Queue Aware Access Prioritization for Massive Machine-type Communication

Mayukh Roy Chowdhury and Swades De

Abstract—**One of the pivotal services of the fifth-generation (5G) of cellular technology is massive Machine-type Communications (mMTC), which is intended to support connecting high density of machine-type devices. Random access channel of long term evolution (LTE)/ LTE advanced needs to be modified in order to support large simultaneous arrival of machine-type devices. 3GPP suggested access class barring (ACB) as a mechanism to inhibit network congestion in mMTC or massive internet of things (IoT) scenario. Consequently, in devices which are repeatedly ignored by ACB, queue of data packets keeps growing. In storage constrained IoT nodes with limited buffer, this may lead to packet drop due to buffer overflow, causing a decline in the overall throughput of the system. To address this issue, a novel queue-aware prioritized access classification (QPAC) based ACB technique is proposed in this paper, where machine-type devices having data queue size close to its buffer limit are dynamically given higher priority in ACB. To study the queue build-up at each MTC device, a node-centric analysis of ACB in buffer-constrained scenario is performed using a two-dimensional Markov chain. It is shown that the proposed QPAC scheme, with optimal model parameters obtained by maximizing overall system utility, offers up to 70% gain in throughput compared to the nearest competitive dynamic ACB scheme.**

Index Terms—**Access class barring, buffer overflow, dynamic priority access classification, internet of things, massive machine-type communication, queue awareness, random access, storage constraint**

I. Introduction

MACHINE type communication (MTC) or machine to machine (M2M) communication is one of the key enablers of the Internet of Things (IoT) paradigm. It enables autonomous exchange of information between machine-type devices (MTDs) which are installed in state-of-the-art infrastructures, like smart city and industry 4.0. Wide variety of applications include industrial IoT, smart metering, smart grid, remote monitoring, e-health, object tracking, and security [1].

Long term evolution (LTE) is an automatic choice as air interface for MTC primarily because of its wide area coverage [2]. It is also well known that existing LTE random access procedure needs to be tuned in order to fit in M2M communication, as it is quite different compared to typical human to human (H2H) communication. IoT or M2M communication has some unique challenges, namely, storage constraint, unique traffic signature, and limited battery power [3]. One of the primary challenges is, it suffers from overload when massive number of MTDs try to access the network simultaneously [4].

LTE random access is based on slotted ALOHA but the difference is that, it is connection based. It has two variants in general, namely, contention-based and contention-free [5]. In this work only the former is considered. In contention-based random access, MTDs that want to transmit a data packet need to choose a preamble and set up a connection with the base station or evolved node B (eNB) first. In massive IoT scenario, large number of MTDs contend for access simultaneously in a random access opportunity (RAO) slot, and each of them randomly chooses one of the orthogonal preambles. As the number of preambles available to each eNB is limited, this massive access leads to congestion or overload in the network.

To mitigate access congestion in LTE random access due to massive arrivals, among different access control schemes proposed by 3rd Generation Partnership Project (3GPP), access class barring (ACB) is one of the most efficient techniques [2]. In ACB, devices seeking access to the eNB are selectively barred based on their predefined access classes. In the process of reducing access attempts, ACB keeps on ignoring packets generated by some of the devices. This leads to queue build-up at the data buffer of those MTDs. As a consequence, not only those packets are delayed, they might also be in danger of being dropped when the buffer is full. Therefore, it is of utmost importance to incorporate queue-awareness in the ACB scheme, particularly in buffer-constrained IoT devices.

A. Related works

In the literature various access control techniques have been proposed to mitigate massive access congestion, e.g., cell shrinking and offloading [6], distribution reshaping [7], and variety of ACB algorithms [2], [8]–[10]. The authors in [11] reviewed different access control mechanisms proposed for mMTC. Some of them are related to the 3GPP specification, namely, ACB, extended access barring (EAB), slotted random access, separated resources for H2H and M2M, pull-based random access [12], and dynamic resource allocation [13]. Apart from these, some other techniques are also proposed to tackle congestion by professional groups other than 3GPP, e.g., code-expanded [14], prioritized [15], spatial-group based [16], non-Aloha based [17], and reliability guaranteed random access [18]. As suggested by 3GPP, in ACB, UEs are grouped in different predefined classes as per their QoS requirements. Separate access class (AC) was proposed for M2M devices. A variant of ACB is extended access barring (EAB), where low priority devices are restricted from accessing the network based on a bitmap [19]. Distributed Queueing (DQ) is another access control scheme where colliding preambles are grouped,

M. Roy Chowdhury and S. De are with the Department of Electrical Engineering and Bharti School of Telecommunication, Indian Institute of Technology Delhi, New Delhi, India.

and a queue of M2M devices that choose preambles belonging to the same group is formed for the subsequent preamble transmission [20].

3GPP has suggested some standard configuration with a set of allowed values for all the system parameters [12]. The problem of choosing the optimum value of those parameters in different scenarios remains an open area of research. A dynamic ACB mechanism, namely D-ACB was proposed in [2] for adaptive congestion control. It dynamically adjusted the barring rate based on the arrival rate to maximize the expected throughput. Modeling and analysis of random access in LTE is hard because of its complex dynamics. Most of the existing works consider the aggregated traffic at the eNB side and ignore the node level analysis at each individual MTDs. Among the first papers to consider node-centric analysis, the work in [21] maximized stable throughput, and the authors in [22] looked into access delay optimization.

While analyzing the random access in LTE or the access control schemes the authors in [21]–[23] considered infinite buffer capacity. This approach may not be realistic in resource-constrained IoT devices [24]. Especially in massive IoT or mMTC scenario, increasing the memory size in each node may not be cost-efficient. In devices with limited buffer constraint, ACB can cause long queues building up at the individual nodes, which eventually may lead to buffer overflow. The authors in [25] studied the problem of buffer overflow in constrained IoT sensors, where buffer size of 256 bits was considered. Performance of heterogeneous MTC was modeled and analyzed in [26] where each MTD was considered to have finite data buffer with queue size of five packets. The authors also assumed the whole buffer to be cleared once the preamble is successfully transmitted, because typical packets of MTC are comparatively of smaller size (a few bytes).

There are some works in the nearly related prior art which considered queueing; some of them considered finite buffer [27], while the others assumed it to be infinite [28]. The authors in [28] proposed queue-aware adaptation of transmission probabilities for two-user and three-user slotted ALOHA network. Analysis of queueing delay as well as characterization of stable throughput region was performed, but not in the context of LTE, and they did not consider ACB or any other congestion control techniques. A downlink packet scheduling technique with both channel and queue awareness was proposed for LTE in [27], where probability of buffer overflow was considered as a performance metric. The authors in [29], [30] considered finite-sized queue and the issue of buffer overflow from device-to-device communication perspective in cellular networks.

B. Motivation and key contributions

ACB helps in congestion control but in the process it discards some of the packets without accounting for the device level constraints. This leads to data queue build-up in some of the MTC nodes. It may cause buffer overflow in storage-constrained IoT devices, and consequently the incoming packets may be dropped when the data buffer is full. The situation worsens as the number of devices contending increases in massive IoT scenario.

In the 3GPP suggested algorithms like individual ACB [20], [31] or EAB [19], or their modified versions, access class of the MTC nodes are predefined and hard-coded in the SIM or the device. D-ACB algorithm in [2] dynamically assigned optimum barring rate so as to maximize throughput. However, all these techniques looked at the aggregated traffic at the eNB and assigned same barring rate to all M2M devices. A node-centric approach was proposed in [21] targeting throughput maximization in stable region. But they focused on the request queue and ignored the data queue build-up at the MTC nodes. Also, they considered an infinite buffer, which may not be realistic in constrained IoT nodes.

In the existing ACB schemes, either the barring rate is fixed throughout, or even if it is updated for different RAOs, it is same for all M2M devices irrespective of their priorities. Moreover, in all these works the access class is predefined and hard-coded in the SIM card of the device. Our closer look at the node level suggests that, different devices may have different queue lengths at different RAOs. There might be some devices at each RAO which are having queue lengths close to their buffer limits, i.e., on the verge of buffer overflow. Hence, to prevent these devices from dropping incoming packets, they must be dynamically assigned higher priority in ACB in the upcoming RAOs. Intuitively, a queue-aware ACB algorithm which dynamically assigns priority classes to MTC nodes based on the queue length at their data buffers is expected to improve the overall system performance. The major contributions of this work are listed below:

(i) A dynamic queue-aware priority access classification (QPAC) based ACB algorithm is proposed for LTE random access to curb access congestion in storage-constrained massive IoT or mMTC scenario.

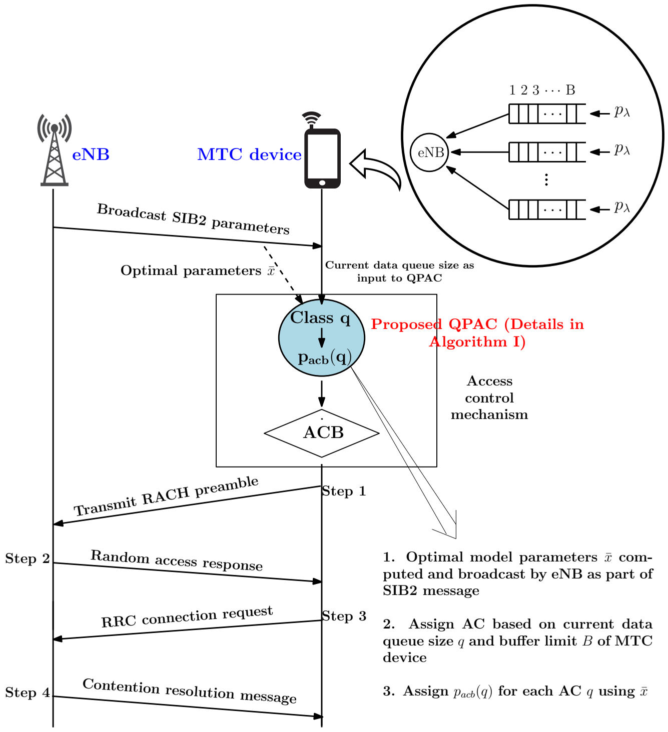
(ii) In the proposed QPAC algorithm, devices with a higher queue size at their respective data buffers are given higher priority in ACB, which is dynamically assigned in each RAO. A parametric model is proposed to assign barring rate that governs success probability in ACB to different MTC devices according to their priority classes.

(iii) A node-centric study of LTE random access is performed considering queue build-up at the data buffers of the individual MTC devices with limited buffer capacity. To the best of our knowledge there is no prior work in the literature where random access with ACB was analyzed in limited buffer scenario.

(iv) The proposed system is analytically modeled using a two-dimensional Markov chain that keeps track of both size of the data queue at each MTC node as well as its status in random access contention. The key performance indicators (KPIs) of the system are expressed in terms of the steady-state probabilities of the Markov chain.

(v) Optimal values of the model parameters of the parametric model in (ii) are obtained by maximizing the overall system utility which comprises of the KPIs derived in (iv).

(vi) Performance of the proposed QPAC-based ACB is evaluated through exhaustive simulation in terms of all the KPIs. It is shown that compared to the nearest com-



1. Optimal model parameters $\bar{x}$ computed and broadcast by eNB as part of SIB2 message

2. Assign AC based on current data queue size q and buffer limit B of MTC device

3. Assign $p_{acb}(q)$ for each AC q using $\bar{x}$

Figure 1: Random access with proposed QPAC based ACB in LTE for M2M communication.

petitive D-ACB approach, QPAC is able to increase the network throughput by up to 70%, while the access delay and blocking probability are brought down by 40% and 26%, respectively.

C. Paper organization

The layout of the paper is as follows. In Section II, a brief overview of the LTE random access protocol and the proposed node-centric system model are presented. Section III contains detailed analysis of the proposed QPAC technique and the overall utility optimization of the proposed system to find the optimal model parameters. Simulation results on performance are discussed in Section IV, followed by the concluding remarks in Section V.

II. System Model

In this section, first, the standard contention based random access in LTE is briefly described, followed by a discussion on the requirement of congestion control. Subsequently, the node-centric approach for performance optimization is explained. The frequently used variables in the system model description and in performance analysis and the corresponding symbols are listed in Table II.

A. Contention-based random access in LTE

In contention-based random access, the active MTDs get access to the eNB using orthogonal preambles [32]. There are 64 preambles available to one eNB in each RAO of t_{rao} duration. Among them 10 are reserved for contention-free random access, remaining 54 preambles can be used for contention-based random access. Before starting the random access procedure, UEs obtain all the related basic configuration parameters which are periodically broadcast by the eNB as part of SystemInformationBlock-Type2 (SIB2) message [33]. In step 1 of contention based random access, each of the active MTDs choose a preamble randomly and send it to the nearest eNB. On receipt of the preamble, in step 2, the eNB sends a random access response consisting of a temporary identifier and time-frequency resource block. This resource is used for sending radio resource control (RRC) connection request in the next step, not for the actual data transmission. In step 3, MTD transmits L2/L3 message (e.g., RRC connection request) to eNB along with its identity. It may be noted that this message is sent over the resource block that was allocated in step 2. In the last step, i.e., step 4, eNB sends a contention resolution message to all those MTDs which are successfully decoded.

In step 1, eNB cannot decode which device is asking for which preamble as it does not know the device ID. As the preambles have orthogonal structure, different MTD may choose different preamble in same RAO without any issue. But if multiple MTDs choose same preamble in a RAO, then all of them are allocated same resource by eNB for transmitting random access data in step 3, resulting in collision. We have taken the assumptions mentioned above in the line of what most of the other researchers in this field [34], [35] have considered. As per the LTE standard, the collision detection probability in the scenario where two devices transmit the same preamble in the same RAO depends mostly on the relative transmission delay between the colliding devices [36]. If the preambles are received at the eNB with very small difference in time, the eNB is expected to decode them as multipath components of the same preamble instead. If the preamble transmission fails due to collision, the MTD waits for a random back-off time $t_{bo} \sim U(0, b_i)$ where b_i is a constant called back-off indicator and can take any value between 0 and 960 ms, which is decided by eNB and broadcast through SIB2 message at the beginning of each RAO.

B. Congestion control - ACB in mMTC

As the preambles are chosen randomly, there is a possibility of multiple devices choosing same preamble in a RAO, resulting in collision. There is a need for efficient congestion control mechanism in LTE random access, specifically in massive IoT scenario, when large number of devices try to access the eNB simultaneously. ACB based congestion control mechanism is employed before the actual random access starts, in order to limit the incoming traffic by restricting some of the active devices from transmitting preamble. Each of the devices is assigned a predefined access class (AC) which is stored in the subscriber identity module (SIM/USIM) of the device. There are total 16 access classes, AC0 to AC15, which might be assigned to the devices based on their QoS requirement as shown in Table I. It is considered in the existing works that the MTDs can be classified as normal UE and hence be assigned one of the classes among AC0 - AC9 [37].

Table I: Standard access classes in ACB

AC	Type
$0-9$	Normal UE
10	Emergency Calls
$11-15$	Higher priority Services (PLMN,Security,Public Utilities,Emergency, PLMN Staff)

Table II: Frequently used variables and the corresponding symbols

Symbols	Variables
t_{rao}	Duration of each RAO
t_{acb}	Barring time
p_{acb}	Barring rate
t_{bo}	Back-off time
n	Number of MTC devices
p_λ	Arrival rate
B	Buffer limit
n_a	Number of active devices
n_p	Number of devices which pass ACB
p_{spt}	Probability of successful preamble transmission
N_{pr}	Number of preambles
b_i	Back-off indicator

There are two primary parameters in ACB: barring rate p_{acb} and barring time t_{acb}, both of which are broadcast in the SIB2 message. Each active device generates a uniform random number and compares it to the barring rate p_{acb}. If the random number is less than p_{acb}, then the device gets permission to transmit preamble. Else, the device is barred and it has to wait for a random period t_{bar} which is calculated using t_{acb}.

C. Node-centric approach in limited buffer scenario

Most of the existing works on random access or ACB look at the aggregated arrival of devices at the eNB [2], [38]. In those works uniform distribution (Traffic model 1) or beta distribution (Traffic model 2) is considered to model activation time of MTDs [12]. In this work we have considered Traffic model 1, i.e., uniformly distributed activation time which is equivalent to Poisson distributed arrival at the eNB. But to look into the problem with focus on the queueing of packets at the buffer of each of the MTDs, we need a node-centric approach. In each of the MTC nodes, packets are assumed to arrive following a Bernoulli trial with success probability $p_\lambda \in \{0, 1\}$ in each slot [21]. When the n i.i.d. Bernoulli trials in n different MTC nodes are combined at the eNB, it gives rise to a Binomial distribution. This Binomial distribution is approximated as Poisson for large n and small p_λ s. The problem we have at hand, can be looked at as a three stage problem.

Stage 1 - Data Generation: At the buffer of each of the n MTC nodes, packets are arriving and getting added to the queue. With probability p_λ a new packet gets added at the data queue of each of the nodes. Let the maximum allowed buffer size be B packets. The packets which find the buffer full on arrival get dropped and do not come back. Hence,

$$\begin{aligned}&\Pr(\text{packets being dropped due to buffer overflow})\\&\quad=\Pr(\text{data queue size}=B)\end{aligned}\tag{1}$$

Once a node has at least one packet in its data queue, it participates in the random access contention in next RAO.

Stage 2 - Access request generation: In each RAO, each of the active nodes, i.e., with non-empty data queue, generates a corresponding access request. Let the number of active nodes, i.e., the number of nodes participating in random access contention, be denoted by n_a. From these n_a nodes, ACB will allow some nodes and restrict others. At the beginning of each RAO slot, each node knows its barring rate p_{acb} (success probability of ACB) and participates in the access contention accordingly. Those devices which pass the ACB, go to the next phase and transmit preamble. The MTC nodes which pass the ACB (n_p in number) move to the next stage i.e., preamble transmission. Remaining $n_a - n_p$ devices which do not pass, are barred for t_{bar} period of time before they can re-attempt preamble transmission. The devices which are in barring period, are not considered as active.

Stage 3 - Preamble transmission: The n_p MTC nodes which pass in ACB are granted permission to proceed further in RA. Each of them chooses a random preamble and transmits it to eNB. The number of preambles available to an eNB, denoted by N_{pr}, is limited. If $n_p > N_{pr}$, some of the devices which passed ACB, will still not get a preamble. Also, as devices choose a preamble randomly, it is possible that multiple nodes choose same preamble, which may lead to collision. Probability of success in preamble transmission is the probability that a device chooses a preamble that is not chosen by any other device. Hence the success probability of an MTD in a trial of preamble transmission is given by:

$$\begin{aligned}p_{spt}=&\sum_{v=1}^{N_{pr}}\Pr(\text{the MTD choosing } v^{th} \text{ preamble}).\Pr(\text{remaining}\\&\quad n_p-1 \text{ MTDs choosing other than } v^{th} \text{ preamble})\\=&N_{pr}\cdot\left(\frac{1}{N_{pr}}\right)\left(1-\frac{1}{N_{pr}}\right)^{n_p-1}=\left(1-\frac{1}{N_{pr}}\right)^{n_p-1}\end{aligned}\tag{2}$$

It may be noted that, in case a node fails in preamble transmission stage, it goes into back-off. In the back-off phase, the node does not participate in contention, i.e., it is not counted in active nodes. Although new packets can arrive at nodes which are in back-off. Hence, it may be noted that packet generation and random access contention are two independent processes. On the backdrop of this system setting, also shown pictorially in Fig. 1, we describe the proposed algorithm in the next section.

III. Proposed Queue-aware Priority Access Classification (QPAC)

The node-centric approach as explained in the previous section gives a glimpse of data queue build-up at the buffer of individual MTDs. In buffer constrained scenarios the status of the queue becomes more critical, as once the buffer of a node is full, it will start discarding any incoming packets. To counter this we propose a QPAC based ACB algorithm which is expected to be more effective in limited buffer scenarios.

A. Proposed modification for limited buffer scenarios

In buffer constrained scenario, those MTDs which have their data buffer almost full, are in danger of their packets being

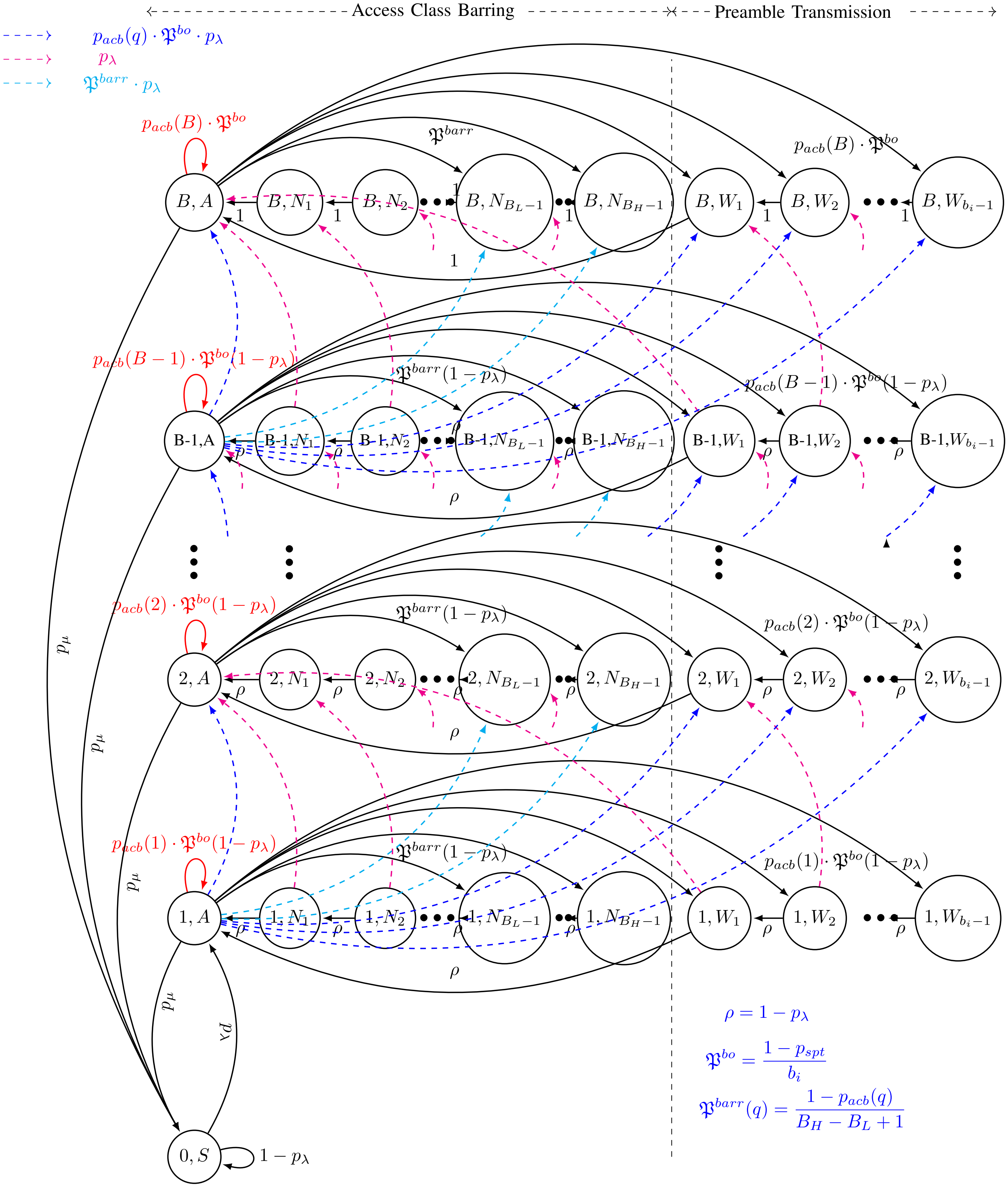

Access Class Barring
Preamble Transmission
$p_{acb}(q)\cdot\mathfrak{P}^{bo}\cdot p_\lambda$
p_λ
$\mathfrak{P}^{barr}\cdot p_\lambda$
$p_{acb}(B)\cdot\mathfrak{P}^{bo}$
$\mathfrak{P}^{barr}$
$p_{acb}(B)\cdot\mathfrak{P}^{bo}$
B, A
B, N_1
B, N_2
B, N_{B_L-1}
B, N_{B_H-1}
B, W_1
B, W_2
B, W_{b_i-1}
1
$p_{acb}(B-1)\cdot\mathfrak{P}^{bo}(1-p_\lambda)$
$\mathfrak{P}^{barr}(1-p_\lambda)$
$p_{acb}(B-1)\cdot\mathfrak{P}^{bo}(1-p_\lambda)$
B-1,A
B-1,N_1
B-1,N_2
B-1,N_{B_L-1}
B-1,N_{B_H-1}
B-1,W_1
B-1,W_2
B-1,W_{b_i-1}
ρ
$p_{acb}(2)\cdot\mathfrak{P}^{bo}(1-p_\lambda)$
$\mathfrak{P}^{barr}(1-p_\lambda)$
$p_{acb}(2)\cdot\mathfrak{P}^{bo}(1-p_\lambda)$
$2, A$
$2, N_1$
$2, N_2$
$2, N_{B_L-1}$
$2, N_{B_H-1}$
$2, W_1$
$2, W_2$
$2, W_{b_i-1}$
$p_{acb}(1)\cdot\mathfrak{P}^{bo}(1-p_\lambda)$
$\mathfrak{P}^{barr}(1-p_\lambda)$
$p_{acb}(1)\cdot\mathfrak{P}^{bo}(1-p_\lambda)$
$1, A$
$1, N_1$
$1, N_2$
$1, N_{B_L-1}$
$1, N_{B_H-1}$
$1, W_1$
$1, W_2$
$1, W_{b_i-1}$
p_μ
p_λ
$0, S$
$1-p_\lambda$
$\rho = 1-p_\lambda$
$\mathfrak{P}^{bo} = \frac{1-p_{spt}}{b_i}$
$\mathfrak{P}^{barr}(q) = \frac{1-p_{acb}(q)}{B_H-B_L+1}$

Figure 2: State diagram of two-dimensional Markov chain at data queue of each MTC node.

dropped due to buffer overflow in upcoming RAO slots. Hence, to prevent packets being dropped due to buffer overflow, devices with data queue size closer to the buffer limit should be given more priority in ACB. Also, this priority should be dynamically assigned as in different RAOs different devices can be in danger of dropping packets.

In the proposed QPAC algorithm, first, different devices are assigned different prioritized AC dynamically in each RAO. A device with data queue size equal to q is assigned AC = q. Subsequently, the barring rate p_{acb}, which governs the success probability of the MTD in ACB, is set. An MTD with longer queue (higher value of q), is given a higher priority in ACB, i.e., a higher value of p_{acb}. Thus, to model the barring rate p_{acb} as a function of queue length q, we need a function which is increasing in q, with values between 0 and 1. We propose the following parametric model to assign p_{acb} based on the current queue size at the buffer of any MTC node:

$$p_{acb}(q) = x_1 q^{x_2} + x_3 \tag{3}$$

This particular form of increasing function for the parametric model in (3) is chosen by trial-and-error method.

The optimal values of the model parameters x_1, x_2, and x_3 are obtained at the eNB by solving an optimization problem to maximize the overall system utility as explained in Section III-F. These optimal parameters are broadcast by eNB at the beginning of each RAO as part of the SIB2 message. Each active MTDs assign themselves an AC based on their respective queue lengths and computes p_{acb} using the AC and the optimal parameters received from the eNB. Once this AC assignment and p_{acb} computation is done, standard ACB mechanism follows. The flow of the proposed QPAC based ACB technique is shown in Algorithm 1.

Algorithm 1: Proposed QPAC based ACB

```
At the beginning of each RAO, eNB broadcasts barring
  time $t_{acb}$, back-off indicator $b_i$, buffer limit $B$, and
  the optimal set of model parameters $x_1, x_2, x_3$ ;
for each RAO do
    for each MTD do
        if Queue length $q < B$ then
            New packet gets added to data queue
              following Bernoulli ($p_\lambda$);
        else
            Drop any incoming packet
    Update the list of active MTDs;
    for each active MTD do
        Assign AC = $q$, $\forall q \in \{1, \cdots, B\}$ ;
        An MTD in class $q$ is assigned barring rate
          $p_{acb}(q) = x_1 q^{x_2} + x_3$;
        Generate random number $r_u \sim U[0,1)$;
        if $r_u \leqslant p_{acb}$ then
            Transmit a randomly chosen preamble;
            if preamble collides then
                Wait for back-off period $t_{bo} \sim U(0, b_i)$ ;
            else
                Preamble transmission is successful;
                All packets in the data queue are
                  transmitted and the buffer is cleared ;
        else
            Generate random number $r_{u_2} \sim U[0,1)$;
            Wait for a period $t_{bar} = t_{acb}(0.7 + 0.6 r_{u_2})$;
```

B. Analysis using two-dimensional Markov chain

To analyze the aforementioned system in node-centric approach, a two-dimensional Markov chain is used where the first dimension refers to the size of the data queue, Q and the second dimension refers to status of the access request R of the MTD. Let us define the steady-state probabilities as $\pi_{q,r} = \Pr(Q = q, R = r)$. As LTE RA procedure is slotted ALOHA based, discrete-time Markov chain (DTMC) is used to model it, where time epoch is same as t_{rao} (in milliseconds). It can also be noted that Bernoulli distribution is used to model the discrete arrival process at each MTD.

Transition probabilities of the two dimensional Markov chain are labeled in its state diagram in Fig. 2. The variate Q can take values from the set $S_Q = \{0, 1, 2, \cdots, B-1, B\}$. The second variate R can take values from the set $S_R = \{S, A, N_1, N_2, \cdots, N_{B_L - 1}, \cdots, N_{B_H - 1}, W_1, W_2, \cdots, W_{b_i - 1}\}$. R taking values N_1 through $N_{B_H - 1}$ indicate the MTD is in barred state and the counter is at 1 through $B_H - 1$. Similarly, R taking values W_1 through $W_{b_i - 1}$ indicate the MTD is in back-off state and the counter is at 1 through $b_i - 1$.

At the beginning, when the data queue is empty, i.e., $q = 0$, we consider the device to be in start or 'S' state. As soon as a packet arrives in the data buffer, we consider it to be in Active or 'A' state (unless it is in barring or back-off states). In states $(q, A) \forall q \in S_Q \setminus \{0\}$, the MTD is active and ready to take part in ACB. If the device fails in ACB, it is barred for $t_{bar} = t_{acb}(0.7 + 0.6 r_u)$, where $r_u \sim U(0, 1)$. Hence t_{bar} can take integer values between B_L and B_H, i.e., $t_{bar} \sim U(B_L, B_H)$, where, $B_L = \left\lceil 0.7 \frac{t_{acb}}{t_{rao}} \right\rceil$ and $B_H = \left\lceil 1.3 \frac{t_{acb}}{t_{rao}} \right\rceil$. Therefore, from state A, if it fails in ACB it can go to one of the barring states between $N_{B_L - 1}$ and $N_{B_H - 1}$. Once it reaches one of those states, it counts down to 1 and then in the next slot it again becomes active and participates in ACB, hence goes to state A.

Similarly when a device succeeds in ACB but fails in preamble transmission stage, it goes to back-off for a random number of slots. Back-off duration is random and can take any value between 1 and $b_i - 1$. Once it reaches one of the back-off states W_1 through $W_{b_i - 1}$, it counts down till 1 and then in the next slot becomes active, i.e., goes to state A. The devices which fail in ACB or in preamble transmission stage, can re-attempt after barring or back-off period. For mathematical tractability it is assumed that there is no limit in number of re-transmission attempts [21].

If a device succeeds in both ACB and preamble transmission, the device get serviced, i.e., it is allocated resource block to transmit data. The probability of getting serviced successfully is given by $p_\mu(q) = p_{acb}(q) \cdot p_{spt}$. Usually in IoT applications, data packets are small in size and one resource block is expected to be enough to transmit all packets in the buffer [21]. We assume that once the preamble transmission is successful, all the packets in the buffer get delivered and the data queue is cleared.

The number of active devices in a RAO is given by,

$$n_a = n \sum_{q=1}^{B} \pi_{q,A}. \tag{4}$$

In the proposed QPAC scheme, each device has different barring rate $p_{acb}(q)$ based on its data queue size q. It may be noted that, while assigning priority, it is not directly taken into consideration whether the packets in the buffer are newly arrived or they are backlogs from previous RAOs. So the average barring rate of all devices is computed as, $\bar{p} = \sum_{q=1}^{B} \pi_{q,A} \cdot p_{acb}(q)$. Hence, the number of MTDs which pass the ACB stage is given by:

$$n_p = n_a \cdot \bar{p} = n_a \sum_{q=1}^{B} \pi_{q,A} \cdot p_{acb}(q). \tag{5}$$

When a device has no packet in its queue its state is identified by the tuple $(0, S)$. Only the devices with non-empty data queue take part in ACB or random access contention. Hence, the steady-state probability is

$$\pi_{0,r} = \Pr(Q = 0, R = r) = 0 \quad \forall r \in S_R \setminus \{S\}. \tag{6}$$

As shown in Fig. 2, the non-zero transition probabilities are expressed and explained in detail in Appendix A.

C. Steady-state probabilities

The detailed balance equations hold for all pairs of adjacent states. Thus, from the state diagram of the two-dimensional Markov chain in Fig. 2,

$$\begin{aligned}\pi_{q,r} &= \sum_{m \in S_Q, n \in S_R} p_{(m,n),(q,r)} \pi_{m,n} \\ &= \sum_{m \in S_Q, n \in S_R} \Pr(q, r | m, n) \pi_{m,n}.\end{aligned} \tag{7}$$

Therefore, using the transition probabilities as shown in the Section III-F, the relation between the steady-state probabilities can be obtained.

Lemma 1. *The steady-state probabilities $\pi_{q,r}$ $\forall q \in S_Q \setminus \{0\}, r \in S_R \setminus \{A\}$ can be expressed in terms of $\pi_{q,A}$.*

States (q, N_1) ***to*** (q, N_{B_L-2})***:***

$$\begin{aligned}\pi_{q,N_{B_L-1-k}} &= \sum_{j=0}^{\min(q-1,k)} \binom{k}{j} p_\lambda^j (1-p_\lambda)^{k-j} \pi_{q-j,N_{B_L-1}} \\ &\qquad \forall q \in \{1, \cdots, B-1\}, \forall k \in \{1, \cdots, B_L - 2\} \\ \pi_{B,N_{B_L-1-k}} &= \pi_{B,N_{B_L-1}} + p_\lambda \pi_{B-1,N_{B_L-1}} + \\ &\qquad p_\lambda \sum_{w=1}^{k-1} \pi_{B-1,N_{B_L-1-w}} \forall k \in \{1, \cdots, B_L - 2\}.\end{aligned} \tag{8}$$

States (q, N_{B_L-1}) ***to*** (q, N_{B_H-1})***:***

$$\begin{aligned}\pi_{q,N_{B_H-1-k}} &= \sum_{j=0}^{q-1} \sum_{l=j}^{k} \binom{l}{j} p_\lambda^j (1-p_\lambda)^{l-j} \pi_{q-j,N_{B_H-1}} \\ &\forall q \in \{1, \cdots, B-1\}, \forall k \in \{1, \cdots, B_H - B_L\} \\ \pi_{B,N_{B_H-1-k}} &= (k+1)\pi_{B,N_{B_H-1}} + p_\lambda \pi_{B-1,N_{B_H-1}} + \\ &p_\lambda \sum_{w=1}^{k-1} \pi_{B-1,N_{B_H-1-w}}, \forall k \in \{1, \cdots, B_H - B_L\}.\end{aligned} \tag{9}$$

States (q, W_1) ***to*** (q, W_{b_i-2})***:***

$$\begin{aligned}\pi_{q,W_{b_i-1-k}} &= \sum_{j=0}^{q-1} \sum_{l=j}^{k} \binom{l}{j} (p_\lambda)^j (1-p_\lambda)^{l-j} \pi_{q-j,W_{b_i}} \\ &\forall q \in \{1, \cdots, B-1\}, \forall k \in \{1, \cdots, b_i - 2\} \\ \pi_{B,W_{b_i-1-k}} &= (k+1)\pi_{B,W_{b_i-1}} + p_\lambda \pi_{B-1,W_{b_i-1}} + \\ &p_\lambda \sum_{w=1}^{k-1} \pi_{B-1,W_{b_i-1-w}} \forall k \in \{1, \cdots, b_i - 2\}.\end{aligned} \tag{10}$$

States (q, W_{b_i-1})***:***

$$\begin{aligned}\pi_{1,W_{b_i-1}} &= p_{acb}(1)\mathfrak{P}^{bo}(1-p_\lambda)\pi_{1,A} \\ \pi_{2,W_{b_i-1}} &= p_{acb}(2)\mathfrak{P}^{bo}(1-p_\lambda)\pi_{2,A} + p_{acb}(1)\mathfrak{P}^{bo} p_\lambda \pi_{1,A} \\ &\vdots \\ \pi_{B,W_{b_i-1}} &= p_{acb}(B)\mathfrak{P}^{bo}\pi_{B,A} + p_{acb}(B-1)\mathfrak{P}^{bo} p_\lambda \pi_{B-1,A}.\end{aligned} \tag{11}$$

States (q, N_{B_H-1})***:***

$$\begin{aligned}\pi_{1,N_{B_H-1}} &= \mathfrak{P}^{barr}(1)(1-p_\lambda)\pi_{1,A} \\ \pi_{2,N_{B_H-1}} &= \mathfrak{P}^{barr}(2)(1-p_\lambda)\pi_{2,A} + \mathfrak{P}^{barr}(1) p_\lambda \pi_{1,A} \\ &\vdots \\ \pi_{B,N_{B_H-1}} &= \mathfrak{P}^{barr}(B)\pi_{B,A} + \mathfrak{P}^{barr}(B-1) p_\lambda \pi_{B-1,A}.\end{aligned} \tag{12}$$

Proof. See Appendix B □

Lemma 2. *The steady-state probabilities $\pi_{q,A}$ $\forall q \in S_Q$ can be computed recursively using the following set of equations:*

$$\begin{aligned}\pi_{B,A} =& \frac{\pi_{B,N_1} + p_\lambda \pi_{B-1,N_1} + \pi_{B,W_1} + p_\lambda \pi_{B-1,W_1}}{1 - p_{acb}(B)\mathfrak{P}^{bo}} + \\ & \frac{p_\lambda p_{acb}(B-1)\mathfrak{P}^{bo}\pi_{B-1,A}}{1 - p_{acb}(B)\mathfrak{P}^{bo}}\end{aligned} \tag{13}$$

$$\begin{aligned}\pi_{q,A} =& \frac{(1-p_\lambda)\pi_{q,N_1} + p_\lambda \pi_{q-1,N_1} + (1-p_\lambda)\pi_{q,W_1}}{1 - p_{acb}(q)\mathfrak{P}^{bo}(1-p_\lambda)} + \\ & \frac{p_\lambda \pi_{q-1,W_1} + p_\lambda p_{acb}(q-1)\mathfrak{P}^{bo}\pi_{q-1,A}}{1 - p_{acb}(q)\mathfrak{P}^{bo}(1-p_\lambda)}, \\ & \forall q \in \{2, \cdots, B-1\}\end{aligned} \tag{14}$$

$$\pi_{1,A} = \frac{(1-p_\lambda)\pi_{1,N_1} + (1-p_\lambda)\pi_{1,W_1} + \sum_{q=2}^{B} p_\mu(q)\pi_{q,A}}{(1 - p_\mu(1) - p_{acb}(1)\mathfrak{P}^{bo}(1-p_\lambda))} \tag{15}$$

$$\pi_{0,S} = \frac{1-p_\lambda}{p_\lambda}\sum_{q=1}^{B} p_\mu(q)\pi_{q,A}. \tag{16}$$

Proof. See Appendix C □

Lemma 1 and Lemma 2 are used to represent all the steady-state probabilities in terms of $\pi_{q,A}\forall q \in S_Q$. Subsequently, in (17) summation of all the steady-state probabilities, in terms of $\pi_{q,A}\forall q \in S_Q$ is equated to 1.

$$\sum_{q\in S_Q}\sum_{r\in S_R}\pi_{q,r} = 1. \tag{17}$$

The solution to the above system of equations is obtained by formulating a nonlinear least squares problem. Levenberg-Marquardt algorithm [39] is used to solve this problem and get the steady-state probabilities of all the states of the two dimensional Markov chain. It may be noted that as a part of this optimization to solve for the steady state probabilities, the system inherently solves for n_a and estimates the number of active devices in the current RAO. The expressions derived in this sub-section will be used next, to design QoS factors.

D. Key performance indicators

In this section the steady-state probabilities of the node level data queues obtained in Section III-C will be used to model the KPIs of the system: throughput, blocking probability and access delay.

1) Blocking probability: New packets arriving at the MTC nodes are dropped when the number of packets in the data queue of the node equals the buffer size. Therefore, the blocking probability, i.e., probability that an incoming data packet in the queue of an MTD is dropped because of buffer overflow, is evaluated as:

$$P_B = \sum_{r\in S_R}\pi_{B,r}. \tag{18}$$

Therefore, the effective arrival rate, i.e., the probability of packets coming into the system is given by: $\Lambda = p_\lambda\cdot(1-P_B)$.

2) Throughput: The MTC nodes participating in the random access are considered to be served when they succeed in the ACB and successfully transmit preamble. Hence the steady-state probability of successful transmission of access requests, is given by:

$$\mathbb{R}_{succ} = p_{spt}\cdot\bar{p} = p_{spt}\sum_{q=1}^{B} p_{acb}(q)\pi_{q,A}. \tag{19}$$

Now, as per our assumption, whenever the access request of an MTD is successful, i.e., its preamble is successfully transmitted, all the packets in its queue are transmitted. If an MTD has q packets in its data queue, then its probability of being active is $\pi_{q,A}$. The probability of access request of an active device with q packets in queue getting serviced is given by $p_\mu(q)$. Therefore the network throughput, i.e., the fraction of packets successfully delivered is given by:

$$\begin{aligned}\mathbb{T} &= \frac{\text{number of packets successfully delivered}}{\text{number of packets arrived}}\\ &= \frac{n\sum_{q=1}^{B} q\cdot p_\mu(q)\pi_{q,A}}{n\cdot\Lambda}\\ &= \frac{\cdot p_{spt}\sum_{q=1}^{B} q\cdot p_{acb}(q)\pi_{q,A}}{p_\lambda(1-P_B)}.\end{aligned} \tag{20}$$

3) Access delay: The total access delay is sum of two components: waiting time and service time. Waiting time is referred to as the time spent by a packet in the data queue, after its arrival till the time it gets an opportunity to take part in the ACB. This duration includes the time it spends in either of the barring or back-off states. Once it becomes active device and takes part in ACB, it's considered to be in service. Service time includes the delay due to ACB and preamble transmission. Expected access delay is expressed as using Little's theorem as:

$$E[\mathbb{D}] = \frac{E[\mathbb{N}]}{\Lambda} \tag{21}$$

where $E[\mathbb{N}]$ expected number of packets in the system (in queue or in service). The reason behind this is, when the buffer of the MTD is full, it blocks any further packet into the data queue. The expected number of packets in the system can be computed using the steady-state probabilities as:

$$E[\mathbb{N}] = \sum_{q=1}^{B} q\cdot\pi_q \tag{22}$$

where π_q is the steady-state probability that there are q packets in the data buffer of the MTD, given by: $\pi_q = \sum\limits_{r\in S_R}\pi_{q,r} \quad \forall q \in S_Q$.

E. Optimal model parameters

To achieve optimum performance, the QPAC algorithm has to use optimal set of model parameters $\bar{x} = \{x_1, x_2, x_3\}$. The optimum values of the parameters can be obtained by maximizing the system utility. The optimization problem formulated to maximize the system utility is:

$$\begin{aligned} &\min_{\bar{x}} f_0 = -\mathbb{T}\\ s.t.:\quad &(i)\quad 0 \leqslant x_1 q^{x_2} + x_3 \leqslant 1, \text{ and}\\ &(ii)\quad x_1x_2 \geqslant 0 \end{aligned} \tag{23}$$

The target of the objective function f_0 in the aforementioned optimization problem is to maximize the network throughput. The first set of constraints makes sure $p_{acb}(q) \quad \forall q \in S_Q\setminus\{0\}$ lies within 0 and 1 because it is a probability value. The second constraint makes sure that $p_{acb}(q)$ is increasing in q, such that an MTD with higher q has higher success probability in ACB. The constrained nonlinear optimization problem in (23) is solved using interior point method [40]. The steps taken in the optimization framework is explained in Algorithm 2.

Algorithm 2: Optimization framework

```
1 Recursively compute expressions of steady state
    probabilities $\pi_{q,A}$ as per Lemma 2 ;
2 Recursively compute expressions of all the steady state
    probabilities $\pi_{q,r}$ in terms of $\pi_{q,A}$ as per Lemma 1 ;
3 for each iteration of interior-point method do
4     Substitute value of $x_1, x_2, x_3$ in expressions
        computed in Step 1 and Step 2
5     Solve (23) to get steady state probabilities $\pi_{q,A}$
        using Levenberg-Marquardt algorithm, starting
        with initial values of all zeros ;
6     Get steady state probabilities of all the states $\pi_{q,r}$
        using $\pi_{q,A}$ as per Lemma 1
7     Compute the value of the objective function $f_0$ in
        the optimization problem in (29) ;
```

F. Analysis using two-dimensional Markov chain

The expressions derived in this section will be used in Section IV to evaluate performance of the proposed algorithm.

IV. Performance Evaluation

In this section we evaluate the performance of the proposed QPAC based ACB in terms of the KPIs: throughput, blocking probability, and access delay, as defined in Section III-D. The performance of QPAC algorithm in terms of the above-mentioned KPIs is evaluated and compared with the nearest competitive D-ACB scheme [2]. We have considered that the RAO occurs in every 5 sub-frames, i.e. $t_{rao} = 5$ ms and barring time is considered to be $t_{acb} = 500$ ms. The performance is evaluated for varying buffer limit B ranging from 2 to 10. Also, to test the scale-ability of the proposed technique, it is tested for varied arrival rates starting from 0.01 to as high as 0.3. Moreover, three different values of number of available preambles, i.e., $N_{pr} = 20, 30, 40$ is considered. The total number of MTC devices is considered to be $n = 1000$.

A. Varying arrival rate

In Figs. 3a, 3b, and 3c, system throughput, blocking probability, and mean access delay are shown for different arrival rates with buffer limit $B = 10$, $t_{acb} = 500$ ms, $N_{pr} = 40$, and $n = 1000$. The analytical and the simulation results are found to have a good match.

The network throughput decreases as the arrival rate is increased, while access delay increases, which is intuitive. As expected, blocking probability of the system increases at the higher arrival rates. With more number of packets getting generated in each MTC node, buffers are filled up and incoming packets are dropped due to buffer overflow. Also, the blocking probability is lower in case of QPAC based ACB compared to D-ACB [2]. The improvement in performance is more prominent at higher arrival rates. Introduction of QPAC results in up to 40% increase in the system throughput compared to the D-ACB, while access delay also is reduced by up to 30%. At lower arrival rates, number of active devices is low and there is no scarcity of preambles. Hence QPAC or D-ACB does not have much impact on the system and the plots corresponding to all the KPIs converge.

B. Varying buffer limit

Performance comparison of the proposed QPAC based ACB with D-ACB for varying buffer size is shown in Figs. 4a, 4b, and 4c in terms of throughput, blocking probability, and access delay, respectively. It can be clearly seen from Fig. 4a that as buffer size B is increased from 1 to 10, QPAC offers a higher throughput gain compared to D-ACB. Fig. 4b and Fig. 4c show that, in all cases blocking probability and delay are reduced by QPAC; the amount of reduction is higher for higher values of B. At very small buffer size, e.g. with $B = 1$, when there is no much room for improvement, performance of D-ACB and QPAC are comparable. However, as B is increased from 2 to 10, QPAC based ACB leads to 14% to 40% higher network throughput compared to D-ACB. Further, compared to D-ACB, introduction of QPAC results in reduced blocking probability and mean access delay by 28% and 30% respectively. Thus, QPAC leads to significant performance improvement in ACB, in terms of all the KPIs and the advantage of QPAC is more visible when buffer limit is comparatively higher.

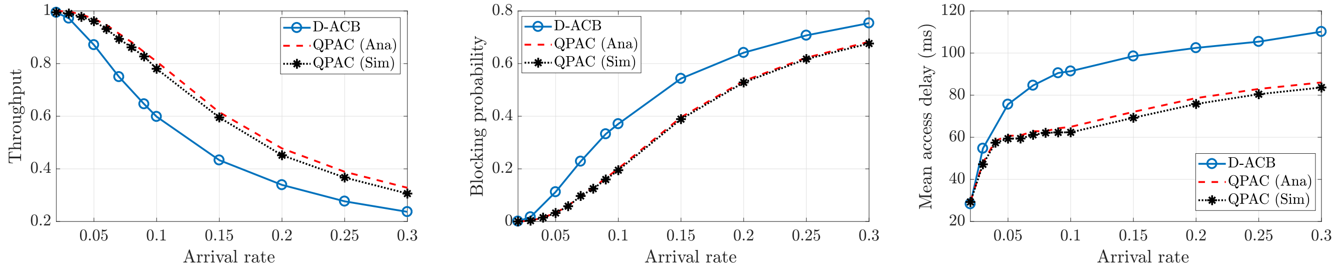
C. Varying number of available preambles

In LTE random access, total 64 preambles are available to each eNB, out of which 54 can be used for contention based RA. These preambles are to be shared among M2M and H2H devices and this distribution is up to the network designer. Therefore, we intend to study how the performance of the QPAC based ACB varies with different number of available preambles for M2M communication.

In Figs. 5, 6, and 7, performance comparison of QPAC based ACB and D-ACB are shown for different number of available preambles for MTDs, in terms of throughput, blocking probability, and mean access delay, respectively. Effect of varying number of available preambles from 20 to 40 is shown in Figs. 5a, 5b, and 5c for arrival rate $p_\lambda = 0.05, 0.10$, and 0.15, respectively. It is clearly visible in the plots that QPAC leads to even higher throughput when lower number of random access resources are available. When arrival rate is higher ($p_\lambda = 0.15$) and number of available preambles is low ($N_{pr} = 20$), QPAC offers 70% increase in network throughput compared to D-ACB. Further, blocking probability and mean access delay are also reduced by 26% and 40% in case of QPAC compared to D-ACB, as shown in Figs. 6 and 7, respectively.

V. Conclusion

In this paper we have proposed a queue-aware dynamic access classification (QPAC) based ACB scheme for access congestion in buffer constrained massive IoT scenario. The proposed QPAC algorithm dynamically assigns access classes to MTDs based on the current length of its data queue and gives higher priority to those devices which are on the verge of dropping packets due to buffer overflow. A two-dimensional Markov chain has been developed for analysis of the proposed

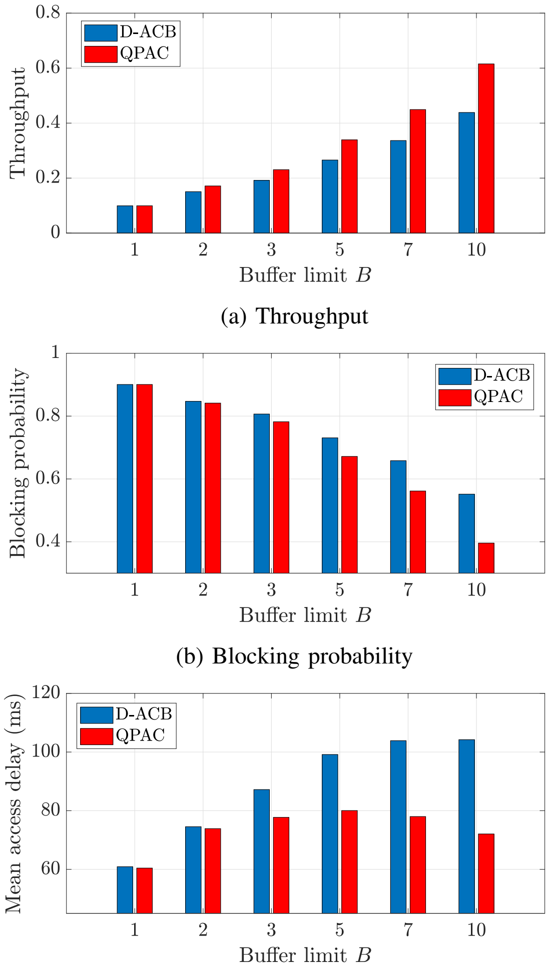


(a) Throughput

(b) Blocking probability

(c) Mean access delay

Figure 3: Performance comparison for different arrival rates; $B = 10, t_{acb} = 500, N_{pr} = 40, n = 1000$.



(a) Throughput

(b) Blocking probability

(c) Mean access delay

Figure 4: KPIs for different buffer limit B; $p_\lambda = 0.15, t_{acb} = 500, N_{pr} = 40, n = 1000$.

scheme and the performance metrics are expressed in terms of the steady-state probabilities. Optimal values of model parameters have been found by forming an optimization problem that maximizes the overall system utility including both throughput and blocking probability. Performance of the proposed QPAC based ACB scheme has been evaluated and compared with nearest competitive D-ACB algorithm in terms of different KPIs for varying arrival rates, buffer limits, and number of available preambles. It has been shown that QPAC improves the network throughput performance by 70% compared to D-ACB, while the blocking probability and mean access delay are also lowered by 26% and 40%, respectively. For the sake of mathematical tractability we have considered homogeneous scenario, i.e., same buffer limit B and arrival rate p_λ at all the MTC nodes. In our future works we would like to look into more complex heterogeneous scenario with diverse QoS requirements.

Appendix A
State-transitions of 2D Markov chain

New packets arrive following Bernoulli (p_λ) distribution. So, with probability p_λ the device gets a transition to $(1, A)$ state and with the leftover probability $1 - p_\lambda$ it remains in the same state, i.e.,

$$\Pr(1, A|0, S) = p_\lambda \text{ and}, \Pr(0, S|0, S) = 1 - p_\lambda. \tag{A.1}$$

When a device is in active state A, it participates in ACB. If it fails, it goes to one of the barring states. The device goes from state (q, A) to state (q, r) if there is no new arrival, and to state $(q + 1, r)$ if there is a new arrival. Hence,

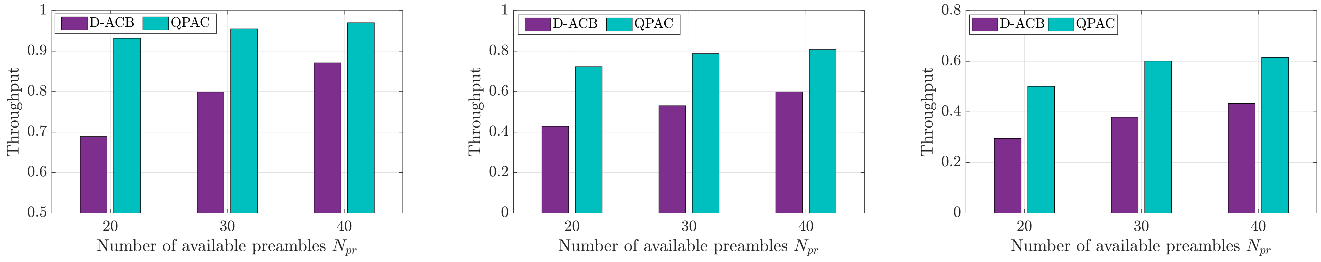
$$\begin{aligned} \Pr(q, r|q, A) &= (1 - p_\lambda)\mathfrak{P}^{barr}(q) \text{ and} \\ \Pr(q + 1, r|q, A) &= p_\lambda \mathfrak{P}^{barr}(q), \quad \forall q \in S_Q \setminus \{0, B\} \\ \Pr(B, r|B, A) &= \mathfrak{P}^{barr}(B), \quad \forall r \in \{N_{B_L-1}, \cdots, N_{B_H-1}\} \\ \text{where } \mathfrak{P}^{barr}(q) &= \frac{1 - p_{acb}(q)}{B_H - B_L + 1}. \end{aligned} \tag{A.2}$$

If a device passes the ACB stage but encounters collision in preamble transmission stage, it goes to one of the back-off states. Therefore,

$$\begin{aligned} \Pr(q, r|q, A) &= p_{acb}(q) \cdot \mathfrak{P}^{bo}(1 - p_\lambda) \text{ and} \\ \Pr(q + 1, r|q, A) &= p_{acb}(q) \cdot \mathfrak{P}^{bo} p_\lambda, \quad \forall q \in S_Q \setminus \{0, B\} \\ \Pr(B, r|B, A) &= p_{acb}(B) \cdot \mathfrak{P}^{bo}, \quad \forall r \in \{A, W_1, \cdots, W_{b_i-1}\} \\ \text{where } \mathfrak{P}^{bo} &= \frac{1 - p_{spt}}{b_i}. \end{aligned} \tag{A.3}$$

While in one of the barring states N_k, a devices counts down till N_1 and then in the next RAO it becomes active again i.e., its status becomes A.

$$\Pr(q, N_{k-1}|q, N_k) = 1 - p_\lambda \text{ and, } \Pr(q + 1, N_{k-1}|q, N_k) = p_\lambda,$$
$$\forall q \in S_Q \setminus \{0, B\}, k \in \{2, \cdots, B_H - 1\}$$

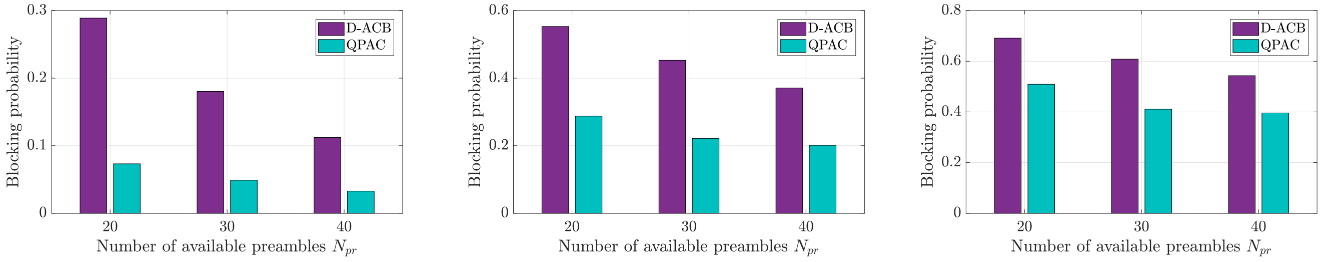


(a) $p_\lambda = 0.05$

(b) $p_\lambda = 0.10$

(c) $p_\lambda = 0.15$

Figure 5: Throughput variation with different number of preambles N_{pr}; $B = 10, t_{acb} = 500$ ms, $n = 1000$.

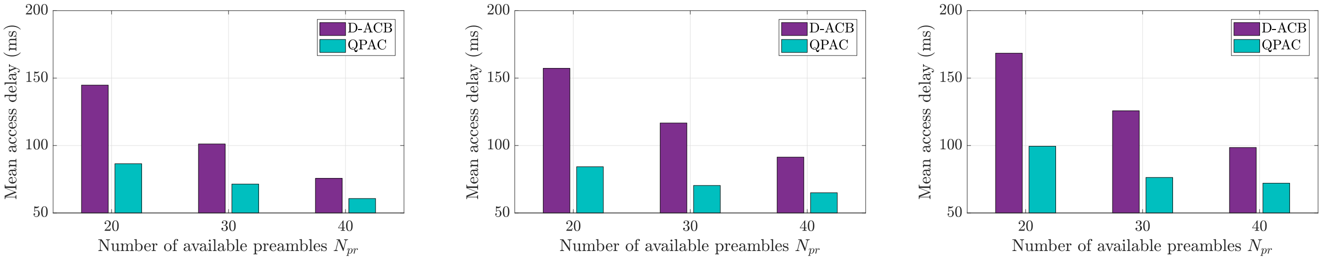


(a) $p_\lambda = 0.05$

(b) $p_\lambda = 0.10$

(c) $p_\lambda = 0.15$

Figure 6: Blocking probability variation with different number of preambles N_{pr}; $B = 10, t_{acb} = 500$ ms, $n = 1000$.



(a) $p_\lambda = 0.05$

(b) $p_\lambda = 0.10$

(c) $p_\lambda = 0.15$

Figure 7: Mean access delay variation with different number of preambles N_{pr}; $B = 10, t_{acb} = 500$ ms, $n = 1000$.

$$\begin{aligned}
\Pr(q, A|q, N_1) &= 1 - p_\lambda \text{ and, } \Pr(q+1, A|q, N_1) = p_\lambda, \\
&\qquad \forall q \in S_Q \setminus \{0, B\} \\
\Pr(B, A|B, N_1) &= 1 \text{ and, } \Pr(B, N_{k-1}|B, N_k) = 1, \\
&\qquad \forall k \in \{2, \cdots, B_H - 1\}. \qquad (A.4)
\end{aligned}$$

In one of the back-off states W_k a devices counts down till W_1 and then in the next step its status becomes A again.

$$\begin{aligned}
\Pr(q, W_{k-1}|q, W_k) &= 1 - p_\lambda, \Pr(q+1, W_{k-1}|q, W_k) = p_\lambda, \\
&\qquad \forall q \in S_Q \setminus \{0\}, k \in \{2, \cdots, b_i - 1\} \\
\Pr(q, A|q, W_1) &= 1 - p_\lambda \text{ and, } \Pr(q+1, A|q, W_1) = p_\lambda, \\
&\qquad \forall q \in S_Q \setminus \{0\} \\
\Pr(B, A|B, W_1) &= 1 \text{ and, } \Pr(B, W_{k-1}|B, W_k) = 1, \\
&\qquad \forall k \in \{2, \cdots, b_i - 1\}. \qquad (A.5)
\end{aligned}$$

From active state A, if an MTD succeeds in ACB and subsequently transmits a preamble successfully, it gets access to the eNB and clears the buffer to send all the packets. Therefore,

$$\Pr(0, S|q, A) = p_\mu(q), \quad \forall q \in S_Q \setminus \{0\} \qquad (A.6)$$

where the probability of getting serviced successfully is given by $p_\mu(q) = p_{acb}(q) \cdot p_{spt}$.

Appendix B
Proof of Lemma 1

From the state diagram in Fig. 2 and the transition probabilities, the steady-state probabilities $\pi_{q,r}$ are derived as follows:

A. Steady-state probabilities π_{q,N_1} to $\pi_{q,N_{B_L-2}}$ in terms of $\pi_{q,N_{B_L-1}}$

1) ***From state $(1, N_1)$ to state $(1, N_{B_L-2})$:***

$$\begin{aligned}
\pi_{1,N_{B_L-2}} &= \sum_{m \in S_Q, n \in S_R} \Pr(1, N_{B_L-2}|m, n)\pi_{m,n} \\
&= \Pr(1, N_{B_L-2}|1, N_{B_L-1})\pi_{1,N_{B_L-1}} = (1 - p_\lambda)\pi_{1,N_{B_L-1}}.
\end{aligned}$$

Likewise, $\pi_{1,N_{B_L-3}} = \Pr(1,N_{B_L-3}|1,N_{B_L-2})\pi_{1,N_{B_L-2}}$

$$= (1-p_\lambda)\pi_{1,N_{B_L-2}} = (1-p_\lambda)^2\pi_{1,N_{B_L-1}}.$$

Generalizing, $\pi_{1,N_{B_L-1-k}} = (1-p_\lambda)^k\pi_{1,N_{B_L-1}}$

$$\forall k \in \{1,\cdots,B_L-2\}. \quad \text{(B.1)}$$

2) From state $(2,N_1)$ to state $(2,N_{B_L-2})$: Here we have,

$$\begin{aligned}\pi_{2,N_{B_L-2}} =& \Pr(2,N_{B_L-2}|2,N_{B_L-1})\pi_{2,N_{B_L-1}}\\ &+\Pr(2,N_{B_L-2}|1,N_{B_L-1})\pi_{1,N_{B_L-1}}\\ =&(1-p_\lambda)\pi_{2,N_{B_L-1}} + p_\lambda\pi_{1,N_{B_L-1}}.\\ \pi_{2,N_{B_L-3}} =&(1-p_\lambda)\pi_{2,N_{B_L-2}} + p_\lambda\pi_{1,N_{B_L-2}}\\ =&(1-p_\lambda)^2\pi_{2,N_{B_L-1}} + 2p_\lambda(1-p_\lambda)\pi_{1,N_{B_L-1}}.\\ \pi_{2,N_{B_L-4}} =&(1-p_\lambda)\pi_{2,N_{B_L-3}} + p_\lambda\pi_{1,N_{B_L-3}}\\ =&(1-p_\lambda)^3\pi_{2,N_{B_L-1}} + 3p_\lambda(1-p_\lambda)^2\pi_{1,N_{B_L-1}}.\end{aligned}$$

Thus $\pi_{2,N_{B_L-1-k}}$ can be generalized as,

$$\pi_{2,N_{B_L-1-k}} = (1-p_\lambda)^k\pi_{2,N_{B_L-1}} + k\cdot p_\lambda(1-p_\lambda)^{k-1}\pi_{1,N_{B_L-1}}. \quad \text{(B.2)}$$

3) From state $(3,N_1)$ to state $(3,N_{B_L-2})$:

We have, $\pi_{3,N_{B_L-2}} = \Pr(3,N_{B_L-2}|3,N_{B_L-1})\pi_{3,N_{B_L-1}}$

$$\begin{aligned}&+\Pr(3,N_{B_L-2}|2,N_{B_L-1})\pi_{2,N_{B_L-1}}\\ &= (1-p_\lambda)\pi_{3,N_{B_L-1}} + p_\lambda\pi_{2,N_{B_L-1}}.\end{aligned}$$

Likewise,

$$\begin{aligned}\pi_{3,N_{B_L-3}} &= (1-p_\lambda)\pi_{3,N_{B_L-2}} + p_\lambda\pi_{2,N_{B_L-2}}\\ &= (1-p_\lambda)^2\pi_{3,N_{B_L-1}} + 2p_\lambda(1-p_\lambda)\pi_{2,N_{B_L-1}}\\ &+(p_\lambda)^2\pi_{1,N_{B_L-1}}.\\ \pi_{3,N_{B_L-4}} &= (1-p_\lambda)\pi_{3,N_{B_L-3}} + p_\lambda\pi_{2,N_{B_L-3}}\\ &= (1-p_\lambda)^3\pi_{3,N_{B_L-1}} + 3p_\lambda(1-p_\lambda)^2\pi_{2,N_{B_L-1}}\\ &+3(1-p_\lambda)p_\lambda^2\pi_{1,N_{B_L-1}}.\\ \pi_{3,N_{B_L-5}} &= (1-p_\lambda)\pi_{3,N_{B_L-4}} + p_\lambda\pi_{2,N_{B_L-4}}\\ &= (1-p_\lambda)^4\pi_{3,N_{B_L-1}} + 4p_\lambda(1-p_\lambda)^3\pi_{2,N_{B_L-1}}\\ &+6(1-p_\lambda)^2(p_\lambda)^2\pi_{1,N_{B_L-1}}.\end{aligned}$$

Following up, $\pi_{3,N_{B_L-1-k}}$ can be generalized as:

$$\begin{aligned}\pi_{3,N_{B_L-k}} =&(1-p_\lambda)^k\pi_{3,N_{B_L}} + k\cdot p_\lambda(1-p_\lambda)^{k-1}\pi_{2,N_{B_L}}\\ &+\frac{k(k-1)}{2}(p_\lambda)^2(1-p_\lambda)^{k-2}\pi_{1,N_{B_L}}. \quad \text{(B.3)}\end{aligned}$$

4) From state $(4,N_1)$ to state $(4,N_{B_L-2})$:

$$\begin{aligned}\pi_{4,N_{B_L-2}} &= \Pr(4,N_{B_L-2}|4,N_{B_L-1})\pi_{4,N_{B_L-1}}\\ &+Pr(4,N_{B_L-2}|3,N_{B_L-1})\pi_{3,N_{B_L-1}}\\ &= (1-p_\lambda)\pi_{4,N_{B_L-1}} + p_\lambda\pi_{3,N_{B_L-1}}.\end{aligned}$$

Next,

$$\begin{aligned}\pi_{4,N_{B_L-3}} &= (1-p_\lambda)\pi_{4,N_{B_L-2}} + p_\lambda\pi_{3,N_{B_L-2}}\\ &= (1-p_\lambda)^2\pi_{4,N_{B_L-1}} + 2p_\lambda(1-p_\lambda)\pi_{3,N_{B_L-1}}\\ &+p_\lambda^2\pi_{2,N_{B_L-1}}.\\ \pi_{4,N_{B_L-4}} &= (1-p_\lambda)\pi_{4,N_{B_L-3}} + p_\lambda\pi_{3,N_{B_L-3}}\\ &= (1-p_\lambda)^3\pi_{4,N_{B_L-1}} + 3p_\lambda(1-p_\lambda)^2\pi_{3,N_{B_L-1}}\\ &+3p_\lambda^2(1-p_\lambda)\pi_{2,N_{B_L-1}} + p_\lambda^3\pi_{1,N_{B_L-1}}\\ \pi_{4,N_{B_L-5}} &= (1-p_\lambda)\pi_{4,N_{B_L-4}} + p_\lambda\pi_{3,N_{B_L-4}}\\ &= (1-p_\lambda)^4\pi_{4,N_{B_L-1}} + 4p_\lambda(1-p_\lambda)^3\pi_{3,N_{B_L-1}}\\ &+6p_\lambda^2(1-p_\lambda)^2\pi_{2,N_{B_L-1}} + 4p_\lambda^3(1-p_\lambda)\pi_{1,N_{B_L-1}}.\\ \pi_{4,N_{B_L-6}} &= (1-p_\lambda)\pi_{4,N_{B_L-5}} + p_\lambda\pi_{3,N_{B_L-5}}\\ &= (1-p_\lambda)^5\pi_{4,N_{B_L-1}} + 5p_\lambda(1-p_\lambda)^4\pi_{3,N_{B_L-1}}+\\ &10p_\lambda^2(1-p_\lambda)^3\pi_{2,N_{B_L-1}} + 10p_\lambda^3(1-p_\lambda)^2\pi_{1,N_{B_L-1}}.\end{aligned}$$

Thus, $\pi_{4,N_{B_L-1-k}}$ can be generalized as:

$$\begin{aligned}\pi_{4,N_{B_L-1-k}} =&(1-p_\lambda)^k\pi_{4,N_{B_L-1}} + k\cdot p_\lambda(1-p_\lambda)^{k-1}\pi_{3,N_{B_L-1}}\\ &+\frac{k(k-1)}{2}p_\lambda^2(1-p_\lambda)^{k-2}\pi_{2,N_{B_L-1}}\\ &+\frac{k(k-1)(k-2)}{2.3}p_\lambda^3(1-p_\lambda)^{k-3}\pi_{1,N_{B_L-1}}.\end{aligned} \quad \text{(B.4)}$$

Hence from (B.1), (B.2), (B.3) and (B.4), $\pi_{q,N_{B_L-1-k}}$ can be obtained as:

$$\pi_{q,N_{B_L-1-k}} = \sum_{j=0}^{q-1}\binom{k}{j}(p_\lambda)^j(1-p_\lambda)^{k-j}\pi_{q-j,N_{B_L-1}}$$

$$\forall q \in \{1,\cdots,B-1\} \text{ and } \forall k \in \{1,\cdots,B_L-2\}. \quad \text{(B.5)}$$

5) From state (B,N_1) to state (B,N_{B_L}):
In this case,

$$\begin{aligned}\pi_{B,N_{B_L-2}} &= \pi_{B,N_{B_L-1}} + p_\lambda\pi_{B-1,N_{B_L-1}}\\ \pi_{B,N_{B_L-3}} &= \pi_{B,N_{B_L-2}} + p_\lambda\pi_{B-1,N_{B_L-2}}\\ &= \pi_{B,N_{B_L-1}} + p_\lambda\pi_{B-1,N_{B_L-1}} + p_\lambda\pi_{B-1,N_{B_L-2}}\\ \pi_{B,N_{B_L-4}} &= \pi_{B,N_{B_L-3}} + p_\lambda\pi_{B-1,N_{B_L-3}} = \pi_{B,N_{B_L-1}}+\\ &p_\lambda\pi_{B-1,N_{B_L-1}} + p_\lambda\pi_{B-1,N_{B_L-2}} + p_\lambda\pi_{B-1,N_{B_L-3}}\\ \pi_{B,N_{B_L-5}} &= \pi_{B,N_{B_L-4}} + p_\lambda\pi_{B-1,N_{B_L-4}} = \pi_{B,N_{B_L-1}}+\\ &p_\lambda\pi_{B-1,N_{B_L-1}} + p_\lambda\sum_{j=1}^{k-1}\pi_{B-1,N_{B_L-1-j}}.\end{aligned}$$

By generalizing,

$$\begin{aligned}\pi_{B,N_{B_L-1-k}} &= \pi_{B,N_{B_L-1}} + p_\lambda\pi_{B-1,N_{B_L-1}}+\\ &p_\lambda\sum_{w=1}^{k-1}\pi_{B-1,N_{B_L-1-w}}\forall k \in \{1,\cdots,B_L-2\}\end{aligned} \quad \text{(B.6)}$$

B. Steady-state probabilities $\pi_{q,N_{B_L-1}}$ to $\pi_{q,N_{B_H-1}}$

1) From state $(1,N_{B_L-1})$ to state $(1,N_{B_H-1})$:

$$\pi_{1,N_{B_H-1}} = \Pr(1,N_{B_H-1}|1,A)\pi_{1,A} = \mathfrak{P}^{barr}(1)(1-p_\lambda)\pi_{1,A}.$$

Likewise,

$$\begin{aligned}\pi_{1,N_{B_H-2}} &= \mathfrak{P}^{barr}(1)(1-p_\lambda)\pi_{1,A} + (1-p_\lambda)\pi_{1,N_{B_H-1}}\\ &= (1+(1-p_\lambda))\pi_{1,N_{B_H-1}}.\end{aligned}$$

$$\pi_{1,N_{B_H-3}} = \mathfrak{P}^{barr}(1)(1-p_\lambda)\pi_{1,A} + (1-p_\lambda)\pi_{1,N_{B_H-2}}$$

$$= [1 + (1-p_\lambda) + (1-p_\lambda)^2]\pi_{1,N_{B_H-1}}.$$

In general, $\pi_{1,N_{B_H-1-k}} = \sum_{j=0}^{k}(1-p_\lambda)^j\pi_{1,N_{B_H-1}},$

$$\forall k \in \{1,\cdots,B_H-B_L\} \qquad \text{(B.7)}$$

2) From state (*$2, N_{B_L-1}$) to state (*$2, N_{B_H-1}$***):***
Now,

$$\begin{aligned}\pi_{2,N_{B_H-1}} &= \Pr(2, N_{B_H-1}|2,A)\pi_{2,A} + \Pr(2, N_{B_H-1}|1,A)\pi_{1,A}\\ &= \mathfrak{P}^{barr}(2)(1-p_\lambda)\pi_{2,A} + \mathfrak{P}^{barr}(1)p_\lambda\pi_{1,A}.\end{aligned}$$

$$\begin{aligned}\pi_{2,N_{B_H-2}} &= \mathfrak{P}^{barr}(2)(1-p_\lambda)\pi_{2,A} + \mathfrak{P}^{barr}(1)p_\lambda\pi_{1,A}\\ &\quad + (1-p_\lambda)\pi_{2,N_{B_H-1}} + p_\lambda\pi_{1,N_{B_H-1}}\\ &= \pi_{2,N_{B_H-1}} + (1-p_\lambda)\pi_{2,N_{B_H-1}} + p_\lambda\pi_{1,N_{B_H-1}}\\ &= (1+(1-p_\lambda))\pi_{2,N_{B_H-1}} + p_\lambda\pi_{1,N_{B_H-1}}.\end{aligned}$$

$$\begin{aligned}\pi_{2,N_{B_H-3}} &= \mathfrak{P}^{barr}(2)(1-p_\lambda)\pi_{2,A} + \mathfrak{P}^{barr}(1)p_\lambda\pi_{1,A}+\\ &(1-p_\lambda)\pi_{2,N_{B_H-1}} + p_\lambda\pi_{1,N_{B_H-1}}\\ &= \pi_{2,N_{B_H-1}} + (1-p_\lambda)(1+(1-p_\lambda))\pi_{2,N_{B_H-1}}+\\ &p_\lambda(1-p_\lambda)\pi_{1,N_{B_H-1}} + p_\lambda(1+(1-p_\lambda))\pi_{1,N_{B_H-1}}\\ &= (1+(1-p_\lambda)+(1-p_\lambda)^2)\pi_{2,N_{B_H-1}}+\\ &p_\lambda(1+2(1-p_\lambda))\pi_{1,N_{B_H-1}}\end{aligned}$$

$$\begin{aligned}\pi_{2,N_{B_H-4}} &= \mathfrak{P}^{barr}(2)(1-p_\lambda)\pi_{2,A} + \mathfrak{P}^{barr}(1)p_\lambda\pi_{1,A}+\\ &(1-p_\lambda)\pi_{2,N_{B_H-3}} + p_\lambda\pi_{1,N_{B_H-3}}\\ &= \left[1+(1-p_\lambda)+(1-p_\lambda)^2+(1-p_\lambda)^3\right]\pi_{2,N_{B_H-1}}\\ &+ p_\lambda\left[1+2(1-p_\lambda)+3(1-p_\lambda)^2\right]\pi_{1,N_{B_H-1}}.\end{aligned}$$

Generalizing, $\pi_{2,N_{B_H-1-k}} = \sum_{j=0}^{k}(1-p_\lambda)^k\pi_{2,N_{B_H-1}}+$

$$p_\lambda\sum_{j=1}^{k} j(1-p_\lambda)^{j-1}\pi_{1,N_{B_H-1}}. \qquad \text{(B.8)}$$

3) From state (*$3, N_{B_L-1}$) to state (*$3, N_{B_H-1}$***):***

$$\begin{aligned}\pi_{3,N_{B_H-1}} &= \mathfrak{P}^{barr}(1-p_\lambda)\pi_{3,A} + \mathfrak{P}^{barr}p_\lambda\pi_{2,A}.\\ \pi_{3,N_{B_H-2}} &= \mathfrak{P}^{barr}(1-p_\lambda)\pi_{3,A} + \mathfrak{P}^{barr}p_\lambda\pi_{2,A}\\ &\quad + (1-p_\lambda)\pi_{3,N_{B_H-1}} + p_\lambda\pi_{2,N_{B_H-1}}\\ &= (1+(1-p_\lambda))\pi_{3,N_{B_H-1}} + p_\lambda\pi_{2,N_{B_H-1}}.\\ \pi_{3,N_{B_H-3}} &= \mathfrak{P}^{barr}(1-p_\lambda)\pi_{3,A} + \mathfrak{P}^{barr}p_\lambda\pi_{2,A}+\\ &(1-p_\lambda)\pi_{3,N_{B_H-2}} + p_\lambda\pi_{2,N_{B_H-2}}\\ &= (1+(1-p_\lambda)+(1-p_\lambda)^2)\pi_{3,N_{B_H-1}}+\\ &p_\lambda(1+2(1-p_\lambda))\pi_{2,N_{B_H-1}} + (p_\lambda)^2\pi_{1,N_{B_H-1}}.\end{aligned}$$

$$\begin{aligned}&\pi_{3,N_{B_H-4}} = \mathfrak{P}^{barr}(1-p_\lambda)\pi_{3,A} + \mathfrak{P}^{barr}p_\lambda\pi_{2,A}+\\ &(1-p_\lambda)\pi_{3,N_{B_H-3}} + p_\lambda\pi_{2,N_{B_H-3}}\\ &= \left[1+(1-p_\lambda)+(1-p_\lambda)^2+(1-p_\lambda)^3\right]\pi_{3,N_{B_H-1}}+\\ &p_\lambda\Big[1+2(1-p_\lambda)+3(1-p_\lambda)^2\Big]\pi_{2,N_{B_H-1}}+\\ &p_\lambda^2(1+3(1-p_\lambda))\pi_{1,N_{B_H-1}}.\end{aligned}$$

$$\begin{aligned}&\pi_{3,N_{B_H-5}} = \mathfrak{P}^{barr}(1-p_\lambda)\pi_{3,A} + \mathfrak{P}^{barr}p_\lambda\pi_{2,A}+\\ &(1-p_\lambda)\pi_{3,N_{B_H-4}} + p_\lambda\pi_{2,N_{B_H-4}}\\ &= \Big[1+(1-p_\lambda)+(1-p_\lambda)^2+(1-p_\lambda)^3+(1-p_\lambda)^4\Big]\\ &\pi_{3,N_{B_H-1}} + p_\lambda\Big[1+2(1-p_\lambda)+3(1-p_\lambda)^2+4(1-p_\lambda)^3\Big]\\ &\pi_{2,N_{B_H-1}} + p_\lambda^2\Big[1+3(1-p_\lambda)+6(1-p_\lambda)^2\Big]\pi_{1,N_{B_H-1}}.\end{aligned}$$

So $\pi_{3,N_{B_H-1-k}}$ can be generalized as,

$$\begin{aligned}\pi_{3,N_{B_H-1-k}} &= \sum_{j=0}^{k}(1-p_\lambda)^j\pi_{3,N_{B_H-1}} + p_\lambda\sum_{j=1}^{k} j(1-p_\lambda)^{j-1}\\ &\pi_{2,N_{B_H-1}} + p_\lambda^2\sum_{j=1}^{k}\frac{j(j-1)}{2}(1-p_\lambda)^{j-2}\pi_{1,N_{B_H-1}}\\ &\forall k \in \{1,\cdots B_H-B_L\} \qquad \text{(B.9)}\end{aligned}$$

Finally, from (B.7), (B.8), and (B.9) a clear pattern is visible, and hence $\pi_{q,N_{B_H-1-k}}$ is generalized as:

$$\pi_{q,N_{B_H-1-k}} = \sum_{j=0}^{q-1}\sum_{l=j}^{k}\binom{l}{j}(p_\lambda)^j(1-p_\lambda)^{l-j}\pi_{q-j,N_{B_H-1}},$$

$$\forall q \in \{1,\cdots,B-1\}, \forall k \in \{1,\cdots,B_H-B_L\} \qquad \text{(B.10)}$$

4) From state (*B, N_{B_L-1}) to state (*B, N_{B_H-1}***):***

$$\begin{aligned}\pi_{B,N_{B_H-1}} =& \mathfrak{P}^{barr}(B)\pi_{B,A} + \mathfrak{P}^{barr}(B-1)p_\lambda\pi_{B-1,A}\\ \pi_{B,N_{B_H-2}} =& \mathfrak{P}^{barr}(B)\pi_{B,A} + \mathfrak{P}^{barr}(B-1)p_\lambda\pi_{B-1,A}+\\ &\pi_{B,N_{B_H-1}} + p_\lambda\pi_{B-1,N_{B_H-1}}\\ =& 2\pi_{B,N_{B_H-1}} + p_\lambda\pi_{B-1,N_{B_H-1}}\\ \pi_{B,N_{B_H-3}} =& \mathfrak{P}^{barr}(B)\pi_{B,A} + \mathfrak{P}^{barr}(B-1)p_\lambda\pi_{B-1,A}+\\ &\pi_{B,N_{B_H-2}} + p_\lambda\pi_{B-1,N_{B_H-2}}\\ =& \pi_{B,N_{B_H-1}} + \pi_{B,N_{B_H-2}} + p_\lambda\pi_{B-1,N_{B_H-2}}\\ =& 3\pi_{B,N_{B_H-1}} + p_\lambda\pi_{B-1,N_{B_H-1}} + p_\lambda\pi_{B-1,N_{B_H-2}}\\ \pi_{B,N_{B_H-4}} =& \mathfrak{P}^{barr}(B)\pi_{B,A} + \mathfrak{P}^{barr}(B-1)p_\lambda\pi_{B-1,A}+\\ &\pi_{B,N_{B_H-3}} + p_\lambda\pi_{B-1,N_{B_H-3}}\\ =& 4\pi_{B,N_{B_H-1}} + p_\lambda\pi_{B-1,N_{B_H-1}} + p_\lambda\pi_{B-1,N_{B_H-2}}\\ &+ p_\lambda\pi_{B-1,N_{B_H-3}} \qquad \text{(B.11)}\end{aligned}$$

Generalizing, $\pi_{B,N_{B_H-1-k}}$ can be expressed as,

$$\begin{aligned}\pi_{B,N_{B_H-1-k}} &= (k+1)\pi_{B,N_{B_H-1}} + p_\lambda\pi_{B-1,N_{B_H-1}}+\\ &p_\lambda\sum_{w=1}^{k-1}\pi_{B-1,N_{B_H-1-w}}\forall k \in \{1,\cdots,B_H-B_L\}\end{aligned} \qquad \text{(B.12)}$$

C. Steady-state probabilities π_{q,W_1} *to* $\pi_{q,W_{b_i-2}}$

1) From state (*$1, W_1$) to state (*$1, W_{b_i-2}$***):***

$$\begin{aligned}\pi_{1,W_{b_i-1}} &= p_{acb}(1)\mathfrak{P}^{bo}(1-p_\lambda)\pi_{1,A}\\ \pi_{1,W_{b_i-2}} &= p_{acb}(1)\mathfrak{P}^{bo}(1-p_\lambda)\pi_{1,A} + (1-p_\lambda)\pi_{1,W_{b_i-1}}\end{aligned}$$

$$\begin{aligned}
&= (1 + (1 - p_\lambda))\pi_{1,W_{b_i-1}} \\
\pi_{1,W_{b_i-3}} &= p_{acb}(1)\mathfrak{P}^{bo}(1 - p_\lambda)\pi_{1,A} + (1 - p_\lambda)\pi_{1,W_{b_i-2}} \\
&= [1 + (1 - p_\lambda) + (1 - p_\lambda)^2]\pi_{1,W_{b_i-1}}
\end{aligned}$$

To generalize, $\pi_{1,W_{b_i-1-k}} = \sum_{j=0}^{k}(1 - p_\lambda)^j \pi_{1,W_{b_i-1}}$

$$\forall k \in \{1 \cdots b_i - 2\} \qquad \text{(B.13)}$$

*2) **From state** $(2, W_1)$ **to state** $(2, W_{b_i-2})$:*

$$\begin{aligned}
\pi_{2,W_{b_i-1}} &= p_{acb}(2)\mathfrak{P}^{bo}(1 - p_\lambda)\pi_{2,A} + p_{acb}(1)\mathfrak{P}^{bo}p_\lambda\pi_{1,A} \\
\pi_{2,W_{b_i-2}} &= p_{acb}(2)\mathfrak{P}^{bo}(1 - p_\lambda)\pi_{2,A} + p_{acb}(1)\mathfrak{P}^{bo}p_\lambda\pi_{1,A}+ \\
&(1 - p_\lambda)\pi_{2,W_{b_i-1}} + p_\lambda\pi_{1,W_{b_i-1}} \\
&= (1 + (1 - p_\lambda))\pi_{2,W_{b_i-1}} + p_\lambda\pi_{1,W_{b_i-1}} \\
\pi_{2,W_{b_i-3}} &= p_{acb}(2)\mathfrak{P}^{bo}(1 - p_\lambda)\pi_{2,A} + p_{acb}(1)\mathfrak{P}^{bo}p_\lambda\pi_{1,A}+ \\
&(1 - p_\lambda)\pi_{2,W_{b_i-2}} + p_\lambda\pi_{1,W_{b_i-2}} \\
&= [1 + (1 - p_\lambda) + (1 - p_\lambda)^2]\pi_{2,W_{b_i-1}}+ \\
&p_\lambda(1 + 2(1 - p_\lambda))\pi_{1,W_{b_i-1}}
\end{aligned}$$

Thus, $\pi_{2,W_{b_i-1-k}}$ can be generalized as,

$$\begin{aligned}
\pi_{2,W_{b_i-1-k}} &= \sum_{j=0}^{k}(1 - p_\lambda)^j \pi_{2,W_{b_i-1}}+ \\
&p_\lambda \sum_{j=1}^{k} j(1 - p_\lambda)^{j-1}\pi_{1,W_{b_i-1}} \qquad \text{(B.14)}
\end{aligned}$$

*3) **From state** $(3, W_1)$ **to state** $(3, W_{b_i-2})$:* Now,

$$\begin{aligned}
\pi_{3,W_{b_i-1}} &= p_{acb}(3)\mathfrak{P}^{bo}(1 - p_\lambda)\pi_{3,A} + p_{acb}(2)\mathfrak{P}^{bo}p_\lambda\pi_{2,A} \\
\pi_{3,W_{b_i-2}} &= p_{acb}(3)\mathfrak{P}^{bo}(1 - p_\lambda)\pi_{3,A} + p_{acb}(2)\mathfrak{P}^{bo}p_\lambda\pi_{2,A} \\
&+ (1 - p_\lambda)\pi_{3,W_{b_i-1}} + p_\lambda\pi_{2,W_{b_i-1}} \\
&= (1 + (1 - p_\lambda))\pi_{3,W_{b_i-1}} + p_\lambda\pi_{2,W_{b_i-1}} \\
\pi_{3,W_{b_i-3}} &= p_{acb}(3)\mathfrak{P}^{bo}(1 - p_\lambda)\pi_{3,A} + p_{acb}(2)\mathfrak{P}^{bo}p_\lambda\pi_{2,A} \\
&+ (1 - p_\lambda)\pi_{3,W_{b_i-2}} + p_\lambda\pi_{2,W_{b_i-2}} \\
&= (1 + (1 - p_\lambda) + (1 - p_\lambda)^2)\pi_{3,W_{b_i-1}}+ \\
&p_\lambda(1 + 2(1 - p_\lambda))\pi_{2,W_{b_i-1}} + p_\lambda^2\pi_{1,W_{b_i-1}} \\
\pi_{3,W_{b_i-4}} &= p_{acb}(3)\mathfrak{P}^{bo}(1 - p_\lambda)\pi_{3,A} + p_{acb}(2)\mathfrak{P}^{bo}p_\lambda\pi_{2,A} \\
&+ (1 - p_\lambda)\pi_{3,W_{b_i-3}} + p_\lambda\pi_{2,W_{b_i-3}} \\
&= (1 + (1 - p_\lambda) + (1 - p_\lambda)^2 + (1 - p_\lambda)^3)\pi_{3,W_{b_i-1}} \\
&+ p_\lambda(1 + 2(1 - p_\lambda) + 3(1 - p_\lambda)^2)\pi_{2,W_{b_i-1}} \\
&+ p_\lambda^2(1 + 3(1 - p_\lambda))\pi_{1,W_{b_i-1}} \\
\pi_{3,W_{b_i-5}} &= p_{acb}(3)\mathfrak{P}^{bo}(1 - p_\lambda)\pi_{3,A} + p_{acb}(2)\mathfrak{P}^{bo}p_\lambda\pi_{2,A}+ \\
&(1 - p_\lambda)\pi_{3,W_{b_i-4}} + p_\lambda\pi_{2,W_{b_i-4}} = (1 + (1 - p_\lambda)+ \\
&(1 - p_\lambda)^2 + (1 - p_\lambda)^3 + (1 - p_\lambda)^4)\pi_{3,W_{b_i-1}}+ \\
&p_\lambda(1 + 2(1 - p_\lambda) + 3(1 - p_\lambda)^2 + 4(1 - p_\lambda)^3)\pi_{2,W_{b_i-1}} \\
&+ p_\lambda^2(1 + 3(1 - p_\lambda) + 6(1 - p_\lambda)^2)\pi_{1,W_{b_i-1}}
\end{aligned}$$

Thus, $\pi_{3,W_{b_i-1-k}}$ is generalized as,

$$\begin{aligned}
\pi_{3,W_{b_i-1-k}} &= \sum_{j=0}^{k}(1 - p_\lambda)^j \pi_{3,W_{b_i-1}} + p_\lambda \sum_{j=1}^{k} j(1 - p_\lambda)^{j-1} \\
&\pi_{2,W_{b_i-1}} + p_\lambda^2 \sum_{j=1}^{k} \frac{j(j-1)}{2}(1 - p_\lambda)^{j-2}\pi_{1,W_{b_i-1}}, \\
&\forall k \in \{1, \cdots b_i - 2\}
\end{aligned}$$

From the pattern in (B.13), (B.14), and (B.15) it is concluded that,

$$\begin{aligned}
\pi_{q,W_{b_i-1-k}} &= \sum_{j=0}^{q-1}\sum_{l=j}^{k}\binom{l}{j}(p_\lambda)^j(1 - p_\lambda)^{l-j}\pi_{q-j,W_{b_i}}, \\
&\forall q \in \{1, \cdots, B - 1\} \text{ and } \forall k \in \{1, \cdots, b_i - 2\} \qquad \text{(B.15)}
\end{aligned}$$

*4) **From state** (B, W_1) **to state** (B, W_{b_i-2}):*

$$\begin{aligned}
\pi_{B,W_{b_i-1}} &= p_{acb}(B)\mathfrak{P}^{bo}\pi_{B,A} + p_{acb}(B - 1)\mathfrak{P}^{bo}p_\lambda\pi_{B-1,A} \\
\pi_{B,W_{b_i-2}} &= p_{acb}(B)\mathfrak{P}^{bo}\pi_{B,A} + p_{acb}(B - 1)\mathfrak{P}^{bo}p_\lambda\pi_{B-1,A}+ \\
&\pi_{B,W_{b_i-1}} + p_\lambda\pi_{B-1,W_{b_i-1}} = \pi_{B,W_{b_i-1}} + \pi_{B,W_{b_i-1}} \\
&+ p_\lambda\pi_{B-1,W_{b_i-1}} = 2\pi_{B,W_{b_i-1}} + p_\lambda\pi_{B-1,W_{b_i-1}} \\
\pi_{B,W_{b_i-3}} &= p_{acb}(B)\mathfrak{P}^{bo}\pi_{B,A} + p_{acb}(B - 1)\mathfrak{P}^{bo}p_\lambda\pi_{B-1,A}+ \\
&\pi_{B,W_{b_i-2}} + p_\lambda\pi_{B-1,W_{b_i-2}} = 3\pi_{B,W_{b_i-1}}+ \\
&p_\lambda\pi_{B-1,W_{b_i-1}} + p_\lambda\pi_{B-1,W_{b_i-2}} \\
\pi_{B,W_{b_i-4}} &= p_{acb}(B)\mathfrak{P}^{bo}\pi_{B,A} + p_{acb}(B - 1)\mathfrak{P}^{bo}p_\lambda\pi_{B-1,A}+ \\
&\pi_{B,W_{b_i-3}} + p_\lambda\pi_{B-1,W_{b_i-3}} = 4\pi_{B,W_{b_i-1}}+ \\
&p_\lambda\pi_{B-1,W_{b_i-1}} + p_\lambda\pi_{B-1,W_{b_i-2}} + p_\lambda\pi_{B-1,W_{b_i-3}}
\end{aligned}$$

To generalize,

$$\begin{aligned}
\pi_{B,W_{b_i-1-k}} &= (k + 1)\pi_{B,W_{b_i-1}} + p_\lambda\pi_{B-1,W_{b_i-1}}+ \\
&p_\lambda \sum_{w=1}^{k-1}\pi_{B-1,W_{b_i-1-w}} \forall k \in \{1, \cdots, b_i - 2\} \qquad \text{(B.16)}
\end{aligned}$$

Appendix C
Proof of Lemma 2

Referring to Fig. 2, the steady-state probabilities of $(0, S)$ and (q, A) $\forall q \in \{1, \cdots, B\}$ are derived as:

$$\begin{aligned}
\pi_{B,A} &= \pi_{B,N_1} + p_{acb}(B)\mathfrak{P}^{bo}\pi_{B,A} + p_{acb}(B - 1)\mathfrak{P}^{bo}p_\lambda\pi_{B-1,A} \\
&+ p_\lambda\pi_{B-1,N_1} + \pi_{B,W_1} + p_\lambda\pi_{B-1,W_1} \\
&= \frac{\pi_{B,N_1} + p_\lambda\pi_{B-1,N_1} + \pi_{B,W_1} + p_\lambda\pi_{B-1,W_1}}{(1 - p_{acb}(B)\mathfrak{P}^{bo})} \\
&+ \frac{p_\lambda p_{acb}(B - 1)\mathfrak{P}^{bo}\pi_{B-1,A}}{(1 - p_{acb}(B)\mathfrak{P}^{bo})}
\end{aligned}$$

$$\begin{aligned}
\pi_{B-1,A} &= (1 - p_\lambda)\pi_{B-1,N_1} + p_\lambda\pi_{B-2,N_1} + (1 - p_\lambda)\pi_{B-1,W_1} \\
&+ p_\lambda\pi_{B-2,W_1} + p_{acb}(B - 1)\mathfrak{P}^{bo}(1 - p_\lambda)\pi_{B-1,A}+ \\
&p_\lambda p_{acb}(B - 2)\mathfrak{P}^{bo}\pi_{B-2,A} \\
&= \frac{(1 - p_\lambda)\pi_{B-1,N_1} + p_\lambda\pi_{B-2,N_1} + (1 - p_\lambda)\pi_{B-1,W_1}}{(1 - p_{acb}(B - 1)\mathfrak{P}^{bo}(1 - p_\lambda))}
\end{aligned}$$

$$+\frac{p_\lambda \pi_{B-2,W_1} + p_\lambda p_{acb}(B-2)\mathfrak{P}^{bo}\pi_{B-2,A}}{(1-p_{acb}(B-1)\mathfrak{P}^{bo}(1-p_\lambda))}.$$

$$\begin{aligned}\pi_{2,A} &= (1-p_\lambda)\pi_{2,N_1} + p_\lambda\pi_{1,N_1} + (1-p_\lambda)\pi_{2,W_1} + p_\lambda\pi_{1,W_1}\\ &+ p_{acb}(2)\mathfrak{P}^{bo}(1-p_\lambda)\pi_{2,A} + p_\lambda p_{acb}(1)\mathfrak{P}^{bo}\pi_{1,A}.\end{aligned}$$

Thus generalizing,

$$\begin{aligned}\pi_{q,A} &= \frac{(1-p_\lambda)\pi_{q,N_1} + p_\lambda\pi_{q-1,N_1} + (1-p_\lambda)\pi_{q,W_1}}{(1-p_{acb}(q)\mathfrak{P}^{bo}(1-p_\lambda))}\\ &+ \frac{p_\lambda\pi_{q-1,W_1} + p_\lambda p_{acb}(q-1)\mathfrak{P}^{bo}\pi_{q-1,A}}{(1-p_{acb}(q)\mathfrak{P}^{bo}(1-p_\lambda))},\\ &\forall q \in \{2,\cdots,B-1\}\end{aligned} \tag{C.1}$$

$$\begin{aligned}\pi_{1,A} &= (1-p_\lambda)\pi_{1,N_1} + (1-p_\lambda)\pi_{1,W_1} + p_\lambda\pi_{0,S}+\\ &p_\lambda\sum_{q=1}^{B} p_\mu(q)\pi_{q,A} + p_{acb}(1)\mathfrak{P}^{bo}(1-p_\lambda)\pi_{1,A}.\end{aligned} \tag{C.2}$$

We also have,

$$\begin{aligned}\pi_{0,S} &= (1-p_\lambda)\pi_{0,S} + (1-p_\lambda)\sum_{q=1}^{B} p_\mu(q)\pi_{q,A}\\ &= \frac{1-p_\lambda}{p_\lambda}\sum_{q=1}^{B} p_\mu(q)\pi_{q,A}.\end{aligned} \tag{C.3}$$

Substituting (C.3) in (C.2) and solving,

$$\pi_{1,A} = \frac{(1-p_\lambda)\pi_{1,N_1} + (1-p_\lambda)\pi_{1,W_1} + \sum_{q=2}^{B} p_\mu(q)\pi_{q,A}}{(1-p_\mu(1) - p_{acb}(1)\mathfrak{P}^{bo}(1-p_\lambda))} \tag{C.4}$$

References

[1] 3GPP, "Service Requirements for Machine-Type Communications (MTC)," Third-Generation Partnership Project, Sophia Antipolis Cedex, France,, Technical Specification (TS), document 3GPP TS 22.368 V13.2.0, Dec. 2016.

[2] S. Duan, V. Shah-Mansouri, Z. Wang, and V. W. S. Wong, "D-ACB: Adaptive Congestion Control Algorithm for Bursty M2M Traffic in LTE Networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 9847–9861, Dec. 2016.

[3] H. Chao, Y. Chen, and J. Wu, "Power saving for Machine to Machine communications in cellular networks," in *Proc. IEEE GLOBECOM Wkshps*, Dec. 2011, pp. 389–393.

[4] M. El Tanab and W. Hamouda, "Machine-to-Machine Communications With Massive Access: Congestion Control," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 3545–3557, 2019.

[5] M. Hasan, E. Hossain, and D. Niyato, "Random access for machine-to-machine communication in LTE-advanced networks: issues and approaches," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 86–93, Jun. 2013.

[6] R. Atat, L. Liu, J. Wu, J. Ashdown, and Y. Yi, "Green massive traffic offloading for cyber-physical systems over heterogeneous cellular networks," *Mobile Netw. Appl.*, vol. 24, no. 4, pp. 1364–1372, 2019.

[7] H. Chao, Y. Chen, J. Wu, and H. Zhang, "Distribution Reshaping for Massive Access Control in Cellular Networks," in *Proc. IEEE Veh. Technol. Conf. (VTC-Fall)*, 2016, pp. 1–5.

[8] Z. Wang and V. W. S. Wong, "Optimal Access Class Barring for Stationary Machine Type Communication Devices With Timing Advance Information," *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, pp. 5374–5387, Oct. 2015.

[9] H. Jin, W. T. Toor, B. C. Jung, and J. Seo, "Recursive Pseudo-Bayesian Access Class Barring for M2M Communications in LTE Systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 8595–8599, Sep. 2017.

[10] S. W. Haider Shah, A. T. Riaz, and K. Iqbal, "Congestion Control Through Dynamic Access Class Barring for Bursty MTC Traffic in Future Cellular Networks," in *Proc. Int. Conf. Frontiers Inf. Technol. (FIT)*, 2018, pp. 176–181.

[11] M. S. Ali, E. Hossain, and D. I. Kim, "LTE/LTE-A Random Access for Massive Machine-Type Communications in Smart Cities," *IEEE Commun. Mag.*, vol. 55, no. 1, pp. 76–83, Jan. 2017.

[12] 3GPP, "Study on RAN Improvements for Machine-Type Communications," Third-Generation Partnership Project, Sophia Antipolis Cedex, France,, Tech. Rep., 3GPP, TR 37.868 V11.0.0, Oct. 2011.

[13] 3GPP TSG RAN WG2 #70bis, "RACH overload solutions," R2-103742, ZTE, Tech. Rep., July 2010, Stockholm, Sweden.

[14] N. K. Pratas, H. Thomsen, C. Stefanović, and P. Popovski, "Code-expanded random access for machine-type communications," in *Proc. IEEE Globecom Wkshps*, Dec. 2012, pp. 1681–1686.

[15] A. C. C. Lo, Y. Wei, M. Jacobsson, and M. Kucharzak, "Enhanced LTE-Advanced Random-Access Mechanism for Massive Machine-to-Machine (M2M) Communications," in *Proc. 27th Wireless World Research Forum*, Oct. 2011.

[16] H. S. Jang, S. M. Kim, K. S. Ko, J. Cha, and D. K. Sung, "Spatial Group Based Random Access for M2M Communications," *IEEE Commun. Lett.*, vol. 18, no. 6, pp. 961–964, 2014.

[17] M. Shirvanimoghaddam, Y. Li, M. Dohler, B. Vucetic, and S. Feng, "Probabilistic Rateless Multiple Access for Machine-to-Machine Communication," *IEEE Trans. Wirel. Commun.*, vol. 14, no. 12, pp. 6815–6826, 2015.

[18] G. C. Madueño, N. K. Pratas, C. Stefanović, and P. Popovski, "Massive M2M access with reliability guarantees in LTE systems," in *Proc. IEEE ICC*, Jun. 2015, pp. 2997–3002.

[19] R. Cheng, J. Chen, D. Chen, and C. Wei, "Modeling and Analysis of an Extended Access Barring Algorithm for Machine-Type Communications in LTE-A Networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 2956–2968, Jun. 2015.

[20] H. Althumali and M. Othman, "A Survey of Random Access Control Techniques for Machine-to-Machine Communications in LTE/LTE-A Networks," *IEEE Access*, vol. 6, pp. 74 961–74 983, 2018.

[21] W. Zhan and L. Dai, "Massive Random Access of Machine-to-Machine Communications in LTE Networks: Modeling and Throughput Optimization," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2771–2785, 2018.

[22] ——, "Access Delay Optimization of M2M Communications in LTE Networks," *IEEE Wirel. Commun. Lett.*, vol. 8, no. 6, pp. 1675–1678, 2019.

[23] N. Jiang, Y. Deng, X. Kang, and A. Nallanathan, "Random Access Analysis for Massive IoT Networks Under a New Spatio-Temporal Model: A Stochastic Geometry Approach," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5788–5803, 2018.

[24] R. Zhu and J. Yang, "Buffer-aware adaptive resource allocation scheme in LTE transmission systems," *EURASIP J. Wireless Commun. Netw.*, vol. 2015, no. 1, p. 176, Jun 2015.

[25] R. Hamidouche, Z. Aliouat, A. A. Abba Ari, and M. Gueroui, "An Efficient Clustering Strategy Avoiding Buffer Overflow in IoT Sensors: A Bio-Inspired Based Approach," *IEEE Access*, vol. 7, pp. 156 733–156 751, 2019.

[26] D. Niyato, P. Wang, and D. I. Kim, "Performance Modeling and Analysis of Heterogeneous Machine Type Communications," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2836–2849, 2014.

[27] F. G. C. Rocha and F. H. T. Vieira, "A Channel and Queue-Aware Scheduling for the LTE Downlink Based on Service Curve and Buffer Overflow Probability," *IEEE Wirel. Commun. Lett.*, vol. 8, no. 3, pp. 729–732, 2019.

[28] I. Dimitriou and N. Pappas, "Stable Throughput and Delay Analysis of a Random Access Network With Queue-Aware Transmission," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3170–3184, 2018.

[29] S. W. H. Shah, M. M. U. Rahman, A. N. Mian, O. A. Dobre, and J. Crowcroft, "Effective capacity analysis of harq-enabled d2d communication in multi-tier cellular networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 9, pp. 9144–9159, 2021.

[30] S. W. H. Shah, M. M. U. Rahman, A. N. Mian, A. Imran, S. Mumtaz, and O. A. Dobre, "On the impact of mode selection on effective capacity of device-to-device communication," *IEEE Wirel. Commun. Lett.*, vol. 8, no. 3, pp. 945–948, 2019.

[31] I. Ahmad and K. Chang, "Mission Critical User Priority-Based Random Access Scheme for Collision Resolution in Coexisting PS-LTE and LTE-M Networks," *IEEE Access*, vol. 7, pp. 115 505–115 517, 2019.

[32] 3GPP, "Physical channels and modulation," Third-Generation Partnership Project, Sophia Antipolis Cedex, France,, Technical Specification (TS), 3GPP, TS 36.211,version 10.0.0 Release 10, Jan. 2011.

[33] I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, and V. Casares-Giner, "On the Accurate Performance Evaluation of the LTE-A Random Access Procedure and the Access Class Barring Scheme," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 7785–7799, Dec. 2017.

[34] O. Arouk and A. Ksentini, "General model for rach procedure performance analysis," *IEEE Commun. Lett.*, vol. 20, no. 2, pp. 372–375, 2016.

[35] T. Lin, C. Lee, J. Cheng, and W. Chen, "PRADA: Prioritized Random Access With Dynamic Access Barring for MTC in 3GPP LTE-A Networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 5, pp. 2467–2472, Jun. 2014.

[36] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the Random Access Channel of LTE and LTE-A Suitable for M2M Communications? A Survey of Alternatives," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 4–16, First quarter 2014.

[37] L. Tello-Oquendo *et al.*, "Performance Analysis and Optimal Access Class Barring Parameter Configuration in LTE-A Networks With Massive M2M Traffic," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3505–3520, Apr. 2018.

[38] M. Roy Chowdhury and S. De, "Delay-aware Priority Access Classification for Massive Machine-type Communication," *IEEE Trans. Veh. Technol.*, vol. 70, no. 12, pp. 13 238–13 254, 2021.

[39] J. Moré, "The Levenberg-Marquardt algorithm: Implementation and theory, Springer Verlag," *Numerical Analysis, ed. G. A. Watson, Lecture Notes in Mathematics*, vol. 630, no. 4, pp. 105–116, 1977.

[40] R. H. Byrd, M. E. Hribar, and J. Nocedal, "An Interior Point Algorithm for Large-Scale Nonlinear Programming," *SIAM J. Optim.*, vol. 9, no. 4, pp. 877–900, 1999.