Traffic Modeling for Aggregated Periodic IoT Data

Tobias Hoßfeld, Florian Metzger
University of Duisburg-Essen
Chair of Modeling of Adaptive Systems
Essen, Germany
Email: tobias.hossfeld@uni-due.de

Poul E. Heegaard
Norwegian University of Science
and Technology (NTNU)
Trondheim, Norway
Email: poul.heegaard@item.ntnu.no

Abstract—**The Internet of Things (IoT) is emerging in the telecommunication sector, and will bring a very large number of devices that connect to the Internet in the near future. The expected growth in such IoT nodes necessitates appropriate traffic models in order to evaluate their impact on different aspects of networking, e.g., on signaling load in the networks, or on processing load of the data in a cloud. In this paper we analyze the characteristics of aggregated periodic IoT data based on related work, and compare them with a Poisson process as approximation for the superposed traffic as assumed in standardization. Such an approximation is crucial in order to investigate the scalability of an IoT network, as it may be impossible in practice to measure or to simulate large-scale IoT deployments. The accuracy and applicability of the Poisson process is investigated for the use case "IoT cloud". The results show that the Poisson process may induce large errors depending on the performance metric of interest. This error must be considered by standardization and requires more sophisticated traffic models. As key contributions, we provide realistic traffic models for periodic IoT data, introduce performance metrics for quantifying the bias, and derive reference values as to when the Poisson process can be assumed for aggregated periodic IoT data.**

I. Introduction

The Internet of Things (IoT) is a networking challenge where billions of new devices for countless different purposes will be interconnected across the digital device mesh (e.g., sensors, appliances, IoT systems). According to BusinessInsider[1], IoT devices may account for 24 billion of the 34 billion devices that will be connected to the Internet by 2020. IoT refers to the inter-networking of objects such as physical devices and items which are embedded with software, sensors, actuators, and network connectivity that enable these objects to collect data from multiple modalities (e.g., sight, sound, tactile). The Internet of Things consists of generic, multipurpose devices that are usually connected to the Internet in some fashion. Information to and from the devices is either data that is 1) collected from the devices, aggregated by a gateway (or aggregator), and processed or stored (a typical client-server approach), 2) pushed to the devices, e.g., in a multicast approach, or 3) exchanged between the devices in a peer-to-peer approach. *Machine-to-Machine (M2M)* communication, which can be seen as a subset of IoT, describes a direct interaction between devices using any form of communications, while Device-to-Device (D2D) communication encompasses direct communication between two mobile devices without traversing a base station or other infrastructure in cellular networks (such as LTE, 5G). In this paper, we solely focus on the client-server variant in an "IoT cloud" use case, where the data is collected from a large number of devices and aggregated in a data center.

[1]http://www.businessinsider.com/top-internet-of-things-trends-2016-1

To understand the relationship between the system performance and the number of devices, the traffic patterns generated by IoT devices must be understood and modeled. In a sensor network, the devices are often sending data packets in a deterministic, periodic manner. Aggregated traffic from a large number of devices can be considered as a superposition (aggregation) of point processes, and by assuming the point processes to be independent (cf., e.g., [6, 16, 25]), the aggregated traffic can be modeled as a Poisson process, which significantly simplifies the modeling of the aggregated arrival process. However, due to the deterministic periodicity of the individual device data this will introduce an error term in the quantification of the models when using a Poisson process approximation. This has previously been addressed by [5] and in works on aggregated periodic cell patterns in ATM networks [30]. In general, [3] discusses how the superposition of processes can be applied to modeling packets, flows, and sessions in access and core networks. Yet, current standardization efforts, e.g. in 3GPP [1], assume a Poisson process without taking a closer look at the resulting error [17]. The objective of the paper is to quantify the error introduced by approximating the aggregated periodic traffic process (APP) by a Poisson process (PP). The paper studies and compares different characteristics of the processes, including the mean, variance, coefficient of variation, probability distribution, and autocorrelation. This knowledge is then applied to the case study of a IoT aggregation cloud server in order to investigate the effects on the performance metrics of approximating the APP by a Poisson process. Based on the analytical and numerical results, the paper suggests actions on how a Poisson process approximation for aggregated IoT traffic can be chosen.

The remainder of this paper is structured as follows. Section II gives some background on IoT. Section III surveys related work for the traffic characteristics of select IoT applications. Section IV analyzes the characteristics of aggregated periodic IoT data and compares them with a PP as approximation for the superposed traffic. Section V demonstrates the accuracy and applicability of the PP on the example of the use case. Finally, Section VI concludes this work.

II. The IoT Environment

The IoT ecosystem sets itself apart from classical computer networks in several key areas. Its general focus lies heavily on the network's edges, with numerous devices spread over wide geographical areas. The devices' tasks often revolve around data either actuation or data acquisition. Speaking of the latter, such data is usually then transferred immediately to a more central processing and aggregation node — in contrast to traditional Wireless Sensor Networks (WSNs), which usually operate in a peer-to-peer fashion. This is done in order to keep the IoT devices cheap and extremely resource-limited, enabling suggested run-times of ten years on one reasonably sized battery charge. This also implies a different kind of information flow: an exchange of data from the edges towards the center, with traffic biased towards the uplink direction. Therefore, the general topology is also leaning more towards stars with interconnected central (or regional) aggregation nodes.

These devices also do not operate on traditional connectivity. Instead low power, narrow-banded interfaces are utilized. This mostly centers around radio communications, which in turn is subdivided into short range (Wireless Personal Area Network (WPAN), often from the IEEE 802.15.4 family) and long range communication (Wide-Area Network (WAN)). The latter offers a wide range of different protocols, both standardized and proprietary in nature, either from the category of standalone Low-Power Wide-Area Network (LPWAN) protocols like LoRa, or using narrow-banded or random access modes in the existing, traditional mobile networks, e.g. through NarrowBand IoT (NB-IoT) and other modes intended to be standardized for 3G and other Third Generation Partnership Project (3GPP) endeavors. The proprietary LoRa and LoRaWAN has indeed already garnered significant public interest through some major provider-backed installations, e.g. a deployment providing nation-wide coverage in South Korea[2], or community-run, interconnected gateway networks[3]. Such gateways aggregate the traffic from devices spread over areas of up to tens of kilometers in diameter.

Not just the lowest networking layers are different from current archetypical networking conditions, but the whole stack can be different: from the low power IP variant 6LoWPAN up to domain-specific, resource constrained application-layer protocols, like Constrained Application Protocol (CoAP) and Message Queue Telemetry Transport (MQTT). All in all, this paints a picture of an in all aspects novel networking ecosystem and traffic patterns, that warrant the creation of entirely new traffic models, both single source models for individual IoT device types, but especially also aggregated models that describe entire sets of communication, i.e. exactly what passes through one of these IoT aggregation nodes.

A number of publications already overview the many already existing and proposed applications of IoT, e.g. in industry automation and supervision [35], cloud-backed at home or in enterprise-settings [11], smart environment scenarios [4].

[2]http://www.sktelecom.com/en/press/press_detail.do?idx=1172
[3]E.g. https://www.thethingsnetwork.org/community

III. Characteristics of Select IoT Applications

IoT traffic can be partitioned by their periodic and event-based nature as their reason for communication (see also, e.g., [17, 24]). Some IoT applications will be always more event-driven. Consider for example an exemplary smart home that is outfitted with the latest intrusion alarm devices. These can consist of simple devices that send an alarm when a window or door is opened, up to motion-activated cameras that will then capture and send images to the home owner. These are triggered by events outside of the domain of influence of these devices. But even then, often some emergent periodicity ensues in these triggered events, e.g. the camera is activated each day when leaving for work and returning, cf. also [26].

On the other hand, IoT devices from other fields of application are often much more periodic in nature. A prominent upcoming example are Smart Grids. This includes not only the measurement and collection of current power usage values from residential and industrial Smart Meters, but also includes the supervision, management, and maintenance of the power generation and distribution network [15]. Once again, these usually operate periodic — with different intervals depending on the type of data — but may switch over to pushing events in case of critical readings. Such devices are currently often supported by low-cost GPRS transceivers. Due to the necessary control plane interactions and the limited available radio resources, GPRS can not support arbitrarily short messaging periods for a large number of devices without modification. In a typical scenario the shortest period is estimated to be $5\,\mathrm{min}$ [23]. Additional work proposes to better utilize the Random Access Channel (RACH) in current and future mobile technologies to allow for more devices and be more resource efficient [18, 34].

Traffic models for modern types of radio and mobile communication devices have been investigated in the literature in the past — for example in the interaction of application and signaling traffic as well as the incurred energy usage [31]. Work on wireless sensor nodes has been conducted ,e.g., in [21] by providing numerical simulation results and investigating aggregate packet counts in which both periodic and event-driven communication appears. [2, 14] attempts to show that in Machine-Type Communications (MTC) the classic Markovian arrival process assumption does not hold anymore due to the burstiness of the traffic. Instead, a Beta distribution should be assumed. On the other hand, the work conducted in [33] strives to verify that a Poisson process can indeed be applicable at least to the general (LTE-A) connection establishment process (without limitation to IoT devices). [32] explores a large-scale mobile network measurement dataset for identifiable and well-known M2M device types and evaluates their traffic characteristics. Of special note are the diurnal patterns as well as the session Inter Arrival Times (IATs) that stand apart from typical mobile phone session arrivals. To give an overall picture of MTC and IoT, Table I depicts a collection

Table I: IoT dimensions and messaging intervals. Values are extracted from the given source, rates are per device.

Type	Density	Rate/Period	Source
IMT-2020 Requirements	10^6 devices per km^2	$10\,(\mathrm{Mbit/s})/\mathrm{m}^2$ (indoor hotspot)	[7]
LTE / smart meter	(per PRB) 7.5×10^4 (urban), 5.6×10^4 (suburban)	$2017\,\mathrm{B}$ every $9000\,\mathrm{s}$	[28]
Scenario: IoT and smart grids in the city			
Water meters	10000 per km^2	$100\,\mathrm{B}$ every $43\,200\,\mathrm{s}$	[19]
Electricity meters	10000 per km^2	$100\,\mathrm{B}$ every $86\,400\,\mathrm{s}$	[19]
Gas meters	10000 per km^2	$100\,\mathrm{B}$ every $1800\,\mathrm{s}$	[19]
Vending machines	150 per km^2	$150\,\mathrm{B}$ every $86\,200\,\mathrm{s}$	[19]
Bike fleet management	200 per km^2	$150\,\mathrm{B}$ every $1800\,\mathrm{s}$	[19]
Pay-as-you-drive	2250 per km^2	$150\,\mathrm{B}$ every $600\,\mathrm{s}$	[19]
3GPP TSG RAN WG2 R2-102340 scenarios			
Central London households / smart meters	4968 per cell	periods of $300\,\mathrm{s}$, $900\,\mathrm{s}$, $3600\,\mathrm{s}$, $43\,200\,\mathrm{s}$, or $86\,400\,\mathrm{s}$	[2]
Urban London households / smart meters	35670 per cell	periods of $300\,\mathrm{s}$, $900\,\mathrm{s}$, $3600\,\mathrm{s}$, $43\,200\,\mathrm{s}$, or $86\,400\,\mathrm{s}$	[2]

of measured, assumed, and modeled traffic characteristics from various works and standards.

IV. Superposition of Periodic IoT Data

In [1] the 3GPP notes that *"[...] for a large amount of users the overall arrival process can be modelled as a Poisson arrival process regardless of the individual arrival process."* The question is, does this statement still hold when considering the periodic traffic characteristics of IoT? In this section, we examine the validity of this argument. For this, Section IV-A uses the well-known Palm-Khintchine theorem for the superposition of independent renewal processes. We apply this theorem to the periodic IoT data in Section IV-B and introduce several performance measures in Section IV-C. These measures are used to compare the resulting Poisson process with an APP and are analytically derived in Section IV-D for a system of asynchronous IoT nodes with constant, identical measurement period. In addition, numerical results are provided and visualized.

A. Definitions and Palm-Khintchine Theorem

The fundamental theorem for the superposition of traffic is the Palm-Khintchine theorem, which shows that the superposition of a large number of independent renewal processes will have Poissonian properties and can therefore be described by a Poisson process.

Theorem 1 (Palm-Khintchine Theorem)**.** *Let $\{N_i(t), t \geq 0\}$ be independent renewal processes for $i = 1, 2, \ldots, n$ with iid interarrival times T_i for each renewal process. The superposition $\{N(t) = \sum_{i=1}^{n} N_i(t), t \geq 0\}$ is asymptotically a Poisson process for $n \to \infty$, if:*

1) *Overall load is finite, $k = n / \sum_{i=1}^{n} E[T_i]$.*
2) *No single process dominates the superposition process, $E[T_i] \ll 1/k$.*

The Palm-Khintchine theorem is based on the assumption that there are a very large number of independent stochastic processes at microlevel, where no single process dominates the aggregated process, i.e., where each individual process relatively rarely generates events when compared to the frequency of events in the aggregated process. Then the superposition of all these processes behaves approximately as a Poisson process on the aggregation level. Therefore, in practice Poisson processes can be observed in cases where the process constitutes of a large number of independent processes.

B. Poisson Approximation for Periodic IoT Data

The Palm-Khintchine theorem suggests that the aggregated periodic IoT traffic from several sources can be approximated by a Poisson process. In this context, we define the aggregated periodic traffic process (APP) as a superposition of n independent arrival processes.

Definition 1 (Periodic IoT Traffic)**.** *A single node i is a traffic source which periodically generates a message every time $t_i = kT_i$ ($k \in \mathbb{N}$). The message period is constant and equal to T_i.*

The traffic from these sources can then be aggregated in three different fashions:

1) Synchronous sources, $t_i = t$, $\forall i$
2) Asynchronous sources with the same sampling period, $t_i \neq t_j$ and $T_i = T$
3) Asynchronous sources with different sampling periods, $t_i \neq t_j$ and $T_i \neq T_j$

Synchronous sources can be modeled through periodic batch arrivals with size n. There is no superposition of traffic and no superimposed Poisson process follows and instead can thus be directly analyzed. In many scenarios with periodic IoT traffic — especially with a large number of nodes n — it is much more likely to assume asynchronicity in their periods (synchronicity can pose significant load problems in the aggregated core, and is not easy to achieve across a large number of distributed devices). In this paper we focus solely on asynchronous sources with the same sampling period, with different sampling periods to be investigated at a later point in time.

Definition 2 (Asynchronous Homogeneous Periodic Traffic)**.** *The system consists of n nodes with the same message sampling period T. In asynchronous mode, the nodes start randomly at time $t_i \sim U(0, T)$. Each node i periodically*

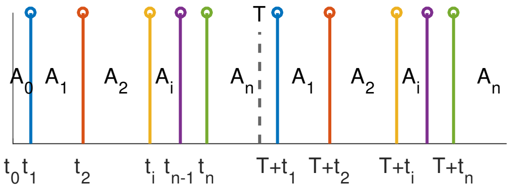

Figure 1: Asynchronous periodic system with n nodes and identical period T is depicted. The interarrival times A_i between the i-th and the $(i+1)$-th $(i = 1, \cdots, n)$ message are identical in each period and $t_i + kT$ for $k \in \mathbb{N}_0$.

generates messages at time $t_i + k \cdot T$ *for* $k = 0, 1, 2, \ldots$. *The interarrival time between the messages of node* i *and node* $i+1$ *is denoted* $A_i = t_{i+1} - t_i$ $(i = 1, \cdots, n-1)$, *with* $t_0 = 0$, *and* $A_n = T + t_1 - t_n$ *(which is the interarrival time between the first message in a window and the last message in the previous window).*

An example of such a sequence of messages t_i with constant period T and the corresponding interarrival times A_i is illustrated in Figure 1. In this paper we focus on homogeneous nodes (with equal $T_i = T$), and in stationary state where each node is sending once (and only once) in the interval $[t_0; t_0 + T]$.

C. Performance Metrics

The Palm-Khintchine theorem holds when n is large, and with independent and identically distributed (iid) interarrival times for each node. This raises two questions: *When is* n *large enough so that the Poisson process is a proper assumption? How much of an error is introduced by this assumption and which traffic characteristics are affected?* The Poisson approximation is described by the interarrival time distribution $A^* \sim \mathrm{Exp}(n/T)$ which is compared to the A_i of the APP. To this end, we introduce several metrics and their ideal outcomes to quantify the error between APP and Poisson process.

(1) Rate of APP (n/T) and Poisson process (λ^*) are identical.

$$\lambda^* = n/T \tag{1}$$

(2) Relative error between average expected interarrival times $\bar{A}$ (APP) and Poisson process should approach zero.

$$r_A = \left| 1 - \frac{\bar{A}}{\mathrm{E}[A^*]} \right| < \epsilon \quad \text{with} \quad \bar{A} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}[A_i] \tag{2}$$

(3) Average expected shift $\bar{S}$ between arrival times of APP and Poisson process should approach zero.

$$\bar{S} = \frac{1}{n} \sum_{i=1}^{n} |\tau_i^* - \tau_i| = \frac{T}{2n} < \epsilon \tag{3}$$

(4) Relative error between the average coefficient of variation $\bar{C}$ of interarrival times of APP and Poisson process ($c_{A^*} = 1$) should approach zero

$$r_C = \left| 1 - \bar{C} \right| < \epsilon \quad \text{with} \quad \bar{C} = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathrm{Std}[A_i]}{\mathrm{E}[A_i]} \tag{4}$$

(5) Number of arrivals N^* within T of Poisson process, $N^* \sim \mathrm{Poiss}(\lambda^* T)$, should be close to n. Thus, the coefficient of variation c_N should approach zero.

$$c_{N^*} = \frac{\mathrm{Std}[N^*]}{\mathrm{E}[N^*]} = \frac{1}{\sqrt{\lambda^* T}} = \frac{1}{\sqrt{n}} < \epsilon \tag{5}$$

These metrics are further defined and discussed in the following section.

D. Asynchronous IoT Data with Same Period

We consider a fixed frequency $1/T$ of measurement data for each of the IoT nodes. We assume that the IoT nodes are not synchronized and start randomly at $t_i \sim U(0, T)$ as given in Def. 2.

1) Expected Time between Arrivals: The expected interarrival time $\mathrm{E}[A_i]$ between two consecutive IoT messages is

$$\mathrm{E}[A_i] = \frac{T}{n+1} \tag{6}$$

for $i = 0, \ldots, n-1$. Note that $\mathrm{E}[T_n] = T - t_n$ is also $\mathrm{E}[T_n] = \frac{T}{n+1}$, and hence

$$\mathrm{E}[A_n] = \mathrm{E}[T_n] + \mathrm{E}[A_0] = 2\frac{T}{n+1}\,. \tag{7}$$

A proof is sketched in the following. Let X be a random variable (RV) uniformly distributed over $(0; T)$, i.e. $X \sim U(0, T)$. The probability density function is $f(x) = 1/T$ if $x \in (0; T)$, otherwise $f(x) = 0$. Then, let X_1 and X_2 be two RVs following X. The distance between the two random points X_1, X_2 in the interval $(0, T)$ is then also a RV, $Y = |X_1 - X_2|$. The joint probability $P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1)P(X_2 = x_2) = \frac{1}{T^2}$ since X_1 and X_2 are independent.

$$\mathrm{E}[Y] = \int_0^T \int_0^T |x_1 - x_2| \frac{1}{T^2} dx_1 dx_2 = \frac{T}{3} \tag{8}$$

Accordingly, we can derive this for larger values of $n > 2$ resulting in Eq. (6). A formal proof is given in [20].

2) Relative Error of Expected Interarrival Time: As explained in the previous section, the expected interarrival of the APP is $\mathrm{E}[A_i] = T/(n+1)$ for $i = 1, \cdots, n-1$ and $\mathrm{E}[A_n] = 2T/(n+1)$. The expected interarrival time over $(0; T)$ is then

$$\bar{A} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}[A_i] = \frac{1}{n} \left(\frac{(n-1)T}{n+1} + \frac{2T}{n+1} \right) = \frac{T}{n} \tag{9}$$

The expected interarrival time of the Poisson process is $\mathrm{E}[A^*] = T/n$. Hence, the relative error between the expected interarrival times in the APP and Poisson process is zero.

3) Expected Traffic Shift: When considering the interarrival times, using Poisson process as an approximation of the APP will introduce a bias τ_i between the arrivals, see Figure 2. The expected shift (bias) τ_i of the i-th arrival is as follows.

$$\tau_i = t_i^* - t_i = i \cdot \left(\frac{T}{n} - \frac{T}{n+1} \right) \tag{10}$$

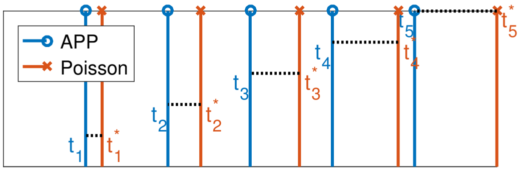

Figure 2: Expected interarrivals $t_i, i = 1, \ldots, n$ of the aggregated periodic process (blue **O**) and expected interarrivals t_i^* of a Poisson process (red **x**) with $\lambda = n/T$ in $[0; T]$.

The expected shift over $(0; T)$ is then

$$\mathrm{E}[S] = \sum_{i=1}^{n} \tau_i = \frac{T}{2}\,. \tag{11}$$

and the expected shift per node in $(0; T)$ is $E[S]/n = \frac{T}{2n}$. For $n > \frac{T}{2\epsilon}$, the average shift becomes negligibly small. For example, with $n > 50$ nodes, the error is smaller than $\epsilon = 1\%$ for $T = 1$.

4) Variance of Periodic Traffic: The first $n-1$ phases follow a Beta distribution $X \sim \mathrm{Beta}(1, n, 0, T)$. Let us consider the *first order statistic* X of the uniform distribution. The t_i are iid and uniformly distributed in $(0; T)$, i.e. $t_i \sim U(0, T)$ with PDF $f(t) = 1/T$ and CDF $F(t) = t/T$ for $0 \le t \le T$.

Let $X = \min\{t_i\}$ be a RV that describes the minimum of the t_i. Then, the CDF of X follows a Beta distribution.

$$\begin{aligned} X(t) &= P(X \le t) = 1 - P(X > t) \\ &= 1 - P(t_1 > t, \ldots, t_n > t) = 1 - \prod_{i=1}^{n} P(t_i > t) \\ &= 1 - (1 - F(t))^n = 1 - (1 - t/T)^n \end{aligned} \tag{12}$$

for which the mean and the coefficient of variation are

$$\mathrm{E}[X] = \frac{T}{n+1}, \qquad c_X = \frac{\sqrt{n}}{\sqrt{n+2}}\,. \tag{13}$$

For the last phase, it is $A_n = X + X$. However, the intervals A_i are *not independently distributed*. The reason is that $\sum_{i=1}^{n} A_i = T$. Therefore, we numerically derive the average coefficient of variation over all n nodes, i.e. $\bar{C} = \frac{1}{n} \sum =_{i=1}^{n} c_i$, which is fitted in Figure 3 leading to

$$\bar{C} = \frac{n-1}{n} = 1 - \frac{1}{n} \quad \text{and} \quad r_C = \left|1 - \bar{C}\right| < \epsilon\,. \tag{14}$$

Thus, for $n > 1/\epsilon$, the relative error of the coefficient of variation is smaller than ϵ, e.g. for $\epsilon = 1\%$ this holds for $n > 100$.

5) Deviation from expected number of arrivals n in $(0; T)$: The APP generates a fixed number of n arrivals in $(0; T)$, while for the Poisson process the expected number of arrivals in $(0; T)$ is $\lambda^* T = n$.

The Poisson distribution yields the probability that exactly n arrivals will occur in $(0; T)$.

$$P_T(n) = \frac{n^n}{n!} e^{-n} \tag{15}$$

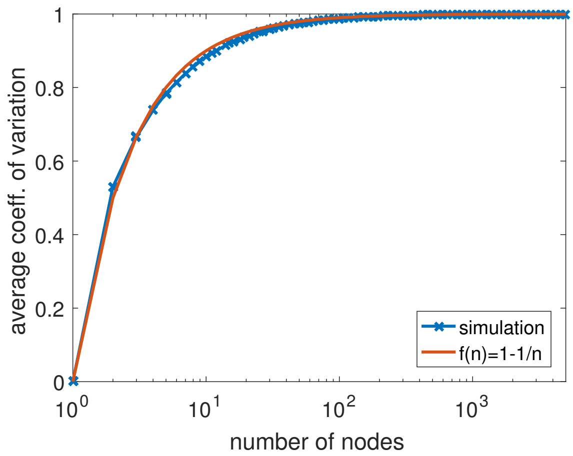

Figure 3: The average coefficient of variation $\bar{c}_i$ of the interarrival times A_i in an asynchronous mode with n nodes converges towards 1.0 for an increasing n, $\bar{C} = 1 - 1/n$.

However, this probability is decreasing with increasing n and even $P_T(1) = 0.3679 \ll 1$. When considering the coefficient of variation of the Poisson process, it is

$$c_{N^*} = \frac{1}{\sqrt{n}} < \epsilon \tag{16}$$

which should be close to zero. For $n > \epsilon^{-2}$, the error is smaller than ϵ.

V. Use Case: Load Balancer at IoT Cloud

To numerically compare the APP with the related Poisson process, we take an IoT cloud as a concrete example. Here, data arrives from sensor nodes and is aggregated at a load balancing gateway which forwards the data to the backend cloud servers. The crucial performance measure for dimensioning the load balancer is the waiting time. The system is modeled as an $nD/D/1$ queue (denoted in Kendall notation) and analyzed in [22, 30]. We further investigate the impact of additional network transmission delays, which leads to an $nG/D/1$ system by means of simulations. The use case shows the limits of the Poisson process approximation.

A. Scenario

We consider a scenario where n nodes are periodically sending messages which arrive at an IoT cloud to be processed. The messages arrive at a load balancer, i.e. the first point of the cloud architecture where the individual traffic flows become aggregated. Due to the large number of IoT devices, such a load balancer is required to distribute the workload across backend servers [12]. The nodes are asynchronous, but have the same sending period T. The processing time S to handle the (typically small) IoT message at the load balancer is constant, see Table I. This system is denoted as $nD/D/1$ waiting system. We consider $S = 1$ time units and express time related measures (like waiting times) relative to S. The system is illustrated in Figure 4.

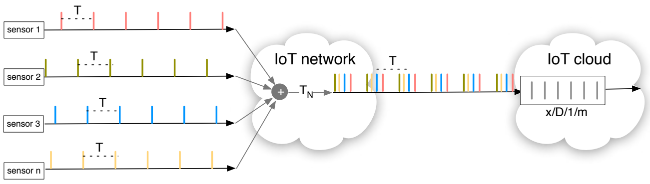

Figure 4: Illustration of the load balancer at an IoT cloud.

B. M/D/1 System State Distribution

An approximation of the system is an $M/D/1$ system in which the arrivals are approximated with a Poisson process with rate $\lambda = n/T$. Iversen and Staalhagen provide an efficient calculation of the $M/D/1$ system state probabilities [13], i.e. the number of customers i in the system. The state probabilities $P(i)$ are recursively computed based on Fry's equation [10]. The load in the system is $\rho = \lambda S$ with constant service time S.

$$P(0) = (1-\rho) \tag{17}$$

$$P(1) = (1-\rho)(e^{\rho}-1) \tag{18}$$

$$P(i+1) = \frac{1}{P(0,\rho)}\Bigg(P(i) - (P(0)+P(1))\cdot P(i,\rho) - \sum_{n=2}^{i} P(n)\cdot P(i-n+1,\rho)\Bigg) \tag{19}$$

Thereby, we define $P(i,\rho)$ as follows.

$$P(i,\rho) = \frac{\rho^i}{i!}\mathrm{e}^{-\rho} \tag{20}$$

C. nD/D/1 System

Roberts and Virtamo analyze the state probability for the $nD/D/1$ queue in [30] based on [9].[4] They compare the system to $M/D/1$ and find that the Poisson approximation can lead to a significant overestimation of buffer requirements, particularly in case of heavy load. For $0 \le r < n$, we have the following complementary cumulative distribution of the number X of customers, while $P(X>0) = 0$ for $r \ge n$.

$$P(X>r) = \sum_{s=1}^{n-r}\binom{n}{r+s}\left(\frac{s}{S}\right)^{r+s}\left(1-\frac{s}{S}\right)^{n-r-s}\frac{S-n+r}{S-s} \tag{21}$$

Good approximation formulae for the waiting time under high load readily exist, as is summarized by Menth and Muehleck in [22].

$$P(W>t) \approx exp\left(\frac{-2t}{S}\left(\frac{t}{(n-1)S}+1-\frac{nS}{\rho}\right)\right) \tag{22}$$

Figure 5 compares the CCDF of the system state X for $nD/D/1$ and $M/D/1$ for high load ($\rho = 0.95$). It can be seen that with a higher number of nodes, the $nD/D/1$ curve approaches the $M/D/1$ curve. Nevertheless, there is a significant gap and the Poisson approximation overestimates the buffer requirements when dimensioning the load balancer according to a certain threshold $P(X > x)$. However, for fairly low load ($\rho = 0.55$), the difference between $nD/D/1$ and $M/D/1$ is negligible, as we argue in the next section.

[4] The report in [29] also provides a comprehensive summary of cost models for periodic traffic.

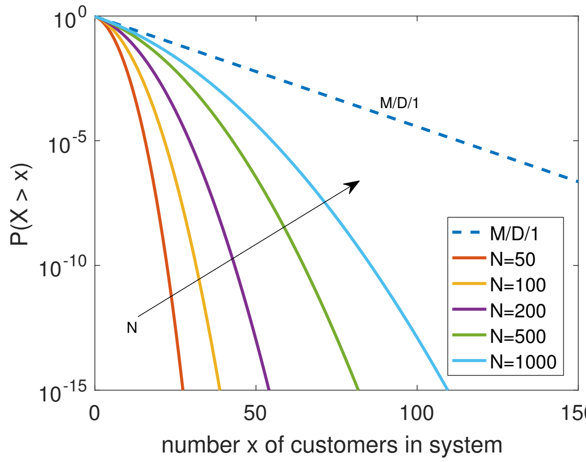

Figure 5: CCDF of the system state $M/D/1$ and $nD/D/1$ under heavy load ($\rho = 0.95$).

D. Mean waiting time and relative error

For $M/D/1$, the expected number of customers X is well known and follows from Eq. (19).

$$\mathrm{E}[X] = \rho + \frac{\rho^2}{2(1-\rho)} \tag{23}$$

With Little's Law, the expected waiting time follows accordingly from $\mathrm{E}[X] = (\mathrm{E}[W]+S)\lambda$.

$$\mathrm{E}\big[W_{M/D/1}\big] = \frac{S\cdot\rho}{2(1-\rho)} \tag{24}$$

For $nD/D/1$, the expected waiting time is derived from a result from Eckberg [8].

$$\mathrm{E}\big[W_{nD/D/1}\big] = \frac{n-1}{2T\cdot B(n-2,T)} \tag{25}$$

For the computation of the Erlang-B formula (B, M, a), the iterative method is used.

$$B(0,a) = 1, \qquad \frac{1}{B(n,a)} = 1 + \frac{n}{a\cdot B(n-1,a)} \tag{26}$$

Figure 6 shows the mean waiting times in relation to the service time S on the y-axis, while the system load is depicted on the x-axis. For heavy load, there are significant differences. The higher the number of nodes the closer the $nD/D/1$ system approaches $M/D/1$. For high load $\rho = 0.95$, the relative error

$$r_W(\rho) = \left|1 - \frac{\mathrm{E}\big[W_{nD/D/1}\big]}{\mathrm{E}\big[W_{M/D/1}\big]}\right| \tag{27}$$

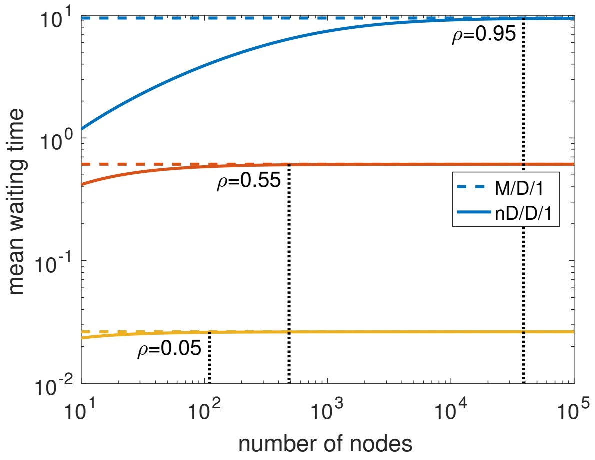

Figure 6: Mean waiting times (normalized by service time S) for $nD/D/1$ and $M/D/1$ reveal significant differences for high load, e.g. $\rho = 0.95$. The higher the number of nodes is, the smaller the difference becomes.

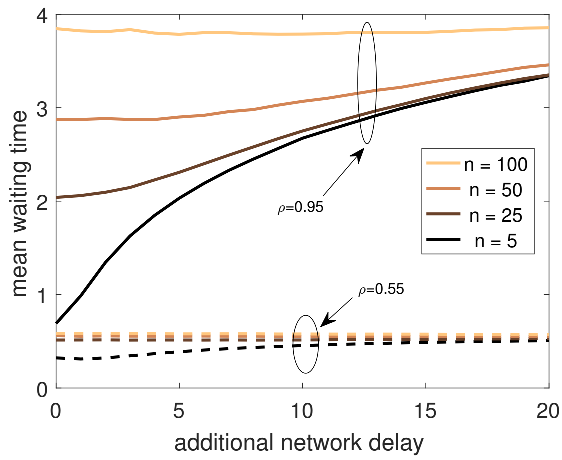

Figure 7: Impact of additional network delay on the mean waiting time for high and medium load.

of the expected waiting time is smaller than epsilon for $n > 38{,}900$. An intuitive explanation of the differences between the mean waiting times of $M/D/1$ and $nD/D/1$ is the boundedness of individual waiting times. In $nD/D/1$, the individual waiting time is bounded by $S(n-1)$, which happens if all n arrivals occur simultaneously. Ramamurthy and Sengupta explain it with the busy period [27]: "*In the $M/D/1$ queue both the waiting time and the busy period is unbounded as the utilization gets close to one* [i.e. heavy load]. *On the other hand, in our $[nD/D/1]$ model the busy period (and consequently the waiting time) is always upper bounded by one. For this reason, it is not surprising that the $M/D/1$ results overestimate those of our $[nD/D/1]$ model.*"

E. Impact of Network Transmission (nGI/D/1)

The $nD/D/1$ model of the IoT load balancer is refined in this section, as we additionally take into account shifts in the arrival pattern, since the sources have to transmit the data to the IoT load balancer. The network transmission delay is a random variable δ and we assume an exponential distribution, $\delta \sim \text{Exp}(\mu)$ with mean delay $\delta_m = 1/\mu$. Thus, any packet sent at time t arrives at the load balancer at time $t+\delta$. As a consequence, the interarrival times I per node do not follow a deterministic distribution, but it is the convolution of the network transmission delay, $I = (i+1)T + \delta - (iT+\delta) = T + \delta - \delta$, for which the following CDF can be derived easily.

$$F(t) = \begin{cases} \frac{1}{2} e^{\mu(t-T)} & , t \le T \\ 1 - \frac{1}{2} e^{-\mu(t-T)} & , t > T \end{cases} \tag{28}$$

The interarrival times between consecutive messages from two nodes is more complex. In particular, messages can overtake other messages and reordering may occur.

Figure 7 presents some simulation results of the mean waiting time for varying network delays and number of nodes.

Table II: Guidelines for the minimum number n of nodes such that the relative error due to Poisson approximation is below a threshold ϵ. We consider $T = 1$, such that n depicts the number of messages per seconds $\lambda = n/T$. 'CoV' abbreviates the coefficient of variation of a random variable.

Measure	**Description**	**Formula**	$\boldsymbol{\epsilon = 0.1}$
Bias of Poisson process to approximate APP arrival pattern			
r_A	mean interarrival time (IAT)	$r_A = 0$	any $n > 0$
$\bar{S}$	avg. shift of IAT	$n > T/2\epsilon$	$n/T > 50$
r_c	CoV of IAT Eq. (14)	$n > 1/\epsilon$	$n > 100$
c_{N^*}	CoV number arrivals in T	$n > 1/\epsilon^2$	$n > 10000$
Example: Waiting times at IoT load balancer			
$r_W(\rho)$	rel. error waiting time	numerically	depends on ρ
		for $\rho = 0.95$	$n > 38{,}899$
		for $\rho = 0.55$	$n > 486$
		for $\rho = 0.15$	$n > 110$

The load in the system is fixed to high load ($\rho = 0.95$) or medium load ($\rho = 0.55$). It can be seen that the additional network delay does not have an impact when n is large. But the mean waiting time still depends on the number n of nodes, see also Figure 6. Therefore, the minimal number of nodes n to keep the error below a threshold is similar. Only — in the (unrealistic) case that the additional network delay δ is much larger than the period T — then the aggregated process leads to a Poisson process, as the sending period T has no significant influence anymore.

In summary, network transmission delays do not significantly influence the waiting times of an IoT load balancer for realistic periodic IoT data. The results imply that an implementation of the IoT stack is not necessary for dimensioning the IoT load balancer. Rather when n is not large, it is more crucial to consider $nD/D/1$ instead of the Poisson process.

VI. Conclusion

IoT traffic characteristics often revolve around periodic communication from asynchronous homogeneous sources with a constant sampling period. For the analysis of such IoT traffic, the superposition of the traffic streams from n such nodes can be approximated by a Poisson process which allows to easily derive results for large-scale IoT systems. However, due to the Poisson approximation, an error is introduced when comparing different characteristics of the Poisson process and the aggregated periodic process. This error is often neglected in literature and standardization, although this may lead e.g. to overdimensioning as in the case of the IoT load balancer.

We introduced several performance metrics and derived this bias analytically and numerically. The results and concrete minimal values for the number of nodes n to keep the relative error below a certain threshold (e.g. 1%) are summarized in Table II. Depending on the concrete use case or characteristic under consideration, the minimal value of n varies significantly. For characteristics like autocorrelation of waiting times, the Poisson process is not able to capture the characteristics of the APP. However, if those characteristics are not relevant (e.g. to dimension the buffer of an IoT gateway and load balancer), then the Poisson process is a good approximation. Especially, in many realistic traffic models (see Table I) the number of nodes is large enough, such that the bias remains small. Nevertheless, when looking at waiting times of the aggregated traffic, e.g. at an IoT load balancer, in high load scenarios, a very large number of nodes is required to keep the bias low. Future work will additionally take into account heterogeneous nodes with random periods (i.e. traffic from asynchronous source with different sampling periods as defined in Sec. IV-B) as well as the superposition in hierarchical IoT networks.

REFERENCES

[1] 3GPP. *GERAN improvements for Machine-Type Communications (MTC)*. TR 43.868. Feb. 2014.

[2] 3GPP. *RAN Improvements for Machine-type Communications*. TR 37.868. Oct. 2011.

[3] M. A. Arfeen et al. "Internet traffic modelling: from superposition to scaling." In: *IET networks* 3.1 (2014).

[4] L. Atzori, A. Iera, and G. Morabito. "The Internet of Things: A survey." In: *Computer Networks* 54.15 (2010).

[5] E. Cinlar and R. A. Agnew. "On the Superposition of Point Processes." In: *Journal of the Royal Statistical Society. Series B (Methodological)* 30.3 (1968).

[6] D. R. Cox and W. L. Smith. "On the superposition of renewal processes." In: *Biometrika* 41.1-2 (1954).

[7] *Draft new Report ITU-R M.[IMT-2020.TECH PERF REQ] - Minimum requirements related to technical performance for IMT-2020 radio interface(s)*. International Telecommunication Union Radiocommunication Sector, Feb. 2017.

[8] A. Eckberg Jr and L. Green. "Response time analysis for pipelining jobs in a tree network of processors." In: *Applied Probability-Computer Science: The Interface Volume 1*. Springer, 1982.

[9] A. Eckberg. "The single server queue with periodic arrival process and deterministic service times." In: *IEEE Transactions on communications* 27.3 (1979).

[10] T. C. Fry et al. *Probability and its engineering uses*. Van Nostrand New York, 1928.

[11] J. Gubbi et al. "Internet of Things (IoT): A vision, architectural elements, and future directions." In: *Future Generation Computer Systems* 29.7 (2013).

[12] L. Hou et al. "Internet of Things Cloud: Architecture and Implementation." In: *IEEE Communications Magazine* 54.12 (Dec. 2016).

[13] V. B. Iversen and L. Staalhagen. "Waiting time distribution in M/D/1 queueing systems." In: *Electronics Letters* 35.25 (1999).

[14] X. Jian et al. "Statistical Description and Analysis of the Concurrent Data Transmission from Massive MTC Devices." In: *International Journal of Smart Home* 8.4 (2014).

[15] R. H. Khan and J. Y. Khan. "A comprehensive review of the application characteristics and traffic requirements of a smart grid communications network." In: *Computer Networks* 57.3 (2013).

[16] A. Y. Khinchin, D. M. Andrews, and M. H. Quenouille. *Mathematical methods in the theory of queuing*. Londin: Griffin, 1960.

[17] M. Laner et al. *M2M Traffic and Models*. 2014.

[18] A. Laya et al. "Massive access in the Random Access Channel of LTE for M2M communications: An energy perspective." In: *IEEE ICC 2015 - Workshop on Smart Communication Protocols and Algorithms*. June 2015.

[19] *Massive IoT in the City*. Ericsson White paper, Nov. 2016.

[20] A. M. Mathai. *An introduction to geometrical probability: distributional aspects with applications*. Vol. 1. CRC Press, 1999.

[21] M. A. Mehaseb, Y. Gadallah, and H. El-Hennawy. "WSN Application Traffic Characterization for Integration within the Internet of Things." In: *2013 IEEE 9th International Conference on Mobile Ad-hoc and Sensor Networks*. Dec. 2013.

[22] M. Menth and S. Muehleck. "Packet waiting time for multiplexed periodic on/off streams in the presence of overbooking." In: *International Journal of Communication Networks and Distributed Systems* 4.2 (2010).

[23] J. J. Nielsen et al. "What can wireless cellular technologies do about the upcoming smart metering traffic?" In: *IEEE Communications Magazine* 53.9 (Sept. 2015).

[24] N. Nikaein et al. "Simple Traffic Modeling Framework for Machine Type Communication." In: *ISWCS 2013; The Tenth International Symposium on Wireless Communication Systems*. Aug. 2013.

[25] C. Palm. "Intensitätsschwankungen im Fernsprechverkehr." In: *Ericsson Technics no.* 44 (1943).

[26] I. Petiz, P. Salvador, and A. Nogueira. "Characterization and modeling of m2m video surveillance traffic." In: *IARIA Int. Conf. on Advances in Future Internet-AFIN*. Citeseer. 2012.

[27] G. Ramamurthy and B. Sengupta. "Delay analysis of a packet voice multiplexer." In: *Telecommunications Symposium, 1990. ITS'90 Symposium Record., SBT/IEEE International*. IEEE. 1990.

[28] R. Ratasuk et al. "Recent advancements in M2M communications in 4G networks and evolution towards 5G." In: *2015 18th International Conference on Intelligence in Next Generation Networks*. Feb. 2015.

[29] J. Roberts, U. Mocci, and J. Virtamo. "Broadband Network Teletraffic: Final Report of Action COST 242." In: (1996).

[30] J. W. Roberts and J. T. Virtamo. "The superposition of periodic cell arrival streams in an ATM multiplexer." In: *IEEE Transactions on Communications* 39.2 (1991).

[31] C. Schwartz et al. "Angry Apps: The Impact of Network Timer Selection on Power Consumption, Signalling Load, and Web QoE." In: *Journal of Computer Networks and Communications* 2013 (2013).

[32] M. Z. Shafiq et al. "A First Look at Cellular Machine-to-machine Traffic: Large Scale Measurement and Characterization." In: *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems*. SIGMETRICS '12. London, England, UK: ACM, 2012.

[33] R. R. Tyagi et al. "Connection Establishment in LTE-A Networks: Justification of Poisson Process Modeling." In: *IEEE Systems Journal* PP.99 (2015).

[34] M. Vilgelm and W. Kellerer. "Impact of request aggregation on machine type connection establishment in LTE-Advanced." In: *Proc. IEEE WCNC*. 2017.

[35] L. D. Xu, W. He, and S. Li. "Internet of Things in Industries: A Survey." In: *IEEE Transactions on Industrial Informatics* 10.4 (Nov. 2014).