots decreases, for a fixed but high value of the arrival rate $\lambda = 0.8$. The average service rate μ and the SCV of service time C_B^2 are both fixed to 1.

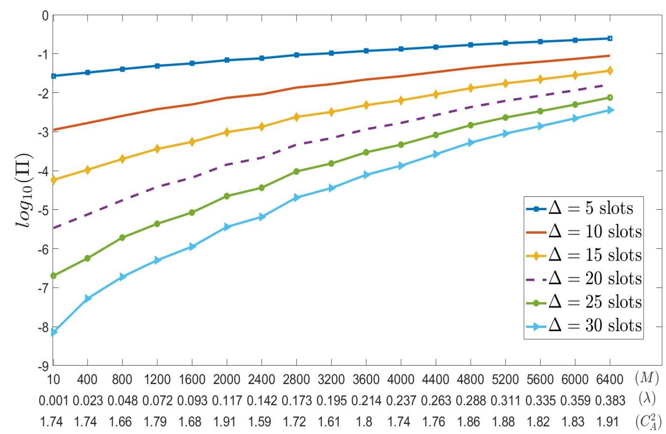

Fig. 5. The probability of missing the deadline (y-axis) in logarithmic scale (to the base ten) estimated with the diffusion approximation, using the traffic statistics of the real data set of [44], is plotted against the number of IoT devices M (x-axis) that are being used. Note that each value of M corresponds to specific measured values of λ and C_A^2 shown along the $x - axis$. The corresponding arrival rate λ normalized against the measured average bit rate from sources is also shown. The real values of the SCV of interarrival times C_A^2 are also indicated.

4) If $a_n \leq t_{n-1} + D$, Send Packet n at $t_n = a_{n-1} + D$,
5) Else if $a_n > t_{n-1} + D$, Send Packet n at a_n.
6) Go to (3).

To implement QDTP in practice, we would need to know λ in advance, which is possible when we have a fixed set of IoT devices, each of which sends data at pre-determined instants. Also, we either need the receiver to know the Packet generation times in advance, and to request all the senders to send their Packets at the instants defined by the QDTP, or the senders can know in advance their sequence number n, "listen" to the successive sending instants and apply the QDTP to determine when they need to send their own data.

Thus the communication channel would need to be two-way or the individual IoT devices should also have the ability to sense the channel.

To evaluate the effectiveness of QDTP, we first conducted measurements of the SCV of interarrival times for the data set [44], both for the raw IoT data from [44], and for the same data using QDTP, for a varying number of active IoT devices M as shown in Figure 6. We see that QDTP substantially reduces the SCV C_A^2 for all values of M.

Then in Figure 7, we show the relative frequency (empirical probability) of missing the deadline – for a very small value of the deadline $\Delta = 2$ – for both the raw data set of [44] and for the case where the QDTP is used with the same data set. QDTP obviously succeeds in considerably reducing the probability that the deadline is missed. Finally, in Figure 8 the data set in [44] with varying numbers M of active IoT devices, and different values of Δ, is used. We see that for all values of Δ above 2, QDTP reduces the empirically measured Π to practically zero. However when the heuristic is not used, Π tends to one as M increases beyond a few hundred devices.

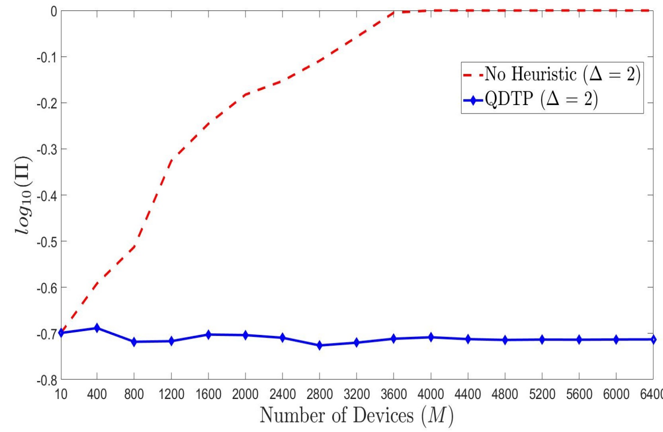

Fig. 7. This figure shows the consequence of the results detailed in Figure 6 for the empirically measured probability of missing the deadline (y-axis). $\log_{10} \Pi$ is shown for both the raw data set of [44] and for the case where the QDTP is used with the same data set, with a very small value of the deadline $\Delta = 2$. We see that QDTP succeeds in considerably reducing the probability that the deadline is missed for all values of M.

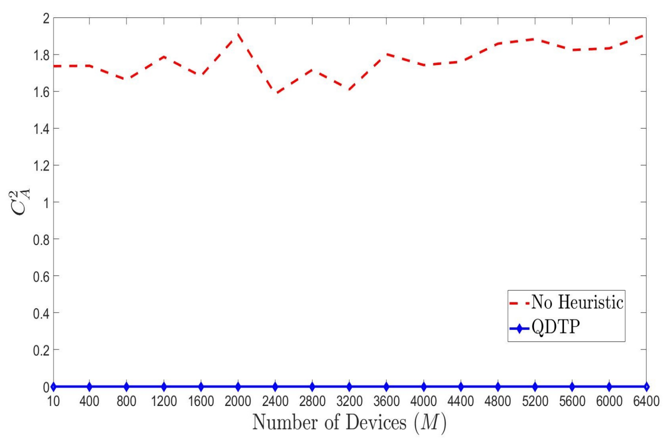

Fig. 6. Measurements of the SCV of interarrival times, both for the raw IoT data from [44], and for the same data using QDTP, for a varying number of active IoT devices M. We observe that QDTP has substantially reduced the empirically measured SCV C_A^2, reducing it to zero for all the distinct numbers of devices M in the data set of [44].

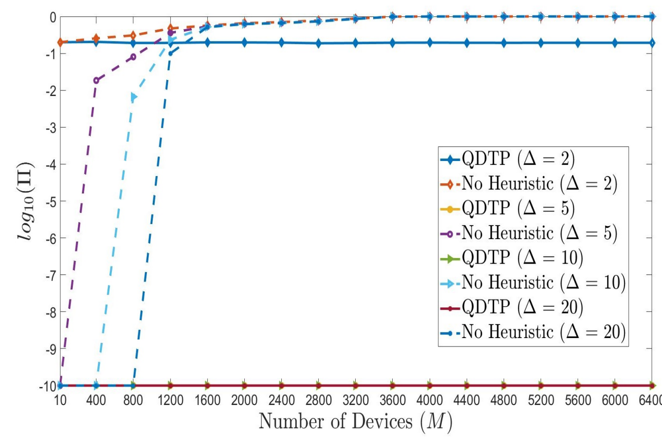

Fig. 8. Here we extend the experimental results in Figure 7 using the data set [44] and varying numbers M of active IoT devices. We compare the logarithm to the base 10 of the (empirically measured) probability Π that the deadline is missed for different values of the deadline Δ. We see that for all values of Δ that are shown above 2, the QDTP reduces Π to practically zero, while when the heuristic is not used then Π tends to one as M increases beyond a few hundred devices.

V. Conclusions

The MAP is one of the biggest challenges for the future of IoT networks and occurs when large numbers of devices access a single channel to reach their gateway, causing congestion at the entry points and leading to deadlines being missed for the data sent from IoT devices. In order to address this problem, predictive network designs have been used, where scheduling modifies priorities between devices and select the instants at which IoT packets are transmitted.

In this paper, we use insight from queueing theory and Diffusion Approximations to design the QDTP scheduling heuristic that improves the scalability of IoT networks. Using a real data set of outputs from a large number of IoT devices, we have examined the relevant statistics and used them to study the effect of the interarrival time statistics on the probability that the IoT packets meet (or do not meet) their deadlines.

The performance of QDTP has then been evaluated extensively using the data set in [44] for a widely varying range of numbers M of IoT devices, resulting in widely varying average arrival rates and many different deadline values. In particular, we have compared the performance of QDTP against the case where the original packet transmission dates (found in the experimental data set) are used.

These evaluations which use real data have demonstrated that QDTP can provide a very large reduction in the empirically measured fraction of packets that miss their deadlines.

Future work will combine QDTP with priority policies to attempt to obtain further improvements in IoT network performance, resulting in further alleviation of the MAP problem.

Furthermore, since queueing systems with deterministic arrivals have been studied by different authors [53]–[55] we expect that the insight provided by our work may lead to further useful interactions between classical queueing theory and the study and optimization of the Internet of Things.

Acknowledgements

The authors thank the anonymous referees for their very helpful recommendations and corrections, which have served to significantly improve this paper.

References

[1] Y. Sahraoui, A. Korichi, C. A. Kerrache, M. Bilal, and M. Amadeo, "Remote sensing to control respiratory viral diseases outbreaks using internet of vehicles," *CoRR*, vol. abs/2103.12512, 2021. [Online]. Available: https://arxiv.org/abs/2103.12512

[2] A. Zaman, Z. Hassan, R. Odarchenko, S. Hassan, S. Ahmed, M. Bilal, I. Ahmad, V. Tiurin, and A. Faheem, "Wireless underground sensor networks: Packet size optimization survey," in *Proceedings of the 2nd International Workshop on Control, Optimisation and Analytical Processing of Social Networks (COAPSN 2020), Lviv, Ukraine, May 21, 2020*, ser. CEUR Workshop Proceedings, S. Fedushko and T. O. Edoh, Eds., vol. 2616. CEUR-WS.org, 2020, pp. 353–365. [Online]. Available: http://ceur-ws.org/Vol-2616/paper30.pdf

[3] M. E. Haque, M. Asikuzzaman, I. U. Khan, I. Ra, M. S. Hossain, and S. B. H. Shah, "Comparative study of iot-based topology maintenance protocol in a wireless sensor network for structural health monitoring," *Remote. Sens.*, vol. 12, no. 15, p. 2358, 2020. [Online]. Available: https://doi.org/10.3390/rs12152358

[4] E. Gelenbe and T. Koçak, "Area-based results for mine detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 1, pp. 12–24, 2000.

[5] G. Fortino, W. Russo, C. Savaglio, M. Viroli, and M. Zhou, "Modeling opportunistic IoT services in open IoT ecosystems." in *WOA*, 2017, pp. 90–95.

[6] O. Bello and S. Zeadally, "Toward efficient smartification of the Internet of Things (IoT) services," *Future Generation Computer Systems*, vol. 92, pp. 663–673, 2019.

[7] A. Zanella, M. Zorzi, A. F. dos Santos, P. Popovski, N. Pratas, C. Stefanovic, A. Dekorsy, C. Bockelmann, B. Busropan, and T. A. Norp, "M2M massive wireless access: Challenges, research issues, and ways forward," in *2013 IEEE Globecom Workshops*. IEEE, 2013, pp. 151–156.

[8] F. Ghavimi and H.-H. Chen, "M2M communications in 3GPP LTE/LTE-A networks: Architectures, service requirements, challenges, and applications," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 2, pp. 525–549, 2015.

[9] E. Gelenbe and K. Sevcik, "Analysis of update synchronization for multiple copy data bases," *IEEE Transactions on Computers*, no. 10, pp. 737–747, 1979.

[10] E. Gelenbe and E. C.-H. Ngai, "Adaptive qos routing for significant events in wireless sensor networks," in *2008 5th IEEE International Conference on Mobile Ad Hoc and Sensor Systems*. IEEE, 2008, pp. 410–415.

[11] E. Gelenbe, E. Ngai, and P. Yadav, "Routing of high-priority packets in wireless sensor networks," in *IEEE Second International Conference on Computer and Network Technology, IEEE*, 2010.

[12] E. Gelenbe and E. Ngai, "Adaptive random re-routing for differentiated qos in sensor networks," *The Computer Journal*, vol. 53, no. 7, pp. 1052–1061, 2010.

[13] E. C.-H. Ngai, E. Gelenbe, and G. Humber, "Information-aware traffic reduction for wireless sensor networks," in *2009 IEEE 34th Conference on Local Computer Networks*. IEEE, 2009, pp. 451–458.

[14] S.-Y. Lien, T.-H. Liau, C.-Y. Kao, and K.-C. Chen, "Cooperative access class barring for machine-to-machine communications," *IEEE Transactions on Wireless Communications*, vol. 11, no. 1, pp. 27–32, 2012.

[15] Y. Liu, C. Yuen, X. Cao, N. U. Hassan, and J. Chen, "Design of a scalable hybrid MAC protocol for heterogeneous M2M networks," *IEEE Internet of Things Journal*, vol. 1, no. 1, pp. 99–111, 2014.

[16] T.-M. Lin, C.-H. Lee, J.-P. Cheng, and W.-T. Chen, "PRADA: Prioritized random access with dynamic access barring for MTC in 3GPP LTE-A networks," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 5, pp. 2467–2472, 2014.

[17] A. Aijaz, S. Ping, M. R. Akhavan, and A.-H. Aghvami, "CRB-MAC: A receiver-based MAC protocol for cognitive radio equipped smart grid sensor networks," *IEEE Sensors Journal*, vol. 14, no. 12, pp. 4325–4333, 2014.

[18] Y.-C. Pang, S.-L. Chao, G.-Y. Lin, and H.-Y. Wei, "Network access for M2M/H2H hybrid systems: A game theoretic approach," *IEEE Communications Letters*, vol. 18, no. 5, pp. 845–848, 2014.

[19] A. Aijaz and A. H. Aghvami, "Cognitive machine-to-machine communications for Internet-of-Things: A protocol stack perspective," *IEEE Internet of Things Journal*, vol. 2, no. 2, pp. 103–112, 2015.

[20] I. Park, D. Kim, and D. Har, "MAC achieving low latency and energy efficiency in hierarchical M2M networks with clustered nodes," *IEEE Sensors Journal*, vol. 15, no. 3, pp. 1657–1661, 2015.

[21] P. Si, J. Yang, S. Chen, and H. Xi, "Adaptive massive access management for QoS guarantees in M2M communications," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 7, pp. 3152–3166, 2015.

[22] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, "Massive non-orthogonal multiple access for cellular IoT: Potentials and limitations," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 55–61, 2017.

[23] H. Jin, W. T. Toor, B. C. Jung, and J.-B. Seo, "Recursive pseudo-Bayesian access class barring for M2M communications in LTE systems," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 9, pp. 8595–8599, 2017.

[24] J. Liu, L. Song *et al.*, "A novel congestion reduction scheme for massive machine-to-machine communication," *IEEE Access*, vol. 5, pp. 18 765–18 777, 2017.

[25] L. Tello-Oquendo, I. Leyva-Mayorga, V. Pla, J. Martinez-Bauset, J.-R. Vidal, V. Casares-Giner, and L. Guijarro, "Performance analysis and optimal access class barring parameter configuration in LTE-A networks with massive M2M traffic," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 4, pp. 3505–3520, 2018.

[26] L. Tello-Oquendo, D. Pacheco-Paramo, V. Pla, and J. Martinez-Bauset, "Reinforcement learning-based ACB in LTE-A networks for handling massive M2M and H2H communications," in *2018 IEEE International Conference on Communications (ICC)*. IEEE, 2018, pp. 1–7.

[27] L. Liang, L. Xu, B. Cao, and Y. Jia, "A cluster-based congestion-mitigating access scheme for massive M2M communications in Internet of Things," *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 2200–2211, 2018.

[28] N. Shahin, R. Ali, and Y.-T. Kim, "Hybrid slotted-CSMA/CA-TDMA for efficient massive registration of IoT devices," *IEEE Access*, vol. 6, pp. 18 366–18 382, 2018.

[29] Z. Alavikia and A. Ghasemi, "Collision-aware resource access scheme for LTE-based machine-to-machine communications," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 4683–4688, 2018.

[30] Y. Edalat, J.-S. Ahn, and K. Obraczka, "Smart experts for network state estimation," *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 622–635, 2016.

[31] V. Petkov and K. Obraczka, "Collision-free medium access based on traffic forecasting," in *2012 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*. IEEE, 2012, pp. 1–9.

[32] ——, "The case for using traffic forecasting in schedule-based channel access," in *2011 IEEE Consumer Communications and Networking Conference (CCNC)*. IEEE, 2011, pp. 208–212.

[33] D. Raca, A. H. Zahran, C. J. Sreenan, R. K. Sinha, E. Halepovic, R. Jana, and V. Gopalakrishnan, "On leveraging machine and deep learning for throughput prediction in cellular networks: Design, performance, and challenges," *IEEE Communications Magazine*, vol. 58, no. 3, pp. 11–17, 2020.

[34] L. Ruan, M. P. I. Dias, and E. Wong, "Machine learning-based bandwidth prediction for low-latency H2M applications," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 3743–3752, 2019.

[35] H.-Y. Kim and J.-M. Kim, "A load balancing scheme based on deep-learning in IoT," *Cluster Computing*, vol. 20, no. 1, pp. 873–878, 2017.

[36] J. Du, E. Gelenbe, C. Jiang, H. Zhang, and Y. Ren, "Contract design for traffic offloading and resource allocation in heterogeneous ultra-dense

networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 11, pp. 2457–2467, 2017.
[37] M. Nakip, B. C. Gül, V. Rodoplu, and C. Güzeliş, "Comparative study of forecasting schemes for IoT device traffic in machine-to-machine communication," in *Proceedings of the 2019 4th International Conference on Cloud Computing and Internet of Things*, 2019, pp. 102–109.
[38] M. Nakip, V. Rodoplu, C. Güzeliş, and D. T. Eliiyi, "Joint forecasting-scheduling for the Internet of Things," in *2019 IEEE Global Conference on Internet of Things (GCIoT)*. IEEE, 2019, pp. 1–7.
[39] V. Rodoplu, M. Nakıp, D. T. Eliiyi, and C. Güzelis, "A multi-scale algorithm for joint forecasting-scheduling to solve the massive access problem of IoT," *IEEE Internet of Things Journal*, 2020.
[40] V. Rodoplu, M. Nakip, R. Qorbanian, and D. T. Eliiyi, "Multi-channel joint forecasting-scheduling for the Internet of Things," *IEEE Access*, vol. 8, pp. 217 324–217 354, 2020.
[41] M. Nakip and E. Gelenbe, "Randomization of data generation times improves performance of predictive IoT networks," in *2021 IEEE World Forum on Internet of Things (WF-IoT)*. IEEEXplore, July 2021, pp. 1–6. [Online]. Available: http://doi.org/10.5281/zenodo.4696170
[42] E. Gelenbe, "Probabilistic models of computer systems Part II: Diffusion approximations, waiting times and batch arrivals," *Acta Informatica*, vol. 12, p. 285–303, 1979.
[43] T. Czachórski, E. Gelenbe, G. S. Kuaban, and D. Marek, "Time-dependent performance of a multi-hop software defined network," *Applied Sciences*, vol. 11, no. 6, p. 2469, 2021.
[44] "IoT Traffic Generation Pattern Dataset," Jan 2021. [Online]. Available: https://www.kaggle.com/tubitak1001118e277/iot-traffic-generation-patterns
[45] L. Takács, *Introduction to the Theory of Queues*. Oxford University Press, 1962.
[46] D. R. Cox and H. D. Miller, *The Theory of Stochastic Processes*. Chapman and Hall, 1965.
[47] G. F. Newell, "Applications of Queueing Theory," June 1971.
[48] L. Kleinrock, *Queueing Systems: Computer Applications*. John Wiley and Sons, 1976.
[49] E. Gelenbe, "Steady-state solution of probabilistic gene regulatory networks," *Physical Review E*, vol. 76, no. 3, p. 031903, 2007.
[50] H. Tarab, *Real Time Performance Testing of LoRa-LPWAN Based Environmental Monitoring UAV System*. University of Windsor, Electronic Theses and Dissertations. 7578, 2018. [Online]. Available: https://scholar.uwindsor.ca/etd/7578
[51] E. Gelenbe, "On approximate computer system models," *Journal of the ACM (JACM)*, vol. 22, no. 2, pp. 261–269, 1975.
[52] H. Kobayashi, "Application of the diffusion approximation to queueing networks i: Equilibrium queue distributions," *Journal of the ACM (JACM)*, vol. 21, no. 2, pp. 316–328, 1974.
[53] G. P. Cosmetatos and G. P. Prastacos, "An approximate analysis of the D/M/1 queue with deterministic customer impatience," *RAIRO Recherche opeérationnelle*, vol. 19, no. 2, pp. 133–142, 1985.
[54] D. Pinotsi and M. A. Zazanis, "Synchronized queues with deterministic arrivals," *Operations Research Letters*, vol. 33, no. 6, pp. 560–566, 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167637705000039
[55] D. Schol, M. Vlasiou, and B. Zwart, "Large fork-join queues with nearly deterministic arrival and service times," 2020. [Online]. Available: https://arxiv.org/pdf/1912.11661.pdf