Fast Uplink Grant for Machine Type Communications: Challenges and Opportunities

Samad Ali, Nandana Rajatheva, and Walid Saad

***Abstract*—The notion of a fast uplink grant is emerging as a promising solution for enabling massive machine type communications (MTCs) in the Internet of Things over cellular networks. By using the fast uplink grant, machine type devices (MTD) will no longer require random access (RA) channels to send scheduling requests. Instead, uplink resources can be actively allocated to MTDs by a base station. In this paper, the challenges and opportunities for adopting the fast uplink grant to support MTCs are investigated. First, the fundamentals of fast uplink grant and its advantages over conventional scheduled and uncoordinated access schemes are presented. Then, the key challenges that include the prediction of the set of MTDs with data to transmit, as well as the optimal scheduling of MTDs, are exposed. To overcome these challenges, a two-stage approach that includes traffic prediction and optimized scheduling is proposed. In particular, various solutions for source traffic prediction for periodic MTC traffic are reviewed and novel methods for event-driven traffic prediction are proposed. For optimal allocation of uplink grants, advanced machine learning techniques are presented. By using the proposed solutions, the fast uplink grant has the potential to enable cellular networks to support massive MTCs and effectively reduce the signaling overhead and overcome the delay and congestion challenges of conventional RA schemes.**

I. Introduction

Realizing the smart cities vision hinges on the introduction of effective wireless solutions for pervasive Internet of Things (IoT) connectivity [1] across both human type devices, such as smartphones, and machine type devices (MTDs), such as drones, sensors, and actuators. While cellular networks provide an appealing solution for IoT connectivity, existing networks were designed with a focus on providing high data rates to a small number of human type devices, in the downlink. However, as shown in Fig. 1, IoT applications will rely on a massive number of MTDs that generate small data packets [2] that are mostly transmitted in the uplink direction, towards a central base station (BS). Beyond its uplink-centered nature, machine type communications (MTCs) in the IoT will also differ from conventional human type communications by the heterogeneous quality-of-service (QoS) requirements of the IoT applications, in terms of latency and reliability, two metrics that are seen as key enablers for IoT applications such as smart grids, autonomous vehicles, factory automation, and e-health. Clearly, supporting such uplink-centric MTCs, with heterogeneous QoS needs will pose major challenges for cellular networks that range from QoS modeling to network optimization and multiple access [2].

One of the main challenges of cellular-enabled MTC in the IoT is the inability of existing random access (RA) protocols to support massive, short-packet transmissions. Moreover, the dense nature of MTCs will inevitably strain the highly-constrained resources of the RA process and, thus, render it inefficient. The RA challenges of MTC are further exacerbated by the massive nature of the IoT which is expected to encompass thousands of MTDs within a geographically constrained area [2], [3]. Recently, there has been a surge in the literature that focuses on optimizing RA process for MTC (e.g., see [3] and references therein). Such works are primarily focused on either reducing signaling overhead to increase efficiency, or developing new backoff mechanisms to reduce collisions. However, solutions that focus on optimizing the signaling overhead fall short in addressing the problem of resource congestion. Moreover, prior art [3] that addresses the RA channel congestion problem typically does so at the cost of increased latency. Such added latency cannot be sustained by mission-critical IoT applications that require reliable packet delivery within stringent deadlines. As a result, without discounting the existing efforts on improving RA for MTC, most of this prior art is still unsuitable to handle massive access due to the associated signaling overhead, collisions, and delays.

Another promising approach to integrating the IoT into cellular systems is to use uncoordinated transmissions in which no RA procedure is performed and the MTDs are not scheduled [4]. For such *uncoordinated access*, MTDs select a random radio resource block (RB) and transmit their data. Even though this method reduces signaling, it still suffers from collisions since many MTDs might select the same RB. Despite some recent promising solutions for this uncoordinated access problem (e.g., see [4]), these existing approaches will still yield high congestion and associated delays.

Clearly, from the radio access point of view, there is a need for new solutions for MTC that can strike a balance between fully scheduled solutions, (that are controlled and reliable but have high signaling overhead, RA congestion, and long delays) and fully uncoordinated solutions (that have low signaling overhead but experience collisions and long delays).

The main contribution of this paper is to develop such a middle-ground multiple access solution by leveraging the idea of a *fast uplink grant*, a method that was proposed by 3GPP in

S. Ali and N. Rajatheva are with the Centre for Wireless Communications (CWC), University of Oulu, Finland. Emails: {samad.ali, nandana.rajatheva}@oulu.fi. W. Saad is with Wireless@VT, Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA, USA, Email: walids@vt.edu.

This research was supported by the Academy of Finland 6Genesis Flagship under Grant 318927 and, in part, by the Office of Naval Research (ONR) under Grant N00014-15-1-2709, and by the U.S. National Science Foundation under Grants CNS-1836802, CNS-1617896, and ACI-1638283.

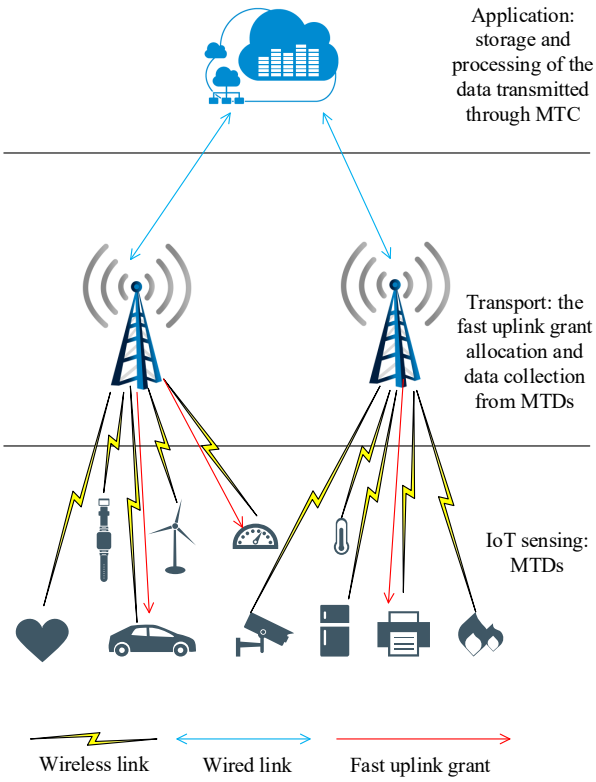

Fig. 1: An illustration of an example IoT environment in which multiple MTDs communicate with BSs connected to a cloud-based gateway.

[5] and later approved to be in the standards [6]. In a fast uplink grant scheme, MTDs do not send RA scheduling requests. Instead, the BS will actively allocate uplink resources to those MTDs. Moreover, in contrast to uncoordinated transmission, MTDs are scheduled by the BS and, hence, collisions can be avoided. The fast uplink grant is suitable for uplink resource allocation in IoT applications with stationary or low-mobility MTDs such as smart grids, smart homes, and environment monitoring. To better understand the potential of this approach for the IoT, first, we present the opportunities provided by the use of the fast uplink grant for MTC. Then, we present an overview of the associated challenges, such as predicting which MTDs have data to transmit and properly scheduling those MTDs. To address these problems, we first exploit the potential of different learning methods for source traffic prediction. In this regard, we discuss a variety of tools and machine learning algorithms that can potentially be used to predict both periodic and event-driven MTC traffic. We then shed light on the use of multi-armed bandit (MAB) theory and deep reinforcement learning (Deep RL) as effective tools for enabling fast uplink grant allocation for massive MTC scenarios. Even though short reviews of the fast uplink grant are provided in [7] and [8], to the best of our knowledge, this is the first work that analyzes how the fast uplink grant can be effectively leveraged to solve the emerging problem of massive MTCs.

The rest of the paper is organized as follows. Section II overviews the cellular RA process and its challenges. The fast uplink grant is overviewed in Section III. A two-stage fast uplink grant approach for MTC is presented in Section IV and conclusions are drawn in Section V.

II. Random Access for MTC: Overview and Challenges

A. Overview of the RA Process

The RA procedure, illustrated in Fig. 2, is the first step needed to establish an uplink connection between any cellular device and a BS [3]. In the LTE/LTE-A RA process, upon having data to transmit, each user selects one RA slot, randomly from a set of available RA slots to send a scheduling request. The number of RBs that are available for RA is limited since RA slots share the same RBs with the uplink channel. In the frequency domain, each RA slot is 1.08 MHz, which is equal to 6 LTE RBs. In LTE, each RB is a frequency-time unit with 180 kHz bandwidth and 1 ms duration. In the time domain, the time intervals for RA availability vary between every 1 ms to every 20 ms, depending on the system configuration. Once RA slots are available, cellular users randomly select one of the available RA slots to send their scheduling request. If the RA process is successful, the BS sends an RA response to the cellular user. However, if more than one device selects the same RA slot for sending a scheduling request, a collision will occur. The BS will attempt to decode the scheduling request in case of collision and will send an RA response which is received by all the users that had used the same RA slot. However, only the MTD whose data is successfully decoded at the BS can continue the RA process while others are barred and have to send a scheduling request in the next RA opportunity. Once an RA response is received, the cellular user will transmit a connection request to the BS. Finally, the BS transmits contention resolutions and, subsequently, the users transmit their data.

B. Challenges of RA in MTC

While the RA process of Section II-A is suitable for conventional human type devices, adopting it for MTC faces several challenges. For instance, the first challenge pertains to the limited number of RA opportunities in a cellular network. In fact, cellular RA resources are often much smaller than the anticipated number of MTDs in the IoT. RA efficiency is maximized when the number of RA opportunities is equal to the number of competing devices. Increasing the number of RA slots is not feasible because RA slots are allocated in the uplink channel which has limited resources, and, there should be a balance between the number of RBs allocated for the RA process and the number of resources left for uplink transmission. Hence, a small number of RA slots relative to the number of contending devices increases the probability of collisions. These collisions will make it impossible for the BS to decode RA pilots which will lead to a waste of resources and long delays since the affected MTD has to wait until the next RA opportunity to send the scheduling request again.

The second key RA challenge pertains to the short data packets size in MTC compared to conventional cellular services. For example, using an RA slot (six RBs) for sending a scheduling request to transmit a short data packet that

might require only one RB, is highly inefficient. The signaling overhead for RA is no longer negligible compared to the actual size of the data packets that will be transmitted by MTDs. Therefore, it is desirable to develop MTC scheduling mechanisms with low signaling overhead.

C. Overview of Existing Access Solutions for MTC

To address the aforementioned challenges of RA for MTC, several recent solutions have been proposed, as extensively reviewed in [3]. The first class of solutions focuses on coordinated transmissions. One solution that is designed for access control in congestion situations is access class barring (ACB) [9]. In ACB, each device has a class and under *special circumstances*, some of those classes are barred from access attempts by broadcasting an ACB parameter by the BS. While ACB solves the problem of RA channel congestion, it can potentially produce excessive delays due to the long waiting times of barred MTDs. Moreover, the ACB scheme can yield a high RA signaling overhead. Another approach to improve RA for MTC is to use an access backoff process. In this method, the BS encourages MTDs to not send a scheduling request for a time duration. However, by doing so, it increases latency. Another alternative solution is the notion of slotted RA in which each MTD is allocated a fixed RA opportunity to transmit only on that slot. However, slotted RA is not suitable for massive MTC access since the periodicity of the RA slots will be large and, hence, incurring long delays. Moreover, if an MTD does not send an RA pilot, the RA slot is wasted. Other methods have also been proposed for improving RA such as pull-based RA in which MTDs wait for permissions from BS to send RA pilots, and, priority-based RA in which specific RA priorities are assigned for devices. A comprehensive list of such methods and their advantages and disadvantages is presented in [3]. However, all of these methods still exhibit high signaling overhead, collisions, and long delays which limits their applicability to MTC.

The second class of solutions for the access problem in MTC focus on the *uncoordinated transmissions* [4], where MTDs do not send scheduling requests. Instead, the MTDs will select a random RB to transmit data [4]. Obviously, this method reduces the signaling level and can potentially increase the efficiency of the system. However, uncoordinated MTCs will still heavily suffer from collisions since the largely constrained uplink resources are shared among a potentially large number of MTDs. Moreover, to realize uncoordinated MTC in practice, there is a need to design complicated receiver structures and retransmission mechanisms. Another major issue is scalability in terms of the number of supported devices since efficiency is maximized when the number of devices is equal to the number of resources. Clearly, in massive MTC, such a requirement is not met, and, hence, the performance of the system will suffer. These drawback of uncoordinated transmissions limit the scope of their applicability to MTC, in general, and massive MTC, in particular.

III. Fast Uplink Grant

A balanced approach between a conventional coordinated access and uncoordinated transmissions can be developed by using the *fast uplink grant*, as shown in Fig. 2.

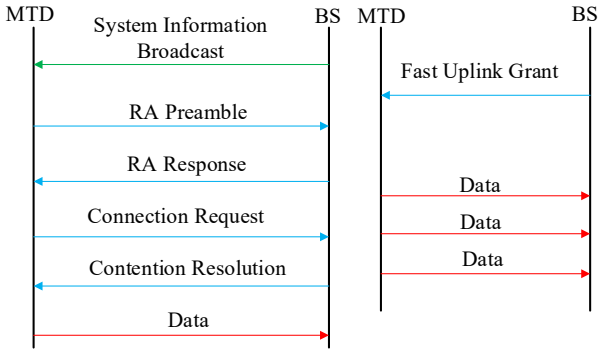

Fig. 2: Comparison of the scheduling process in conventional RA (left) and in a fast uplink grant process (right).

A. Fast Uplink Grant: Definition and Opportunities

The fast uplink grant was introduced in [5] as an effective process that a cellular BS can use to select an MTD and allocate uplink resources to it. As such, by using the fast uplink grant, the MTDs will no longer need to perform an RA process. Instead, whenever an MTD has data to transmit, it can simply wait for obtaining a fast uplink grant.

The fast uplink grant presents several benefits compared to the previously discussed approaches. First, the amount of signaling that is required is much less than RA. This is due to the fact that, by using the fast uplink grant:

- Only one level of signaling is performed.
- The amount of signaling is minimal since the fast uplink grant for the entire system can be sent in one broadcast message.

Second, in a system with a large number of devices, collisions of RA pilots in coordinated access and packet collisions in uncoordinated transmission can be overcome by using the fast uplink grant. The benefits of the fast uplink grant can be summarized as follows:

- RA congestion is mitigated.
- RA radio resources can be used to transmit uplink data and, hence, a larger number of devices can be supported at each time, this will also decrease the latency of the entire system.
- Packet collisions of uncoordinated transmission are avoided.
- The BS can satisfy the heterogeneous QoS requirements of MTDs by performing an active allocation of the fast uplink grant to MTDs with stricter latency requirements.
- MTDs can save energy by eliminating the need for an RA process and for retransmission of scheduling requests in case of RA failure.
- The RA process delay and delay of waiting for next RA opportunity in case of collisions are eliminated.

A summary of the differences between the fast uplink grant and conventional schemes is given in Table I.

B. Challenges of the Fast Uplink Grant

The first drawback of the fast uplink grant is the possibility of wasting resources whenever an MTD that receives a fast

TABLE I: Coordinated vs uncoordinated vs fast uplink grant

	Signaling	**Collisions**	**Latency**
Coordinated	High (6 RBs) + messages 2 to 4	High (number of MTDs >> number of RA slots)	Waiting for RA slot + RA signaling
Uncoordinated	Zero	High(number of MTDs >> number of RBs)	High (when number of MTDs >> number of RBs)
Fast uplink grant	Small (one broadcast message for entire cell)	Zero	Small (a few ms if grant is allocated on time)

uplink grant does not have data to transmit [7]. Moreover, if the fast uplink grant is not received within the maximum tolerable access delay of the data packets, packets will be dropped, yielding transmission failures. This can potentially lead to unreliable and high latency MTD transmissions. Hence, to adopt the fast uplink grant for MTC, one of the main challenges that must be overcome is the optimal selection of MTDs by the BS. This MTD selection process, in turn, faces two key challenges. First, there is a need to predict the set of MTDs that will likely have data to transmit, at any given time. By doing accurate predictions, the BS can solve the problem of allocating the fast uplink grant to silent MTDs. Once predictions are properly implemented, the BS must also be able to determine the scheduling sequence of MTDs. This challenge is particularly pronounced when the number of devices significantly exceeds the number of resources. Hence, sophisticated scheduling algorithms are needed to enable fast uplink grant allocation. Next, we propose a two-stage approach for leveraging the fast uplink grant for MTCs.

IV. Proposed Two-Stage Solution

A. Source Traffic Prediction in MTC

As stated previously, if an MTD is selected for the fast uplink grant and does not have data to transmit, uplink radio resources are wasted. To address this challenge, the BS must implement advanced traffic prediction mechanisms to predict the set of MTDs that have data to transmit. Most of the prior art on traffic modeling for MTC is focused on aggregate traffic modeling at the BS. Such traffic modeling only estimates the number of devices or the number of packets arriving in the system. However, source traffic modeling is a fundamentally different problem since we are interested in precisely predicting which MTDs will enter the network. In essence, at each time slot, the BS needs to predict which MTDs will have traffic to send and, hence, need uplink resources. Since the majority of MTDs are in *idle mode* most of the time, such a source traffic prediction becomes more challenging. However, such predictions for idle mode MTDs are generally feasible in an IoT due to two facts: a) most of the MTDs are stationary or exhibit low mobility and therefore, we can assume that the BS has location information of the idle mode MTDs and, b) the set of MTDs communicating with a BS is often fixed. Alternatively, this prediction can also be done across multiple BSs in a central controller that keeps track of the BSs and their associated MTDs.

For traffic predictions, one must distinguish between two types of MTC traffic: periodic reporting and event-driven transmissions. In periodic reporting, MTDs periodically transmit data packets at specific, pre-determined times. In event-driven traffic, often, a large number of MTDs will initiate a transmission request to provide reports on a certain IoT event. Clearly, the prediction of event-driven traffic is much harder than periodic traffic. Next, we present mathematical tools that can be used to develop algorithms for prediction of *both MTC traffic types.*

1) Prediction of Periodic Traffic: Many IoT applications, such as smart metering and environment sensing, rely on MTDs that *periodically* transmit sensory data generated from the observations of the physical environment. Different applications generate heterogeneous data sizes in various durations. These durations could be as low as a few milliseconds and up to once a month. An MTD might also transmit data pertaining to multiple IoT applications. This results in different data packets with different transmission intervals. Hence, the BS must learn the exact time instances at which any given MTD will generate its data, as well as the associated packet size. Clearly, the BS must collect data from the past transmissions of all MTDs and subsequently use machine learning algorithms to predict the source traffic for each MTD. This prediction must be precise, since some IoT applications generate data with very strict latency requirements, as low as 10 ms. Mathematical methods such as a non-homogeneous Poisson process (NHPP) could be used to model the arrival rate of packets to the queue of each MTD at different times. In an NHPP, arrivals follow a Poisson distribution, however, at each time, the rate of arrival is different. Such pattern analysis is called calendar-based periodic pattern mining and models such as the sequential association rule and the calendar association rule exist for analyzing them (e.g, see [10]).

2) Prediction of Event-Driven Traffic: In IoT applications that rely on event-driven MTC, whenever an event occurs, several MTDs that detect the event must initiate data transmission to the BS. This leads to a burst of RA scheduling requests from a large number of MTDs. Such event-driven MTC traffic will exacerbate the challenges pertaining to scheduling a large number of MTDs (identified in Section II-B). Hence, effective traffic prediction in event-driven MTCs is critical. Naturally, predicting an IoT event that was never observed is not possible[1]. However, it is possible to detect an event based on unusual traffic generated by MTDs. If the BS, based on the data gathered from previous IoT events, can calculate the likelihood with which other MTDs face the same event, it will be possible to design algorithms to predict event-driven traffic.

Here, we present novel methods that can be used for source traffic prediction in event-driven MTCs. First, we assume that, during past events, the BS has collected sufficient data about the transmission of MTDs. That is, the BS knows which

[1]The BS can use these uncoordinated events as side information for better aggregate scheduling or potential future predictions.

devices were transmitting during each event along with their order of transmission. Second, we assume that the set of MTDs with periodic traffic is predicted and MTDs do not send scheduling requests for periodic reporting. Hence, any scheduling request can be considered as an event trigger and used for detection of events. We could also consider that, once an event happens, MTDs wait for a short period of time for an uplink grant, if they do not receive it, they use RA. A flowchart of decision making at MTD for RA is given in Fig. 4. Now, once an event happens, some MTDs will report it earlier than others. The BS considers the first RA request as an event trigger. The event-driven traffic prediction problem is now reduced to the following question: *Once a specific MTD detects an event, which other MTDs will experience the same event with a high probability?*

Answering this question requires analysis of the data collected from previous events. One natural solution is to use probabilistic models from machine learning. Using previously collected data, a probabilistic relationship between two MTDs facing the same event can be calculated. Another possible solution is to use the paradigm of *causality*. Causality deals with the following problem: Given that an MTD detects an event, which other MTDs that specific MTD *statistically causes* to detect the same event. A novel method for causal inference is based on the concept of *directed information* [11] which can be used to infer causality between sequences of random variables. Considering two sequences of random variables, past and present values of the first sequence, and past values of the second sequence can be used to evaluate the present value of the second sequence. Directed information is a powerful method that is used for prediction of seizure in epilepsy patients and causality between neurons of the human brain. We can model the transmission history of each MTD with a sequence of binary random variables, where transmitting at each time is presented by 1 and being silent with 0. Fig. 3 presents the values of directed information between two sequences of binary random variables with event length 12. Directed information is calculated between pairs of length two of the sequence X to sequence Y and vice versa. Fig. 3 presents the amount of the flow of information between two sequences of transmission history of MTDs which can be used for inferring the causality and hence, source traffic prediction. Details of this method are outlined in [12]. It is clear from Fig. 3 that flow of information from X to Y is higher than Y to X. Once causality is inferred, one can predict which MTDs face the same IoT event and start allocating the fast uplink grant to them. In Fig. 4, we present the flowchart of an algorithm that can use event-detection in the BS for fast uplink grant allocation. We note that the data analytics for source traffic prediction can be performed *in an offline manner* and, hence, the complexity of the source traffic prediction algorithms is not a major concern.

B. Optimal Fast Uplink Grant Allocation

Once the set of MTDs that have data to transmit is predicted, if the number of devices is smaller than the number of available RBs, all the MTDs can be scheduled to transmit.

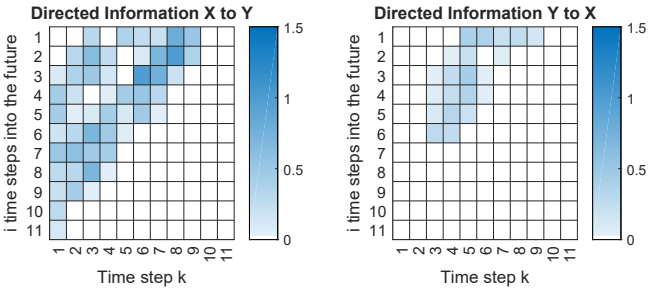

Fig. 3: Directed information between sequences of length two of two binary random sequences. Transmission history of each MTD is presented by a sequence of binary random variables. In sequence X, binary value randomly takes values of 1 at times $\{1, 2, 3, 4, 6, 7, 8, 9\}$ and in sequence Y at times $\{4, 5, 6, 8, 9, 10, 11\}$. $I(X_{k,k+1} \rightarrow Y_{k+i-1,k+i})$ is presented in element (i, k) of $I(X \rightarrow Y)$ [12].

However, if the number of MTDs exceeds the number of available RBs at any given time, the network must select which MTDs can be granted access. MTDs should be scheduled based on their QoS requirements, namely, their maximum tolerable delay. If fast uplink grants are allocated randomly, it is possible that the network may prioritize the scheduling of delay-tolerant MTD data, thus jeopardizing the performance of delay-sensitive MTD data. If the BS has full knowledge of the QoS requirements of all MTDs, this scheduling can be performed in a centralized manner. However, in a realistic scenario, such information might not be available to the BS and any fast uplink grant allocation algorithm should be able to select MTDs in an uncertain environment. Therefore, the design of sophisticated algorithms for optimal fast uplink grant allocation is needed. Here, we present some initial directions toward building such scheduling algorithms that exploit recent advances in machine learning and artificial intelligence to optimize the allocation of fast uplink grants to MTDs [13]. Each MTD should be allowed to transmit in a given frequency band until they finish their transmission. This can help in dealing with various data packets sizes of IoT applications.

1) Multi-Armed Bandit Theory: MABs are a class of reinforcement learning (RL) problems that deal with decision making in uncertain environments with limited or no prior information [14]. The basic MAB problem consists of a set of arms (available actions) that can be chosen by a decision-making agent that plays an arm at each time and receives a reward. The rewards are drawn from an unknown probability distribution. The agent has no prior information about the rewards of each arm and has to randomly select arms, observe the rewards, and, then, try to find the best possible arm. In MAB, the notion of *regret* – defined as the difference between the best possible arm that could have been played and the arm that is selected – is used as a measure of performance. The main goal of any MAB algorithm is to minimize the cumulative regret over time. To solve conventional MAB problems, algorithms such as ϵ-greedy and upper confidence bound (UCB) are often adopted [14]. There are also special MAB problems such as *sleeping MAB* problems in which, at each time, only a subset of arms is available for the agent, or *contextual MAB* where at each time there is some side infor-

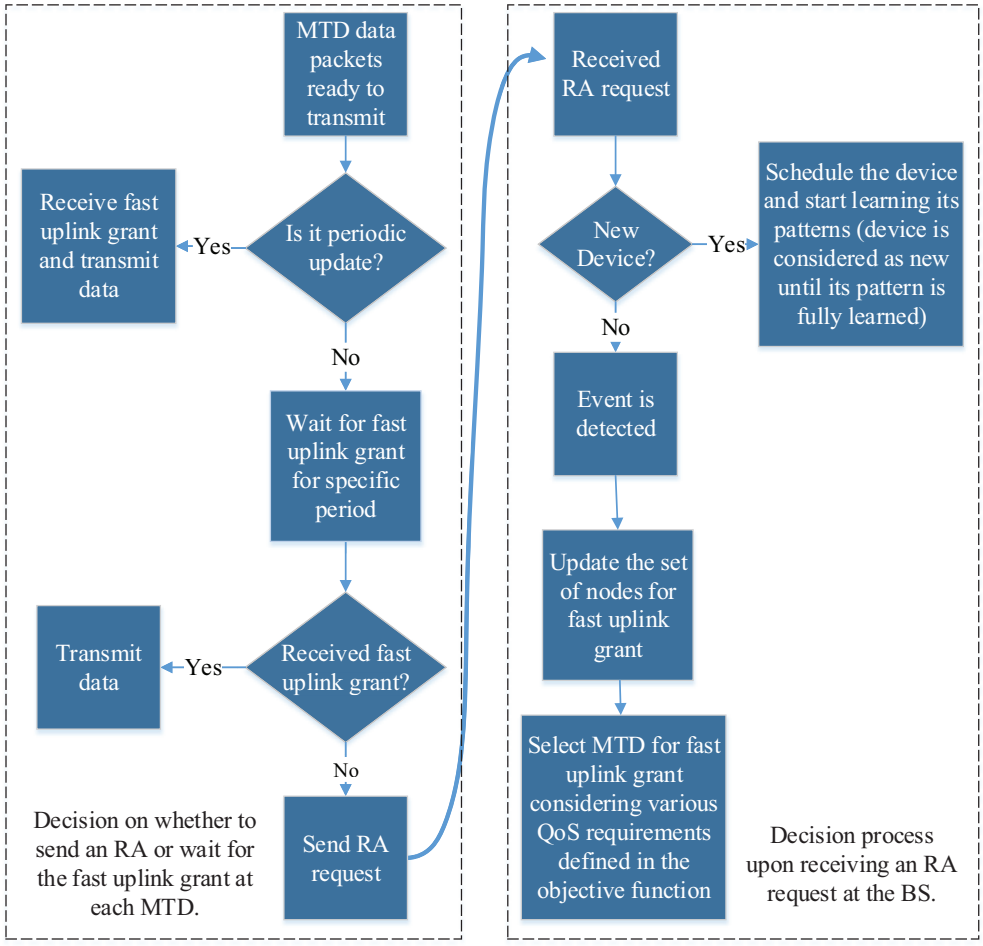

Fig. 4: Flowchart of the proposed algorithm for RA decision making as implemented by any given MTD (left) and event detection and uplink grant allocation at the BS (right).

mation provided to the decision maker. Clearly, such problems are apropos for addressing the MTD selection problem. For example, the sleeping MAB framework is particularly suitable to select MTDs for the fast uplink grant, since the availability of MTDs can change at each time. Here, we note that the MAB reward functions can be defined in terms of the various QoS metrics that the system seeks to optimize. A sleeping MAB fast uplink grant allocation is presented for MTC in [15]. In Fig. 5, we show the performance of the probabilistic sleeping MAB method that is used to learn the maximum tolerable access delay requirements of the MTDs in a system. The proposed method in [15] can achieve up to three-fold performance gain in terms of achieved access delay compared to a random fast uplink grant allocation policy.

2) Deep Reinforcement Learning: Deep RL is used in RL problems having an extremely large action-state space where it is not possible to explore all the possible states and actions. In deep RL, neural networks are used to approximate the environment, and, for the states that were not seen before, the neural network output determines the action [13]. To use deep RL for MTD selection using the fast uplink grant, one can first formulate the problem using a Markov decision process (MDP) [14]. In this MDP formulation, each state is a combination of the set of available MTDs and their

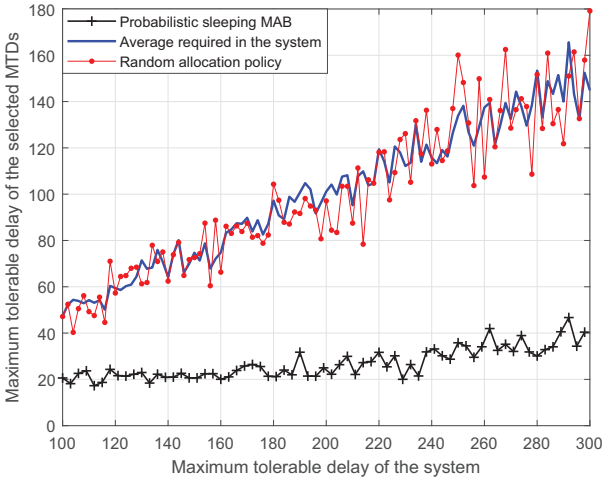

Fig. 5: Sleeping MABs for learning the maximum tolerable access delay of the MTDs and optimal allocation of fast uplink grants in terms of satisfying the QoS requirements of the IoT applications [15].

associated QoS requirements and each MDP action is a subset of the set of available MTDs. For each selected subset, with some probability, some of the MTDs with have successful

transmission, and some will fail. QoS requirements of the failed and not selected MTDs might change due to time. In the next time step, some new MTDs will be active. Therefore, each action will move the system to a new state, that is the new set of available MTDs with different QoS requirements. Clearly, MTC scheduling problems with a large number of MTDs will have to deal with very large action and state spaces and one can use deep RL to find the optimal action for each given set.

V. Conclusion

In this paper, we have studied the potential of incorporating the fast uplink grant as an enabler for massive MTCs in the IoT. First, we have reviewed the challenges faced by conventional access schemes in MTC and discussed the merits of the fast uplink grant. Then, we have presented the shortcomings of the fast uplink grant, and outlined solutions to address them. In particular, we have elaborated on the methods for source traffic prediction, for both periodic and event-driven traffic. Then, we have proposed machine learning techniques for the optimal selection of MTDs to be used in the fast uplink grant. In a nutshell, this work can be thought of as a stepping stone towards a better understanding of how the fast uplink grant can be effectively leveraged for massive MTCs.

References

[1] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of things for smart cities," *IEEE Internet of Things Journal*, vol. 1, no. 1, pp. 22–32, Feb 2014.

[2] Z. Dawy, W. Saad, A. Ghosh, J. G. Andrews, and E. Yaacoub, "Toward massive machine type cellular communications," *IEEE Wireless Communications*, vol. 24, no. 1, pp. 120–128, February 2017.

[3] 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 Communications Surveys and Tutorials*, vol. 16, no. 1, pp. 4–16, January 2014.

[4] D. Zucchetto and A. Zanella, "Uncoordinated access schemes for the IoT: Approaches, regulations, and performance," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 48–54, September 2017.

[5] 3GPP, "Study on latency reduction techniques for LTE," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 36.881.

[6] ——, "Medium access control (MAC) protocol specification, release 15," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 36.321.

[7] C. Hoymann, D. Astely, M. Stattin, G. Wikstrom, J.-F. Cheng, A. Hoglund, M. Frenne, R. Blasco, J. Huschke, and F. Gunnarsson, "LTE release 14 outlook," *IEEE Communications Magazine*, vol. 54, no. 6, pp. 44–49, 2016.

[8] P. Schulz, M. Matthe, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S. A. Ashraf, B. Almeroth, J. Voigt, I. Riedel, A. Puschmann, A. Mitschele-Thiel, M. Muller, T. Elste, and M. Windisch, "Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 70–78, February 2017.

[9] 3GPP, "Technical specification group services and system aspects; service accessibility," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 22.011.

[10] J. Adhikari and P. Rao, "Identifying calendar-based periodic patterns," *Emerging paradigms in machine learning*, pp. 329–357, 2013.

[11] J. Massey, "Causality, feedback and directed information," in *Proc. Int. Symp. Inf. Theory Applic. (ISITA-90)*, Waikiki, Hawaii, 1990, pp. 303–305.

[12] S. Ali, W. Saad, and N. Rajatheva, "A directed information learning framework for event-driven M2M traffic prediction," *IEEE Communications Letters*, vol. 22, no. 11, pp. 2378–2381, Nov 2018.

[13] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Machine learning for wireless networks with artificial intelligence: A tutorial on neural networks," *arXiv preprint arXiv:1710.02913*, 2017.

[14] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[15] S. Ali, A. Ferdowsi, W. Saad, and N. Rajatheva, "Sleeping multi-armed bandits for fast uplink grant allocation in machine type communications," in *Proc. IEEE Global Communications Conference (GLOBECOM), Workshop on Ultra-High Speed, Low Latency and Massive Connectivity Communication for 5G/B5G*, Abu Dhabi, UAE, Dec 2018, pp. 1–6.