On the Capacity Region of ALOHA Protocol for the Internet of Things

Moslem Noori, Samira Rahimian, Masoud Ardakani
Department of Electrical and Computer Engineering, University of Alberta, Canada
Email: \{moslem, srahimia, ardakani\}@ualberta.ca

Abstract—Accommodating the needs of a large number of diverse users in the Internet of Things (IoT), notably managing how the users access the common channel, has posed unique challenges to the network designers. In this paper, we study a heterogeneous IoT network consisting of multiple classes of users who may have different service requirements. For this network, we consider the application of irregular repetition slotted ALOHA (IRSA) that is shown to offer large throughput for single-class networks. Then, we focus on finding the network performance boundaries by studying the set of feasible throughput values for each class, called the capacity region. To this end, we first introduce the concept of dual network of a multi-class network meaning a homogeneous network with the same number of users. We then prove that finding the capacity region of the assumed multi-class network boils down to finding the maximum achievable throughput of its dual network. Using this finding, we then discuss how any given point of the capacity region can be achieved. Further, a delay performance study is conducted to evaluate the average and maximum packet transmission delay experienced by the users of each class.

I. INTRODUCTION

In the future Internet of Things (IoT), a variety of transmitting devices ranging from patient monitoring sensors and traffic control devices to smart cars and appliances will coexist [1]. In such a heterogeneous environment, the nature of data communicated by each device dictates a level of priority of using the communication resources for that user in the network. For instance, in a hospital environment, a patient monitoring device that measures the vital signs of a patient should be given priority to transmit its data over a sensor which reports the temperature of a medicine fridge. That said, novel network protocols should be devised to accommodate the needs of such multi-class IoT networks with different service priorities for each class.

Medium access control (MAC) protocols play a critical role in addressing the needs of a multi-class network by managing how users from different classes share a common medium for communication. Since the MAC protocol should deal with a massive number of uncoordinated and dynamic devices with sporadic traffic loads in an IoT network, random access protocols stand as a promising candidate. Despite their potential, only few prior studies [2]–[7] have considered the application of these protocols for a multi-class network scenario. This work is an attempt to further explore the application of random MAC protocols for heterogeneous multi-class IoT networks.

ALOHA [8], [9] is one of the existing random access protocols that is well applicable to IoT scenarios. While the original ALOHA protocol offers a small throughput, recent advances [10]–[12] have allowed for a significant throughput increase over the original ALOHA. For instance, it is shown in [10] that through sending several packet replicas by the users and using successive interference cancellation (SIC) to resolve some of the packet collisions at the receiver, the throughput of the network can be markedly improved. The idea of transmitting packet replicas at the users and exploiting SIC at the receiver for achieving higher throughputs was further developed in [11] where irregular repetition slotted ALOHA (IRSA) is introduced and in [12] where the authors propose frameless slotted ALOHA. Although these improvements for the ALOHA protocol have been suggested, they do not address all requirements demanded in the IoT era, importantly different service requirements of the users.

One of the few studies to consider ALOHA for a heterogeneous network is [7] where the authors study a network with different importance classes of users. For each class, a utility function is defined to reflect the requirements of the users within that class. The authors then formulate an optimization problem whose goal is to maximize a weighted sum of the class utility functions. Due to the difficulty of solving this optimization problem, the authors propose an approximate solution to maximize the sum of the utility functions in a network setup where larger probabilities of access by higher priority classes are enforced.

In this work, we further advance the existing results by identifying the throughput performance boundaries of a heterogeneous network. For this, we first formally define the capacity region of the network [13] and identify an outer bound for it. Later, we introduce the concept of dual network of the considered heterogeneous multi-class network referring to a homogeneous single-class network with the same number of users. For the case of IRSA as the MAC protocol, we then analytically find the capacity region of the multi-class network using the achievable throughputs of its dual network. Further, it is discussed how any given point within this capacity region can be achieved by carefully activating specific number of users from each class and utilizing the optimal IRSA scheme for the dual network. In addition, we provide analytical results on how the user activation strategy affects the average and maximum packet transmission delay in the
network. Numerical example are also presented.

II. BACKGROUND AND SYSTEM MODEL

We consider a network consisting of $k$ disjoint sets (classes) of users, denoted by $C_i$ for $i \in K = \{1, 2, \ldots, k\}$, who share a common channel to transmit their packets to a base station (BS). Each class $C_i$ has $|C_i| = N_i$ users and the total number of users in the network is $N = \sum_{i \in K} N_i$. It is assumed that all users within different classes always have a packet for transmission, however, different classes may have different service needs and requirements making the network heterogeneous. We call such a network a $k$-class network.

To share the communication medium, irregular repetition slotted ALOHA (IRSA) [11] is used as the MAC protocol. To this end, channel access time is divided into equal-duration slots. The length of each slot is equal to the time needed to transmit a packet plus a possible guard time to counteract propagation delays [14]. Also, $M$ slots are grouped together to form a frame. Here, it is assumed that only $L_i$ out of the $N_i$ users in $C_i$ are active and send their packets within a frame. Thus, the total number of active users in the network is $L = \sum_{i \in K} L_i$.

By adopting IRSA as the access strategy by the users, each user may transmit several replicas of its packet within a frame according to a repetition distribution. Here, it is assumed that all active users within the same class employ the same repetition distribution. To this end, a user within $C_i$ transmits $l$ replicas of its packet within a frame with probability $\Lambda_{i,l}$ where $1 \leq l \leq M$. We call $l$ the user degree. To send the replicas, the user randomly selects $l$ of the $M$ available time slots and send each replica in one of these slots. Following the above, the user degree distribution of class $i$ is defined as

$$\Lambda_i(x) \triangleq \sum_{l=1}^{M} \Lambda_{i,l} x^l$$

where $\sum_{l=1}^{M} \Lambda_{i,l} = 1$. Further, the average packet repetition by users within $C_i$ is

$$\bar{\Lambda}_i = \sum_{l=1}^{M} l \Lambda_{i,l} = \Lambda_i(1).$$

Following the users’ transmissions, each of the frame’s time slots has one of the following statuses: a) no transmission (idle slot), b) only one transmission (singleton slot), c) more than one transmission (collision slot). To better describe the slots’ status, we use $\Psi_m$ showing the probability of having exactly $m$ transmissions within a time slot. Now, similar to the user degree distribution, the slot degree distribution is defined as

$$\Psi(x) \triangleq \sum_{m=0}^{L} \Psi_m x^m.$$  

After receiving the users’ signals, the BS performs successive interference cancellation (SIC) to resolve the packets [11]. For this purpose, the BS stores the received signals within a frame and first resolves the collision-free packets. Then, the BS cancels the corresponding interference of the resolved packets from the collision slots of the frame. By doing such, some other singleton slots may appear making resolving some other packets possible. This process is iteratively repeated until no more packets can be resolved. At this point, the BS sends a feedback to the users informing them about the resolved packets. Users whose transmissions were not successful will try to send their packets in the upcoming frames.

As shown in [11], the performance of SIC heavily depends on the input traffic load to the network. For class $i$, the input traffic load is

$$G_i = \frac{L_i}{M}.$$  

We also define the traffic load vector as $G = [G_i]_{i \in K}$ where the total network traffic is $G_t = \sum_{i=1}^{k} G_i$. To define the throughput, we denote the number of users from class $i$ whose packets are successfully received at the BS by $S_i(G)$. That said, the throughput of $C_i$ for the traffic load vector $G$, denoted by $T_i(G)$ is defined as

$$T_i(G) = \frac{S_i(G)}{M}.$$  

Now, we introduce the tuple presentation of the described $k$-class network. According to this presentation, a $k$-class network with class $i$ having $N_i$ users whose active users transmit according to a user degree distribution $\Lambda_i(x)$ is described by

$$\mathbb{N}_k = (k, \{N_i\}_{i \in K}, \{\Lambda_i(x)\}_{i \in K}).$$

**Definition 1:** For the considered $k$-class network $\mathbb{N}_k$, we define a dual homogeneous network, namely $\mathbb{N}_1$, as a 1-class network with $N$ users and a user degree distribution $\Lambda(x)$. That is,

$$\mathbb{N}_1 = (1, N, \Lambda(x)).$$

**Definition 2:** We call a dual homogeneous network with maximum throughput as the optimal dual network denoted by $\mathbb{N}_1^*$ where

$$\mathbb{N}_1^* = (1, N, \Lambda^*(x))$$

and the maximum throughput $T^*$ is achieved at the optimal traffic load $G^*$.

For more details on how $G^*$ and $\Lambda^*(x)$ are found for a single-class (homogeneous) network, please see [11].

**Remark 1:** From the definition of the optimal dual network, for any arbitrary dual network $\mathbb{N}_1$ with load $G$, if $G \neq G^*$ or $\Lambda(x) \neq \Lambda^*(x)$, then $T(G) < T^*$ where $T(G)$ is the throughput of $\mathbb{N}_1$. Also, for $G < G^*$ and $\Lambda(x) = \Lambda^*(x)$, $T(G) = G$ for asymptotically large $N$ and $M$ [11]. That is, the packet loss probability asymptotically approaches 0 for $G < G^*$. Furthermore, $T(G) \leq G$ regardless of $G$ and $\Lambda(x)$.

As we discuss later, the throughput of a class depends on the traffic load of the class as well as other classes’ traffic loads.
III. Capacity Region of the MAC Protocol

In this section, we identify the capacity region of the IRSA for the considered k-class network \( \mathbb{N}_k \). The capacity region of the network determines the throughput performance boundaries of the system [13]. The formal definition of the capacity region is presented in the following.

Definition 3: For \( \mathbb{N}_k \), a throughput k-tuple \( T(\mathcal{G}) = (T_1(\mathcal{G}), T_2(\mathcal{G}), \ldots, T_k(\mathcal{G})) \) is said to be achievable if for a given traffic load vector \( \mathcal{G} \), there exists a set of probability distributions \( \{\Lambda_i(x)\}_{i \in K} \) resulting in the throughput \( T_i(\mathcal{G}) \) for class \( C_i \).

Definition 4: The closure of the set of all achievable \( T(\mathcal{G}) \), taken over all possible \( \mathcal{G} \), is called the capacity region of the network.

Lemma 1: The following constitutes an outer bound to the capacity region of \( \mathbb{N}_k \)

\[
T_i(\mathcal{G}) \leq \min\{1, \frac{N_i}{M}\},
\]

\[
\sum_{i \in K} T_i(\mathcal{G}) \leq \min\{1, \sum_{i \in K} \frac{N_i}{M}\}.
\]

Now, before finding the capacity region of the network, we state the following lemma.

Lemma 2: For a k-class network \( \mathbb{N}_k \) with traffic vector \( \mathcal{G} \), the slot degree distribution is similar to the one for its dual homogenous network \( \mathbb{N}_1 \) with traffic load \( G = G \) and

\[
\Lambda(x) = \frac{1}{G} \sum_{i \in K} G_i \Lambda_i(x).
\]

Proof: To prove the lemma, we start by deriving \( \Psi_m \) for an arbitrary \( m \). The degree of an arbitrary slot in the frame is \( m \) if exactly \( m \) out of the all active \( G_i \) users in \( \mathbb{N}_k \) transmit within this slot. Thus,

\[
\Psi_m = \left(\frac{G_i M}{m}\right) \rho^m (1 - \rho)^{L-m}
\]

where \( \rho \) is the probability of a packet transmission by an arbitrary user \( u \) in the considered slot. On the other hand,

\[
\rho = \frac{1}{M} \sum_{i \in K} \mathbb{P}[u \in C_i] \bar{\Lambda}_i = \frac{1}{M} \sum_{i \in K} \frac{G_i}{G} \Lambda_i(1) = \frac{\Lambda'(1)}{M}.
\]

Thus,

\[
\Psi_m = \left(\frac{G_i M}{m}\right) \left(\frac{\Lambda'(1)}{M}\right)^m \left(1 - \frac{\Lambda'(1)}{M}\right)^{L-m}
\]

that is basically the probability of having a degree \( m \) slot in \( \mathbb{N}_1 \) with load \( G_i = G \).

In other words, Lemma 2 indicates that from the viewpoint of SIC, the effective user degree distribution of \( \mathbb{N}_k \) is the weighted average of user degree distributions of all classes. Thus, in the asymptotic situation as \( N \) and \( M \) become large, we expect SIC to perform similarly for \( \mathbb{N}_k \) and \( \mathbb{N}_1 \). Now, since a successful recovery of one of the packets from any class of \( \mathbb{N}_k \) by SIC can be mapped to a successful packet recovery for \( \mathbb{N}_1 \), throughput of \( \mathbb{N}_1 \) in Lemma 2 is \( T(G_i) = \sum_{i \in K} T_i(\mathcal{G}) \).

Using this fact, we are able to state the following theorem on the capacity region of the assumed k-class network.

Theorem 1: For a k-class network \( \mathbb{N}_k \) with asymptotically large classes, the capacity region is the closure of the convex hull of all \( T(\mathcal{G}) = (T_1(\mathcal{G}), T_2(\mathcal{G}), \ldots, T_k(\mathcal{G})) \) satisfying

\[
T_i(\mathcal{G}) \leq \min\{T^*, \frac{N_i}{M}\},
\]

\[
\sum_{i \in K} T_i(\mathcal{G}) \leq \min\{T^*, \sum_{i \in K} \frac{N_i}{M}\}
\]

for all possible \( \mathcal{G} \) where \( T^* \) is the throughput of the optimal dual network.

Proof: To prove Theorem 1, we first show the converse of the theorem and then the achievability of the region.

Using the result of Lemma 2, we model the the considered network \( \mathbb{N}_k \) with its dual network \( \mathbb{N}_1 \) with an active load of \( G = G \) and a user degree distribution \( \Lambda(x) \) defined in (10). Let us first consider the case where only (some) users from \( C_i \) are active and all other classes are silent. The associated load vector in this case is \( \mathcal{G} = [0, 0, \ldots, G_i, 0, \ldots, 0] \). Now, if \( G_i < G^* \) and using Remark 1, we have

\[
T(G_i) = T_i(\mathcal{G}) \leq G_i \leq \frac{N_i}{M}.
\]

On the other hand, if \( G_i > G^* \), the throughput of \( \mathbb{N}_1 \) is capped by the throughput of the optimal dual network, i.e. \( T^* \). Thus,

\[
T(G_i) = T_i(\mathcal{G}) \leq T^*
\]

and as a consequence, (14a) holds. Now, consider a general load vector \( \mathcal{G} = [G_i]_{i \in K} \) for \( \mathbb{N}_k \). If \( G_i = \sum_{i \in K} G_i \leq G^* \), then from Remark 1

\[
T(G_i) = \sum_{i \in K} T_i(\mathcal{G}) \leq T^*.
\]

On the other hand, if \( G_i > G \), the optimal dual network determines the throughput upper bound and

\[
T(G_i) = \sum_{i \in K} T_i(\mathcal{G}) \leq T^*.
\]

This completes the proof of the theorem’s converse.

Now, we focus on proving the achievability of the capacity region. To achieve any throughput tuple \( T = (T_1, T_2, \ldots, T_k) \) within the described capacity region, we need to determine the set of the user degree distributions and load vector achieving this throughput point. To this end, we enforce a load vector \( \mathcal{G} = [G_i]_{i \in K} \) by picking \( MT_i \) users from \( C_i \) and activating them. We assign \( \Lambda^*(x) \) as the user degree distribution to all of the active users from all classes. As a result of this load and degree distribution assignment, the dual homogeneous network of \( \mathbb{N}_k \) is the optimal dual network \( \mathbb{N}_1^* \). Furthermore, since \( \sum_{i \in K} T_i \leq T^* \), we have

\[
G_i = \sum_{i \in K} G_i = \sum_{i \in K} T_i \leq T^* = G^*.
\]

Now, from Remark 1 since \( G_i \leq G^* \), the throughput of the dual network \( \mathbb{N}_1^* \) is \( T(G_i) = G_i \). This means that all active
users can successfully send their packets to the BS and the desired throughput tuple is achieved.

**Corollary 1:** For a non-asymptotic network, (14a) and (14b) define an outer bound for the capacity region.

**Remark 2:** To achieve any throughput point within the capacity region, the set of active users are chosen by the BS. At the end of each frame, the BS sends a feedback message to the users informing them if their packets have been successfully received and what users should transmit in the next frame. Later we will discuss how the choice of active users by the relay affects the delay performance of the system.

To help the reader to clearly understand the concept of capacity region, here, we present some numerical examples for the capacity region of a multi-class network. To this end, we consider $N$ users where $N = M$. Before plotting the capacity region for this network, we need to find $\Lambda^*(x)$, $G^*$, and $T^*$ for its optimal dual network. To this end, we use the results in [11] where for a maximum transmission degree of eight, it is shown that $\Lambda^*(x) = 0.5x^2 + 0.28x^3 + 0.22x^8$ is the optimal user degree distribution for a single-class network. Asymptotically when the number of users approaches infinity, the maximum throughput is achieved at $G^* = 0.938$. However, for finite number of users, $\Lambda^*(x)$ achieves lower throughputs [11]. Before presenting the capacity regions, we depict the throughput versus the traffic load of this distribution for the optimal dual network of $N_2$ in Figure 1 for different values of $N$. As seen in this figure, the maximum achievable throughput $T^*$ increases with $N$. For $N = 100$ and $N = 200$, $T^* = 0.72$ and $T^* = 0.77$ that are respectively achieved at $G^* = 0.76$ and $G^* = 0.81$. For $N = 300$, the maximum throughput is $T^* = 0.79$ achieved at $G^* = 0.82$.

Using these numbers and Theorem 1 we present the capacity region for two different network setups in Figure 2 and Figure 3. In Figure 2 the capacity region is depicted for different values of $N$ where $M = N$ and the network has two classes each with $N_1 = N_2 = N$ users. As seen in Figure 2 since the two classes have equal number of users, the capacity region is symmetric. For this setup, the capacity region is

\[
T_1(G), T_2(G) \leq \frac{1}{2},
\]

\[
T_1(G) + T_2(G) \leq T^*.
\]

Figure 3 present the results for a two-class network where $N_1 = 0.8N$ and $N_2 = 0.2N$. Due to the asymmetry of the number of users within the classes, the capacity region is asymmetric.

**IV. DELAY PERFORMANCE ANALYSIS**

In the previous section, we identified the capacity region of a multi-class network $N_k$ and proposed an approach to achieve any given point within this capacity region. However, we did not discuss how the active users within each class are selected and its effect on the delay performance of the network. The goal of this section is to further elaborate on...
Definition 5: The average transmit delay of class $i$, denoted by $D_{a,i}$, denotes the average number of slots that takes for a user in $\mathcal{C}_i$ to successfully transmit its packet. The average transmit delay of the whole network is defined as

$$D_a = \frac{1}{k} \sum_{i \in K} D_{a,i}. \quad (19)$$

Definition 6: The maximum transmit delay of class $i$, denoted by $D_{w,i}$, refers to the maximum number of slots that takes for a user in $\mathcal{C}_i$ to successfully transmit its packets. Similarly, the maximum transmit delay of the network is defined as

$$D_w = \max_i D_{w,i}. \quad (20)$$

### A. Random Selection of the Active Users

Now that we have defined our delay measures, we study the delay performance of the system. Assume that to achieve a throughout vector $T = (T_1, T_2, \ldots, T_k)$ within the capacity region, $MG_i$ users in $\mathcal{C}_i$ should be activated. One way to select these users is to randomly pick them from the $N_i$ users in the class. While the desired throughout vector is achieved through this selection, we are interested in knowing how this selection approach affects the delay measures of the network. In this case, it can be shown that

$$D_{a,i} = \frac{N_i}{MT_i}. \quad (21)$$

Now, using (19) and (21), $D_a$ can be found. On the other hand, it is easy to see that $D_{w,i}$, and as a result $D_w$, are unbounded and could approach infinity albeit with a probability approaching zero.

### B. Deterministic Selection of the Active Users

Another approach for choosing the active users from $\mathcal{C}_i$ at the base station is to first map the class users to numbers from 1 to $N_i$. The base station then puts these numbers in a circular queue and starts user selection by choosing the users mapped from 1 to $MG_i$. Out of these users, a number of them, say $f$ which is about $M(G_i - T_i)$, fail to successfully transmit their packet within the frame. In the next frame, the base station chooses $MG_i$ active users as the failed $f$ users from the previous frame and the next $MG_i - f$ users in the circular queue and continues so on. For this approach, it is easy to show that

$$D_{a,i} = \frac{N_i}{MT_i}. \quad (22)$$

Further, for large networks where $T_i = G_i$ and thus $f = 0$,

$$D_{w,i} = \left\lceil \frac{N_i}{MT_i} \right\rceil. \quad (23)$$

Having $D_{a,i}$ and $D_{w,i}$, $D_a$ and $D_w$ are found accordingly.

Following the above, one can see that the BS approach to choose active users influences the delay measures. While both aforementioned schemes offer the same average transmit delay, their maximum transmit delays differ drastically. Note that the improvement seen in $D_w$ for the deterministic selection comes at the price of implementing $k$ circular queues at the base station to keep the track of the selected users.

### V. Conclusion

In this paper, we studied the capacity region as well as the delay performance of a multi-class IoT network when IRSA is applied. To achieve any given point within the capacity region, we proposed an IRSA-based scheme by carefully activating a specific number of users within each class. Here, the active user selection process was completed in a centralized manner at the base station. As a future research direction, we would like to study the effect of employing a decentralized user selection strategy, where the decision of becoming active is made locally at the users, on the performance of the network.

### ACKNOWLEDGMENT

The authors would like to thank Alberta Innovates Technology Futures (AITF), Natural Sciences and Engineering Research Council of Canada (NSERC), and TELUS Corporation for supporting their research.

### REFERENCES

[1] M. Zorzi, A. Gluhak, S. Lange, and A. Bassi, “From today’s INTRANet to a future INTERnet of things: a wireless- and mobility-related view,” IEEE Wireless Communications, vol. 17, no. 6, pp. 44–51, 2010.

[2] X. Wang, J. Kanjiyl, Y. Onozato, and S. Noguchi, “Heterogeneous ALOHA networks: A sufficient condition for all equilibrium states to be stable,” Computer networks and ISDN systems, vol. 22, no. 3, pp. 213–224, 1991.

[3] Y. Jin and G. Kesidis, “Equilibria of a noncooperative game for heterogeneous users of an ALOHA network,” IEEE Commun. Lett., vol. 6, no. 7, pp. 282–284, 2002.

[4] M. S. Hefeida, T. Canili, and A. Khokhar, “CL-MAC: A cross-layer MAC protocol for heterogeneous wireless sensor networks,” Ad Hoc Networks, vol. 11, no. 1, pp. 213–225, 2013.

[5] Y. Liu, C. Yuen, X. Cao, N. Hassan, and J. Chen, “Design of a scalable hybrid MAC protocol for heterogeneous M2M networks,” IEEE Internet of Things Journal, vol. 1, no. 1, pp. 99–111, 2014.

[6] Y. Liu, Z. Yang, R. Yu, Y. Xiang, and S. Xie, “An efficient MAC protocol with adaptive energy harvesting for machine-to-machine networks,” IEEE Access, vol. 3, pp. 358–367, 2015.

[7] L. Toni and P. Frossard, “Prioritized random MAC optimization via graph-based analysis,” IEEE Trans. Commun., vol. 63, no. 12, pp. 5002–5013, 2015.

[8] N. Abramson, “The ALOHA system: another alternative for computer communications,” in Proceedings of ACM fall joint computer conference, November 1970, pp. 281–285.

[9] L. G. Roberts, “ALOHA packet system with and without slots and capture,” ACM SIGCOMM Computer Communication Review, vol. 5, no. 2, pp. 28–42, 1975.

[10] E. Casini, R. De Gaudenzi, and O. R. Herrero, “Contention resolution diversity slotted ALOHA (CRDSA): An enhanced random access scheme for satellite access packet networks,” IEEE Trans. Wireless Commun., vol. 6, no. 4, pp. 1408–1419, 2007.

[11] G. Liva, “Graph-based analysis and optimization of contention resolution diversity slotted ALOHA,” IEEE Trans. Commun., vol. 59, no. 2, pp. 477–487, 2011.

[12] C. Stefanovic, P. Popovski, and D. Vukobratovic, “Frameless ALOHA protocol for wireless networks,” vol. 16, no. 12, pp. 2087–2090, 2012.

[13] V. Navare, G. Mergen, and L. Tong, “Stability and delay of finite-user slotted ALOHA with multipacket reception,” IEEE Trans. Inform. Theory, vol. 51, no. 7, pp. 2636–2656, 2005.

[14] J. Ahn, A. Syed, B. Krishnamachari, and J. Heidemann, “Design and analysis of a propagation delay tolerant ALOHA protocol for underwater networks,” Ad Hoc Networks, vol. 9, no. 5, pp. 752–766, 2011.