Physical Layer Authentication in Mission-Critical MTC Networks: A Security and Delay Performance Analysis

Henrik Forssell, Ragnar Thobaben, Hussein Al-Zubaidy, James Gross

***Abstract*—We study the detection and delay performance impacts of a feature-based physical layer authentication (PLA) protocol in mission-critical machine-type communication (MTC) networks. The PLA protocol uses generalized likelihood-ratio testing based on the line-of-sight (LOS), single-input multiple-output channel-state information in order to mitigate impersonation attempts from an adversary node. We study the detection performance, develop a queueing model that captures the delay impacts of erroneous decisions in the PLA (i.e., the false alarms and missed detections), and model three different adversary strategies: data injection, disassociation, and Sybil attacks. Our main contribution is the derivation of analytical delay performance bounds that allow us to quantify the delay introduced by PLA that potentially can degrade the performance in mission-critical MTC networks. For the delay analysis, we utilize tools from stochastic network calculus. Our results show that with a sufficient number of receive antennas (approx. 4-8) and sufficiently strong LOS components from legitimate devices, PLA is a viable option for securing mission-critical MTC systems, despite the low latency requirements associated to corresponding use cases. Furthermore, we find that PLA can be very effective in detecting the considered attacks, and in particular, it can significantly reduce the delay impacts of disassociation and Sybil attacks.**

***Index Terms*—Delay performance, low-latency machine-type communication, wireless physical layer security, physical layer authentication.**

I. Introduction

As mission-critical machine-type communication (MTC) emerges as a new approach to interconnect cyber-physical infrastructures, also new requirements on security features arise. Mission-critical machine-type communication targets at low latencies and high transmission reliabilities, in order to realize new use cases for instance arising in industrial automation. Thus, while in human-oriented communication data confidentiality followed by integrity form the utmost priorities (while service availability and security overhead typically have less relevance), the priorities change in the mission-critical setting. In detail, the order of concern is reversed [1]: Service availability has highest priority since automation applications are typically supposed to run uninterrupted over long time spans. The second highest priority has message integrity, as in a closed control loop it is of vital importance that sensor and actuation information is not altered during transmission, while it also must be assured that the received data indeed stems from the claiming source. Finally, confidentiality is of least importance, as in automation applications the reading of sensor and actuation information poses only little threat to the controlled plant. Paired with the general requirement for low transmission latencies, these inverted security priorities are challenging, as traditionally integrity is assured through crypto schemes on the higher layers, which comes with significant computational complexities.

Physical layer authentication (PLA) has been proposed as a lightweight alternative for crypto security for authentication in reliable MTC communications [2]. In general, PLA schemes perform hypothesis testing based on dedicated features of the communication pair like, e.g., the location-specific channel frequency response [3] or a device-specific local oscillator offset [4] to determine if transmissions originate from legitimate sources. The advantage of this method is that messages can be authenticated quickly at the physical layer, without relying on cryptographic methods at higher layers and with slim-to-none security overhead. However, such schemes also come with drawbacks. First of all, due to the hypothesis testing PLA inevitably results in false alarms from time to time (i.e., some legitimate messages will be erroneously rejected) which can necessitate a retransmission. Furthermore, missed detections (i.e., accepting messages from an adversary) can occur if communication is subject to an impersonation attack. Thus, despite the complexity advantages, PLA also comes with costs which potentially can be significant in the context of mission-critical MTC. This raises the question how these costs (i.e., false positives and missed detections) potentially impact in particular the delay performance of a mission-critical MTC system.

Related work so far has largely not been addressing this question. PLA for mission-critical MTC is proposed for instance in [2, 5] but without considering the impact on the delay. In [6], the reduction in delay from removing authentication-induced processing delays in cell-handovers by using PLA is simulated. However, this paper does not focus on MTC and additionally does not take false alarms of PLA into account. Ozmen *et al.* considers the delay-sensitive performance of a communication system under information-theoretic secrecy [7, 8]. Delay in these works is characterized through the concept of effective capacity, which essentially allows for the approximation of queuing-related performance metrics like the backlog or latency. Furthermore in [9], the

Manuscript received June 22, 2018; revised December 9, 2018; accepted January 25, 2019. This work was supported in part by the Swedish Civil Contingencies Agency, MSB, through the CERCES project.

H. Forssell, R. Thobaben, H. Al-Zubaidy, and J. Gross are with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden (e-mail: hefo@kth.se; ragnart@kth.se; hzubaidy@kth.se; james.gross@ee.kth.se).

delay performance of a Rayleigh fading wiretap channel is studied using *stochastic network calculus* for queueing analysis. All papers [7–9] apply queueing analysis tools to study the delay impacts of different physical layer security techniques; however, none of them considers PLA.

In this work, we are interested in delay analysis that quantifies the cost of authenticating messages at the physical layer in a mission-critical communication system. To this end, we develop and analyze queueing models that capture the erroneous decisions of the PLA, and furthermore, include attacker behaviors such as data injection, disassociation, and Sybil attacks. The queueing performance in the developed models is analytically guaranteed in terms of upper bounds on the delay violation probability, derived using tools from stochastic network calculus [10], that allow us to quantify the authentication-induced delays and attacker impacts. The considered system setup consists of a centralized MTC network with a multiple-antenna access point that employs PLA to authenticate multiple single-antenna devices. The PLA scheme studied in this paper is based on generalized likelihood-ratio hypothesis testing (e.g., similar to [3]) and extended to allow multiple-message authentication. However, the developed queueing models are general and apply to other PLA approaches as well. This work thereby significantly extends the analysis in our previous work [11] where we considered the delay impacts of single-antenna PLA without an active attacker.

The contributions of this paper are the following: We develop queueing models for PLA equipped wireless systems to study how erroneous authentication decisions impact the delay performance, in particular when under attacks such as disassociation, and Sybil attacks, and use these to study the performance of a particular PLA scheme based on the channel state of the single-input/multiple-output channel. Furthermore, we derive delay performance bounds for the developed queueing model by using the stochastic network calculus framework. With respect to stochastic network calculus, we provide an approximation to a previously unsolved mathematical problem: an upper bound on the delay violation probability over a Rice fading single-input multiple-output channel. From our results, we conclude that PLA, under relatively strong line-of-sight conditions and with sufficient number of receive antennas, can indeed provide high security in a mission-critical application. We also show that PLA effectively reduces the impact of disassociation and Sybil attacks at the cost of an approximately constant increase in delay violation probability. Thus, our results show that despite some costs, PLA promises to be an effective scheme in ensuring message integrity even in mission-critical MTC systems.

The rest of the paper is organized as follows: Section II introduces the system assumptions and our problem formulation. In Section III, we describe the attacker models and their impact on the queueing model. Section IV is devoted to deriving the delay performance bounds using tools from stochastic network calculus. In Section V, we present our numerical results, and Section VI concludes the paper.

Notation: Matrices are represented by bold capital symbols $\mathbf{X}$, and $\mathbf{X}^T$ and $\mathbf{X}^\dagger$ denote the matrix transpose and conjugate transpose, respectively. We let $\mathrm{tr}(\mathbf{X})$ denote the trace of a matrix. Bold symbols $\mathbf{x}$ represents vectors with entries x_i and $\mathbf{I}_N$ denotes the $(N \times N)$ identity matrix. We let $\|\mathbf{x}\| = \sqrt{|x_1|^2 + ... + |x_n|^2}$ be the Euclidian norm. For an event E, we let $\mathbb{P}(E)$ and $\mathbb{I}(E)$ denote the probability and indicator function, respectively. For a random variable X, $\mathbb{E}[X]$ denotes its expected value and $f_X(x)$ and $F_X(x)$ its probability density and cumulative distribution function, respectively. We let $\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ represent the multivariate complex Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the corresponding real-valued Gaussian distribution, χ^2_k a central χ^2 distribution with k degrees of freedom, and $\chi^2_k(\lambda)$ a non-central χ^2 distribution with k degrees of freedom and non-centrality parameter λ.

II. Preliminaries

In this section, we present a centralized MTC network model consisting of K_d wireless devices communicating uplink data to an access point, as depicted in Fig. 1. The network is assumed to run a mission-critical application in which the MTC devices buffer data (e.g., sensor measurements) that need to be delivered reliably to the access point with minimal delay, as for example in motion control or generally in factory automation. Furthermore, as depicted in Fig. 1, we assume that there is an adversary present in the vicinity of the network, attempting to disturb the system using stealthy wireless impersonation attacks that are compliant with the typical behavior of legitimate devices within the network (e.g., sending payloads, data or disconnection requests). For protection against such attacks, the access point is using a feature-based physical layer authentication (PLA) protocol that compares the channel-state information associated with each transmission to a pre-stored feature bank. The access point is assumed to be equipped with N_Rx antennas, both in order to improve the PLA detection performance and to improve capacity, while the MTC devices (e.g., small sensors) are assumed to have single antennas. The stationary feature bank consists of the statistics of the phased-array antenna responses from each device to the multiple-antenna access point, and we assume that the devices are deployed such that a line-of-sight (LOS) path to the access point is available.

A. Medium Access and Physical Layer

We assume that the MTC devices access the wireless medium in a frame-based structure, each beginning with a beacon transmitted by the access point for synchronization, followed by a management (MGMT) period where devices can make various requests[1]. A device can request connecting to the access point (CN), disconnecting (DCN), or resources for transmission of data payload (DTA). The allocation of resources is then communicated to the devices in a broadcast period (BP), followed by the data transmission period (DTP) where devices transmit buffered data. We let $\mathcal{I}_\mathrm{MGMT}(k)$ denote the set of request messages received in

[1]The MGMT phase is based on contention access; however, we assume that collisions are handled appropriately such that we can neglect their impact.

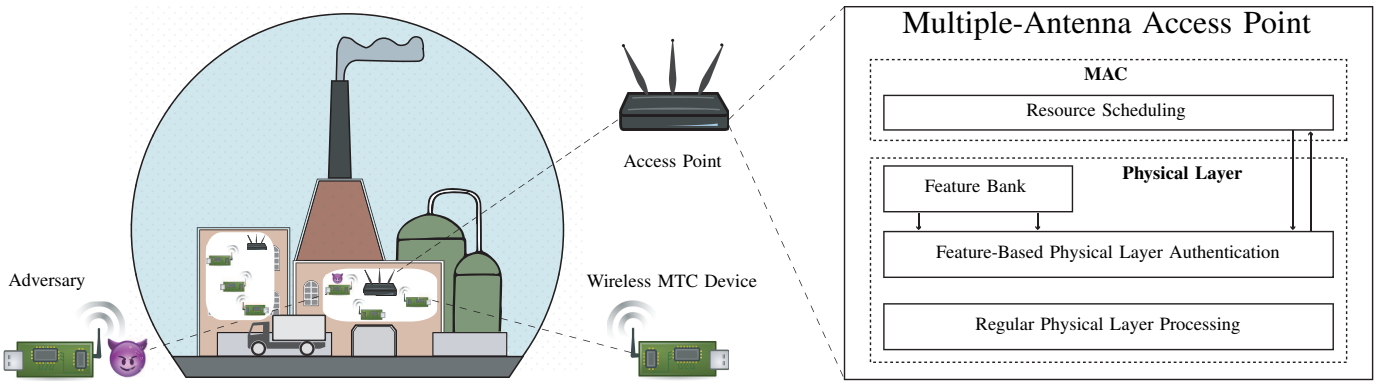


Fig. 1. Single-antenna MTC devices (e.g., wireless sensors in a critical monitoring application) communicating in uplink to a multiple-antenna access point. The access point is equipped with a feature-based PLA protocol.

the MGMT period in frame k, each associated with a request $\text{REQ}(m) \in \{\text{DTA}, \text{CN}, \text{DCN}\}$ and a device identifier $\text{ID}(m) \in \{1, \cdots, K_d\}$ (e.g., an identification code such as a MAC address). We denote by $\mathcal{I}_{\text{DTP}}(k)$ the set of devices that are granted DTP resources in frame k and we assume that the access point expects at most one request from each device. This setup could be extended to model periodic transmission of MTC data by, for example, considering that certain devices automatically get DTP resources every N^{th} frame. However, such a scenario is not considered in this work.

We assume the DTP has a fixed length of N_{Frame} complex symbols that are divided by TDMA to the devices in $\mathcal{I}_{\text{DTP}}(k)$. A fair division of resources is assumed, where the number of symbols each device gets allocated in frame k is denoted by[2]

$$N_k = \left\lfloor \frac{N_{\text{Frame}}}{|\mathcal{I}_{\text{DTP}}(k)|} \right\rfloor, \tag{1}$$

where $\lfloor x \rfloor$ denotes the largest integer smaller than x. We denote the $(N_{\text{Rx}} \times N_{\text{Frame}})$ complex symbols received at the access point in frame k by $\mathbf{Y}_k = [\mathbf{Y}_{k,i_1} \cdots \mathbf{Y}_{k,i_{|\mathcal{I}_{\text{DTP}}|}}]$ and let $\mathbf{y}_{k,i}(n)$ denote the nth column of $\mathbf{Y}_{k,i}$ (i.e., the observation of the nth symbol received from device i in frame k). The single-input multiple-output (SIMO) channel is modeled according to

$$\mathbf{y}_{k,i}(n) = \mathbf{h}_{k,i} x_{k,i}(n) + \mathbf{w}_{k,i}(n), \tag{2}$$

for $n \in \{1, \cdots, N_k\}$, where $\mathbf{h}_{k,i}$ represent the channel vector between device i and the access point in frame k, $x_{k,i}(n)$ are the transmitted data symbols, and $\mathbf{w}_{k,i}(n) \sim \mathcal{CN}(\mathbf{0}, N_0 \mathbf{I}_{N_{\text{Rx}}})$ is the additive noise represented by a circular symmetric complex Gaussian random vector. We assume that $\mathbb{E}[\|\mathbf{h}_{k,i}\|^2] = P_i N_{\text{Rx}}$ where P_i represents the average power received per antenna from device i. We model the channel $\mathbf{h}_{k,i}$ as a narrowband SIMO Rice fading channel, i.e., $\mathbf{h}_{k,i} \sim \mathcal{CN}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ with $\boldsymbol{\mu}_i$ representing the LOS component and covariance matrix $\boldsymbol{\Sigma}_i$ representing the fading. The covariance matrix is given by $\boldsymbol{\Sigma}_i = \frac{P_i}{K_{\text{Rice}}+1} \boldsymbol{\Lambda}$, where $[\boldsymbol{\Lambda}]_{i,j} = \rho^{|i-j|}$ is an $(N_{\text{Rx}} \times N_{\text{Rx}})$ matrix, ρ is a correlation coefficient, and K_{Rice} is a common Rice factor (e.g., ranging from 2-10 dB [12] in indoor channels) experienced by all antennas and all devices in the network. Furthermore, we assume that the frame period is shorter than the coherence time of the channel so that the channel realizations $\mathbf{h}_{k,i}$ can be assumed to be constant within a frame, independent from frame to frame, and independent among the MTC devices.

For device i, positioned at distance d_i and with angle of arrival (AoA) Φ_i relative to the receiver antenna array, the channel mean (i.e., the LOS component) is modeled as a phased-array antenna $\boldsymbol{\mu}_i = a e^{-\frac{j2\pi d_i}{\lambda_c}} \mathbf{e}(\Omega_i)$, where λ_c is the carrier wavelength, $\Omega_i = \cos(\Phi_i)$ is the directional cosine, $a = \|\boldsymbol{\mu}_i\|$, and $\mathbf{e}(\Omega_i)$ is the unit spatial signature given by

$$\mathbf{e}(\Omega_i) = \frac{1}{\sqrt{N_{\text{Rx}}}} \left[z^0, z^{\Omega_i}, \cdots, z^{(N_{\text{Rx}}-1)\Omega_i} \right] \tag{3}$$

in terms of the complex number $z = e^{-j2\pi\Delta_r}$, where Δ_r is the antenna spacing (normalized by the wavelength) [13]. From normalization of $\mathbb{E}[\|\mathbf{h}_{k,i}\|^2]$ we get $a = \sqrt{\frac{P_i N_{\text{Rx}} K_{\text{Rice}}}{K_{\text{Rice}}+1}}$, and we assume the received power follows as $P_i = P_0 d_i^{-\beta/2}$ where β is a path-loss exponent, P_0 is the transmit power, and d_i is the distance. Additionally, in the following we normalize the noise power spectral density $N_0 = 1$ such that $P_i N_{\text{Rx}}$ also represents the average received signal-to-noise ratio (SNR) on the ith link.

Adversary Assumptions: In this paper, we assume that a single attacker is present in the system, referred to as Eve, having a single antenna, located at distance d_E and with AoA Φ_E relative to the access point. We model Eve's channel similarly to the legitimate channels with Rice factor $K_{\text{Rice},E}$ and denote Eve's channel realization in frame k by $\mathbf{h}_{k,E} \sim \mathcal{CN}(\boldsymbol{\mu}_E, \boldsymbol{\Sigma}_E)$, where $\boldsymbol{\mu}_E = a_E e^{-\frac{j2\pi d_E}{\lambda_c}} \mathbf{e}(\Omega_E)$ with the normalized spatial signature given in (3). With this representation, we can model both the case when Eve is an external device or when the attack is launched from a compromised device within the network by letting $\boldsymbol{\mu}_E = \boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_E = \boldsymbol{\Sigma}$ for some legitimate device i. The power received from Eve's transmissions is assumed to be $P_E = P_0 d_E^{-\beta/2}$.

The physical layer channel model is mainly motivated by LOS scenarios with low device mobility. For instance, deployments carefully planned for using the LOS channel for security purposes (e.g., LOS beamforming for physical layer security [14]). At a first glance, the fact that our model represents a narrow-band single-carrier TDMA system with no explicit assumptions on the bandwidth may be perceived as

[2]Note that N_k in general is a random variable depending on the number of users allocated in the frame, and that N_k can get very small if many devices request resources at the same time.

a limitation. However, the allocated resources could straightforwardly be distributed over multiple frequencies as long as the assumption on i.i.d. fading across frames holds. That is to say that the results obtained based on this model also apply to frequency hopping systems, e.g., as in state-of-the-art industrial sensor network protocols or to systems like LTE that have the possibility of exploiting resource scheduling in order to benefit from frequency diversity.

B. Feature-Based Physical Layer Authentication

The access point performs hypothesis testing based on the observed channel states in order to verify the legitimacy of received MGMT and DTP transmissions. When a single message is received from device i, the hypothesis $\mathcal{H}_1$[3] represents that the observed channel state is from the legitimate distribution $\mathcal{CN}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, and hypothesis $\mathcal{H}_0$ represents that it originates from the adversary's distribution $\mathcal{CN}(\boldsymbol{\mu}_E, \boldsymbol{\Sigma}_E)$. We assume that the access point has access to a feature bank containing the legitimate distribution parameters $\boldsymbol{\mu}_i, \boldsymbol{\Sigma}$. In order to learn such distributions, an initial trust must be established between the device and the access point, for instance, by initially using cryptographic authentication whenever a new device joins the network. However, the process by which this is done is considered to be outside the scope of this work.

For a received set of messages $\mathcal{I} = \{m_1, \cdots, m_M\}$, we denote by $\tilde{\mathbf{h}}_{m_i} = \text{CSI}(m_i)$ the observed SIMO channel state associated with each message m_i. In general, this channel state is an estimate with limited precision. However, to simplify the analysis we assume perfect channel-state knowledge in the following. Furthermore, we assume that PLA is applied to $\mathcal{I} \in \{\mathcal{I}_{\text{MGMT}}, \mathcal{I}_{\text{DTP}}\}$, i.e., MGMT requests and DTP data payloads are authenticated separately. We consider now the case when L messages share the same ID (e.g., due to multiple impersonated messages injected by an adversary). The PLA procedure divides the set $\mathcal{I}$ into subsets $\mathcal{I}_i = \{m \in \mathcal{I} : \text{ID}(m) = i\}$ of messages with the same ID, each authenticated independently. To test the legitimacy of the messages in the set $\mathcal{I}_i$, the access point constructs a $L+1$-ary hypothesis test. We here denote by $\mathcal{H}_l$ for $l \in \{1, \cdots, L\}$, the disjoint hypotheses that message m_l is authentic, i.e., that we believe $\tilde{\mathbf{h}}_{m_l} \sim \mathcal{CN}(\boldsymbol{\mu}_{\text{ID}(m_l)}, \boldsymbol{\Sigma}_{\text{ID}(m_l)})$, and by $\mathcal{H}_0$ the hypothesis that no message in $\mathcal{I}_L$ is authentic. The decision of $\mathcal{H}_l$ results in accepting m_l and rejecting the rest, while the decision of $\mathcal{H}_0$ results in rejecting all messages in $\mathcal{I}_i$, since the authentication is predicated on that the access point expects only one message per legitimate device. The access point decides between the L messages through

$$d_i(\tilde{\mathbf{h}}_{m_l}) \underset{\mathcal{H}_l}{\overset{\mathcal{H}_0}{\gtrless}} T, \quad \text{with } m_l = \underset{m=m_1,\cdots,m_L}{\arg\min}\ d_i(\tilde{\mathbf{h}}_m), \qquad (4)$$

where $d_i(\cdot)$ is a discriminant function associated with the channel feature of the device with ID i, given by $d_i(\tilde{\mathbf{h}}_m) = 2(\tilde{\mathbf{h}}_m - \boldsymbol{\mu}_i)^\dagger \boldsymbol{\Sigma}^{-1}(\tilde{\mathbf{h}}_m - \boldsymbol{\mu}_i)$. The minimization of the righthand side of (4) is to be viewed as choosing the maximum-likelihood (ML) decision (the discriminant function $d_i(\cdot)$ is also the log-likelihood of the observation given the legitimate distribution) while the threshold decision in the lefthand side determines if the ML decision is authentic.

Single message authentication ($L = 1$): The L message hypothesis test in (4) is an extension of the standard generalized likelihood-ratio test (GLRT), used for PLA when deciding upon a multi-dimensional complex Gaussian feature such as a multi-carrier frequency response [15] or a channel impulse response [16]. Note that when (4) is reduced to $L = 1$ (i.e., only a single message with ID $= i$ is received), the hypothesis test becomes $d_i(\tilde{\mathbf{h}}_m) \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\gtrless}} T$, where $\mathcal{H}_1$ represents that the message is legitimate and $\mathcal{H}_0$ represents that the message stems from an adversary.

C. Adversarial Strategies

Given Eve's ability to send messages with fraudulent IDs, we differentiate four cases of adversary behavior:

a) Baseline: Eve is present, but inactive, and the performance of the system is only affected by false alarms. The baseline scenario models the impact of introducing the PLA protocol in the system when no attacks are attempted.

b) Data Injection Attack: Eve is sending DTA requests impersonating a legitimate MTC device. Once successful, Eve gets DTP resources and transmits false data with the aim of harming the underlying application (e.g., drive a control system into a dangerous state by introducing fake sensor or actuation signals). In our work, we do not model the impact of the data injection attack on the application; however, metrics like missed detection rate (see Section II-D and III-C) measure Eve's success-rate under such attacks, and the number of resources N_k each device gets scheduled will be affected.

c) Sybil Attack: Eve transmits multiple DTA requests with fraudulent IDs, referred to as Sybil IDs/devices, with the goal of depleting resources available to the other legitimate devices [17]. In a Sybil attack, we assume that Eve targets a set of inactive devices $D_{\text{Sybil}} \subset \{1, \cdots, K_d\}$ that are not transmitting in the frame and sends DTA requests with the corresponding IDs. Note that it does not make sense for Eve to target active devices in this attack since they will already transmit DTA requests. With each successful Sybil ID, N_k in (1) is reduced which degrades the performance of the other links in the network.

d) Disassociation Attack: Eve targets a particular device and sends fraudulent requests to disassociate from the access point (DCN) with the corresponding device's ID. If successful, Eve disconnects the legitimate device which needs to reconnect, a process we model as being disconnected for K_{RC} frames (e.g., due to management processes such as generating session keys).

The impersonation attacks that we consider can be launched by external entities (e.g., an attacker positioned in close proximity to the system, using a stolen MTC device or a software defined radio unit) or internal devices whose behavior has been hijacked by malicious code. Our attacker model allows us to model both cases by modifying the assumptions

[3] In related litterature, typically $\mathcal{H}_0$ represents the hypothesis that the message is legitimate. In this work, however, the notation is reversed in order to be consistent with the L-message authentication where $\mathcal{H}_1$ represents that the first message is legitimate.

on Eve's channel. We note, however, that Sybil attacks are generally assumed to originate from internal devices that are compromised [17].

D. False Alarm and Missed Detection Rates

Here, we summarize the error events and corresponding probabilities for the single message authentication, which are standard results (c.f., [15] for proofs). In the $L = 1$ message case, two error events can occur: (i) a *false alarm* when a legitimate message is rejected; and (ii) a *missed detection* when an illegitimate message is accepted. Under the legitimate hypothesis $\mathcal{H}_1$, we have $d_i(\tilde{\mathbf{h}}_m) \sim \chi^2_{2N_{\text{Rx}}}$ and the false alarm rate is

$$p_{\text{FA}}(T) = \mathbb{P}(d_i(\tilde{\mathbf{h}}_m) > T|\mathcal{H}_1) = 1 - F_{\chi^2_{2N_{\text{Rx}}}}(T), \quad (5)$$

where $F_{\chi^2_{2N_{\text{Rx}}}}(\cdot)$ is the cumulative distribution function (CDF) of a χ^2 distribution with $2N_{\text{Rx}}$ degrees of freedom. Observe that for a given choice of threshold T, the false alarm rate is equal across all device IDs i, independently of our assumptions on Eve. In practice, the PLA could be designed with different thresholds T_i for different devices. However, in order to simplify the analysis we assume a constant threshold T. Under $\mathcal{H}_0$ (i.e., Eve is sending the message m with $\text{ID}(m) = i$), given that Eve's channel covariance-matrix is of the form $\mathbf{\Sigma}_E = \frac{P_E}{1+K_{\text{Rice},E}}\mathbf{\Lambda}$, we have $d_i(\tilde{\mathbf{h}}_m) \sim \lambda_i \chi^2_{2N_{\text{Rx}}}(\nu_i)$, where $\lambda_i = \frac{P_E(1+K_{\text{Rice}})}{P_i(1+K_{\text{Rice},E})}$ and ν_i is the non-centrality parameter. Hence, the missed detection rate is

$$p_{\text{MD}}(i,T) = \mathbb{P}(d_i(\tilde{\mathbf{h}}_m) < T|\mathcal{H}_0) = F_{\chi^2_{2N_{\text{Rx}}}(\nu_i)}(T/\lambda_i), \quad (6)$$

where $F_{\chi^2_{2N_{\text{Rx}}}(\nu_i)}(\cdot)$ is the CDF of a non-central χ^2 distribution with $2N_{\text{Rx}}$ degrees of freedom and non-centrality parameter $\nu_i = 2(\boldsymbol{\mu}_E - \boldsymbol{\mu}_i)^\dagger \mathbf{\Sigma}_E^{-1}(\boldsymbol{\mu}_E - \boldsymbol{\mu}_i)$. From this we can note that the missed detection rate varies with the device i that Eve tries to impersonate. Error analysis for PLA with $L > 1$ has to our knowledge not been studied before. In Section III-C, we provide bounds on the missed detection rate for $L = 2$ and show that this case will suffice for the delay performance analysis under the considered attack strategies.

E. Problem Formulation

The presented PLA scheme and adversary strategies impact delay and reliability in the considered system in several ways: Firstly, the inevitable false alarms for a given authentication threshold will cause legitimate messages to be dropped. Secondly, disassociation attacks will in the case of missed detections cause service dropouts for legitimate devices. Thirdly, Sybil attacks will impact the number of symbols allocated to legitimate devices. The problem we address in the rest of this paper is the development and analysis of queueing models that capture the behavior of the PLA scheme in the baseline scenario and under the considered attacks. While we derive some results specific to the PLA scheme presented in this section, the queueing model in Section III and delay performance bounds in Section IV are effectively only dependent on the scheduled resources N_k and the probability of service dropout. Therefore, our presented models can easily encompass other PLA schemes and attacker models as long as closed form expressions for false alarm and missed detection probabilities exists.

III. Queueing Modeling of Authentication Delays and Attacker Impacts

In this section, we develop and analyze models of how erroneous PLA decisions impact the system delay performance under each of the adversarial strategies.

A. Delay Performance Metric

As mentioned in Section I, the use of PLA for improved security might have unintended consequences on the system's ability to meet delay requirements. To study such delay performance issues, we introduce infinite-buffer queues that model the flow of data from each MTC device to the access point. The queueing model is described by the bivariate stochastic processes

$$A_i(\tau,t) = \sum_{k=\tau}^{t} a_k^{(i)}, \; D_i(\tau,t) = \sum_{k=\tau}^{t} d_k^{(i)},$$

representing the cumulative arrivals to and departures from the queue in the time interval $[\tau, t)$ for all $0 \leq \tau \leq t$. In frame k, $a_k^{(i)}$ represents the instantaneous arrivals to the ith MTC device buffer measured in bits (e.g., incoming sensor measurements), and $d_k^{(i)}$ represent the instantaneous departures from the ith queue (i.e., information successfully transmitted to the access point). The ability to transfer data from the buffer queue to the destination at the access point is characterized by the cumulative service process $S_i(\tau,t) = \sum_{k=\tau}^{t} s_k^{(i)}$. Considering that a device is assigned resources, we assume that the transmitter chooses a coding rate $R_k^{(i)}$, and transmits $s_k^{(i)} = N_k R_k^{(i)}$ encoded information bits over the SIMO channel. Furthermore, we introduce the Bernoulli random variable $X_k^{(i)}$, indicating if resources are scheduled to device i. This results in the general service model

$$s_k^{(i)} = \begin{cases} N_k R_k^{(i)}, & \text{if } X_k^{(i)} = 1 \\ 0 & \text{if } X_k^{(i)} = 0. \end{cases} \quad (7)$$

We use the Shannon capacity $R_k^{(i)} = \log_2(1+\gamma_{k,i})$ as a proxy for the amount of bits per channel use that can be transmitted over the channel (i.e., the rate $R_k^{(i)}$ is representing the capacity of the discrete time channel model given in (2)). Assuming the access point has perfect channel state information and uses maximum-ratio combining for the channel model (2), the instantaneous SNR is given by $\gamma_{k,i} = \frac{\|\mathbf{h}_{k,i}\|^2}{N_0}$.

A widely used measure on the queueing system's ability to meet delay requirements is the *delay violation probability* [18]. The queueing delay at time point t is defined as

$$W_i(t) \triangleq \inf\{u > 0; A_i(0,t) \leq D_i(0,t+u)\}, \quad (8)$$

representing the frames required to serve the bits in the queue at time t. This delay is randomly varying due to the random service process and the delay violation probability is defined as $p_i(w) = \mathbb{P}(W_i(t) > w)$, i.e., the probability that a bit is not

received within a defined deadline w. In many cases, an exact expression for the delay violation probability is complicated to derive. However, queueing analysis can give statistical bounds on this function. In particular, the stochastic network calculus framework, introduced in Section IV, contains tools that are appropriate for deriving an upper bound on $p_i(w)$ given the underlying service process in (7). Such delay bounds are particularly suitable for performance evaluation in mission-critical networks since they provide upper limits on the delay violation probability, i.e., a real system operating under the assumed conditions will certainly achieve a better delay performance.

B. Baseline Scenario

In the baseline scenario, the adversary is inactive and the queueing model is affected only by dropped messages due to false alarms. We assume that a set $D_{\text{Active}} \subseteq \{1, \cdots, K_d\}$ of devices are active and that each has a constant arrival rate α_i, which means that each of the active devices will request DTA resources in each frame. Considering one of the active devices i, it will request resources with a DTA request in the MGMT period. Since the adversary is inactive, the access point will receive only one request with the ID of device i and the message will be authenticated based on the single message authentication $d_i(\tilde{\mathbf{h}}_m) \gtrless_{\mathcal{H}_1}^{\mathcal{H}_0} T$ (see Section II-B). The observed channel state $\tilde{\mathbf{h}}_m$ will in this case be the authentic channel $\mathcal{CN}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ and the false alarm rate is given by $p_{\text{FA}}(T)$ in (5). Since we assume perfect channel-state information and a frame period shorter than the coherence time of the channel, the observed channel state will remain constant during the frame. Hence, if the DTA request is accepted, so will the following data payload message in the DTP[4]. Since the requests independently get rejected by PLA with $p_{\text{FA}}(T)$, the number of scheduled devices follows a binomial distribution

$$p_{|\mathcal{I}_{\text{DTP}}|}(k) = \binom{|D_{\text{Active}}|}{k} (1 - p_{\text{FA}}(T))^k p_{\text{FA}}(T)^{|D_{\text{Active}}|-k}, \quad (9)$$

and the distribution of N_k follows as $p_{N_k}(n) = p_{|\mathcal{I}_{\text{DTP}}|}(\frac{N_{\text{Frame}}}{n})$. The threshold T is ideally set such that $p_{\text{FA}}(T)$ is low, giving a possible approximation $|\mathcal{I}_{\text{DTP}}| \approx |D_{\text{Active}}|$. For a particular device i, the distribution of $X_k^{(i)}$ is given by

$$\Pr(X_k^{(i)} = 0) = p_{\text{FA}}(T). \quad (10)$$

That is, in case of a false-alarm in frame k, the data buffer observes zero service.

C. Detection of Data Injection Attacks

In a data injection attack, Eve transmits a DTA request in the MGMT period with the aim of getting DTP resources for transmitting a false data message. Either Eve impersonates an inactive device i that is not requesting resources in the current frame, in which case the DTA requests undergoes single-message authentication and is undetected with probability $p_{\text{MD}}(i, T)$, or Eve impersonates an active device, in which case the message is authenticated by $L = 2$ message authentication. In the latter case, denoting by m_i and m_E the messages from device i and Eve, respectively, a missed detection occurs in the union of events $\left\{ \underset{m=m_i,m_E}{\arg\min}\, d_i(\tilde{\mathbf{h}}_m) = m_E \right\}$ and $\{d_i(\tilde{\mathbf{h}}_{m_E}) < T\}$. In this case, the probability of missed detection, denoted by $p_{\text{MD}}^{L=2}(i, T)$, can be written as

$$\begin{aligned} p_{\text{MD}}^{L=2}(i,T) &= \mathbb{P}(d_i(\mathbf{h}_E) < d_i(\mathbf{h}_i), d_i(\mathbf{h}_E) < T) \\ &= \mathbb{P}(d_i(\mathbf{h}_E) < T)\mathbb{P}(d_i(\mathbf{h}_i) > T) \\ &\quad + \mathbb{P}(d_i(\mathbf{h}_E) < d_i(\mathbf{h}_i) | d_i(\mathbf{h}_i) < T). \end{aligned} \quad (11)$$

We now use the notation $d_i = d_i(\mathbf{h}_i)$ and $d_E = d_i(\mathbf{h}_E)$ to discuss the probability (11). The second line of (11) is simply $p_{\text{FA}}(T)p_{\text{MD}}(i, T)$. However, for the second term $\mathbb{P}(d_E < d_i | d_i < T)$ an exact expression can only be obtained in integral form. Instead, by noting that $\mathbb{P}(d_E < d_i, d_E < T) \leq \mathbb{P}(d_E < T) = p_{\text{MD}}(i, T)$, we can provide upper and lower bounds

$$p_{\text{FA}}(T)p_{\text{MD}}(i,T) \leq p_{\text{MD}}^{L=2}(i,T) \leq p_{\text{MD}}(i,T). \quad (12)$$

Additionally, we can observe that $\mathbb{P}(d_E < d_i, d_E < T) \leq \mathbb{P}(d_E < d_i)$ and provide an upper bound $\mathbb{P}(d_E < d_i) \leq p_d(i)$ in the following lemma:

Lemma 1. *The probability $\mathbb{P}(d_E < d_i)$ can be upper bounded by*

$$p_d(i) = \min_{-\lambda/2 < t < 1/2} (1 + 2(\lambda - 1)t - 4\lambda t^2)^{-N_{\text{Rx}}} e^{-\frac{\nu_i \lambda t}{1+2\lambda t}}, \quad (13)$$

where $\lambda_i = \frac{P_E(1+K_{Rice})}{P_i(1+K_{Rice,E})}$, and $\nu_i = 2(\boldsymbol{\mu}_E - \boldsymbol{\mu}_i)^\dagger \boldsymbol{\Sigma}_E^{-1}(\boldsymbol{\mu}_E - \boldsymbol{\mu}_i)$.

Proof. We rewrite $\mathbb{P}(d_i(\mathbf{h}_E) < d_i(\mathbf{h}_i)) = \mathbb{P}(d_i(\mathbf{h}_i) - d_i(\mathbf{h}_E) > 0)$ and use the Chernoff bound to get that for every $t > 0$

$$\begin{aligned} &\mathbb{P}(d_i(\mathbf{h}_i) - d_i(\mathbf{h}_E) > 0) = \mathbb{P}(e^{t(d_i(\mathbf{h}_i) - d_i(\mathbf{h}_E))} > 1) \\ &\leq \mathbb{E}\left[e^{t(d_i(\mathbf{h}_i) - d_i(\mathbf{h}_E))}\right] = \mathbb{E}\left[e^{td_i(\mathbf{h}_i)}\right] \mathbb{E}\left[e^{-td_i(\mathbf{h}_E)}\right], \end{aligned} \quad (14)$$

where we have applied the Markov inequality and used the independence of $\mathbf{h}_i$ and $\mathbf{h}_E$. Now since $d_i(\mathbf{h}_i) \sim \chi^2_{2N_{\text{Rx}}}$ and $d_i(\mathbf{h}_E) \sim \lambda\chi^2_{2N_{\text{Rx}}}(\nu_i)$ (see Section II-D), we get $\mathbb{E}[e^{td_i(\mathbf{h}_i)}] = (1 - 2t)^{-N_{\text{Rx}}}$ and $\mathbb{E}[e^{-td_i(\mathbf{h}_E)}] = (1 + 2\lambda t)^{-N_{\text{Rx}}} \exp\left(\frac{-\nu_i \lambda t}{1+2\lambda t}\right)$ for $-\lambda/2 < t < 1/2$ from the standard moment generating functions for the corresponding distributions. Plugging these expressions into (14) and minimizing over t yields (13) which completes the proof. □

The upper bound that is tightest out of (12) and (13) depends on the authentication threshold T (clearly $p_{\text{MD}}(i, T) \to 1$ as $T \to \infty$ and $p_{\text{MD}}(i, T) \to 0$ as $T \to 0$ while $p_d(i)$ is independent of T). Hence, we tighten our bound on the missed detection probability when Eve is launching a data injection attack against active device i by

$$p_{\text{MD}}^{L=2}(i,T) \leq p_{\text{MD,Upper}}(i) = \min\{p_{\text{MD}}(i,T), p_d(i)\}. \quad (15)$$

[4]This is a consequence of our previous assumptions. However, if the coherence time is shorter, or estimation errors are present, modeling of this as a two independent authentication decisions would be straightforward.

This bound will additionally later prove useful when analyzing the disassociation attack in Section III-E.

Remark 1. *In a data injection attack, the delay performance of legitimate devices will be affected since accepted DTA requests from Eve will reduce the amount of resources scheduled to other devices. However, this impact is principally the same as under the Sybil attack discussed in Section III-D. Therefore, we only use the data injection scenario to study the detection performance of PLA, leaving questions regarding queueing performance to be answered by the study of Sybil attacks.*

D. Queueing Impacts of Sybil Attacks

Recall that in a Sybil attack, Eve targets a set of inactive devices D_{Sybil} and sends DTA requests with the corresponding IDs. Consequently, the access point receives messages from $D_{\text{Active}} \cup D_{\text{Sybil}}$ in the MGMT period and needs to differentiate which ones are legitimate. If Eve successfully gets many DTA requests through, N_k given by (1) decreases and legitimate devices get less resources which can result in growing queue backlogs. Under a Sybil attack, we assume that active legitimate devices D_{Active} experience service dropouts modeled by $X_k^{(i)}$ the same way as in the baseline case (10); however, the distribution of N_k is different due to Sybil IDs launched by Eve.

Assuming all IDs in $D_{\text{Active}} \cup D_{\text{Sybil}}$ are distinct (Eve is assumed to target only inactive devices in the Sybil attack), the number of devices that get resources in the data transmission period is

$$|\mathcal{I}_{\text{DTP}}| = \sum_{i \in D_{\text{Active}} \cup D_{\text{Sybil}}} \mathbb{I}(d_i(\tilde{\mathbf{h}}_{m_i}) < T). \tag{16}$$

We decompose $|\mathcal{I}_{\text{DTP}}| = K_{\text{Active}} + K_{\text{Sybil}}$ where K_{Active} is characterized by the baseline distribution of (9) (i.e., requests rejected by false alarms) and

$$K_{\text{Sybil}} = \sum_{i \in D_{\text{Sybil}}} \mathbb{I}(d_i(\mathbf{h}_E) < T) \tag{17}$$

represents the number of Sybil IDs successfully launched by Eve. For a moderate number of Sybil IDs (< 30), the distribution of K_{Sybil} can be combinatorially approximated as

$$p_{K_{\text{Sybil}}}(k) \approx \sum_{B \in A_k} \prod_{i \in B} p_{\text{MD}}(i,T) \times \prod_{j \in B^c} (1 - p_{\text{MD}}(j,T)), \tag{18}$$

where A_k denotes the set of all size k subsets of D_{Sybil}. This approximation stems from an assumption that the events $\{d_i(\mathbf{h}_E) < T\}_{i \in D_{\text{Sybil}}}$ can be approximated as independent, in which case K_{Sybil} is Poisson-binomial distributed. Now the distribution of $|\mathcal{I}_{\text{DTP}}|$ under a Sybil attack can be written as the convolution

$$p_{\mathcal{I}_{\text{DTP}},\text{Sybil}}(k) = \sum_{l=0}^{k} p_{K_{\text{Active}}}(l) p_{K_{\text{Sybil}}}(l-k), \tag{19}$$

from which the distribution of N_k follows as $p_{N_k}(n) = p_{\mathcal{I}_{\text{DTP}},\text{Sybil}}(\frac{N_{\text{Frame}}}{n})$.

The impact of the Sybil attack depends on the system's available resources N_{Frame}: If N_{Frame} by design allows all devices to communicate simultaneously, the Sybil IDs will not have a substantial impact on the service of the legitimate devices. However, if the system is optimized to only have a subset of devices communicating at a time (e.g., in order to reduce latency or if only a subset of devices is involved in a particular sensing tasks), the result of launching multiple additional Sybil IDs might have severe impacts on the active legitimate devices. An alternative counter-strategy is to only accept requests from devices that are expected to transmit (e.g., sensors carrying relevant measurements for the running application). However, such application-layer information might not be available at the physical and MAC layers.

E. Queueing Impacts of Disassociation Attacks

In a disassociation attack, Eve targets an active legitimate device and sends DCN request with the corresponding ID. In an attacked frame, the access point will observe two messages m_1 and m_2 with the same ID (i.e., $\text{ID}(m_1) = \text{ID}(m_2)$) and uses (4) to decide which one is authentic. If the access point accepts the DCN request from Eve, the legitimate device will need to reconnect in order to continue its data transfer which results in a disruption of the communication (i.e., $s_k = 0$) for K_{RC} consecutive frames which can lead to growing backlogs and increased delay. In principle, Eve could launch disassociation attacks against multiple links within the network. However, here we model the queueing impact when Eve targets a single device i.

In the disassociation attack, the frame-level service process s_k in (7) follows the same model as in the baseline scenario (10) (i.e., frames are dropped with the false alarm rate and N_k is given by its baseline distribution). We consider independent Bernoulli attack attempts from Eve with probability p_{Attack} and to model the impact on the queueing performance, we divide the data flow from device i to the access point into blocks consisting of K_{RC} frames each and define the aggregated arrival process as $a'_l = \sum_{k=K_{\text{CN}}l}^{K_{\text{RC}}(l+1)-1} a_k$ and service process as

$$s'_l = \begin{cases} \sum_{k=K_{\text{RC}}l}^{K_{\text{RC}}(l+1)-1} s_k, & \text{if } D_l = 0 \\ 0 & \text{if } D_l = 1, \end{cases} \tag{20}$$

where D_l is a Bernoulli random variable indicating a successful disassociation attack in the block. The distribution of D_l is then given by

$$\mathbb{P}(D_l = 1) = 1 - (1 - p_{\text{MD}}^{L=2}(i,T) p_{\text{Attack}})^{K_{\text{RC}}}, \tag{21}$$

where $p_{\text{MD}}^{L=2}(i,T)$ is the probability of accepting Eve's DCN message (i.e., the same situation as in the data injection attack and hence $p_{\text{MD}}^{L=2}(i,T)$ is given by (11)). We recall from Section III-C that a closed form solution for $p_{\text{MD}}^{L=2}(i,T)$ is not available. However, since $\Pr(D_l = 1)$ is monotonically increasing with $p_{\text{MD}}^{L=2}(i,T) \in [0,1)$, an upper bound on $p_{\text{MD}}^{L=2}(i,T)$ suffices to upper bound $\mathbb{P}(D_l = 1)$. An upper bound on $p_{\text{MD}}^{L=2}(i,T)$ is given by (15) and hence we get an upper bound $\mathbb{P}(D_l = 1) \leq 1 - (1 - p_{\text{MD,Upper}}(i) p_{\text{Attack}})^{K_{\text{RC}}}$.

Our analysis of the disassociation attack serves as a worst-case model due to the upper bound on $\mathbb{P}(D_l = 1)$. However,

since in the next Section IV we aim to upper bound the delay violation probability, an upper bound on the disassociation probability suffices for this purpose. Additionally, we acknowledge that other methods could be used for reducing the impact of disassociation attacks. For example, one could always choose DTA over DCN requests, which would render the disassociation attack harmless as long as an active device is targeted. However, we see this as an issue of protocol design and include disassociation attacks in our studies in the following.

IV. Delay Performance Analysis

In this section, we derive delay performance bounds for the considered system using tools from stochastic network calculus. We begin by introducing necessary results and notation from the stochastic network calculus framework:

A. Stochastic Network Calculus

Stochastic network calculus is a mathematical framework that allows us to analyze input-output relationships of stochastic queueing systems through, for example, performance bounds on delay or backlog given arrival and service distributions. For a complete overview of stochastic network calculus, we refer to [10]. The work in [18] developed the stochastic network calculus framework for wireless fading links by observing that the analysis is simplified by converting the bivariate stochastic processes $A(\tau,t)$, $S(\tau,t)$ and $D(\tau,t)$ into $\mathcal{A}(\tau,t) \triangleq e^{A(\tau,t)}$, $\mathcal{S}(\tau,t) \triangleq e^{S(\tau,t)}$ and $\mathcal{D}(\tau,t) \triangleq e^{D(\tau,t)}$. This transformation allows the characterization of the random service process in terms of the varying instantaneous SNR due to fading. This is referred to as transforming the bit-domain processes into the SNR-domain since the processes become linear in the instantaneous SNR γ_k instead of logarithmic. Arrival processes in the SNR-domain can then be seen as instantaneous SNR demands. In bit-domain, stochastic network calculus is based on a $(\min,+)$ dioid algebra over $\mathbb{R}^+$. Stochastic network calculus in the SNR-domain, on the other hand, is instead based on the $(\min,\times)$ dioid algebra since processes in the SNR-domain become multiplicative instead of additive. The performance bounds, which can be seen as variations of moment bounds, are derived in terms of Mellin transforms of the involved queueing processes. The Mellin transform of a random variable X, closely related to the moment-generating function (MGF), is defined as $\mathcal{M}_X(s) = \mathbb{E}[X^{s-1}]$.

The upper bound on the delay violation probability we utilize in this paper is given by the following lemma:

Lemma 2. *For $s>0$,*

$$p(w) \leq \mathcal{K}(s,t+w,t), \tag{22}$$

where $\mathcal{K}(s,\tau,t)$ is called the kernel and given by

$$\mathcal{K}(s,\tau,t) \triangleq \sum_{u=0}^{\min(\tau,t)} \mathcal{M}_{\mathcal{A}}(1+s,u,t)\mathcal{M}_{\mathcal{S}}(1-s,u,\tau), \tag{23}$$

and $\mathcal{M}_{\mathcal{S}}(s,\tau,t) = \mathbb{E}[\mathcal{S}(\tau,t)^{s-1}]$ and $\mathcal{M}_{\mathcal{A}}(s,\tau,t) = \mathbb{E}[\mathcal{A}(\tau,t)^{s-1}]$ are Mellin transforms of the independent SNR-domain service and arrival processes.

Proof. See Theorem 1 in [18]. □

With i.i.d. instantaneous arrivals and service, we can write $\mathcal{M}_{\mathcal{S}}(s,\tau,t) = \mathcal{M}_{\mathcal{S}}(s)^{t-\tau}$ and $\mathcal{M}_{\mathcal{A}}(s,\tau,t) = \mathcal{M}_{\mathcal{A}}(s)^{t-\tau}$, where $\mathcal{M}_{\mathcal{S}}(s) \triangleq \mathbb{E}[e^{s_k(s-1)}]$ and $\mathcal{M}_{\mathcal{A}}(s) \triangleq \mathbb{E}[e^{a_k(s-1)}]$ due to the independence of the instantaneous service and arrivals s_k and a_k. Then, assuming stationarity of the underlying queueing processes, we let $t \to \infty$ in the righthand side of (22) and get

$$\lim_{t\to\infty} \mathcal{K}(s,t+w,t) = \frac{\mathcal{M}_{\mathcal{S}}(1-s)^w}{1-\mathcal{M}_{\mathcal{A}}(1+s)\mathcal{M}_{\mathcal{S}}(1-s)}, \tag{24}$$

under the stability condition $\mathcal{M}_{\mathcal{A}}(1+s)\mathcal{M}_{\mathcal{S}}(1-s) < 1$ required for the sum in (23) to converge. Since Lemma 2 holds for all $s>0$, it follows that minimization of (24) over $s>0$ gives us an asymptotic upper bound on the delay violation probability. Hence, for the stable and stationary queueing system, the upper bound on the delay violation probability can be compactly written as $p(w) \leq \inf_{s>0}\{\lim_{t\to\infty}\mathcal{K}(s,t+w,t)\}$ with the objective function to be minimized given by the steady-state kernel in (24). This function can be shown to be a convex function for every s in the stability interval $\mathcal{M}_{\mathcal{A}}(1+s)\mathcal{M}_{\mathcal{S}}(1-s) < 1$ (see Theorem 1 in [19]). However, no analytical tools from convex optimization can be applied, and therefore, one typically resorts to a numerical grid search for the minimization over s.

Since in this paper we assume constant arrivals of α bits per frame, independently of the service process, the arrival process is deterministic and the Mellin transform of the SNR-domain arrival process can easily be found to be $\mathcal{M}_{\mathcal{A}}(s) = e^{\alpha(s-1)}$. The service process, following the SIMO channel service model (7), has a more complicated Mellin transform which we derive in the following subsections for the considered attack scenarios.

It is worth noting that alternative stochastic network calculus approaches exist that may be used for this analysis including *effective capacity* [20] and *MGF-based analysis* [21]. Nevertheless, the usefulness of the approach in [19] that we employ is most apparent when applied to wireless fading channels as the Mellin transform $\mathcal{M}_{\mathcal{S}}$ is already derived for many fading channels in the literature, e.g., [9, 22, 23]. This makes the approach particularly attractive for wireless networks analysis.

B. Baseline Analysis

Recall that in the baseline scenario no active attacker is present and frames are dropped with the false alarm rate, i.e, $\Pr(X=0) = 1-p_X = p_{\text{FA}}$. The service model is given by (7) with $R_k = \log_2(1+\mathbf{h}_k^\dagger\mathbf{h}_k)$ where we now, for ease of notation, have dropped the user index i. Note that in this section we assume the allocated resources N_k to be deterministic, something we will later generalize when deriving the Sybil attack bound. To simplify the derivation, we define the functions $h(\gamma_k,X_k,N_k) \triangleq e^{s_k}$ and $g(\gamma_k) = 1+\gamma_k$ in terms of the instantaneous SNR γ_k so that

$$h(\gamma_k,X_k,N_k) = \begin{cases} g(\gamma_k)^{\frac{N_k}{\ln(2)}}, & \text{if} \quad X_k = 1 \\ 1, & \text{if} \quad X_k = 0. \end{cases} \tag{25}$$

In the following, we provide our main analytical result, which is an approximate expression for the Mellin transform of $g(\gamma_k)$ in Theorem 1. From this result, the Mellin transform of the service process in steady-state easily follows, as stated in Corollary 1.

Theorem 1. *For the Rice fading SIMO channel with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, the Mellin transform of $g(\gamma_k)$ can be approximated by*

$$\mathcal{M}_{g(\Gamma_k)}(s) \approx \frac{e^{1/2\alpha_g}(2\alpha_g)^{s-1}}{\Gamma(k_g/2)} \times \sum_{m=0}^{\infty} \binom{\frac{k_g-2}{2}}{m} \frac{1}{(-2\alpha_g)^m} \Gamma\left[s-m+\frac{k_g-2}{2}, \frac{1}{2\alpha_g}\right] \tag{26}$$

where $\Gamma(s,x) = \int_x^\infty t^{s-1}e^t dt$ denotes the upper incomplete gamma function and α_g and k_g are parameters of the approximate distribution of γ_k given by

$$\alpha_g = \frac{\frac{1}{2}(\mathrm{tr}(\boldsymbol{\Sigma}^2) + \boldsymbol{\mu}^\dagger\boldsymbol{\Sigma}\boldsymbol{\mu})}{1+\mathrm{tr}(\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\dagger\boldsymbol{\mu}} \quad \textit{and} \quad k_g = \frac{(1+\mathrm{tr}(\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\dagger\boldsymbol{\mu})^2}{\frac{1}{2}(\mathrm{tr}(\boldsymbol{\Sigma}^2) + \boldsymbol{\mu}^\dagger\boldsymbol{\Sigma}\boldsymbol{\mu})}. \tag{27}$$

Proof. We begin by using the fact that γ_k is a sum of independent non-central χ^2 distributed random variables with $\mathbb{E}[\gamma_k] = \mathrm{tr}(\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\dagger\boldsymbol{\mu}$ and $\mathrm{Var}[\gamma_k] = \mathrm{tr}(\boldsymbol{\Sigma}^2) + \boldsymbol{\mu}^\dagger\boldsymbol{\Sigma}\boldsymbol{\mu}$. Now we use that the sum of non-central χ^2 random variables can be approximated as a scaled central χ^2 [24]. That is, we write $\gamma_k \approx \alpha_g X$, where $X \sim \chi^2_{k_g}$. Transferred to the Mellin transform, the approximation becomes $\mathcal{M}_{g(\mathbf{h}_k)}(s) \approx \mathcal{M}_{1+\alpha_g X}(s)$. Now, we seek

$$\begin{aligned}\mathcal{M}_{1+\alpha X}(s) &= \int_0^\infty (1+\alpha x)^{s-1} \frac{1}{2^{\frac{k_g}{2}}\Gamma(k_g/2)} x^{\frac{k_g}{2}-1} e^{-\frac{x}{2}} dx \\ &= \frac{e^{\frac{1}{2\alpha}}}{(2\alpha)^{\frac{k_g}{2}}\Gamma(k_g/2)} \underbrace{\int_1^\infty u^{s-1}(u-1)^{\frac{k_g}{2}-1} e^{-\frac{u}{2\alpha}} du}_{\triangleq I}, \end{aligned} \tag{28}$$

where in the second line we have used the change of variable $u = 1+\alpha x$ and defined the integral I which remains to be solved. To solve it, we can use the binomial expansion $(u-1)^{\frac{k_g}{2}-1} = \sum_{m=0}^\infty \binom{\frac{k_g-2}{2}}{m}(-1)^m u^{\frac{k_g}{2}-1-m}$, which plugged into the integral I yields

$$\begin{aligned} I &= \int_1^\infty u^{s-1} \sum_{m=0}^\infty \binom{\frac{k_g-2}{2}}{m} (-1)^m u^{\frac{k_g}{2}-1-m} e^{-\frac{u}{2\alpha_g}} du \\ &= \sum_{m=0}^\infty \binom{\frac{k_g-2}{2}}{m} \frac{(2\alpha_g)^{s-m+\frac{k_g-2}{2}}}{(-1)^m} \int_{1/2\alpha_g}^\infty t^{s-m+\frac{k_g-2}{2}-1} e^{-t} dt \\ &= (2\alpha_g)^{k'+s} \sum_{m=0}^\infty \binom{k'}{m} \frac{1}{(-2\alpha_g)^m} \Gamma\left[s-m+k', \frac{1}{2\alpha_g}\right], \end{aligned} \tag{29}$$

where we in the second to third line have used the change of variable $t = u/2\alpha_g$ and introduced $k' = \frac{k_g-2}{2}$. Plugging (29) into (28) yields (26). Finally, since $\mathbb{E}[\alpha_g X] = \alpha_g k_g$ and $\mathrm{Var}[\alpha_g X] = \alpha_g^2 2k_g$, we need $\alpha_g = \frac{\mathrm{Var}[\gamma_k]}{2\mathbb{E}[\gamma_k]}$ and $k_g = \frac{2\mathbb{E}[\gamma_k]^2}{\mathrm{Var}[\gamma_k]}$ in order to match the two first moments of the approximation, which completes the proof. □

With the result of Theorem 1 in place, we get the service-process Mellin transform through Corollary 1:

Corollary 1. *For the baseline scenario, with Bernoulli frame drops with probability p_{FA} due to PLA, the Mellin transform of the service process is given by*

$$\begin{aligned}&\mathcal{M}_{\mathcal{S},\textit{Baseline}}(s) = \\ &(1-p_{FA})\mathcal{M}_{g(\gamma_k)}\left[1+\frac{N_k(s-1)}{\ln 2}\right] + p_{FA}. \end{aligned} \tag{30}$$

Proof. This result follows by taking the expectation of $h(\gamma_k, X_k, N_k)^{s-1}$ where N_k is deterministic, utilizing that X_k is independent of γ_k and Bernoulli distributed with $p_X = \mathbb{P}(X_k = 1) = 1 - p_{\mathrm{FA}}$ in the baseline scenario. For mathematical details, we refer to [11, 22]. □

C. Analysis for Sybil Attacks

In a Sybil attack, the number of resources each device gets assigned $N_k \sim p_N(n)$ is varying depending on the success of the adversary. We provide the Mellin transform of the service process in this generalized case in the following corollary (following from Theorem 1):

Corollary 2. *Under Sybil attack, with scheduled resources distributed according to $p_N(n)$ and frame-drops with the false alarm rate p_{FA}, the service-process Mellin-transform is given by*

$$\begin{aligned}&\mathcal{M}_{\mathcal{S},\textit{Sybil}}(s) = \\ &(1-p_{FA})\sum_n \left[\mathcal{M}_{g(\gamma_k)}\left[1+\frac{n(s-1)}{\ln 2}\right] p_N(n)\right] + p_{FA}. \end{aligned} \tag{31}$$

Proof. We first note that in the context of the Sybil attack scenario, N_k is a random variable which requires us to take its distribution into consideration in the expectation of $h(\gamma_k, X_k, N_k)$. Following the same logic as in the proof of Corollary 1, we find that

$$\begin{aligned}&\mathcal{M}_{h(\gamma_k,X_k,N_k)}(s) = \mathbb{E}_{\gamma_k,X_k,N_k}\left[h(\gamma_k,X_k,N_k)^{s-1}\right] \\ &= p_X \sum_n \left[\mathbb{E}_{\gamma_k}\left[h(\gamma_k,1,n)^{s-1}\right] p_N(n)\right] + (1-p_X). \end{aligned} \tag{32}$$

Similarly to Corollary 1, we have

$$\mathbb{E}_{\gamma_k}\left[h(\gamma_k,1,n)^{s-1}\right] = \mathcal{M}_{g(\gamma_k)}\left[1+\frac{n(s-1)}{\ln 2}\right], \tag{33}$$

and by plugging this into (32) and again noting that $p_X = 1 - p_{\mathrm{FA}}$, the proof of (31) follows. □

D. Analysis for Disassociation Attacks

The modifications to the queueing model for disassociation attacks are described in Section III-E. The delay bound in Section IV-A applies to the block-aggregated service and arrival processes s'_l and a'_l. However, with the redefinition of the time scale, we now have $p(w) = \Pr(W(t) > K_{\mathrm{RC}} w)$. In the following, we present the Mellin transforms of the aggregated

arrival and service process under these assumptions. Since we assume constant arrivals of α bits per frame, we simply have $a'_l = K_{\text{CN}}\alpha$ and $\mathcal{M}_{\mathcal{A},\text{Disassociation}}(s) = e^{K_{\text{CN}}\alpha(s-1)}$. What remains is the Mellin transform of the aggregated service process, provided in the following corollary:

Corollary 3. *For the K_{CN} aggregated service process under a disassociation attack with success probability p_d, the Mellin transform is given by*

$$\mathcal{M}_{\mathcal{S},\textit{Disassociation}}(s) = (1-p_d)\left[\mathcal{M}_{\mathcal{S},\textit{Baseline}}(s)\right]^{K_{CN}} + p_d, \quad (34)$$

where $\mathcal{M}_{\mathcal{S},\textit{Baseline}}(s)$ is the Mellin transform for the baseline scenario given by (30)*.*

Proof. We note that $\mathcal{M}_{\mathcal{S},\text{Disassociation}}(s)$ is given by

$$\mathbb{E}[e^{s'_l(s-1)}] = (1-p_d)\mathbb{E}\left[\left(\prod_{k=K_{\text{CN}}l}^{K_{\text{CN}}(l+1)-1} h(\gamma_k, X_k)\right)^{s-1}\right] + p_d$$
$$= (1-p_d)\left[\mathcal{M}_{\mathcal{S},\text{Steady}}(s)\right]^{K_{\text{CN}}} + p_d,$$

where we have used that $h(\gamma_k, X_k)$ is independent for each k. □

V. Numerical Results

In this section, we use our analytical results to study a network consisting of $K_d = 24$ MTC devices deployed in a square 20 m $\times$ 20 m grid, one access point placed at the origin, and the adversary Eve positioned either outside the network or representing a compromised device within the network (see Fig. 2(a) for an example deployment). The network is operating at carrier frequency $f_c = 2.4$ GHz and the access point antenna array has normalized antenna separation $\Delta_r = 0.5$ and is oriented parallel to the line $y = -x$. We assume that all channel covariance matrices are of the form $[\mathbf{\Lambda}]_{i,j} = \rho^{|i-j|}$ where ρ is a correlation coefficient. The delay performance bounds are upper bounds on the delay violation probability $p_i(w) < p_{i,\text{Bound}}(w)$ computed as described in Section IV-A, where the minimization over s is carried out by a grid search. To specify the arrival rates α_i, we compute the rate corresponding to a fixed server utilization, defined as $u = \frac{\mathbb{E}(a_k)}{\mathbb{E}(s_k)}$.

A. Bound Validation

In Fig. 2(b), we show the delay violation probability for device D12, evaluated through link-level simulations, together with the corresponding bounds. In this figure, $K_{\text{Rice}} = 6$ dB, $\rho = 0$, $N_{\text{Rx}} = 4$ and we have included results with PLA in the baseline scenario, under Sybil attack, and under disassociation attack, as well as results without PLA in the baseline scenario. For the Sybil attack, we assume Eve launches $|D_{\text{Sybil}}| = 4$ Sybil IDs, and for the disassociation attack we assume the reconnection time is $K_{\text{CN}} = 4$ frames. To validate the bounds for varying channel conditions, we show the bound compared to Monte Carlo simulation results for varying Rice factors and a delay target of $w = 2$ frames in Fig, 2(c). In the link level simulations, the parameters of the SIMO channel distributions are computed based on the spatial positions depicted in Fig. 2(a) and delay is evaluated using a Monte Carlo approach where the stochastic service process is generated based on the complex channel realizations. Results in Fig. 2(b) and 2(c) are computed based on $\sim 10^8$ simulated frames. We can observe that in each scenario the analytical bounds and simulation results follow the same slope with a gap of 1-3 orders of magnitude between the curves; hence, our analysis can validly upper bound the performance of the modeled system. Typically, the gap between simulation and bound increases with the slope of the curves. Therefore, the derived bounds will clearly overestimate the true delay violation probability as seen in Fig. 2(b). However, they provide us with an efficient (i.e., in the sense that computing the bounds is significantly less computationally demanding than simulating the system) and conservative (i.e., the true system will perform considerably better than the bounds predict) way of evaluating the system's delay performance.

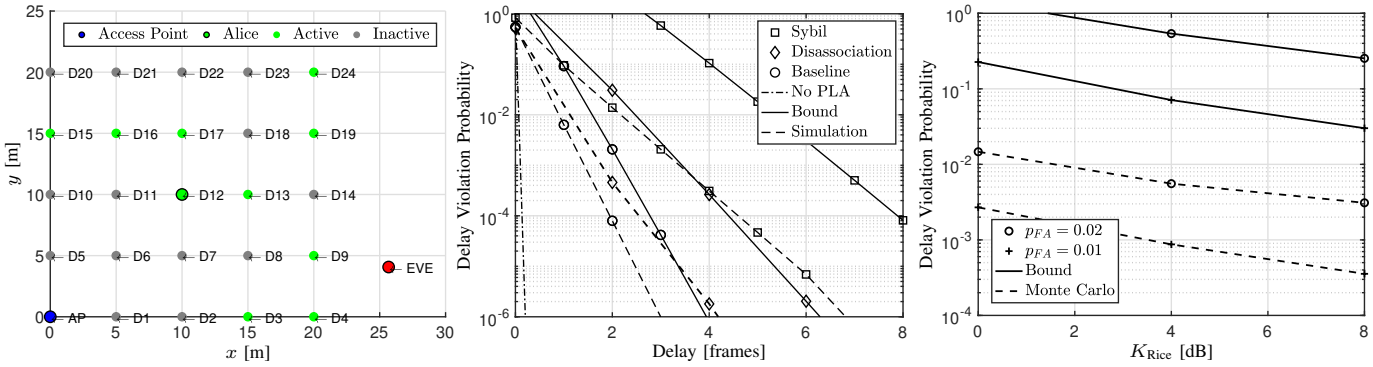
B. Baseline Performance

Fig. 3(a) illustrates the delay w_ϵ that can be analytically guaranteed with a violation probability of $\epsilon = 10^{-6}$, i.e., $p_{i,\text{Bound}}(w_\epsilon) = \epsilon$, for a given false alarm rate. For illustration, we consider only a subset of devices covering the full range of AoAs and distances. The value of w_ϵ varies little between devices. This is beacuse arrival rates are adapted differently for each device to fix the utilization u. Fig. 3(a) also illustrates how PLA impacts the system: to get a low missed detection rate we typically want to have a low threshold T. However, decreasing T increases the false alarm rate, which clearly impacts the delay performance guarantee w_ϵ. For this particular scenario, we see that a false alarm rate approaching 10^{-2} can have an impact of 2-5 frames on the delay guarantee. Higher utilization in Fig. 3(a) means that the arrival rates are higher, resulting in an increased delay guarantee. However, we observe that the delay shows a similar behavior with the false alarm rate for both $u = 0.5$ and $u = 0.9$.

C. Data Injection Attacks

Here, we consider the detection performance of PLA in the data injection attack. Delay impacts of data injection are not studied in this case since these are similar to the Sybil attack, as discussed in Section III-C.

a) Inactive device targeted: Fig. 3 shows the analytical missed detection rate $p_{\text{MD}}(i, T)$ when Eve launches a data injection attack against an inactive device. In these figures, we assume that Eve is positioned at $(x, y) = (25, 0)$ [m], targeting devices $\{\text{D4}, \text{D8}, \text{D12}, \text{D16}, \text{D20}\}$, and that the PLA threshold is fixed at a false alarm rate $p_{\text{FA}} = 10^{-2}$. Fig. 3(b) depicts the missed detection rate for varying K_{Rice} with fixed $K_{\text{Rice},E} = 0$ dB. As expected, the missed detection rate improves with stronger LOS component. We observe that the detection performance for D4 is worse due to its location at $(0, 20)$ close to Eve. Additionally, we can observe that higher antenna correlation has a positive effect on the missed detection rate performance. Fig. 3(c) shows the influence of $K_{\text{Rice},E}$ on the missed detection rate for fixed $K_{\text{Rice}} = 5$ dB. For PLA of devices far from Eve, a stronger LOS component from Eve allows the access point to better differentiate messages



(a) (b) (c)

Fig. 2. (a) Considered MTC deployment. Comparison of link-level simulations to derived bounds for device D12: (b) for all considered attack strategies and (c) delay violation probability for delay target $w = 2$ frames in baseline scenario for varying Rice factors.

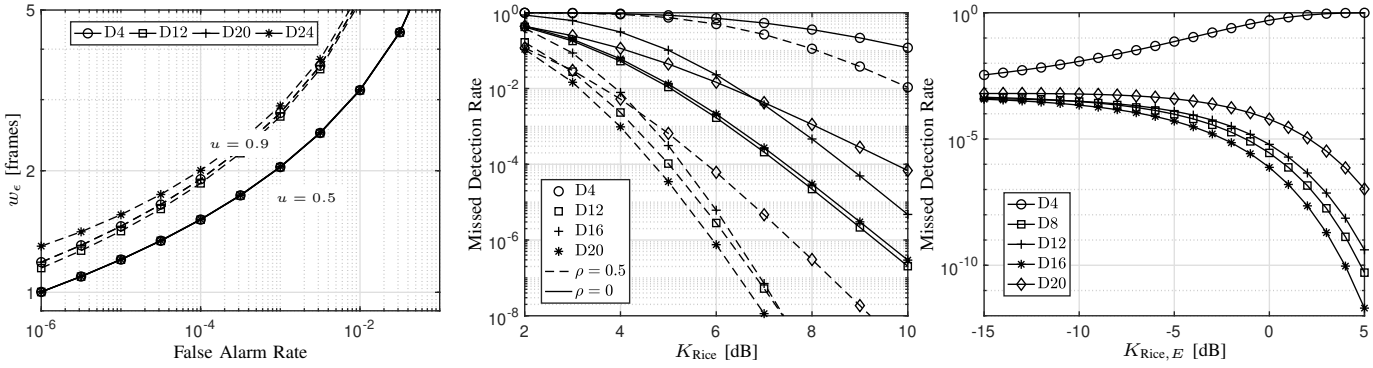
from Eve. However, for device D4 the missed detection rate shows the opposite behavior since Eve's channel more and more resembles the legitimate channel. We can also see that for low $K_{\text{Rice},E}$ (i.e, Eve's channel is approaching NLOS), the missed detection rate approaches the same value for all choices of devices to impersonate. In Fig. 4(a), we plot the missed detection rate for varying N_{Rx} showing that the missed detection rate follows an approximately log-linear decrease with N_{Rx}.

b) Active device targeted: Fig. 4(b) shows the missed detection rate $p_{\text{MD}}^{L=2}$ when Eve targets an active device from $d_E = 30$ m and varies her AoA from $\pi/4$ to $3\pi/4$. The first upper bound corresponds to $p_{\text{MD}}(T, i)$, the second to $p_d(i)$, the lower bound correspond to $p_{\text{MD}}(T, i)p_{\text{FA}}(T)$ (see (12) and (15) in Section III-C), and the solid curve is generated by Monte Carlo simulation. We can observe that the gap between the tightest upper bound and the true value is around 1 order of magnitude. Additionally, this figure illustrates that there is an optimal AoA for Eve to impersonate this particular device with the highest success rate. In Fig. 4(c), we depict the upper bound on $p_{\text{MD}}^{L=2}$, for each device, when Eve is choosing the optimal AoA. Note that this is the upper bound and that the actual detection performance is around one order of magnitude lower. We observe that for devices D1, D5 and D6, the missed detection rate is very low ($< 10^{-8}$), while the upper bound can approach values higher than 10^{-1} for some poorly positioned devices. Also, we observe that generally, the missed detection rate is improved when Eve only has a NLOS channel (i.e., $K_{\text{Rice},E} = -\infty$ dB). As a single-antenna attacker, we note that it is impossible for Eve to estimate the optimal AoA through eavesdropping communications. Though through knowledge of the deployment, Eve can position herself at a similar LOS path as the legitimate device to optimize her chances of success. However, if Eve's objective is to impersonate several devices simultaneously, the optimal AoA becomes conflicting as illustrated by Fig. 4(c).

These results highlight two variables affecting the detection performance of PLA for a given false alarm rate: (i) network deployment and environment affecting LOS strengths for legitimate channels and for Eve; and (ii) access point design in terms of amount and placement of antennas. It is clear that we can improve the missed detection rate by adding more antennas and placing them such that antenna-correlation is high. Moreover, by designing the deployment and the immediate environment such that devices have a strong LOS path to the access point, while Eve is unable to get a strong LOS path (e.g., through deployment of the system in a closed environment), we can improve detection performance. Influencing channel characteristics for improved PLA performance might be feasible in some scenarios (e.g., in a factory deployment). Moreover, deployments with strong LOS components might be desirable for pure communication reasons as well.

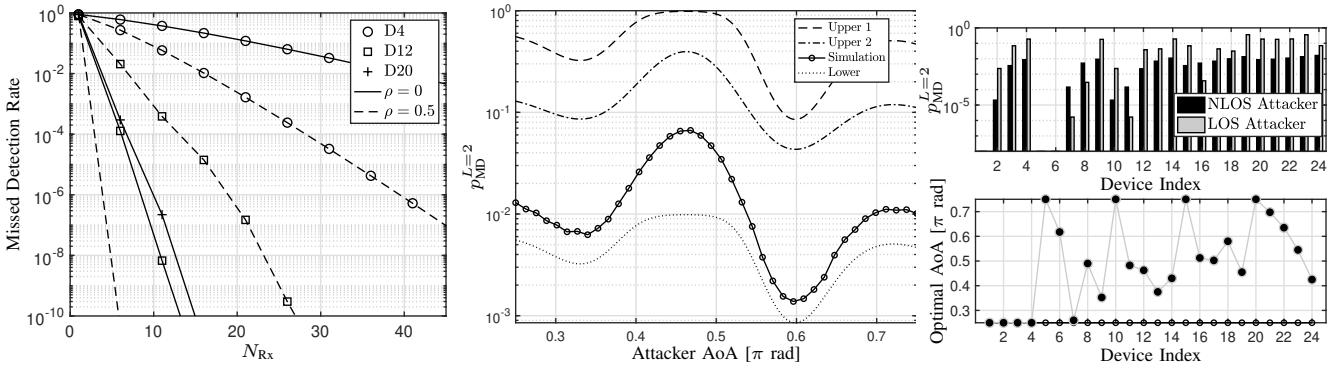
D. Sybil Attacks

Here, we assume that devices $D_{\text{Active}} = \{\text{D12}, \text{D13}, \text{D14}, \text{D17}, \text{D18}, \text{D19}, \text{D22}, \text{D23}, \text{D24}\}$ (i.e., the upper-right quadrant of the deployment) are active and that device D4 has been compromised and is launching a Sybil attack. Fig. 5(a) shows $\mathbb{E}[K_{\text{Sybil}}]$, the average number of Sybil nodes successfully launched by Eve, as a function of the number of targeted devices $|D_{\text{Sybil}}|$. The solid lines are computed according to our approximate distribution (18), while the dashed lines correspond to simulation results, showing that our approximation is accurate. We see that with no PLA, Eve successfully gets every Sybil ID accepted. With PLA and lower p_{FA}, the number of successful Sybil nodes is kept lower. For instance, when $p_{\text{FA}} = 10^{-2}$, the expected number of Sybil nodes does not exceed $\mathbb{E}[K_{\text{Sybil}}] > 2$ even though Eve can launch up to $|D_{\text{Sybil}}| = 14$ Sybil IDs, which means that PLA is effective against the attack. However, it is apparent from Fig. 5(a) that there are Sybil IDs that cannot be detected by the PLA. The reason is that in the particular scenario that we have investigated, Eve is device D4, and hence, more easily impersonates devices $\{\text{D1}, \text{D2}, \text{D3}\}$ due to having the same AoA. Fig. 5(b) shows the delay guarantee w_ϵ for D12 and $\epsilon = 10^{-6}$ under the Sybil attack. We can observe that without PLA, the increasing number of Sybil IDs launched by Eve has a severe effect on the delay performance. For example, when the link utilization is high ($u = 0.7$), Eve only has to introduce 4-5 Sybil IDs to cause the delay in the queue to grow towards infinity. On the other hand, by effectively detecting the Sybil IDs with PLA with $p_{\text{FA}} = 10^{-2}$, the delay performance can be made almost independent of the number of Sybil IDs, at a cost of a constant higher delay of around 3 frames.



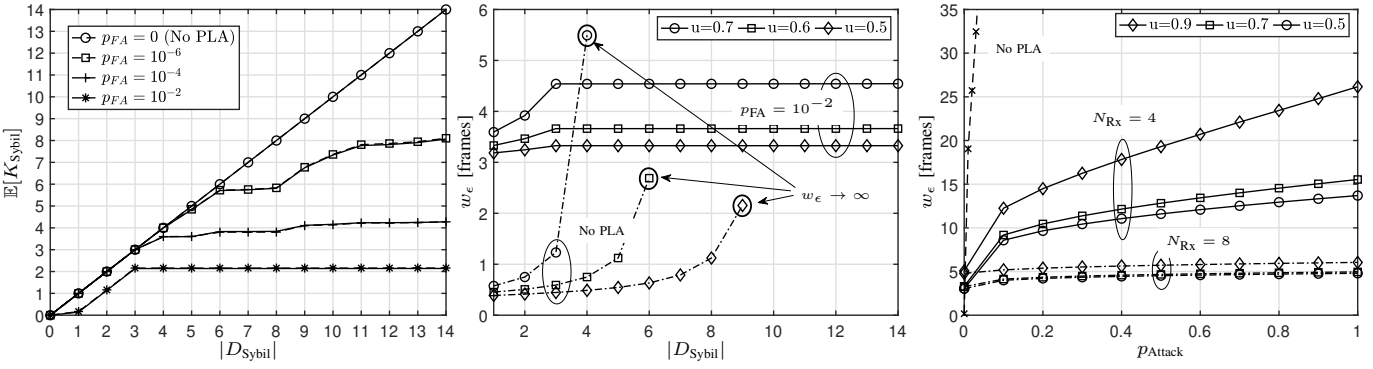
(a) (b) (c)

Fig. 3. (a) Delay guarantee w_ϵ with $\epsilon = 10^{-6}$ for device D12 in baseline scenario and PLA detection performance under data injection attack with $N_{\text{Rx}} = 8$ when Eve impersonates $\{\text{D4}, \text{D8}, \text{D12}, \text{D16}, \text{D20}\}$: (b) For varying LOS strength, (c) for varying attacker LOS strength, and (c) for varying number of receive antennas.



(a) (b) (c)

Fig. 4. (a) Detection performance under data injection attack for varying number of receive antennas, (b) Missed detection rate during data injection attack vs. attacker AoA when D12 is targeted, $K_{\text{Rice}} = 6$ dB and $K_{\text{Rice},E} = 0$ dB, (c) Upper bound on missed detection rate during data injection attack and Eve's optimal AoA.



(a) (b) (c)

Fig. 5. (a) Expected number of successful Sybil IDs $\mathbb{E}[K_{\text{Sybil}}]$ for various choices of p_{FA}. (b) Delay performance impacts for D12 under Sybil attack. (c) Delay performance impacts for D12 under disassociation attack.

E. Disassociation Attacks

Here, we assume that Eve is an external entity, positioned at $d_E = 25$ m and $\Phi_E = \pi/3$, launching a disassociation attack targeting device D12. Fig. 5(c) shows the delay guarantee w_ϵ for $\epsilon = 10^{-6}$ for the targeted device as a function of the attack probability. In the results without PLA, we have assumed that the access point performs random guessing whenever two requests are received at the same time. For this case, we can clearly see that the attack causes the delay to increase for very low attack intensities, simply because the device gets disconnected 50% of the times Eve sends a DCN request. With PLA the impact is reduced and the system is able to give delay guarantees even when $p_{\text{Attack}} \to 1$ and Eve is sending DCN requests in every frame. For $N_{\text{Rx}} = 4$, we however see an increase in w_ϵ with p_{Attack} due to the occasional missed detections. Fig. 5(c) also illustrates that this increase can be mitigated by increasing the number of receive antennas to $N_{\text{Rx}} = 8$.

F. Discussion

The PLA scheme studied in this paper can achieve missed detection rates of around 10^{-6} and even as low as 10^{-10}

given Rice factor in the order 5-10 dB which is reasonable compared to LOS measurements at 2.4 GHz [25]. These values can certainly meet security requirements even in applications where message integrity is of critical importance. Our results also indicate that these security enhancements come at a limited cost in terms of delay: In the baseline scenario, given reasonable false alarm rates $p_{\text{FA}} < 10^{-2}$, our results show that a delay $w_\epsilon < 5$ frames can be guaranteed with a reliability of $\epsilon = 10^{-6}$. Throughout the paper and in these results, we keep the delay in terms of the number of frames in order to make the results general and not tailored to a specific system parameterization. The particular impact in a real application would obviously depend on the frame duration, the transmission technology and its relation to the coherence time of the channel, and the latency requirements of the application which can range from 5-10 ms in intelligent transportation down to 0.5 ms in certain industrial applications [26]. In relation to standard LTE with frame periods of 1-10 ms, coherence time around the same order could be expected in certain scenarios [12]. However, in ultra-low latency communications with very short frame periods the assumption of independent fading across the frames that underlies our results might not hold. In such scenarios, the queueing analysis would need to be extended to time-correlated fading channels; something that is considered outside the scope of this paper. Alternatively, as discussed in Section II-A, our analysis would in such scenarios apply to systems that utilize frequency hopping or resource scheduling in order to combat the impact of long coherence times. As a side note, our results are also valid for a multi-carrier system with strongly correlated fading carriers since the number of subcarriers can be included in $N_k = N_{k,\text{time}} N_{\text{Subcarriers}}$ (i.e., $N_{\text{Subcarriers}}$ represents bandwidth and becomes a multiplicative factor in front of the Shannon capacity). However, for the results associated with a fixed utilization such constant scaling of the service process has no impact on the delay distribution since the arrival rate will be equivalently re-scaled.

Introducing multiple-antenna access points appear beneficial from both a PLA security and a delay perspective. Already with 4-8 receive antennas, we observe large benefits in terms of missed detection rate that continues to improve in a log-linear fashion. The benefits of introducing more antennas at the access point can be interpreted in two ways: (i) improved detection performance for a given false alarm rate, or (ii), improved false alarm rate for a given detection performance. That is, we can utilize extra antennas either to strengthen the integrity of communication, or to reduce the delay impacts (i.e., decrease false alarm rate). Our results also provide insight into deployment strategies: We have seen that if many devices are deployed along a straight line resulting in similar AoA profiles, an external or internal attacker will be more effective in impersonating this set of devices simultaneously. Hence, if the deployment of MTC devices can be influenced for security purposes, this can be used to make sure that Sybil attacks targeting many devices are unlikely to succeed. Furthermore, if certain devices transmit particularly sensitive information, these can be placed in positions such that Eve's success-rate when impersonating is minimized.

VI. Conclusions

We have studied delay impacts of a feature-based PLA protocol in order to investigate the viability of PLA for mission-critical MTC applications. Based on a MTC network model consisting of multiple devices and a multi-antenna access point we have derived delay performance bounds that quantify the delay impacts of PLA. Evaluation of the derived bounds for a network with a square-grid deployment of 24 MTC devices shows that PLA can, under good LOS conditions, be used without introducing excessive delays. Additionally, we have found that PLA allows low-latency high-reliability communication even under hostile attack scenarios such as Sybil and disassociation attacks. As a means of improving detection and delay performance, one could consider multiple antenna-arrays deployed at separate locations in a distributed manner. Additionally, in this paper we have limited the analysis to a single-antenna adversary; however, this could be extended to several adversaries with multiple antennas. Moreover, channel estimation techniques and their effect on the queueing model and authentication performance is still an open problem. Finally, our analysis could easily be modified to encompass other authentication schemes (e.g., based on other features or fingerprinting tags) and through this be used to compare different PLA schemes from a delay perspective.

References

[1] 3GPP, "Study on communication for automation in vertical domains (CAV)," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 22.804, 2017. [Online]. Available: https://portal.3gpp.org/ngppapp/CreateTdoc.aspx?mode=view&contributionUid=SP-180332

[2] A. Weinand, M. Karrenbauer, J. Lianghai, and H. D. Schotten, "Physical layer authentication for mission critical machine type communication using Gaussian mixture model based clustering," in *IEEE Vehicular Technology Conference*, June 2017, pp. 1–5.

[3] L. Xiao, L. Greenstein, N. Mandayam, and W. Trappe, "Fingerprints in the ether: Using the physical layer for wireless authentication," in *IEEE International Conference on Communications*, June 2007, pp. 4646–4651.

[4] W. Hou, X. Wang, J.-Y. Chouinard, and A. Refaey, "Physical layer authentication for mobile systems with time-varying carrier frequency offsets," *IEEE Transactions on Communications*, vol. 62, no. 5, pp. 1658–1667, May 2014.

[5] A. Weinand, M. Karrenbauer, R. Sattiraju, and H. Schotten, "Application of machine learning for channel based message authentication in mission critical machine type communication," in *European Wireless Conference*, May 2017, pp. 1–5.

[6] X. Wang, P. Hao, and L. Hanzo, "Physical-layer authentication for wireless security enhancement: current challenges and future developments," *IEEE Communications Magazine*, vol. 54, no. 6, pp. 152–158, June 2016.

[7] M. Ozmen and M. C. Gursoy, "Secure transmission of delay-sensitive data over wireless fading channels," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 9, pp. 2036–2051, Sept 2017.

[8] ——, "Energy-delay-secrecy tradeoffs in wireless communications under channel uncertainty," in *IEEE Wireless Communications and Networking Conference*, April 2018, pp. 1–6.

[9] F. Naghibi, S. Schiessl, H. Al-Zubaidy, and J. Gross, "Performance of wiretap Rayleigh fading channels under statistical delay constraints," in *IEEE International Conference on Communications*, May 2017, pp. 1–7.

[10] M. Fidler and A. Rizk, "A guide to the stochastic network calculus," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 92–105, Firstquarter 2015.

[11] H. Forssell, R. Thobaben, H. Al-Zubaidy, and J. Gross, "On the impact of feature-based physical layer authentication on network delay performance," in *IEEE Global Communications Conference*, Dec 2017, pp. 1–6.

[12] H. MacLeod, C. Loadman, and Z. Chen, "Experimental studies of the 2.4-GHz ISM wireless indoor channel," in *3rd Annual Communication*

Networks and Services Research Conference (CNSR'05), May 2005, pp. 63–68.
[13] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.
[14] A. Abdelaziz, R. Burton, and C. E. Koksal, "Message authentication and secret key agreement in VANETs via angle of arrival," *CoRR*, Sep. 2016. [Online]. Available: http://arxiv.org/abs/1609.03109
[15] L. Xiao, L. Greenstein, N. Mandayam, and W. Trappe, "Using the physical layer for wireless authentication in time-variant channels," *IEEE Transactions on Wireless Communications*, vol. 7, no. 7, pp. 2571–2579, July 2008.
[16] A. Mahmood, W. Aman, M. O. Iqbal, M. M. U. Rahman, and Q. H. Abbasi, "Channel impulse response-based distributed physical layer authentication," in *IEEE Vehicular Technology Conference*, June 2017, pp. 1–5.
[17] L. Xiao, L. Greenstein, N. Mandayam, and W. Trappe, "Channel-based detection of sybil attacks in wireless networks," *IEEE Transactions on Information Forensics and Security*, vol. 4, no. 3, pp. 492–503, Sept 2009.
[18] H. Al-Zubaidy, J. Liebeherr, and A. Burchard, "Network-layer performance analysis of multihop fading channels," *IEEE/ACM Transactions on Networking*, vol. 24, no. 1, pp. 204–217, Feb 2016.
[19] N. Petreska, H. Al-Zubaidy, R. Knorr, and J. Gross, "Power-minimization under statistical delay constraints for multi-hop wireless industrial networks," *CoRR*, vol. abs/1608.02191, 2016. [Online]. Available: http://arxiv.org/abs/1608.02191
[20] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Transactions on Wireless Communications*, vol. 2, no. 4, pp. 630–643, July 2003.
[21] M. Fidler, "WLC15-2: A network calculus approach to probabilistic quality of service analysis of fading channels," in *IEEE Globecom 2006*, Nov 2006, pp. 1–6.
[22] S. Schiessl, J. Gross, and H. Al-Zubaidy, "Delay analysis for wireless fading channels with finite blocklength channel coding," in *Int. Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, 2015, pp. 13–22.
[23] H. Al-Zubaidy, V. Fodor, G. Dán, and M. Flierl, "Reliable video streaming with strict playout deadline in multihop wireless networks," *IEEE Transactions on Multimedia*, vol. 19, no. 10, pp. 2238–2251, Oct 2017.
[24] E. S. Pearson, "Note on an approximation to the distribution of non-central χ^2," *Biometrika*, vol. 46, no. 3/4, pp. 364–364, 1959. [Online]. Available: http://www.jstor.org/stable/2333533
[25] T. A. Wysocki and H. J. Zepernick, "Characterization of the indoor radio propagation channel at 2 . 4 GHz," in *Journal of Telecommunications and Information Technology*, 2000, pp. 84–90.
[26] H. Chen, R. Abbas, P. Cheng, M. Shirvanimoghaddam, W. Hardjawana, W. Bao, Y. Li, and B. Vucetic, "Ultra-Reliable Low Latency Cellular Networks: Use Cases, Challenges and Approaches," *ArXiv e-prints*, Sep. 2017.

Henrik Forssell received the B.Sc degree in 2013 and the M.Sc degree in 2015 from KTH Royal Institute of Technology, Stockholm. In 2015, he joined the Division of Information Science and Engineering, KTH Royal Institute of Technology, Stockholm, where he is currently pursuing the Ph.D degree. He is also a part of the CERCES research project which is concerned with the analysis and design of resilient critical infrastructures. His research focus is on physical layer security in the context of wireless communication in critical infrastructures. In particular, his topics of interest are authentication, key-agreement, and jamming resilience at the physical layer.

Ragnar Thobaben (M'07) received the Dr.-Ing. (Ph.D.) degree in electrical engineering from the Christian-Albrechts-University of Kiel, Germany, in 2007. In 2006, he joined the KTH Royal Institute of Technology, Stockholm, Sweden, as a Post-Doctoral Researcher, where he now serves as an Associate Professor. He has researched various topics such as cognitive radio, cooperative communication, coordination, and security. His current research focuses on physical-layer security, simultaneous wireless energy and information transfer, as well as signal processing, information theory, and coding theory applied to problems in communications and learning. Dr. Thobaben has served on the Technical Program Committee for several IEEE conferences, such as GLOBECOM, ICC, and PIMRC. He has also served as the publicity chair for the 2011 IEEE Swedish Communication Technologies Workshop, the 2012 International Symposium on Turbo Codes and Iterative Information Processing, and the 2015 IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), and he is currently serving as the local-arrangement chair for the 2019 IEEE Information Theory Workshop (ITW). Since 2016, he has been serving as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS.

Hussein Al-Zubaidy (S'07–M'11–SM'16) received the Ph.D. degree in electrical and computer engineering from Carleton University, Ottawa, ON, Canada, in 2010. He was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada, from 2011 to 2013. In the Fall of 2013, he joined as a Postdoctoral Fellow the School of Electrical Engineering (EES), Royal Institute of Technology (KTH), Stockholm, Sweden, and as a Senior Researcher in the fall of 2015. He is the recipient of many honors and awards, including the Ontario Graduate Scholarship (OGS), NSERC Visiting Fellowship, NSERC Summer Program in Taiwan, OGSST, and NSERC Post-Doctoral Fellowship.

James Gross received his Ph.D. degree from TU Berlin in 2006. From 2008-2012 he was Assistant Professor and head of the Mobile Network Performance Group at RWTH Aachen University, as well as a member of the DFG-funded UMIC Research Centre of RWTH. Since November 2012, he has been with the Electrical Engineering and Computer Science School, KTH Royal Institute of Technology, Stockholm, as an Associate Professor. He also serves as Director for the ACCESS Linnaeus Centre and is a member of the board of KTHs Innovative Centre for Embedded Systems. His research interests are in the area of mobile systems and networks, with a focus on critical machine-to-machine communications, cellular networks, resource allocation, as well as performance evaluation methods. He has authored about 150 (peer-reviewed) papers in international journals and conferences. His work has been awarded multiple times, including the Best Paper Award at ACM MSWiM 2015, the Best Demo Paper Award at IEEE WoWMoM 2015, the Best Paper Award at IEEE WoWMoM 2009, and the Best Paper Award at European Wireless 2009. In 2007, he was the recipient of the ITG/KuVS dissertation award for his Ph.D. thesis. He is also co-founder of R3 Communications GmbH, a Berlin-based start-up in the area of ultrareliable low-latency wireless networking for industrial automation.