TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

Optimal activation rates in ultra-dense wireless networks with intermittent traffic sources

Citation for published version (APA):
Cecchi, F., Borst, S. C., Van Leeuwaarden, J. S. H., & Whiting, P. A. (2018). Optimal activation rates in ultra-dense wireless networks with intermittent traffic sources. In *INFOCOM 2018 - IEEE Conference on Computer Communications* (pp. 2672-2680). Article 8485889 Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/INFOCOM.2018.8485889

DOI:
10.1109/INFOCOM.2018.8485889

Document status and date:
Published: 08/10/2018

Document Version:
Accepted manuscript including changes made at the peer-review stage

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Download date: 16. Nov. 2023

Optimal Activation Rates in Ultra-Dense Wireless Networks with Intermittent Traffic Sources

F. Cecchi*, S.C. Borst*†, J.S.H. van Leeuwaarden*, P.A. Whiting‡,
* Eindhoven University of Technology, Eindhoven, The Netherlands
† Nokia Bell Labs, Murray Hill, NJ, USA ‡ Macquarie University, North Ryde, NSW, Australia
Email: f.cecchi@tue.nl, s.c.borst@tue.nl, j.s.h.v.leeuwaarden@tue.nl, philip.whiting@mq.edu.au

***Abstract*—As the Internet-of-Things (IoT) emerges, connecting immense numbers of sensors and devices, the continual growth in wireless communications increasingly manifests itself in terms of a larger and denser population of nodes with intermittent traffic patterns. A crucial issue that arises in these conditions is how to set the activation rates as a function of the network density and traffic intensity. Depending on the scaling of the activation rates, dense node populations may either result in excessive activations and potential collisions, or long delays that may increase with the number of nodes, even at low load.**

Motivated by the above issues, we examine optimal activation rate scalings in ultra-dense networks with intermittent traffic sources. We establish stability conditions, and provide closed-form expressions which indicate that the mean delay is roughly inversely proportional to the nominal activation rate. We also discuss a multi-scale mean-field limit, and use the associated fixed point to determine the buffer content and delay distributions. The results provide insight in the scalings that minimize the delay while preventing excessive activation attempts. Extensive simulation experiments demonstrate that the mean-field asymptotics yield highly accurate approximations, even when the number of nodes is moderate.

I. Introduction

A. Background and motivation

In the present paper we examine optimal activation rate scalings in large-scale wireless networks with intermittent traffic flows. The sustained growth in wireless communications will increasingly manifest itself in terms of a larger number and denser population of nodes, driven by a proliferation of low-cost sensors and machine-type devices in the emerging Internet-of-Things (IoT). Forecasts indicate that the number of IoT nodes will reach into the tens of billions over the next few years, and outgrow the number of human-operated devices by an order-of-magnitude [1], [2].

With such a massive number of nodes, each of which individually may only be sporadically active, any form of dedicated spectrum allocation or scheduled medium access is impractical. Instead these large-scale networks will typically rely on the individual nodes to dynamically share the medium in a distributed fashion. A popular mechanism for distributed medium access control is provided by the Carrier-Sense Multiple-Access (CSMA) protocol. In the CSMA protocol each node attempts to access the medium after a certain back-off time, but nodes that sense activity of interfering nodes freeze their back-off timer until the medium is sensed idle.

While the CSMA protocol is fairly easy to understand at a local level, the interaction among interfering nodes gives rise to quite intricate behavior on a macroscopic scale. As it turns out though, in saturated-buffer scenarios (so nodes always have packets to transmit), the joint activity process of the various nodes has a product-form stationary distribution [22], providing useful throughput estimates for persistent traffic flows [4], [24], [31]. However, these results do not capture the relevant performance metrics in unsaturated-buffer scenarios, which in particular arise in an IoT context with highly intermittent traffic sources. In such situations, buffers will frequently be empty, and nodes will refrain from competition for the medium during these periods. The resulting two-way interaction between the activity states and the buffer contents produces extremely complex behavior, and even basic throughput characteristics and stability conditions have remained largely elusive so far [9], [23].

A particularly crucial issue that arises in the above scenarios, with vast numbers of intermittently active sources, is how to set the back-off rates as a function of the network density and traffic intensity. In order to avoid collisions, the value of the back-off rate should first of all account for the maximum signal propagation delay between interferers, which is mostly governed by the physical attributes of the network. However, as networks grow increasingly dense, the sheer number of interferers can grow extremely large as well. Thus the aggregate back-off rate of the nodes within interference range can be correspondingly large, which may also give rise to spurious collisions.

The above issue can be countered by lowering the value of the back-off rate in dense networks and for example setting it inversely proportional to the number of nodes. Such a rule of thumb would be somewhat reminiscent of the well-known fact that in a slotted Aloha system with N nodes a transmission probability $1/N$ maximizes the throughput [28]. However these considerations implicitly rely on the assumption that nodes have saturated buffers and always packets to transmit. Dense networks with a huge number of nodes can only sustain the collective load if they each individually have low traffic rates and are only sporadically active. In these cases setting the back-off rate inversely proportional to the number of nodes could result in unnecessarily long delays that may increase with the number of nodes, even when the collective load is not particularly high.

B. Key contributions and implications

We will examine optimal activation rate scalings in ultra-dense wireless networks with intermittent traffic sources. Specifically, we assume that the mean nominal back-off rate at each node is scaled by a factor $f(N)$ as a function of the total number of nodes. We then investigate the impact of $f(N)$ on key performance metrics such as the buffer content and delay. We further investigate how the stationary aggregate back-off rate behaves as a function of $f(N)$. The key contributions and findings may be summarized as follows:

1) We establish that, for any fixed subcritical aggregate traffic intensity, the system is stable as long as the mean nominal back-off period scales sublinearly with the number of nodes, i.e., $f(N)$ falls off slower than $1/N$.

2) We provide closed-form expressions for the mean stationary buffer content and delay, which illuminate the impact of the function $f(N)$. In particular, for a fixed traffic intensity, the mean stationary delay and system-wide backlog scale roughly inversely proportional with $f(N)$, i.e., they grow approximately linearly with the mean nominal back-off period.

3) We discuss a multi-scale mean-field limit for the overall buffer occupancy process. In contrast to the standard mean-field set-up, we do not simply normalize the entire process by the number of nodes N as a common factor, but scale the number of nodes with k or more packets in their buffer by $N(Nf(N))^{-k}$. We identify the fixed point of the mean-field limit, and use that to determine the buffer content and delay distributions in the many-sources regime. In particular, the probability that the buffer of an individual node contains k or more packets scales as $(Nf(N))^{-k}$ as $N \to \infty$.

4) The mean-field limit further reveals that the aggregate back-off rate settles around a constant value as the number of nodes grows large, which is independent of the exact back-off scaling, in contrast to saturated-buffer scenarios. This reflects a self-regulating property, where the number of backlogged nodes that is actively competing for the medium is roughly inversely proportional to the nominal back-off rate. For example, a more aggressive back-off scaling indirectly results in a proportional reduction of the number of backlogged nodes.

5) In order to further sharpen the above-mentioned 'concentration' property, we establish the central limit behavior of the system-wide backlog, and leverage that to show that the aggregate back-off rate is bounded in the many-sources regime with high probability. We use the mean-field asymptotics and central limit result to provide insight in the scalings that minimize the delay while preventing excessive activation attempts.

6) The various analytical results will be corroborated by extensive simulation experiments, which confirm that the mean-field asymptotics yield highly accurate approximations for the relevant performance metrics, even when the number of nodes is moderate.

C. Related work

As noted earlier, the issue of how to set activation probabilities was extensively studied in slotted Aloha systems, see for instance [18], [28], [30], [32]. However, these studies either consider static optimization as a function of the number of nodes in saturated buffer scenarios, or dynamic optimization for infinite-source models, and the collision dynamics render the problem fundamentally different from that for collision avoidance schemes.

The line of work that has pursued optimization of activation rates for CSMA schemes, has mainly been concerned with optimizing some fairness criterion or global throughput utility metric in scenarios with saturated buffers [21], [25]. Scenarios with packet arrivals have been considered in the context of queue-based CSMA strategies [17], [20], [27], [29], but these involve given queue-dependent activation probabilities, whereas we focus on optimization of the nominal back-off rates as a function of the total number of nodes.

Mean-field analysis has emerged as a powerful approach to obtain tractable performance estimates in random-access networks. Mean-field concepts were already leveraged to derive throughput estimates in saturated-buffer scenarios in the seminal paper [3], with further results in [12], [15], [26]. Mean-field techniques have also proved useful in examining stability issues [5] and obtaining expressions for queue length distributions and delay metrics in unsaturated-buffer scenarios [10], [11]. However, none of these papers have pursued *multi-scale* mean-field limits, or specifically addressed the performance impact of the activation rates and how to set these as a function of the network density and traffic intensity.

D. Paper organization

The remainder of the paper is organized as follows. In Section II we present a detailed model description. In Section III we establish stability conditions and provide closed-form expressions for the mean stationary buffer content and delay as a function of the activation rate. We discuss a multi-scale mean-field limit for the overall buffer occupancy process in Section IV, and use the fixed point to determine the buffer content and delay distributions and demonstrate the concentration property of the aggregate back-off rate in the many-sources regime. Section V establishes the central limit behavior of the total buffer content and bounds for the aggregate back-off rate. In Section VI we discuss the simulation experiments that we have conducted to corroborate the analytical results. In Section VII we make some concluding remarks and describe possible topics for further research.

II. Model description

Consider a random-access network of N mutually interfering nodes sharing a wireless medium. Packets are generated at the various nodes as independent Poisson processes of rate λ/N. When a node gains access to the medium, it transmits a single packet, which takes an exponentially distributed time with parameter μ. Before initiating a transmission, a node obeys a back-off period. This period is frozen whenever

another node is transmitting and immediately resumed when the medium is sensed free again. When a node with packets to transmit completes a back-off period, it gains access to the medium and starts a transmission. The back-off period at every node is exponentially distributed with parameter $\nu f(N)$. Denote the aggregate traffic intensity of the network by $\rho = \lambda/\mu$.

The network evolution is described by the queue length process $\boldsymbol{Q}^{(N)}(t)$ and the activity process $Y^{(N)}(t)$. The queue length $Q_n^{(N)}(t)$ represents the number of packets in the buffer of node n at time t (excluding the one possibly in transmission). The activity state at time t is $Y^{(N)}(t) = 1$ if any of the nodes is active (transmitting) and $Y^{(N)}(t) = 0$ otherwise. Because all the nodes are exchangeable, the queue length process may be equivalently represented by the population process $\boldsymbol{X}^{(N)}(t)$, where $X_k^{(N)}(t)$ denotes the number of nodes having k packets in their buffer at time t, i.e,

$$X_k^{(N)}(t) = \sum_{n=1}^{N} \mathbb{1}\Big\{Q_n^{(N)}(t) = k\Big\}.$$

The process $(\boldsymbol{X}^{(N)}(t), Y^{(N)}(t))$ is Markovian with the following transitions:

- A packet arrives at a node having k packets in its buffer with rate $\lambda X_k^{(N)}/N$, generating the transition

$$(\boldsymbol{X}^{(N)}, Y^{(N)}) \to (\boldsymbol{X}^{(N)} + \boldsymbol{e}_{k+1} - \boldsymbol{e}_k, Y^{(N)}),$$

where $\boldsymbol{e}_k$ is the k-th N-dimensional unit vector.

- A transmission is completed with rate μ and only if a node is transmitting, i.e., $Y^{(N)} = 1$, generating the transition

$$(\boldsymbol{X}^{(N)}, Y^{(N)}) \to (\boldsymbol{X}^{(N)}, Y^{(N)} - 1).$$

- A back-off is completed by a node having $k > 0$ packets in its buffer with rate $\nu f(N) X_k^{(N)}$ and only if no node is transmitting, i.e., $Y^{(N)} = 0$, generating the transition

$$(\boldsymbol{X}^{(N)}, Y^{(N)}) \to (\boldsymbol{X}^{(N)} + \boldsymbol{e}_{k-1} - \boldsymbol{e}_k, Y^{(N)} + 1).$$

Similarly define the 'cumulative' population process $\boldsymbol{Z}^{(N)}(t)$, where $Z_k^{(N)}(t)$ denotes the number of nodes having at least k packets in their buffer at time t, i.e.,

$$Z_k^{(N)}(t) = \sum_{m \geq k} X_m^{(N)}(t),$$

and let $L^{(N)}(t)$ be the total number of packets in the system at time t, i.e.,

$$L^{(N)}(t) = \sum_{k=1}^{\infty} k X_k^{(N)}(t) = \sum_{k=1}^{\infty} Z_k^{(N)}(t).$$

We now introduce some convenient notation that will be used later to characterize the behavior of various quantities as functions of the back-off scaling $f(N)$ in the regime where N grows large. Given two functions $f_1, f_2 : \mathbb{N} \to \mathbb{R}_+$ we say that f_1 is asymptotically bounded from below by f_2, denoted by $f_1(N) \geq_N f_2(N)$, if $f_1(N) = \Omega(f_2(N))$, i.e.,

$$f_1(N) \geq_N f_2(N) \qquad \Longleftrightarrow \qquad \lim_{N\to\infty} \frac{f_1(N)}{f_2(N)} > 0.$$

Similarly, we say that f_1 asymptotically dominates f_2, denoted by $f_1(N) >_N f_2(N)$, if $f_1(N) = \omega(f_2(N))$, i.e.,

$$f_1(N) >_N f_2(N) \qquad \Longleftrightarrow \qquad \lim_{N\to\infty} \frac{f_1(N)}{f_2(N)} = \infty.$$

III. Stability and mean stationary performance

In this section we establish stability conditions and provide closed-form expressions for the mean stationary queue length and waiting time which reveal the impact of the back-off scaling factor $f(N)$. We will leverage results from the polling literature, harnessing a connection between a CSMA network with complete interference and a polling system with a 1-limited service discipline and a uniform random routing policy [13]. In order to explain the connection, suppose that not each individual node has an exponential back-off clock with rate $\nu f(N)$, but that there is a fictitious global exponential back-off clock with rate $\nu N f(N)$. When the global back-off clock ticks, one of the N nodes is selected uniformly at random, and granted the opportunity to transmit one packet, if it has any. The latter operation is statistically identical to that in the CSMA network, and at the same time equivalent to the evolution of a system with N queues and a single server, which selects queues uniformly at random, takes an exponential time with parameter $\nu N f(N)$ to move to the selected queue, and then serves exactly one packet there (if present). This in turn corresponds exactly to a polling system with a 1-limited service discipline, a uniform random routing policy, and exponential switch-over times with parameter $\nu N f(N)$.

The next proposition follows from a direct application of the stability conditions derived in [14], [16] for the latter polling system to the CSMA network.

Proposition 1. *The queue length process $\boldsymbol{Q}^{(N)}(t)$ is positive-recurrent if and only if $S(N) > 0$, where*

$$S(N) := 1 - \rho - \frac{\lambda}{\nu N f(N)}. \tag{1}$$

In particular, observe that if $f(N) <_N \frac{1}{N}$, then the queue length process $\boldsymbol{Q}^{(N)}(t)$ is transient for all N sufficiently large. Conversely, the queue length process is positive-recurrent for all N sufficiently large if

$$\rho < 1 \quad \text{ and } \quad \tilde{\xi} = \limsup_{N\to\infty} \frac{\xi}{f(N)N} < 1, \tag{2}$$

where $\xi := \frac{\lambda}{\nu(1-\rho)}$. Condition (2) entails $f(N) \geq_N \frac{1}{N}$, and we will focus on the case where $f(N) >_N \frac{1}{N}$, so that $\tilde{\xi} = 0$. The case with $\tilde{\xi} > 0$ was analyzed in [10], [11].

We henceforth assume that condition (2) is satisfied, so that the queue length process is positive-recurrent for all sufficiently large values of N. Denote by $\boldsymbol{Q}^{(N)}$ the random vector with the stationary distribution of the queue length process $\boldsymbol{Q}^{(N)}(t)$, tracking the buffer contents of the various nodes. Note that each component $Q_n^{(N)}$ has the same distribution since all the nodes are exchangeable, and let $Q^{(N)}$ be a random variable with that common distribution. Similarly we define the stationary versions of the processes $\boldsymbol{X}^{(N)}$, $\boldsymbol{Z}^{(N)}$, and

$L^{(N)}$, introduced in Section II. We further denote by $W^{(N)}$ the stationary waiting time of an arbitrary packet, i.e., the amount of time spent in the buffer of a node until the start of transmission.

In view of the connection with a polling system with a 1-limited service discipline and a uniform random routing policy described above, we can use the so-called pseudo-conservation law for the latter system in [7, Eqn. (5.31)] to obtain the expected stationary waiting time $\mathbb{E}[W^{(N)}]$ as presented in the next proposition. Because of Little's law, this also immediately yields the expected stationary queue lengths as

$$\mathbb{E}[Q^{(N)}] = \frac{\lambda \mathbb{E}[W^{(N)}]}{N}, \quad \mathbb{E}[L^{(N)}] = \lambda \mathbb{E}[W^{(N)}] = N\mathbb{E}[Q^{(N)}].$$

Proposition 2. *If condition* (1) *holds, then*

$$\mathbb{E}[W^{(N)}] = \frac{1}{S(N)}\Big(\frac{\rho}{\mu} + \frac{1}{\nu f(N)}\Big), \tag{3}$$

$$\mathbb{E}[L^{(N)}] = \frac{1-\rho}{S(N)}\Big(\frac{\rho^2}{1-\rho} + \frac{\xi}{f(N)}\Big). \tag{4}$$

Since $S(N)$ is increasing in $f(N)$ when condition (1) is satisfied, it directly follows from (3) that the expected stationary waiting time and queue length are both decreasing in the back-off scaling factor $f(N)$, as expected. More specifically, we can observe that, for a fixed traffic intensity ρ, the expected stationary waiting time and total backlog in the system scale both roughly inversely proportional with $f(N)$, i.e., they grow approximately linearly with the mean nominal back-off period.

IV. Multi-scale mean-field limit

The analysis in the previous section revealed the impact of the back-off scaling on stationary performance metrics like the expected waiting time and expected total backlog. In order to gain deeper insight in the impact of the back-off scaling, we investigate in this section a multi-scale mean-field limit for the overall buffer occupancy process. In contrast to the conventional mean-field framework, we do not simply normalize the entire process by the number of nodes N as a common factor, but scale the number of nodes with a queue length of k or larger by $N(Nf(N))^{-k}$ to obtain a refined view of the buffer occupancy dynamics on the relevant scales.

A. Mean-field limit path

Let $\tilde{\boldsymbol{Z}}^{(N)}(t)$ be the rescaled cumulative population process,

$$\tilde{Z}_k^{(N)}(t) = \frac{\big(Nf(N)\big)^k}{N} Z_k^{(N)}(t), \qquad k \geq 1. \tag{5}$$

As mentioned above, the scaling factor $(Nf(N))^k/N$ may be interpreted as the inverse of the typical number of nodes with a queue length of k or larger, meaning that $\tilde{Z}_k^{(N)}(t)$ should typically be $O(1)$. Accordingly, define

$$\bar{k} = \sup_{k\in\mathbb{N}} \Big\{ \frac{\big(Nf(N)\big)^k}{N} <_N 1 \Big\} \geq 1 \tag{6}$$

as the largest value of k for which the typical number of nodes with a queue length of k or larger grows large as $N \to \infty$. Let $\boldsymbol{\zeta}^{(N)}(t)$ be the $\bar{k}$-dimensional process where $\zeta_k^{(N)}(t) = \tilde{Z}_k^{(N)}(t)$ for every $k = 1, \ldots, \bar{k}$.

Even though the typical numbers of nodes with a queue length $k = 1, \ldots, \bar{k}$ grow large as $N \to \infty$, we observe that whenever $f(N) >_N 1/N$, the number of nodes with a queue length of $k+1$ or larger is of a smaller order-of-magnitude than the number of nodes with a queue length exactly k. Hence, the variable $\tilde{Z}_k^{(N)}(t)$ may equivalently be thought of as the rescaled number of nodes with a queue length exactly k. In particular, taking $k = 0$, the number of nodes with a non-zero queue length is of a smaller order-of-magnitude than the number of nodes with a zero queue length. In other words, the overwhelming majority of nodes have empty buffers.

The next proposition states that, for suitable initial conditions, as $N \to \infty$, the rescaled and accelerated cumulative population process $\boldsymbol{\zeta}^{(N)}(t/f(N))$ converges to a deterministic limit $\boldsymbol{\zeta}(t)$ described in terms of a set of differential equations. This is referred to as a multi-scale mean-field limit, since the components of the original population process $\boldsymbol{Z}^{(N)}(t)$ are of different orders-of-magnitude and have been scaled differently.

Proposition 3. Multi-Scale Mean-Field Limit *Assume that* $\frac{1}{N} <_N f(N) <_N 1$ *and that* $\boldsymbol{\zeta}^{(N)}(0) \Rightarrow \boldsymbol{\zeta}(0)$ *as* $N \to \infty$. *Then*

$$\boldsymbol{\zeta}^{(N)}\Big(\frac{t}{f(N)}\Big) \Rightarrow \boldsymbol{\zeta}(t) \qquad \text{as } N \to \infty, \tag{7}$$

where $\boldsymbol{\zeta}(t)$ *satisfies the system of differential equations*

$$\frac{\partial \boldsymbol{\zeta}(t)}{\partial t} = \boldsymbol{H}(\boldsymbol{\zeta}(t)), \tag{8}$$

with

$$H_1(\boldsymbol{\zeta}(t)) = \lambda - \nu \pi^0_{\boldsymbol{\zeta}(t)} \zeta_1(t), \tag{8.a}$$

$$H_k(\boldsymbol{\zeta}(t)) = \lambda \zeta_{k-1}(t) - \nu \pi^0_{\boldsymbol{\zeta}(t)} \zeta_k(t), \tag{8.b}$$

for $k = 2, \ldots, \bar{k}$ *and* $\pi^0_{\boldsymbol{\zeta}(t)} = \mu/(\mu + \nu\zeta_1(t))$.

B. Discussion of mean-field limit behavior

The differential equations that arise in the above multi-scale mean-field limit may be interpreted as follows. The function $H_k(\boldsymbol{\zeta}(t))$ represents the change in the rescaled number of nodes with a queue length of k or larger in the state $\boldsymbol{\zeta}(t)$, which is increasing at rate $\lambda\zeta_{k-1}(t)$ due to packet arrivals at nodes with queue length exactly $k-1$ and decreasing at rate $\nu\pi^0_{\boldsymbol{\zeta}(t)}\zeta_k(t)$ due to successful back-off completions at nodes with queue length exactly k. The term $\pi^0_{\boldsymbol{\zeta}(t)}$ captures the fraction of time that successful back-off completions can occur in state $\boldsymbol{\zeta}(t)$, which corresponds to the fraction of time that no node is active. The function $H_1(\boldsymbol{\zeta}(t))$ similarly represents the change in the rescaled number of nodes with a non-zero queue length in the state $\boldsymbol{\zeta}(t)$, but the term corresponding to arrivals at nodes with zero queue length takes a slightly simpler form since by definition $\tilde{Z}_0 = 1$.

The proof of Proposition 3 is lengthy and technically involved. For the details we refer to [8, Sect. 5.3]. The key argument involves a stochastic averaging principle similar to [11], [19], capturing a separation of time scales between the

'fast' activity process $Y^{(N)}(t)$ and the 'slow' rescaled cumulative population process $\tilde{\boldsymbol{Z}}^{(N)}(t)$. To illustrate the separation of time scales, it is useful to consider the drifts of these two processes. When the state is $(\tilde{\boldsymbol{z}}, y) = \big(\tilde{\boldsymbol{Z}}^{(N)}(t), Y^{(N)}(t)\big)$, the k-th component of $\tilde{\boldsymbol{Z}}^{(N)}(t)$ experiences a drift

$$\begin{aligned} F_k(\tilde{\boldsymbol{z}}, y) =& f(N)\lambda\Big(\tilde{z}_{k-1} - \frac{\tilde{z}_k}{Nf(N)}\Big) \\ & - f(N)\nu\mathbb{1}\{y=0\}\Big(z_k - \frac{\tilde{z}_{k+1}}{Nf(N)}\Big), \end{aligned} \tag{9}$$

while the activity process is subject to a drift

$$G(\tilde{\boldsymbol{z}}, y) = \nu\mathbb{1}\{y=0\}\tilde{z}_1 - \mu\mathbb{1}\{y=1\}. \tag{10}$$

Equations (9) and (10) reflect that the rescaled population process $\tilde{\boldsymbol{Z}}^{(N)}(t)$ changes at a 'slow' rate $f(N)$ as $N \to \infty$, while the activity process $Y^{(N)}(t)$ evolves at a 'fast' O(1) rate. (Equation (9) further reflects that although all the components $Z_k^{(N)}(t)$ are of different orders-of-magnitude, the rescaled versions $\tilde{Z}_k^{(N)}(t)$, $k = 1, \ldots, \bar{k}$, all evolve on the common time scale $1/f(N)$). Hence, by rescaling time by $1/f(N)$ and letting N grow large, the activity process instantaneously converges to its stationary distribution, which only depends on the current state of the rescaled population process. The latter stationary distribution is given by $\pi_{\tilde{\boldsymbol{z}}} = (\pi^0_{\tilde{\boldsymbol{z}}}, \pi^1_{\tilde{\boldsymbol{z}}})$ such that

$$\nu\pi^0_{\tilde{\boldsymbol{z}}}\tilde{z}_1 - \mu\pi^1_{\tilde{\boldsymbol{z}}} = 0, \qquad \pi^0_{\tilde{\boldsymbol{z}}} + \pi^1_{\tilde{\boldsymbol{z}}} = 1,$$

yielding $\pi^0_{\tilde{\boldsymbol{z}}} = \mu/(\mu + \nu\tilde{z}_1)$.

This then implies that the sum of the first and third term in Equation (9) converges to the function $H_k(t)$ as $N \to \infty$. It can further be shown that the other two terms scaled by $Nf(N)$ vanish as $N \to \infty$, yielding Proposition 3.

C. Fixed point

We now turn attention to the fixed point of the mean-field limit as stated in Proposition 3. The fixed point will be used to determine the buffer content and delay distributions at individual nodes and demonstrate the concentration property of the aggregate back-off rate.

The next proposition identifies the fixed point of the system of differential equations (7), and establishes that it is globally stable. The proof is provided in Appendix A.

Proposition 4. Fixed Point of Mean-Field Limit *For any initial state $\boldsymbol{\zeta}(0)$, there exists a unique solution $\boldsymbol{\zeta}(t)$ of the system of differential equations* (7)*, with*

$$\lim_{t\to\infty} \boldsymbol{\zeta}(t) = \boldsymbol{\zeta}^*, \qquad \zeta^*_k = \xi^k, \qquad \forall\, k = 1, \ldots, \bar{k},$$

with $\xi = \frac{\lambda}{\nu(1-\rho)}$ as defined earlier.

• *Stationary performance at individual nodes.* The fixed point can be leveraged to obtain stationary performance measures at individual nodes. Specifically, if we assume the large-scale ($N \to \infty$) and stationary ($t \to \infty$) limits to commute, then we obtain an asymptotic approximation for the buffer content distribution at an individual node

$$\mathbb{P}\{Q^{(N)} \geq k\} = \frac{\mathbb{E}[Z_k^{(N)}]}{N} \sim \frac{\zeta^*_k}{(Nf(N))^k} = \Big(\frac{\xi}{Nf(N)}\Big)^k, \tag{11}$$

for all $k = 1, \ldots, \bar{k}$. Applying a similar observation as used in the the distributional form of Little's law, this also yields an asymptotic approximation for the waiting-time distribution

$$\mathbb{P}\{W^{(N)} \geq t\} \sim \mathrm{e}^{-\left(\nu(1-\rho)f(N) - \frac{\lambda}{N}\right)t}, \tag{12}$$

and thus $\mathbb{P}\{f(N)W^{(N)} \geq t\} \to \mathrm{e}^{-\nu(1-\rho)t}$ as $N \to \infty$. This reflects that asymptotically the queue length and waiting time at each individual node behave as the *total* number of jobs and the *sojourn* time in an M/M/1 system with arrival rate λ/N and service rate $\nu(1-\rho)f(N)$, and thus evolve as in an M/M/1 system with arrival rate $\lambda/(Nf(N))$ and service rate $\nu(1-\rho)$ when viewed on a time scale $1/f(N)$.

• *Implications for aggregate back-off rate.* The mean-field limit in Proposition 3 implies that the rescaled number $f(N)Z_1^N(t)$ of backlogged nodes at time t converges to $\zeta_1(t)$ as $N \to \infty$. In particular, the rescaled stationary number of backlogged nodes $f(N)Z_1^{(N)}$ settles around its mean value $\zeta_1 = \xi = \frac{\lambda}{\nu(1-\rho)}$ as $N \to \infty$. It follows that the stationary aggregate back-off rate $V^{(N)} = \nu f(N) Z_1^{(N)}$ concentrates around its mean value $\nu\zeta_1 = \frac{\lambda}{1-\rho}$. Thus, the mean stationary aggregate back-off rate multiplied with the stationary fraction of time $1-\rho$ that no node is active, i.e., $Y^{(N)} = 0$, equals λ, which makes sense as summing the balance equations for the population process $\boldsymbol{X}^{(N)}(t)$ in fact yields the identity relation $\mathbb{E}\left[V^{(N)}\mathbb{1}\left\{Y^{(N)} = 0\right\}\right] = \lambda$.

It is interesting to observe that the mean stationary aggregate back-off rate does not depend on the scaling factor $f(N)$. This reflects a self-regulating property, where the number of backlogged nodes that is actively competing for the medium is roughly inversely proportional to the nominal back-off rate. For example, a more aggressive back-off scaling indirectly results in a proportional reduction of the number of backlogged nodes. Thus, the improvement of the packet delay performance observed in Section III does not come at the expense of an increase in the aggregate back-off rate and excessive activations or collisions.

The mean-field limit also suggests that each individual node is backlogged with probability $\frac{\xi}{Nf(N)}$ as $N \to \infty$. Assuming independence among nodes, this would imply that $Z_1^{(N)}$ is roughly distributed as a binomial random variable $\mathrm{Bin}(m, p)$ where $m = N$ is the number of trials and $p = \frac{\xi}{Nf(N)}$ is the probability of success. This would imply that $Z_1^{(N)}$, when properly centered and normalized, is approximately distributed as a standard normal random variable

$$\sqrt{\frac{f(N)}{\xi}}\left(Z_1^{(N)} - \frac{\xi}{f(N)}\right) \sim \mathcal{N}(0,1).$$

However, independence only holds among finite subsets of nodes in the mean-field limit, and not among all nodes. In the next section we will establish nevertheless a central limit result for $L^{(N)}$ which allows us to show that the above normal approximation provides a conservative estimate for $Z_1^{(N)}$ and that the stationary aggregate back-off rate is bounded with high probability.

V. Aggregate back-off rate

The mean-field limit established in the previous section provides detailed insight in the dynamics of the buffer occupancy process and revealed a remarkable self-regulating property. Specifically, the rescaled stationary number of backlogged nodes $f(N)Z_1^{(N)}$ that actively compete for the medium, settles around its mean value ξ, and hence the aggregate back-off rate $V^{(N)} = \nu f(N) Z_1^{(N)}$ concentrates around $\nu\xi = \lambda/(1-\rho)$. In this section, we seek to obtain further insights in the fluctuation of $V^{(N)}$ around $\lambda/(1-\rho)$. Specifically, we show that the variation in the rescaled total system backlog around ξ is governed by a central limit behavior which yields concentration bounds for the aggregate back-off rate.

A. Central-limit behavior of the total system backlog

The next theorem shows that the total system backlog satisfies a central limit law.

Theorem 1. *Assume $\rho < 1$ and $N^{-\frac{2}{3}} <_N f(N) <_N 1$. Then*

$$\frac{f(N)L^{(N)} - \xi}{\sqrt{f(N)\sigma}} \Rightarrow \mathcal{N}(0,1), \qquad \sigma := \big(1 + \frac{\rho^2}{1-\rho}\big)\xi. \quad (13)$$

To prove Theorem 1, we first introduce a variation of the CSMA network, and establish a central limit theorem for this auxiliary model. The core of the proof consists in showing that the CSMA network and its variation are strongly related and that the CSMA network therefore obeys the same central limit law when $N^{-2/3} <_N f(N) <_N 1$. Specifically, it can be shown that the total system backlog in the CSMA network is stochastically bounded from above by that in the auxiliary model. This, together with proving that the stationary total system backlogs of the two models are relatively close in expectation, concludes the proof.

• *A variation of the CSMA network.* In the CSMA network under investigation each backlogged node attempts back-offs at rate $\nu f(N)$, independent of its exact buffer content. Now imagine a variation where a node with buffer content k attempts back-offs at rate $k\nu f(N)$. Intuitively, in the variation of the CSMA network an exponential back-off clock is associated with every packet, while in the original CSMA network only with those packets at the head of their queue. This corresponds to a 1-limited vacation queueing model analyzed in [6], where vacations are interrupted at the instants of a time-inhomogeneous Poisson process with a rate that is $\nu f(N)$ times the number of packets in the queue.

Let $L^{A,(N)}$ be the stationary number of waiting packets and $\tilde{L}^{A,(N)}$ the stationary total number of packets (including the one in service) in the auxiliary model. In [6, Cor. 3.5] the probability generating function of $\tilde{L}^{A,(N)}$ is shown to be

$$G_{\tilde{L}^{A,(N)}}(r) = \Big(\frac{1-\rho}{1-\rho r}\Big)^{1+\frac{\lambda}{\nu f(N)}} \Big(e^{(r-1)\frac{\lambda}{\nu f(N)}}\Big). \quad (14)$$

The next theorem shows the central limit theorem for $L^{A,(N)}$. The proof is presented in Appendix B.

Theorem 2. *Assume $\rho < 1$ and $f(N) <_N 1$. Then*

$$\frac{f(N)L^{A,(N)} - \xi}{\sqrt{f(N)\sigma}} \Rightarrow \mathcal{N}(0,1), \quad (15)$$

where σ is as defined in (13).

• *Proof of Theorem 1.* In order to deduce (13) from (15), it remains to be shown that the original CSMA network and its variation are strongly similar in a certain sense, and that the central limit law for the auxiliary model hence carries over to the original CSMA network.

First of all, we observe that $L^{A,(N)}$ is stochastically bounded from above by $L^{(N)}$. This result is plausible since given the same total buffer content in the two models, the aggregate back-off rate in the auxiliary model is larger than that in the original network. The proof of this fact involves stochastic coupling arguments, and is omitted because of page constraints. At the same time, $L^{A,(N)}$ and $L^{(N)}$ are relatively close in expectation. Indeed, noting that

$$\mathbb{E}[\tilde{L}^{A,(N)}] = G'_{\tilde{L}^{A,(N)}}(1), \qquad \mathbb{E}[L^{A,(N)}] = \mathbb{E}[\tilde{L}^{A,(N)}] - \rho,$$

we obtain from (4) and (14), that

$$\mathbb{E}\big[L^{(N)} - L^{A,(N)}\big] = \frac{\xi}{Nf(N) - \xi}\Big(\frac{\rho^2}{1-\rho} + \frac{\xi}{f(N)}\Big), \quad (16)$$

which scales as $\frac{1}{Nf(N)^2}$ when $N \to \infty$ and $f(N) \leq_N 1$.

We are now ready to present the following proposition, which allows us to prove the central limit law for the original CSMA network, exploiting Theorem 2 and the closeness with the auxiliary model. The proof of the proposition is omitted here due to page limitations and the details are presented in [8, Prop. 5.3].

Proposition 5. *Consider two sequences of random variables $\{X_n^U\}_{n\geq 1}$, $\{X_n^L\}_{n\geq 1}$, a sequence $\{(c_n, d_n)\}_{n\geq 1} \subseteq \mathbb{R}_+ \times \mathbb{R}$, and a random variable $X > 0$ such that $X_n^L \leq_s X_n^U$ for every $n \geq 0$, and*

$$c_n X_n^L + d_n \Rightarrow X, \qquad \lim_{n\to\infty} c_n\Big(\mathbb{E}\big[X_n^U\big] - \mathbb{E}\big[X_n^L\big]\Big) = 0.$$

Then, it holds that $c_n X_n^U + d_n \Rightarrow X$.

To prove Theorem 1, we apply Proposition 5 to the sequences given by $X_N^U = L^{(N)}$, $X_N^L = L^{A,(N)}$, and

$$(c_N, d_N) = \Big(\sqrt{\frac{f(N)}{\sigma}}, \frac{-\xi}{\sqrt{f(N)\sigma}}\Big).$$

Observe that Theorem 2, for $f(N) <_N 1$, yields

$$c_N L^{A,(N)} + d_N \Rightarrow \mathcal{N}(0,1).$$

In order to apply Proposition 5, we only need to show that

$$\lim_{N\to\infty} \sqrt{\frac{f(N)}{\sigma}}\mathbb{E}\big[L^{(N)} - L^{A,(N)}\big] = 0,$$

which is the case due to (16) if and only if $f(N) >_N N^{-\frac{2}{3}}$.

• *Discussion of Theorem 1.* As discussed above, the proof of Theorem 1 is based on the connection between the CSMA network and the auxiliary model. However, while the central limit law for the auxiliary model also holds for $f(N) \leq_N N^{-\frac{2}{3}}$, the same does not for the CSMA network. The difference between the two scenarios is due to the number of nodes $Z_2^{(N)}$ with two or more packets in the buffer. In Section IV, we proved that, since $\bar{k} > 2$ for $f(N) <_N N^{-\frac{2}{3}}$, we have $Z_2^{(N)} \Rightarrow \xi^2/(Nf(N)^2)$, which does not vanish when the central limit scaling is applied.

B. Concentration bounds for the aggregate back-off rate

We now aim to understand the implications of Theorem 1 for the stationary aggregate back-off rate. We proved that, assuming $\rho < 1$ and $N^{-\frac{2}{3}} <_N f(N) <_N 1$,

$$\lim_{N\to\infty} \mathbb{P}\{f(N)L^{(N)} > \tau(N)\} = \mathbb{P}\{N_1 > \bar{\tau}\}, \qquad (17)$$

where

$$N_1 \sim \mathcal{N}(0,1), \qquad \bar{\tau} = \lim_{N\to\infty} \frac{\tau(N) - \xi}{\sqrt{f(N)\sigma}}.$$

Since $Z_1^{(N)} \leq L^{(N)}$ by definition, we obtain by taking $\tau(N) = \xi + \frac{\kappa\sqrt{f(N)}}{\nu}$ that

$$\lim_{N\to\infty} \mathbb{P}\{V^{(N)} > \frac{\lambda}{1-\rho} + \kappa\sqrt{f(N)}\} \leq 1 - \Phi\big(\frac{\kappa}{\sqrt{\sigma}}\big), \quad (18)$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution and σ is as defined in (13). In particular, the aggregate back-off rate $V^{(N)}$ is asymptotically bounded from above by $\nu\xi + \kappa\sqrt{f(N)} = \frac{\lambda}{1-\rho} + \kappa\sqrt{f(N)}$ with high probability for sufficiently large κ values.

C. Optimal back-off rates

Consider $f(N) = -\log(N)$, i.e., $N^{-\epsilon} <_N f(N) <_N 1$ for every $\epsilon > 0$. Let us examine the performance in terms of the stationary waiting time and aggregate back-off rate. From (11) and (12) we deduce that

$$Q^{(N)} \sim \text{Geom}\Big(\frac{\xi\log(N)}{N}\Big), \quad W^{(N)} \sim \text{Exp}\Big(\frac{\nu(1-\rho)}{\log(N)} - \frac{\lambda}{N}\Big),$$

and from (13) we obtain

$$\frac{L^{(N)} - \xi\log(N)}{\sqrt{\log(N)\sigma}} \Rightarrow \mathcal{N}(0,1).$$

From (17), we obtain

$$\lim_{N\to\infty} \mathbb{P}\Big\{V^{(N)} > \frac{\lambda}{1-\rho} + \kappa\frac{1}{\sqrt{\log(N)}}\Big\} \leq 1 - \Phi\big(\frac{\kappa}{\sqrt{\sigma}}\big),$$

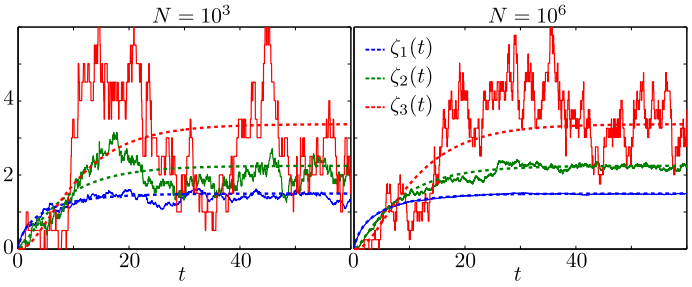

Fig. 1. Example 1 - Sample path of $\zeta_k^{(N)}(t)$ for $k = 1$ (blue), for $k = 2$ (green), and for $k = 3$ (red). From left to right, we consider $N = 10^3, 10^6$.

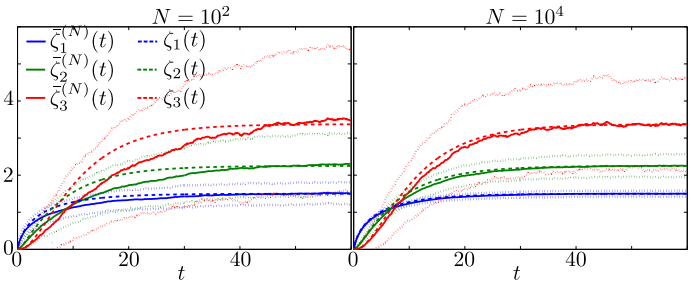

Fig. 2. Example 1 - Empirical average $\bar{\zeta}_k^{(N)}(t)$ for $k = 1$ (blue), for $k = 2$ (green), and for $k = 3$ (red). The empirical variance is displayed with a dotted line. From left to right, we consider $N = 10^2, 10^4$.

and in particular

$$\mathbb{P}\Big\{V^{(N)} > \frac{\lambda}{1-\rho} + \kappa\frac{1}{\sqrt[2+\epsilon]{\log(N)}}\Big\} \to 0$$

for every $\epsilon, \kappa > 0$ as $N \to \infty$. This demonstrates that as the network dimension grows large, on the one hand the waiting time increases only logarithmically in the dimension, on the other hand the system self-regulates so that the large majority of the nodes are empty most of the time and therefore storms of back-off completions are rare. As is illustrated by this example, a mean waiting time of the order $1/f(N)$ corresponds to a variation in the aggregate back-off rate of the order $\sqrt{f(N)}$. As a general guideline, this advocates a scaling factor that decays slowly to optimize the waiting-time performance, unless the variation in the aggregate back-off rate is a major concern.

VI. Numerical and simulation experiments

A. Accuracy of the multi-scale mean-field approximation

We consider Example 1 with $\lambda = 0.75$, $\nu = 2$, $\mu = 1$, and $f(N) = N^{-7/10}$, so that $\xi = 1.5$, $\rho = 0.75$, and $\bar{k} = 3$. In Figure 1 we show the sample path of $\boldsymbol{\zeta}^{(N)}(t)$ obtained from simulations with $N = 10^3, 10^6$. Observe that as N grows large, the sample path of $\zeta_k^{(N)}(t)$ gets closer to the numerical solution of (7) which is displayed as a dashed line. The rate of convergence depends on how fast $f(N)^k N^{k-1}$, the magnitude of the process, tends to zero. Note that for $k = 3$ we have $f(N)^k N^{k-1} = N^{1/10}$, hence the convergence is slow and the process is noisy even for $N = 10^6$.

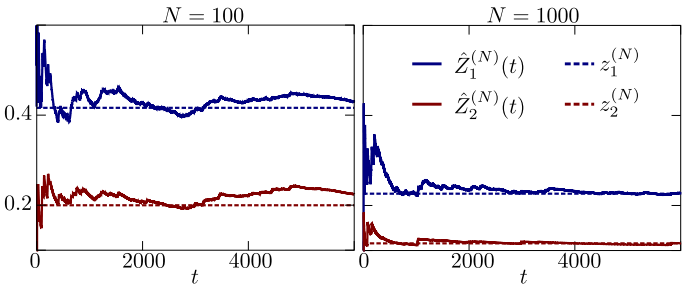

Fig. 3. Example 2 - Evolution of the buffer occupancy at node $n = 1$. From left to right, we consider $N = 100$ and $N = 1000$.

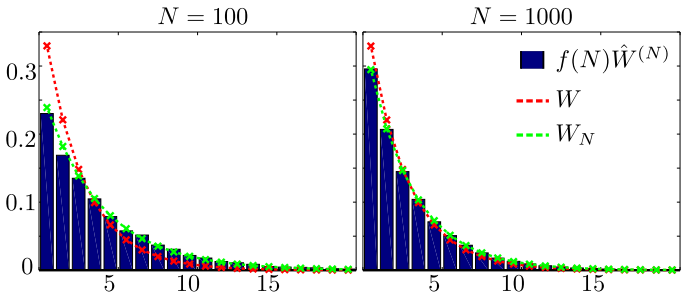

Fig. 4. Example 2 - Waiting-time distribution. From left to right, we consider $N = 100$ and $N = 1000$.

Although sample path convergence may be slow for large k, the mean-field limit captures the average behavior very accurately, even for low values of N. We simulated 1000 independent instances for $N = 10^2, 10^4$, and considered

$$\bar{\zeta}_k^{(N)}(t) = \frac{1}{1000} \sum_{j=1}^{1000} \zeta_k^{(N),j}(t), \qquad k = 1, 2, 3,$$

where $\zeta_k^{(N),j}(t)$ describes the cumulative population process observed in the j-th simulation with N nodes. In Figure 2 we present the comparison between $\bar{\zeta}_k^{(N)}(t)$ and $\zeta_k^{(N)}(t)$. Observe that already for $N = 10^4$ these are almost indistinguishable.

B. Performance of individual nodes

Consider Example 2 with $\lambda = 0.8$, $\nu = 2$, $\mu = 1$, and $f(N) = N^{-3/5}$, so that $\xi = 2$, $\rho = 0.8$, and $\bar{k} = 2$. In Section IV-C, we discussed how to leverage the mean-field limit to approximate the stationary queue length distribution at an individual node. In particular, equation (11), in our example, yields

$$\mathbb{P}\{Q^{(N)} \geq 1\} \approx 2N^{-2/5} =: z_1^{(N)},$$
$$\mathbb{P}\{Q^{(N)} \geq 2\} \approx 4N^{-4/5} =: z_2^{(N)}.$$

In Figure 3 we display the evolution of the queue length at a specific node (node $n = 1$) both in case $N = 100$ and $N = 1000$, and observe the striking accuracy of the proposed approximation. In particular, we computed

$$\hat{Z}_k^{(N)}(t) = \frac{1}{t} \int_0^t \mathbb{1}\Big\{Q_1^{(N)}(t) \geq k\Big\},$$

N \ κ	Example 3			Example 4		
	$\frac{\nu\sqrt{\sigma}}{4}$	$\nu\sqrt{\sigma}$	$\nu\sigma$	$\frac{\nu\sqrt{\sigma}}{4}$	$\nu\sqrt{\sigma}$	$\nu\sigma$
10^2	21.12%	0.15%	–	36.77%	2.46%	0.02%
10^3	10.43%	–	–	15.83%	0.02%	–
10^4	0.61%	–	–	2.88%	–	–

TABLE I
FRACTION OF TIME THAT $V^{(N)}(t)$ SPENDS ABOVE THE THRESHOLD $\frac{\lambda}{1-\rho} + \kappa$. WHEN THE FRACTION IS BELOW 0.01% WE INDICATE IT WITH –.

and show that $\lim_{t\to\infty} \hat{Z}_k^{(N)}(t)$ is well approximated by $z_k^{(N)}$ already for $N = 100$.

In Section IV-C we also claimed that the properly scaled stationary waiting-time distribution at an individual node is accurately approximated by W_N, an exponential random variable with parameter $\nu(1-\rho) - \lambda/Nf(N)$, and eventually converges to $W \sim \text{Exp}\big(\nu(1-\rho)\big)$. We simulated the scenario in Example 2 both for $N = 100$ and $N = 1000$ until $\sum_{n=1}^{N} T_n^{(N)}(t) = 200N$, where $T_n^{(N)}(t)$ is the cumulative number of back-off completions of node n by time t. We kept track of $\hat{W}_j^{(N)}$, the time spent in the buffer by the j-th packet starting a transmission for $j = 1, \ldots, 200N$. In Figure 4 we compare the empirical distribution of $f(N)\hat{\boldsymbol{W}}^{(N)}$ with that of W_N and W. As expected, the approximation W_N is extremely accurate for $N = 1000$, and remarkably sharp already for $N = 100$. As N grows large, W_N clearly converges to W as well as the empirical distribution.

C. Bounded aggregate back-off rate

In Section V we observed that the total system backlog scales as $1/f(N)$ as $N \to \infty$, and then used (18) to bound the probability that the aggregate back-off rate $V^{(N)}$ exceeds a certain threshold. We now aim to understand how frequently $V^{(N)}$ exceeds $\frac{\lambda}{1-\rho} + \kappa$ for different values of N and κ. We empirically show that $V^{(N)}$ rarely grows large as N gets larger, thus storms of back-off completions are rare.

Let us consider the following examples. In Example 3 we set $\lambda = 0.6$, $\nu = 1$, $\mu = 1$, and $f(N) = N^{-1/2}$, so that $\xi = 1.5$ and $\sigma = 2.85$. In Example 4 we set $\lambda = 0.8$, $\nu = 8$, $\mu = 1$, and $f(N) = N^{-1/2}$, so that $\xi = 0.5$ and $\sigma = 2.1$.

For both examples, we let simulations run for different choices of N and κ from an initially empty configuration, and report the results in Table I. Observe that already for small values of N, the fraction of time that the aggregate back-off rate exceeds $\frac{\lambda}{1-\rho} + \nu\sigma$ is negligible.

VII. CONCLUSION

We investigated how back-off rates impact the performance of ultra-dense random-access networks with intermittent traffic sources. We presented closed-form stability conditions and expressions for the expected stationary queue length and waiting time as a function of the back-off rate. We performed a multi-scale mean-field analysis and observed that the number of backlogged nodes in stationarity is inversely proportional

to the nominal back-off rate as N grows large. This self-regulating property hints that the aggregate back-off rate remains bounded with high probability thus precluding storms of back-off completions. This latter property is formally established by proving a central limit theorem for the stationary total number of backlogged packets.

To the best of our knowledge, this is the first work to examine the performance impact of the back-off rates in ultra-dense random-access networks. We have only considered full interference conditions, and aim to extend the results to more complicated interference scenarios in future work.

REFERENCES

[1] Cisco (2011). Cisco visual networking index: Global mobile data traffic forecast update.

[2] Ericsson (2011). More than 50 billion connected devices. White paper.

[3] G. Bianchi (2000). Performance analysis of the IEEE 802.11 distributed coordination function. *IEEE J. Sel. Areas Commun.* **18 (3)**, 535–547.

[4] R.R. Boorstyn, A. Kershenbaum, B. Maglaris, V. Sahin (1987). Throughput analysis in multihop CSMA packet radio networks. *IEEE Trans. Commun.* **35**, 267–274.

[5] C. Bordenave, D. McDonald, A. Proutiere (2008). Performance of random medium access control, an asymptotic approach. In: *Proc. ACM SIGMETRICS* **36 (1)**, 1–12.

[6] N. Bouman, S.C. Borst, O.J. Boxma, J.S.H. van Leeuwaarden (2014). Queues with random back-offs. *Queueing Syst.* **77 (1)**, 33–74.

[7] O.J. Boxma, J.A. Weststrate (1989). Waiting times in polling systems with Markovian server routing. In: *Proc. MMB Rechensys. Net.* 89–105.

[8] F. Cecchi (2018). Mean-field limits for ultra-dense random-access networks. PhD Thesis. Eindhoven University of Technology.

[9] F. Cecchi, S.C. Borst, J.S.H. van Leeuwaarden (2014). Throughput of CSMA networks with buffer dynamics. *Perf. Eval.* **79**, 216–234.

[10] F. Cecchi, S.C. Borst, J.S.H. van Leeuwaarden, P.A. Whiting (2016). CSMA networks in a many-sources regime: A mean-field approach. In: *Proc. IEEE Infocom*.

[11] F. Cecchi, S.C. Borst, J.S.H. van Leeuwaarden, P.A. Whiting (2016). Mean-field limits for large-scale random-access networks. *arXiv:1611.09723*.

[12] J. Cho, J.Y. Le Boudec, Y. Jiang (2012). On the asymptotic validity of the decoupling assumption for analyzing 802.11 MAC protocol. *IEEE Trans. Inf. Theory* **58 (11)**, 6879–6893.

[13] J.L. Dorsman, S.C. Borst, O.J. Boxma, M. Vlasiou (2015). Markovian polling systems with an application to wireless random-access networks. *Perf. Eval.* **85**, 33–51.

[14] D. Down (1998). On the stability of polling models with multiple servers. *J. Appl. Prob.* **34 (4)**, 925–935.

[15] K.R. Duffy (2010). Mean field Markov models of wireless local area networks. *Markov Proc. Rel. Fields* **16 (2)**, 295–328.

[16] C. Fricker, M.R. Jaibi (1998). Stability of multi-server polling models. *INRIA Res. Rep., No. 3347*.

[17] J. Ghaderi, R. Srikant (2010). On the design of efficient CSMA algorithms for wireless networks. In: *Proc. CDC*.

[18] B. Hajek, T. van Loon (1982). Decentralized dynamic control of a multiaccess broadcast channel. *IEEE Trans. Aut. Contr.* **27**, 559–569.

[19] P.J. Hunt, T.G. Kurtz (1994). Large loss networks. *Stoch. Proc. Appl.* **53 (2)**, 363–378.

[20] L. Jiang, D. Shah, J. Shin, J. Walrand (2010). Distributed random access algorithm: scheduling and congestion control. *IEEE Trans. Inf. Theory* **56 (12)**, 6182–6207.

[21] L. Jiang, J. Walrand (2008). A distributed CSMA algorithm for throughput and utility maximization in wireless networks. In: *Proc. Allerton*.

[22] F.P. Kelly (1985). Stochastic models of computer communication systems. *J. Roy. Stat. Soc. B* **47 (3)**, 379–395.

[23] R. Laufer, L. Kleinrock (2016). The capacity of wireless CSMA/CA networks. *IEEE/ACM Trans. Netw.* **24 (3)**, 1518–1532.

[24] S.C. Liew, C.H. Kai, H.C. Leung, P. Wong (2010). Back-of-the-envelope computation of throughput distributions in CSMA wireless networks *IEEE Trans. Mob. Comput.* **9 (9)**, 1319–1331.

[25] P. Marbach, A. Eryilmaz (2008). A backlog-based CSMA mechanism to achieve fairness and throughput-optimality in multihop wireless networks. In: *Proc. Allerton*.

[26] M. Michalopoulou, P. Mähönen (2017). A mean field analysis of CSMA/CA throughput. *IEEE Trans. Mob. Comput.* **16 (8)**, 2093–2104.

[27] S. Rajagopalan, D. Shah, J. Shin (2009). Network adiabatic theorem: an efficient randomized protocol for content resolution. In: *Proc. ACM SIGMETRICS/Perf.*

[28] R.L. Rivest (1987). Network control by Bayesian broadcast. *IEEE Trans. Inf. Theory* **33 (3)**, 323–328.

[29] D. Shah, J. Shin, P. Tetali (2010). Medium access using queues. In: *Proc. FOCS*.

[30] J.N. Tsitsiklis (1987). Analysis of a multiaccess control scheme. *IEEE Trans. Aut. Contr.* **32 (11)**, 1017–1020.

[31] X. Wang, K. Kar (2005). Throughput modeling and fairness issues in CSMA/CA based ad-hoc networks. In: *Proc. IEEE Infocom*.

[32] H. Wu, C. Zhu, R.J. La, X. Liu, Y. Zhang (2013). FASA: Accelerated S-ALOHA using access history for event-driven M2M communications. *IEEE/ACM Trans. Netw.* **21 (6)**, 1904–1917.

APPENDIX

A. Proof of Proposition 4

Existence and uniqueness of a solution $\boldsymbol{\zeta}(t)$ easily follows by observing that $\boldsymbol{H}(\cdot)$ in (8) is a Lipschitz continuous function. So as to show that $\boldsymbol{\zeta}^*$ is a globally stable fixed point we use induction on k. For $\zeta_1(t)$, it follows from (8a) that

$$\frac{\partial \zeta_1(t)}{\partial t} = \lambda - \frac{\nu\mu\zeta_1(t)}{\mu + \nu\zeta_1(t)}.$$

Note that if $\zeta_1(0) \geq 0$, then $\zeta_1(t) \geq 0$ for all $t \geq 0$. Hence, given that $\zeta_1(t) \geq 0$, it holds that $\frac{\partial \zeta_1(t)}{\partial t} > 0$ if and only if $\zeta_1(t) < \zeta_1^*$, where $\zeta_1^* = \xi$. Let us now consider $k \leq \bar{k}$ and assume the inductive hypothesis, i.e., $\zeta_j(t) \to \zeta_j^* = \xi^j$ for every $j < k$. It follows from (8b) that

$$\frac{\partial \zeta_k(t)}{\partial t} = \lambda\zeta_{k-1}(t) - \frac{\nu\mu\zeta_k(t)}{\mu + \nu\zeta_1(t)},$$

with $\zeta_1(t) \to \xi$ and $\zeta_{k-1}(t) \to \xi^{k-1}$. Hence, $\frac{\partial \zeta_k(t)}{\partial t} \to \lambda\xi^{k-1} - \nu(1-\rho)\zeta_k(t)$, which is positive if and only if $\zeta_k(t) < \xi^k$. Thus we conclude that $\zeta_k(t) \to \xi^k$.

B. Proof of Theorem 2

It follows from (14) that $\tilde{L}^{A,(N)} \sim P^{(N)} + B^{(N)}$, where

$$P^{(N)} \sim \text{Poiss}\Big(\frac{\lambda}{\nu f(N)}\Big), \quad B^{(N)} \sim \text{NB}\Big(1 + \frac{\lambda}{\nu f(N)}, 1-\rho\Big),$$

and $\text{NB}(n,p)$ is a Negative Binomial random variable counting the number of failures before the n-th success in a sequence of Bernoulli trials with success probability p. As N grows large, $P^{(N)}$ and $B^{(N)}$, suitably centered and scaled, converge to normal random variables:

$$\hat{P}^{(N)} \Rightarrow \mathcal{N}_P\big(0, \frac{\lambda}{\nu}\big), \qquad \hat{B}^{(N)} \Rightarrow \mathcal{N}_B\big(0, \frac{\rho}{(1-\rho)^2}\frac{\lambda}{\nu}\big),$$

where

$$\hat{P}^{(N)} := \sqrt{f(N)}\Big(P^{(N)} - \frac{\lambda}{\nu f(N)}\Big),$$

$$\hat{B}^{(N)} := \sqrt{f(N)}\Big(B^{(N)} - \frac{\rho}{1-\rho}\big(\frac{\lambda}{\nu f(N)} + 1\big)\Big).$$

The proof is completed by observing that $\tilde{L}^{A,(N)} - 1 \leq L^{A,(N)} \leq \tilde{L}^{A,(N)}$, $\frac{\lambda}{\nu} + \frac{\rho}{1-\rho}\frac{\lambda}{\nu} = \xi$, and $\frac{\lambda}{\nu} + \frac{\rho}{(1-\rho)^2}\frac{\lambda}{\nu} = \sigma$.