IoT Traffic Shaping and the Massive Access Problem*

Erol Gelenbe
Inst. of Theoretical & Applied Informatics
Polish Ac. Sci. 44-100 Gliwice, PL
CNRS I3S Lab., Université Côte d'Azur, FR
& CNRS A. de Moivre Lab., Imperial College
& Yaşar University, Izmir, TR
ORCID:0000-0001-9688-2201

Karl Sigman
Dept. Industrial Eng. & Operations Res.
& Center for Applied Probability
Columbia University
New York, NY 10027, USA
ORCID:0000-0002-0126-3895

***Abstract*—IoT gateways aim to meet the deadlines and QoS needs of packets from as many IoT devices as possible, though this can lead to a form of congestion known as the Massive Access Problem (MAP). While much work was conducted on predictive or reactive scheduling schemes to match the arrival process of packets to the service capabilities of IoT gateways, such schemes may use substantial computation and communication between gateways and IoT devices. This paper proves that the recently proposed "Quasi-Deterministic-Transmission-Policy (QDTP)" traffic shaping approach which delays packets at IoT devices, substantially alleviates the MAP: QDTP does not increase overall end-to-end delay and reduces gateway queue length. We then introduce the Adaptive Non-Deterministic Transmission Policy (ANTP) that requires only one packet buffer at the gateway, offering substantial QoS improvement over FIFO scheduling.**

***Index Terms*—Internet of Things (IoT), Traffic Shaping, Quasi-Deterministic Transmission Policy (QDTP), Adaptive Non-Deterministic Transmission Policy (ANTP), Quality of Service, Massive Access Problem, Queueing Analysis**

I. Introduction

The number of devices on the Internet may reach $30Bn$ by 2023 [1] with the majority being low-cost machine-type devices [20] communicating with base stations or IoT Gateways (IoTGW). Such large systems can cause a form of congestion [19] known as the "Massive Access Problem" (MAP) which has has often been addressed with *reactive techniques* that attempt to adapt to incoming traffic [23], [27], [31]–[33]. Despite the difficlty of managing distributed accesses [5]–[7], the cooperation among transmitters to improve channel usage efficiency and QoS has also been considered [16], [17].

Another approach [37]–[39] uses *proactive prediction* of IoT traffic patterns, such as Joint Forecasting-Scheduling (JFS) and Priority based on Average Load (PAL), allocating channel resources to IoT devices based on traffic characteristic [42], [43]. "Randomization of Generation Times" (RGT) [44] can be implemented at each device to improve JFS, with PAL and the Earliest-Deadline First algorithm. However, scheduling requires additional computation and time consuming Machine Learning (ML) is needed to analyze arrival and service characteristics, and the best schedules can incur computation and communication costs.

The simpler traffic shaping, widely used in networks [2], can reduce latency, and optimize the bandwidth available to certain packets by delaying other packets. Typically applied at the source or edge, it is defined by the International Telecommunication Union [3] as a scheme which *"alters the traffic characteristics of a stream of cells ... to achieve a desired modification of those traffic characteristics, in order to achieve better network efficiency whilst meeting the QoS objectives or to ensure conformance ... with the consequence of increasing the mean cell transfer delay."* Though traffic shaping is accomplished by delaying packets, it is sometimes confused with "traffic policing" which includes preventive packet dropping [4], while traffic shaping can result in more delay for some packets that may cause loss of data in finite buffers. Both techniques have been discussed for ATM [9], [12], IP [11], [13], [14], and Sensor Networks where traffic shaping with adaptive routing was also studied [21]. Recent work [22] addresses traffic shaping for large numbers of IoT Devices (IoTD) that forward packets to a single Gateway (IoTGW) with the Quasi-Deterministic-Transmission-Policy (QDTP) which delays packets at the IoTD at most D (fixed) time units, obtaining more deterministic arrivals within given time slots at the IoTGW; experiments with IoT data [47] show QDTP's effectiveness to alleviate the MAP and improve QoS.

In the present paper we prove that ANTP does **not increase the overall end-to-end delay of packets as compared to an ordinary FIFO policy, including the traffic shaping delay and the delay at the IoTGW**. We modify QDTP to propose the Adaptive Non-Deterministic Transmission Policy (ANTP) which shapes packet delay to match service times at the IoTGW, and prove that **ANTP reduces the delay and reduces the packet queue length at the IoTGW to no more than one packet, hence also reduceing packet loss probabilities due to finite buffers**.

In the sequel, Section II discusses the notation that is used and recalls Lindley's equation regarding the waiting time of the First-In-First-Out (FIFO) queue. Then, Section II-A details the

*Research supported by the EU H2020 Program under the IoTAC Research and Innovation Action, with Grant Agreement No. 952684.

QDTP algorithm [22], and develops an equation that resembles Lindley's equation for the total time spent by a packet in the IoTD (device) before being forwarded to the IoTGW. Assuming a Poisson process for the generation of IoT packets across all the IoTDs, in Section II-B we also derive new results concerning the probability distribution (and the average) of the number of IoTD's (i.e. IoT devices) that are witholding packets in the QDTP delay in steady-state.

Section III introduces the Adaptive Non-Deterministic Transmission (ANTP) policy in which the traffic shaping delay depends on individual packets and their service times at the gateway, which is practically feasible when service time is a function of (e.g., proportional to) known packet length. We prove that the ANTP delay also satisfies an equation structurally similar to Lindley's equation, and establish the stability condition. A formal analysis of ANTP end-to-end-delay is conducted, showing that each sample path of ANTP does not increase total delay as compared to a FIFO system that does not use ANTP. Also ANTP strongly reduces the need for buffer space at the gateway. The paper ends with conclusions and a discussion of further topics for research.

II. The Ordinary FIFO Queue and QDTP

We consider a sequence of packets (or customers $C_n,\ n \geq 0$ characterized by an infinite sequence of variables:

$$\{(a_n,\ S_n:\ n \geq 0\}, \tag{1}$$

which are the intrinsic characteristics of each successive packet that is created at an IoT node or source of packets. For each successive packet:

- The $0 = a_0 \leq a_1 \leq a_2,\ ...$ are the packet creation instants instants at the IoT nodes being considered; we also define the intervals between the creation times of packets $A_{n+1} = a_{n+1} - a_n$. Note that the packets are created at any of the IoT nodes, so that each a_n is a time stamp and successive packets may or may not originate at the same IoT node.
- In order to shape the traffic, the IoT node may delay the transmission of a packet for some time using the QDTP or A-QDTP policy.
- After this traffic shaping delay, the $n-th$ packet is transmitted and arrives instantaneously at the IoTGW. Thus we are assuming that the transmission time from its source IoT node to the IoTGW is negligible, as compared to the other times of interest, such as the service (or processing and forwarding) time S_n.
- S_n is the processing time of the $n-th$ packet by the IoTGW, and at the IoTGW all packets are served in First-In-First-Out (FIFO) order. Thus Figure 1 describes the timing at the IoTGW when there is **no traffic shaping**. In certain cases, for instance when the packet processing and forwarding time at the input gateway is proportional to the length of the packet, S_n may be proportional to packet length.

Assuming a simple, the well-known "Lindley's Equation" [45], [46], [48] gives us a recursive formula for computing the pure waiting time (before processing) of the successive packets when IoTD forwards each packet to the IoTGW as soon as it assembled (at time a_n), and the IoTGW processes packets in FIFO order:

$$L_{n+1} = [L_n + S_n - A_{n+1}]^+,\ L_0 = 0,\ n \geq 0. \tag{2}$$

where for a real number X, we denote $[X]^+ = X\ if X > 0,\ and [X]^+ = 0\ if\ X \leq 0$. Thus if the IoT gateway acts as a FIFO processor and forwarder, the packet that arrives to it at time a_n will have been processed and forwarded downstream out of the IoTGW, towards (for instance) a local edge server or a Cloud server, at time $d_n = a_n + L_n + S_n$.

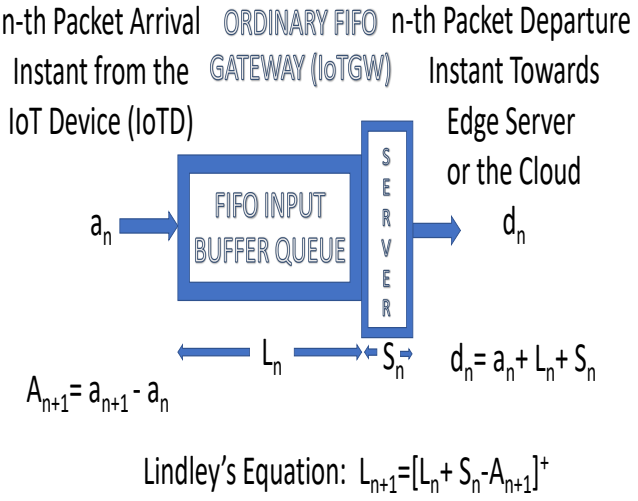

Fig. 1. Schematic representation of a FIFO forwarding node. The packet that arrives at instant a_n waits in the buffer until it arrives at the head of the queue. It then receives service of duration S_n and leaves the server at time d_n. Its waiting time is L_n, so that $d_n = a_n + L_n + S_n$.

For the system without the QDTP policy:

$$R_n = L_n + S_n, \tag{3}$$

is the ordinary "response time" which includes both the waiting time and the service time.

A. The QDTP Policy

On the other hand, the *Quasi-Deterministic Transmission Policy (QDTP)* introduced in [22], is defined via a sequence of forwarding or release times t_n for each C_n, so that the customer arriving at a_n is only released at some time t_n into the FIFO queue for servicing, with $t_0 = a_0 = 0$, and:

$$t_{n+1} = \max\{t_n + D, a_{n+1}\},\ n \geq 1, \tag{4}$$

where $D \geq 0$ is a constant. Obviously if $D = 0$ we are back at the ordinary FIFO service.

Lemma 1 The waiting times of packets at the IoTD (devices) using QDTP satisfy an expression similar to Lindey's equation:

$$W_{n+1} = [W_n + D - A_{n+1}]^+,\ n \geq 1. \tag{5}$$

Furthermore, if the inter-arrival times $\{A_n,\ n \geq 1\}$ are a sequence of independent and identically distributed random variables and $E[A_n] > D$, then we have the "stability result":

$$\lim_{n\to\infty} W_n = W,\ in\ probability\ distribution. \quad (6)$$
$$where\ Prob[W < \infty] = 1.$$

Proof Note that the the departure instant of customer C_{n+1} from the QDTP delay unit is given by:

$$\begin{aligned} t_{n+1} &= t_n + D\ if\ a_{n+1} < t_n + D,\ and \\ &= a_{n+1}\ if\ a_{n+1} \geq t_n + D_n,\ or \\ a_{n+1} + W_{n+1} &= a_n + W_n + D \\ &\quad if\ a_{n+1} < a_n + W_n + D,\ and \\ a_{n+1} + W_{n+1} &= a_{n+1}\ if\ a_{n+1} \geq a_n + W_n + D, \\ or\ W_{n+1} &= [W_n + D - A_{n+1}]^+, \end{aligned}$$

completing the proof of (5). On the other hand, the result (6) follows from the well known property of the Lindley equation for the GI/D/1 queue (deterministic service) [45] which is identical to (5).

B. QDTP with Poisson Arrivals and Deterministic Delays

As an interesting and illustrative special case, suppose that the incoming traffic is Poisson with arrival rate λ, and that D is constant as in [22]. Now W_{n+1} in (13) is identical to the waiting time of the $n-th$ customer of an M/D/1 queue that has Poisson arrivals and constant service times, i.e. the $\{A_{n+1},\ n \geq 1\}$ are independent and exponentially distributed random variables with parameter λ. However the key difference is that QDTP only has a waiting time, and no "service time" since as soon as the delay W_n ends, the packet is released into the second queue for servicing, as shown in Figure 2 which covers the case where each of the QDTP delay D_n may be distinct, which we call the Adaptive QDTP traffic shaping scheme which will be discussed in Section III.

Thus we can now compute the distribution of the random variable N, which is the number of packets waiting in steady-state at all of the IoT devices (the IoTDs) which are using the QDTP traffic shaping policy. We do this by relating N to the number M of customers in the M/D/1 queue with Poisson arrivals and constant service times in steady-state [48], see [46] eqn. (2.8), since the $n-th$ packet leaves an IoTD to enter directly into the IoTGW input (see Figure 2), it does not incur the service time of the M/D/1 queue. Hence N corresponds to the number of packets which are stored in the waiting time of an M/D/1 queue, to the exclusion of its service time. Therefore:

$$\begin{aligned} Prob[N=0] &= Prob[M=1] + Prob[M=0], \quad (7) \\ Prob[N=k] &= Prob[M=k+1],\ k \geq 1. \quad (8) \end{aligned}$$

Denoting $\pi_k = Prob[N=k],\ p_k = Prob[M=k],\ k \geq 0$, and defining the generating functions:

$$P(z) = \sum_{k=0}^{\infty} p_k z^k,\ \Pi(z) = \sum_{k=0}^{\infty} \pi_k . z^k,\ for\ |z| \leq 1, \quad (9)$$

which after some calculations yield:

$$\Pi(z) = \frac{(p_0 + P(z))(z-1)}{z},\ P(z) = p_0 \frac{e^{\lambda(z-1)D}(z-1)}{z - e^{\lambda(z-1)D}}.$$

From queueing theory we know that for the distribution $\{\pi_k,\ k \geq 0\}$ to exist we need the condition that $0 \leq 1 - \pi_0 < 1$. We also know that when the distribution exists it follows that $1 - \pi_0 = \lambda D$ [48], which guaranties that the distribution $\{p_k,\ k \geq 0\}$ also exists, since p_k is obtained from (7).

C. Average Number of Packets Waiting at their IoTDs

With QDTP some of the packets will first wait at their "home" IoT devices before they are forwarded to the IoTGW. In the case of Poisson arrivals, the above analysis allows us to estimate the average number $< N >$ that wait at their different IoTDs, from the average $< M >$ of the number M in the M/D/1 queue at steady state. The computation is simple, because from (7) we have:

$$\begin{aligned} < N > &= \sum_{k=1}^{\infty} k.\pi_k = \sum_{k=1}^{\infty} k.p_{k+1}, \\ &= \sum_{k=1}^{\infty} (k+1).p_{k+1} - \sum_{k=1}^{\infty} p_{k+1}, \\ &= < M > -p_1 - [1 - p_0 - p_1], \\ &= < M > -[1 - p_0]. \end{aligned}$$

$< M >$ is available from the well-known Pollaczek-Khintchine formula [46], [48] applied to the M/D/1 queue, hence:

$$< N > = \lambda D[1 + \frac{\lambda D}{2(1-\lambda D)}] - \lambda D = \frac{(\lambda D)^2}{2(1-\lambda D)},$$

and we know that **we must have** $D < \frac{1}{\lambda}$. However more than that, we know that the average interdeparture times $\overline{\mathbf{D_n}} = E[t_{n+1} - t_n]$ of packets from the QDTP delay are given by:

$$\begin{aligned} \overline{D_n} &= t_{n+1} - t_n = a_{n+1} - a_n + W_{n+1} - W_n, \\ &= E[A_n] + E[W_{n+1} - W_n], \\ &= \frac{1}{\lambda} + E[W_{n+1} - W_n],\ hence \\ \lim_{n\to\infty} \overline{D_n} &= E[A] = \frac{1}{\lambda}, \quad (10) \end{aligned}$$

since when $\lambda E[A] < 1$, both W_{n+1} and W_n converge to the same W in distribution, and hence have the same mean.

III. The Adaptive QDTP Policy: ANTP

Let us now pursue the ANTP case where the individual QDTP delays may depend on other system parameters, and hence we allow D_n to vary with each value of n, so that the departure instants from the IoTDs to the gateway IoTGW are:

$$t_{n+1} = \max\{t_n + D_n, a_{n+1}\},\ n \geq 1, \quad (11)$$

and each D_n can be chosen as a function of the parameter S_n. Indeed, when we examine the data shown in Figure 3 from a publicly available source [47], we see that the length of IP packets emanating from IoT devices have substantial

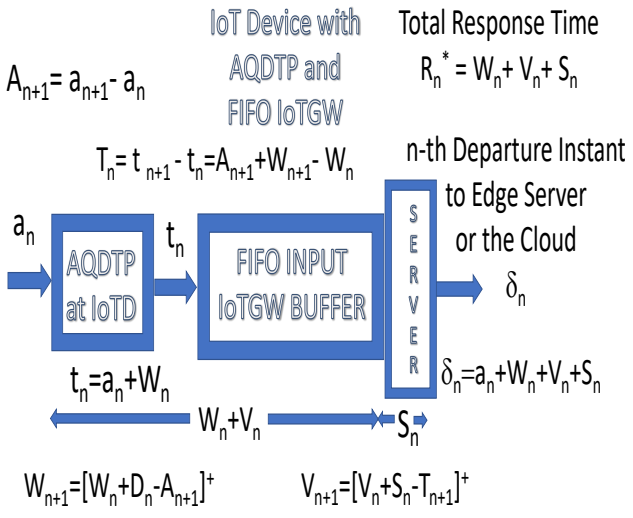

Fig. 2. IoTD nodes operating with the ANTP policy which forward traffic to a IoTGW gateway. The resulting timing issues and end-to-end delays are detailed.

regularity so that the service time parameter S_n can be estimated accurately when the IoTGW forwards the arriving packets to Edge Servers or the Cloud. The ANTP delay D_n at the IoTDs can then be chosen to be smaller than the time needed to process the packes at the IoTGW.

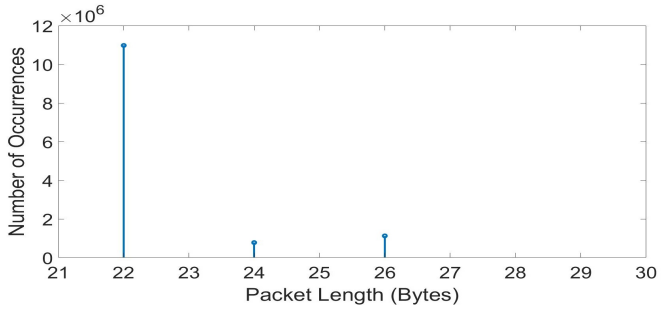

Fig. 3. The Open Source IoT Data Set [47] provides precise information regarding IP packet lengths from the IoTDs including the headers. The histogram shows the resulting number of Bytes transmitted for each packet by the IoTDs indicating that most packets are constant in length.

When D_n is the random variable adefined in (1), and the waiting time from a_n until the packet C_n is actually placed in the gateway's input queue is W_n:

$$W_n = t_n - a_n, \ for \ ANTP \ we \ have: \tag{12}$$

Lemma 2 The successive waiting times of packets in the ANTP delay unit satisfy the relation:

$$W_{n+1} = [W_n + D_n - A_{n+1}]^+, \ n \geq 1. \tag{13}$$

Proof The proof is identical to that of Lemma 1, since the departure instant of C_{n+1} from the ANTP delay unit is given by:

$$\begin{aligned} t_{n+1} &= t_n + D_n \ if \ a_{n+1} < t_n + D_n, \ and \\ &= a_{n+1} \ if \ a_{n+1} \geq t_n + D_n, \ or \\ a_{n+1} + W_{n+1} &= a_n + W_n + D_n \\ &\quad if \ a_{n+1} < a_n + W_n + D_n, \ and \\ a_{n+1} + W_{n+1} &= a_{n+1} \ if \ a_{n+1} \geq a_n + W_n + D_n, \\ or \ W_{n+1} &= [W_n + D_n - A_{n+1}]^+. \end{aligned}$$

A. Total Response Time with ANTP

From the analysis in the previous section, the total response type of the $n-th$ packet C_n with QDTP can now be computed as

$$R_n^* = W_n + V_n + S_n, \tag{14}$$

where W_n is given by equation (13), and V_n is the waiting time in the FCFS queue that is entered by the packet after the delay in the QDTP.

Defining $T_{n+1} = t_{n+1} - t_n, \ n \geq 0$ and applying Lindley's equation to the FCFS queue that is entered by each successive packet $\{C_n\}$, at the instants $\{t_n = a_n + W_n, \ n \geq 0\}$ so that $T_{n+1} = A_{n+1} + W_{n+1} - W_n$, we obtain:

$$\begin{aligned} V_{n+1} &= [V_n + S_n - T_{n+1}]^+, & (15) \\ &= [V_n + (S_n - (W_{n+1} - W_n)) - A_{n+1}]^+. & (16) \end{aligned}$$

QDTP does not increase the delay experienced by packets, if we can show that R_n^* defined in (14) is less than or equal to R_n defined in (3). Since S_n is an additive term in both expressions, we only need to compare the waiting time L_n with $W_n + V_n$, namely:

$$\begin{aligned} &L_{n+1} = [L_n + S_n - A_{n+1}]^+, \ and \\ &W_{n+1} + V_{n+1} = \\ &W_{n+1} + [V_n + (S_n - (W_{n+1} - W_n)) - A_{n+1}]^+, \\ &where \ W_{n+1} = [W_n + D_n - A_{n+1}]^+. \end{aligned}$$

Theorem 1 If we $D_n \leq S_n$ for all $n \geq 1$, then it follows that $W_n + V_n \leq L_n$. Hence the ANTP policy does not increase the total response time for each customer or packet as long as $D_n \leq S_n$ since

$$R_n^* = W_n + V_n + S_n \leq L_n + S_n. \tag{17}$$

Proof: The proof is by induction:

- The **base** of the induction is $0 = W_1 + V_1 = 0 \leq L_1 = 0$.
- The **step** of the induction is to assume that the statement is true for some $n > 1$:

$$W_n + V_n \leq L_n, \ for \ some \ n \geq 1, \ and \tag{18}$$

and prove that it is true for $n + 1$,
i.e. we must prove that:

$$W_{n+1} + V_{n+1} \leq L_n. \tag{19}$$

To prove (19) we use $V_n \leq L_n - W_n$ and have:

$$\begin{aligned}
&L_{n+1} = [L_n + S_n - A_{n+1}]^+, \ and \\
&W_{n+1} + V_{n+1} \leq [W_n + D_n - A_{n+1}]^+ \\
&\qquad + [L_n - W_n + S_n - A_{n+1} - W_{n+1} + W_n]^+, \\
&\leq [W_n + D_n - A_{n+1}]^+ \\
&+[L_n + S_n - A_{n+1} - W_{n+1}]^+ \ .
\end{aligned}$$

There are two cases to consider, A) and B):

- A) If $W_n + D_n - A_{n+1} \leq 0$, which implies that $W_{n+1} = 0$, we have:

$$\begin{aligned}
W_{n+1} + V_{n+1} &= [V_n + S_n - A_{n+1} + W_n]^+ \\
&= [L_n + S_n - A_{n+1}]^+ = L_{n+1},
\end{aligned}$$

  so that using the induction step $W_n + V_n \leq L_n$, we have proved for case A) that $W_{n+1} + V_{n+1} \leq L_{n+1}$.
- B) On the other hand if $W_n + D_n - A_{n+1} > 0$ then

$$\begin{aligned}
&W_{n+1} + V_{n+1} = W_n + D_n - A_{n+1} \\
&+[V_n + S_n - A_{n+1} - W_{n+1} + W_n]^+, \\
&= W_n + D_n - A_{n+1} + [V_n + S_n - D_n]^+.
\end{aligned}$$

  Since $V_n \geq 0$ and $S_n \geq D_n$ we know that:

$$V_n + S_n - D_n > 0,$$

  and as a consequence:

$$\begin{aligned}
&W_{n+1} + V_{n+1} = \\
&W_n + D_n - A_{n+1} + V_n + S_n - D_n, \\
&= L_n + S_n - A_{n+1} \\
\leq \quad &L_{n+1} = [L_n + S_n - A_{n+1}]^+.
\end{aligned}$$

In addition, since the delay W_n at the IoTD is non-negative and the total delay at IoTD plus the IoTGW (V_n) is no greater than the delay of an ordinary FIFO gateway L_n, the delay and buffer queue length at the IoTGW will be reduced. **This completes the proof of Theorem 1.**

B. Delay and Queue Length Bounds for the IoTGW

Let us now define $G_n = V_n + S_n$, the **total delay incurred at the IoTGW** by the $n-th$ packet that was generated by the IoTD device at time a_n, and let $G = \lim_{n \to G_n}$ when the limit exists. From *Theorem 1*, we know that if $D_n \leq S_n$ then:

$$\begin{aligned}
W_n + V_n \quad &\leq \quad L_n, \ hence \\
G_n = V_n + S_n \quad &\leq \quad L_n - W_n + S_n. \qquad (20)
\end{aligned}$$

Since $W_1 = L_1 = 0$, if $D_n = S_n$ we obviously have:

$$\begin{aligned}
W_n \quad &= \quad [W_{n-1} + D_n - A_{n+1}] \\
&= \quad L_n = [L_{n-1} + S_n - A_{n+1}], \ \forall \ n \geq 0.
\end{aligned}$$

As a consequence, the following follows from (20) and (21):

Theorem 2: If $D_n = S_n$, and $S = \lim_{n \to} S_n$, then

$$G_n \quad = \quad V_n + S_n \leq L_n - W_n + S_n \leq S_n, \ G \leq S, (21)$$

meaning that the IoTGW buffer contains at most one packet at a time, showing the effectiveness of ANTP traffic shaping.

C. ANTP in the Poisson Case

As a consequence we obtain rigorous performance estimates for the case of Poisson arrivals of rate λ using the results of Section II-B, when the service times S_n are independent and identically distributed with general distribution with mean $E[S_n]$ and squared coefficient of variation C_S^2.

By Little's Law [46], B_{ANTP} the average number of packets at steady-state in the IoTGW buffer is $B_{ANTP} \leq \lambda E[S_n] < 1$. It can be compared with the steady-state average number B_{FIFO} of packets in the IoTGW when the ANTP algorithm is not used and all packets arrive directly to the buffer from the IoTD without delay. B_{FIFO} is given by the Pollaczek-Khintchine expression for the average number of packets in the IoTGW buffer acting as a a First-In-First-Out queue, leading to the figure of merit F_{ANTP} for ANTP:

$$\begin{aligned}
F_{ANTP} \quad &= \quad 1 - \frac{B_{ANTP}}{B_{FIFO}} \leq 1 - \frac{\rho}{\rho(1 + \frac{1+\rho C_S^2}{2(1-\rho)})}, \\
&\leq \quad \frac{1 + \rho C_S^2}{1 + \rho C_S^2 + 2(1-\rho)}, \qquad (22)
\end{aligned}$$

and $F_{ANTP} \to 1$ when $\rho \to 1$.

IV. Conclusions

The ANTP traffic shaping policy for IoT devices that transmit packets to an IoTGW has been introduced, and its stabiity conditions have been obtained. More importantly, we have shown that both QDTP and ANTP shape the traffic flowing from IoT devices to a IoT gateway by delaying packets at the IoTD without increasing the overall end-to-end packet delay.

We also show that ANTP will limit the gateway buffer occupancy to at most one packet, reducing significantly buffer queue lengths and delay at the IoTGW itself.

It will be interesting to consider how traffic streams from large numbers of independent IoTDs can interact efficiently in this framework.

References

[1] CISCO, "Cisco Annual Internet Report (2018–2023) White Paper", *Cisco Systems*, March 9, 2020. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

[2] IETF, "An Architecture for Differentiated Services" , *ETF RFC 2475*, December 1988.

[3] ITU, "Traffic control and congestion control in B-ISDN", *ITU-T Recommendation I.371*, March 2004.

[4] CISCO, "Comparing Traffic Policing and Traffic Shaping for Bandwidth Limiting," *Cisco Tech Notes*, Document ID: 19645, Cisco Systems, August 2005.

[5] E. Gelenbe and K. C. Sevcik, "Analysis of Update Synchronization for Multiple Copy Data Bases," *EEE Transactions on Computers*, vol. C-28, no. 10, pp. 737-747, Oct. 1979, doi: 10.1109/TC.1979.1675241.

[6] A. Chesnais, E. Gelenbe and I. Mitrani, 'On the modeling of parallel access to shared data", *Comm. ACM*, vol. 26, no. 3, pp. 196-202, 1983.

[7] G. Hebrail and E. Gelenbe, "A probability model of uncertainty in data bases," *1986 IEEE Second International Conference on Data Engineering*, pp. 328-333, IEEE, New York.

[8] Anwar Elwalid, Debasis Mitra and Robert H. Wentworth, "A New Approach for Allocating Buffers and Bandwidth to Heterogeneous, Regulated Traffic in an ATM Node," *EEE J. on Selected Areas in Comm.*, vol. 13, No. 6, AUGUST 1995 pp. 1115-1127, 1995.

[9] V. Srnivasan, A. Ghanwani and E. Gelenbe, "Block loss reduction in ATM networks", *Computer Communications*, vol. 19, no. 13, pp. 1077-1091, 1996.

[10] A. Elwalid and D. Mitra, "Traffic shaping at a network node: theory, optimum design, admission control," *Proceedings of INFOCOM '97*, vol. 2, pp. 444-454, 1997. doi: 10.1109/INFCOM.1997.644493.

[11] D. O. Awduche, "MPLS and Traffic Engineering in IP Networks", *IEEE Comm. Mag.*, vol. 37, no. 12, pp. 42-27, December 1999.

[12] H. Brandt, "ATM", John Wiley & Sons, Chichester, England, 2001.

[13] C. Barakat, E. Altman and W. Dabbous, "On TCP Performance in a Heterogeneous Network: A Survey", *IEEE Comm. Mag.*, vol. 38, no. 1, pp. 40-46, January 2000.

[14] J. Helzer and L. Xu, "Congestion control for mult-media streaming with self-limiting sources", *13TH IEEE International Conference on Network Protocols (ICNP'05), Nov. 6-9, 2005*, Boston, MA. [Online]. Available: http://csr.bu.edu/icnp2005/posters/helzer.pdf

[15] A. Zaman, *et al.*, "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*. [Online]. Available: http://ceur-ws.org/Vol-2616/paper30.pdf

[16] J. Du, *et al.*, "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.

[17] N. Li, *et al.*, "Cooperative Wireless Edges with Composite Resource Allocation in Hierarchical Networks", *2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM)*, pp. 1-6, 2021, doi: 10.1109/HEALTHCOM49281.2021.9398997

[18] 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.

[19] A. Zanella, *et al.*, "M2M massive wireless access: Challenges, research issues, and ways forward," in *2013 IEEE Globecom Workshops*. IEEE, 2013, pp. 151–156.

[20] 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.

[21] 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.

[22] E. Gelenbe, M. Nakip, D. Marek and T. Czachorski, "Diffusion Analysis Improves Scalability of IoT Networks to Mitigate the Massive Access Problem", *2021 MASCOTS International Symposium on Modelling, Analysis and imulation of Computer and Telecommunication Systems*, pp. 182-189, Houston, Texas USA, IEEEXpress, November 3-5, 2021.

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

[24] 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.

[25] 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 Trans. on Vehicular Technology*, vol. 63, no. 5, pp. 2467–2472, 2014.

[26] A. Aijaz, *et al.*, "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.

[27] 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.

[28] 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.

[29] 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.

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

[31] 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 Trans. on Vehicular Technology*, vol. 66, no. 9, pp. 8595–8599, 2017.

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

[33] L. Tello-Oquendo, *et al.*,"Performance analysis and optimal access class barring parameter configuration in LTE-A networks with massive M2M traffic," *IEEE Trans. on Vehicular Technology*, vol. 67, no. 4, pp. 3505–3520, 2018.

[34] 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.

[35] 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.

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

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

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

[39] D. Raca, et al., "On leveraging machine and deep learning for throughput prediction in cellular networks: Design, performance, and challenges," *IEEE Comm. Mag.*, vol. 58, no. 3, pp. 11–17, 2020.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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

[45] K. Sigman, "Stationary Marked Point Processes: An Intuitive Approach," Chapman and Hall, CRC Press, London, UK, May 1995, ISBN:978-0412984310

[46] E. Gelenbe and I. Mitrani, "Anaysis and Synthesis of Computer Systems", World Scientific Ltd., Singapore & London, 2010.

[47] "IoT Traffic Generation Pattern Dataset," Jan 2021. [Online]. Available: https://www.kaggle.com/tubitak1001118e277/iot-traffic-generation-patterns

[48] L. Takács, *Introduction to the Theory of Queues*. Oxford University Press, 1962.