Hybrid Scheduling in Heterogeneous Half- and Full-Duplex Wireless Networks

Tingjun Chen*, Jelena Diakonikolas†, Javad Ghaderi*, and Gil Zussman*
*Electrical Engineering, Columbia University, New York, NY
†Computer Science, Boston University, Boston, MA, USA
Email: {tingjun, jghaderi, gil}@ee.columbia.edu, jelenad@bu.edu

***Abstract*—Full-duplex (FD) wireless is an attractive communication paradigm with high potential for improving network capacity and reducing delay in wireless networks. Despite significant progress on the physical layer development, the challenges associated with developing medium access control (MAC) protocols for *heterogeneous* networks composed of both legacy half-duplex (HD) and emerging FD devices have not been fully addressed. Therefore, we focus on the design and performance evaluation of scheduling algorithms for infrastructure-based *heterogeneous* networks (composed of HD and FD users). We develop the *hybrid* Greedy Maximal Scheduling (H-GMS) algorithm, which is tailored to the special characteristics of such heterogeneous networks and combines both centralized GMS and decentralized Q-CSMA mechanisms. Moreover, *we prove that H-GMS is throughput-optimal*. We then demonstrate by simple examples the benefits of adding FD nodes to a network. Finally, we evaluate the performance of H-GMS and its variants in terms of throughput, delay, and fairness between FD and HD users via extensive simulations. We show that in heterogeneous HD-FD networks, H-GMS achieves 5–$10\times$ better delay performance and improves fairness between HD and FD users by up to 50% compared with the fully decentralized Q-CSMA algorithm.**

***Index Terms*—Full-duplex wireless, scheduling, distributed throughput maximization**

I. Introduction

Full-duplex (FD) wireless – an emerging wireless communication paradigm in which nodes can simultaneously transmit and receive on the same frequency – has attracted significant attention [1]. Recent work has demonstrated physical layer FD operation [2]–[5], and therefore, the technology has the potential to increase network capacity and improve delay compared to legacy half-duplex (HD) networks. Based on the advances in integrated circuits-based implementations that can be employed in mobile nodes (e.g., [4], [5]), we envision a gradual but steady replacement of existing HD nodes with the more advanced FD nodes. During this gradual penetration of FD technology, the medium access control (MAC) protocols will need to be carefully redesigned to not only support a *heterogeneous* network of HD and FD nodes but also to guarantee fairness to the different node types.

Therefore, we focus on the design and performance evaluation of scheduling algorithms for heterogeneous HD-FD networks. In particular, we consider infrastructure-based random-access networks (e.g., IEEE 802.11) consisting of an FD access point (AP) and both HD and FD users in a single collision domain. Further, we consider a single channel which is shared by all the uplinks (ULs) and downlinks (DLs) between the AP and the users. To focus on fundamental limits due to the incorporation of FD nodes and to expose the main features of our scheduling algorithms, we assume perfect self-interference cancellation (SIC) at FD nodes. Yet, we expect that the results can be extended to more realistic settings by incorporating imperfect SIC.

Traditionally, three approaches have been used for the design of wireless scheduling algorithms that guarantee maximum throughput:

Maximum Weight Scheduling (MWS) [6], which relies on the queue-length information and schedules non-conflicting links with the maximum total queue length. In contrast to the all-HD networks where only a single link can be scheduled at a time, in the considered setting the UL and the DL of any FD user can be scheduled simultaneously. Thus, to implement MWS, queue-length information needs to be shared between each FD user and the AP, which requires significant overhead.

Greedy Maximal Scheduling (GMS) [7], which is a centralized policy that greedily selects the link with the longest queue, disregards all conflicting links, and repeats the process. Typically, GMS has better delay performance than MWS and Q-CSMA. Although GMS is equivalent to MWS in an all-HD network, in general, it is not equivalent to MWS and is not throughput-optimal in general topologies.

Queue-based random-access algorithms (Q-CSMA) (e.g., [8], [9]), which are fully distributed and do not require sharing of the queue length information between the users and the AP. These algorithms have been shown to achieve throughput optimality. However, they generally suffer from excessive queue lengths that lead to long delays.

In this paper, we show that a combination of the two latter approaches guarantees maximum throughput and provides good delay performance in heterogeneous HD-FD networks. Specifically, we first show by using the notion of Local Pooling [7], [10] that in the considered networks GMS guarantees maximum throughput. However, since GMS still requires some level of centralization, we leverage ideas from distributed Q-CSMA to develop the Hybrid-GMS (H-GMS) algorithm. *Hybrid* represents the combination of centralized GMS and distributed Q-CSMA, and instead of approximating MWS (as done in "traditional" Q-CSMA), it approximates GMS.

Moreover, the design of H-GMS leverages the fact that in infrastructure-based networks, the AP has access to all the DL

queues and can resolve the contention among the DL queues (e.g., using longest-queue-first). In contrast, the users do not have access to all DL queues or to other UL queues, and therefore, must share the medium in a distributed manner, while ensuring FD operation when possible.

We prove the throughput optimality of H-GMS (namely that it can support any rate vector in the capacity region of heterogeneous HD-FD networks) by using the fluid limit technique. In contrast to the classical Q-CSMA, the contention resolution of DL queues at the AP under the H-GMS algorithm can force a schedule that is not with maximum weight. Hence, we make a connection to GMS in fluid limits (which, as mentioned above, is throughput-optimal in heterogeneous HD-FD networks). We also present variants of H-GMS with different degrees of centralization that affect the delay performance.

Before thoroughly evaluating H-GMS and its variants, we evaluate the benefits of introducing FD-capable users into an all-HD network in terms of both network and individual throughput gains. Compared to the all-HD network, the considered heterogeneous HD-FD network can potentially have doubled throughput for certain rate vectors within the capacity region, while the network throughput gain generally depends on both the number of FD users and the particular rate vector in which the network operates. Using simple examples, we show that when all links have equal rate, the throughput gain of the HD-FD network over the all-HD network increases as the number of FD users increases. Moreover, when all users are FD-capable, the network throughput gain is exactly two. We also demonstrate that it is generally possible for all users to experience improved individual throughput at the cost of lowering the priority of FD users, revealing an interesting *fairness-efficiency tradeoff* in HD-FD networks.

Finally, we present extensive simulation results to evaluate the different variants of the H-GMS algorithm and compare them to the classical Q-CSMA algorithm. We primarily focus on delay performance and fairness between FD and HD users, but also show throughput gains. We consider a wide range of arrival rates and varying number of FD users. The results show that in heterogeneous HD-FD networks, H-GMS achieves 5–$10\times$ better delay performance and improves fairness between HD and FD users by up to 50% compared to the fully distributed Q-CSMA algorithm. This delay and fairness improvement results from the different degrees of centralization at the AP. Further, we discuss the different variants and how different degrees of centralization at the AP affect the delay performance, and show that a higher degree of centralization at the AP (e.g., H-GMS-E) can result in better fairness between the FD and HD users.

To summarize, *the main contribution of this paper is the design and evaluation of a distributed scheduling algorithm for infrastructure-based heterogeneous HD-FD networks that guarantees maximum throughput*. The algorithm has a relatively good delay performance and to the best of our knowledge is the first such algorithm with rigorous performance guarantees in HD-FD networks.

II. Related Work

There has been extensive work dedicated to physical layer FD radio/system design [2]–[4], [11] (see also the review in [1] and references therein). Recent research also focused on characterizing and quantifying achievable throughput improvements and rate regions of FD networks in both single-channel and multi-channel cases with realistic imperfect SIC [12]–[14]. However, these papers consider only simple network scenarios consisting of up to two links.

Most of the existing MAC layer studies focused on *homogeneous* networks [15]–[19] considering signal-to-noise ratio (SNR) or a specific standard (e.g., IEEE 802.11). For example, [16] considered an IEEE 802.11 network with an FD-capable AP and HD users, and proposed an SNR-based distributed MAC protocol. As another example, [15] considered an all-FD network and proposed a distributed MAC protocol. Most relevant to our work is [20], which proposed a MAC layer algorithm for a heterogeneous HD-FD network based on IEEE 802.11 and analyzed its throughput. However, to the best of our knowledge, the only previous work that provided MAC design with provable performance guarantees is [19], which focused on scheduling in multi-hop random-access all-FD networks. Moreover, to the best of our knowledge, the fairness between users that have different HD/FD capabilities was not considered before.

III. Model and Preliminaries

A. Network Model

We consider a single-channel, *heterogeneous* wireless network consisting of one AP and N users, in which there is a UL and a DL between each user and the AP. The set of users is denoted by $\mathcal{N}$. The AP is FD, while N_F of the users are FD and $N_H = N - N_F$ are HD. Without loss of generality, we index the users by $[N] = \{1, 2, \cdots, N\}$ in which the first N_F indices correspond to FD users and the remaining N_H indices correspond to HD users. The sets of FD and HD users are denoted by $\mathcal{N}_F$ and $\mathcal{N}_H$, respectively. We consider a collocated network where the users are within the communication range of each other and the AP. The network can be represented by a directed star graph $G = (\mathcal{V}, \mathcal{E})$ with the AP at the center and two links between AP and each user in both directions. Thus, we have $\mathcal{V} = \{\text{AP}\} \cup \mathcal{N}$ (with $|\mathcal{V}| = 1 + N$) and $|\mathcal{E}| = 2N$.

B. Traffic Model, Schedule, and Queues

We assume that time is slotted and packets arrive at all UL and DL queues according to some stochastic process. For brevity, we will use superscript $j \in \{\text{u}, \text{d}\}$ to denote the UL and DL of a user. Let l_i^j denote link j (UL or DL) of user i, each of which is associated with a queue Q_i^j. We use $A_i^j(t) \leq A_{\max} < \infty$ to denote the number of packets arriving at link j (UL or DL) of user i in slot t. The arrival process is assumed to have a well-defined long-term rate of $\lambda_i^j = \lim_{T\to+\infty} \frac{1}{T} \sum_{t=1}^{T} A_i^j(t)$. Let $\boldsymbol{\lambda} = [\lambda_i^{\text{u}}, \lambda_i^{\text{d}}]_{i=1}^N$ be the arrival rate vector on the ULs and DLs.

All the links are assumed to have capacity of one packet per time slot and the SIC at all the FD-capable nodes is *perfect*.[1] A *schedule* at any time slot t is represented by a vector

$$\mathbf{X}(t) = [X_1^{\mathrm{u}}(t), X_1^{\mathrm{d}}(t), \cdots, X_N^{\mathrm{u}}(t), X_N^{\mathrm{d}}(t)] \in \{0,1\}^{2N},$$

in which $X_i^{\mathrm{u}}(t)$ (resp. $X_i^{\mathrm{d}}(t)$) is equal to 1 if the UL (resp. DL) of user i is scheduled to transmit a packet in time slot t and $X_i^{\mathrm{u}} = 0$ (resp. $X_i^{\mathrm{d}} = 0$), otherwise. We denote the set of all feasible schedules by $\mathcal{S}$. Let $\mathbf{e}_i \in \{0,1\}^{2N}$ be the i^{th} basis vector (i.e., an all-zero vector except the i^{th} element being one). Since a pair of UL and DL of the same FD user can be activated at the same time, we have:

$$\mathcal{S} = \{\mathbf{0}\} \cup \{\mathbf{e}_{2i-1}, \mathbf{e}_{2i}, \forall i \in \mathcal{N}\} \cup \{\mathbf{e}_{2i-1} + \mathbf{e}_{2i}, \forall i \in \mathcal{N}_F\}.$$

Choosing $\mathbf{X}(t) \in \mathcal{S}$, the queue dynamics are described by:

$$Q_i^j(t) = [Q_i^j(t-1) + A_i^j(t) - X_i^j(t)]^+, \ \forall t \geq 1,$$

in which $[\cdot]^+ = \max(0, \cdot)$. We use $\mathbf{Q}(t) = [Q_i^{\mathrm{u}}(t), Q_i^{\mathrm{d}}(t)]_{i=1}^N$ to denote the queue vector. We also use $\mathbb{1}(\cdot)$ to denote the indicator function.

C. Capacity Region and Throughput Optimality

The capacity region of the network is defined as the set of all arrival rate vectors for which there exists a scheduling algorithm that can stabilize the queues. It is known that in general, the capacity region is the convex hull of all feasible schedules [6]. Therefore, the capacity region of the heterogeneous HD-FD network is given by $\Lambda_{\mathrm{HD-FD}} = \mathrm{Co}(\mathcal{S})$, where $\mathrm{Co}(\cdot)$ is the convex hull operator. It is easy to see that this capacity region can be equivalently characterized by the following set of linear constraints[2]:

$$\Lambda_{\mathrm{HD-FD}} = \{\boldsymbol{\lambda} \in [0,1]^{|\mathcal{E}|} : \sum_{i\in\mathcal{N}_F} \max\{\lambda_i^{\mathrm{u}}, \lambda_i^{\mathrm{d}}\} + \sum_{i\in\mathcal{N}_H} (\lambda_i^{\mathrm{u}} + \lambda_i^{\mathrm{d}}) \leq 1\}. \quad (1)$$

Let a network in which all the users and the AP are only HD-capable be the *benchmark all-HD network*, whose capacity region is given by $\Lambda_{\mathrm{HD}} = \mathrm{Co}(\mathbf{e}_1, \cdots, \mathbf{e}_{2N})$, or equivalently

$$\Lambda_{\mathrm{HD}} = \{\boldsymbol{\lambda} \in [0,1]^{|\mathcal{E}|} : \textstyle\sum_{i\in\mathcal{N}} (\lambda_i^{\mathrm{u}} + \lambda_i^{\mathrm{d}}) \leq 1\}. \quad (2)$$

A scheduling algorithm is called *throughput-optimal* if it can keep the network queues stable for all arrival rate vectors $\boldsymbol{\lambda} \in \mathrm{int}(\Lambda)$ in which $\mathrm{int}(\Lambda)$ denotes the interior of Λ.

To compare $\Lambda_{\mathrm{HD-FD}}$ with Λ_{HD} and quantify the network throughput gain when a certain number of HD users become FD-capable, similar to [12], we define the *capacity region expansion function* $\gamma(\cdot)$ as follows. Given $\boldsymbol{\lambda}_0$ on the Pareto boundary of Λ_{HD}, the capacity region expansion function at point $\boldsymbol{\lambda}_0$, denoted by $\gamma(\boldsymbol{\lambda}_0)$, is defined as

$$\gamma(\boldsymbol{\lambda}_0) = \sup\{\zeta > 0 : \ \zeta \cdot \boldsymbol{\lambda}_0 \in \Lambda_{\mathrm{HD-FD}}\}. \quad (3)$$

$\gamma(\cdot)$ can be interpreted as a function that scales an arrival rate vector on the Pareto boundary of Λ_{HD} to a vector on the Pareto boundary of $\Lambda_{\mathrm{HD-FD}}$, as N_F users become FD-capable. It is not hard to see that $\gamma : \Lambda_{\mathrm{HD}} \to [1,2]$.

[1]We remark that imperfect SIC can also be incorporated into the model by letting the corresponding link capacity be $c_i^j \in (0,1)$. For simplicity and analytical tractability, we assume $c_i^j = 1, \ \forall i \in \mathcal{N}$, throughout this paper.

[2]It is straightforward to only use linear inequalities, by replacing $\max\{\lambda_i^{\mathrm{u}}, \lambda_i^{\mathrm{d}}\}$ with λ_i and adding linear inequalities $\lambda_i^{\mathrm{u}} \leq \lambda_i, \lambda_i^{\mathrm{d}} \leq \lambda_i$.

Algorithm 1 GMS for HD-FD Networks (in slot t)

1. Initialize $\mathbf{X}(t) = \mathbf{0}$.
2. Select link $l^\star \in \mathcal{E}$ with the largest queue length (i.e., $l^\star = \arg\max_{i\in\mathcal{N},\ j\in\{\mathrm{u},\mathrm{d}\}} \{Q_i^j(t)\}$). If the longest queue is not unique, break ties uniformly at random.
3. • If $l^\star = l_i^{\mathrm{u}}$ or l_i^{d} for some $i \in \mathcal{N}_F$, set $X_i^{\mathrm{u}}(t) = X_i^{\mathrm{d}}(t) = 1$;
   • If $l^\star = l_i^j$ for some $i \in \mathcal{N}_H$ and $j \in \{\mathrm{u}, \mathrm{d}\}$, set $X_i^j(t) = 1$.
4. Use $\mathbf{X}(t)$ as the transmission schedule in slot t.

IV. Scheduling Algorithms and Main Result

In this section, we develop a hybrid scheduling algorithm tailored for heterogeneous HD-FD networks. We first use Local Pooling [7], [10] to prove that GMS is throughput-optimal in the considered networks, and therefore, MWS [6] is unneeded. Based on that, we present the H-GMS algorithm which is a decentralized version of GMS, that leverages ideas from distributed Q-CSMA [8], [9]. H-GMS uses information about the DL queues, available at the AP, but does not require global information about the UL queues. We state the main result (Theorem 1) about the throughput optimality of H-GMS and describe various implementations of H-GMS with different levels of centralization. We later show (in Section VII) that they have different delay performance.

A. Centralized Greedy Maximal Scheduling (GMS)

We first show that a (centralized) GMS, as described in Algorithm 1, is throughput-optimal in *any* collocated heterogeneous HD-FD network, independent of the values of N_F and N_H. In this algorithm, a pair of FD UL and DL is always scheduled at the same time, as such a schedule yields a higher throughput than scheduling only the UL or only the DL.

Proposition 1. *The greedy maximal scheduling (GMS) algorithm is throughput-optimal in any collocated heterogeneous HD-FD network.*

The proof is based on [7, Theorem 1], [10], and the fact that the interference graph of *any* collocated heterogeneous HD-FD network satisfies the Overall Local Pooling (OLoP) conditions, which guarantee that GMS is throughput-optimal. The proof is omitted and appears in [21].

B. Hybrid GMS (H-GMS) Algorithm

We now present a hybrid scheduling algorithm which combines the concepts of GMS and Q-CSMA [8], [9]. Instead of approximating MWS [6] in a decentralized manner (as in "traditional" Q-CSMA), we aim to approximate GMS, which is easier to decentralize. Further, we leverage the existence of an AP to resolve the contention among the DL queues, since the AP has explicit information about these queues and can select one of them (e.g., the longest queue). Thus, effectively at most one DL queue needs to perform Q-CSMA in each time slot. On the other hand, since users are unaware of the UL and DL queue states of other users and at the AP, every user needs to perform Q-CSMA in order to share the channel distributedly. Therefore, the number of possible participants in Q-CSMA in each slot is at most $(N+1)$.

We present the H-GMS algorithm (see Algorithm 2) that operates in such a *hybrid* fashion (combines the centralization at the AP and the distributed Q-CSMA). As shown in Section VII, this hybrid approach yields delay performance that is much better than that of the pure Q-CSMA approach, while still achieving maximum throughput.

The H-GMS algorithm operates as follows. Each slot t is divided into a short control slot and a data slot. The control slot contains only two control mini-slots, independently of the number of users. We refer to the first mini-slot as the *initiation mini-slot* and to the second one as the *coordination mini-slot*. H-GMS has three steps: (1) Initiation, (2) Coordination, and (3) Data transmission, as explained below.

(1) Initiation. By the end of slot $(t-1)$, the AP knows $\mathbf{X}(t-1)$ since every packet transmission has to be sent from or received by the AP. If $\mathbf{X}(t-1)=\mathbf{0}$ (i.e., idle channel), then the AP starts an initiation in slot t using the initiation mini-slot as follows. First, the AP centrally finds the index of the user with the longest DL queue, i.e., $i^\star(t) = \arg\max_{i\in\mathcal{N}} Q_i^{\mathrm{d}}(t)$. If multiple DLs have equal (largest) queue length, it breaks ties according to some deterministic rule. Then, the AP randomly selects an initiator link $\mathrm{IL}(t)$ from the set $\mathcal{L}(t) = \{l_1^{\mathrm{u}}, \cdots, l_N^{\mathrm{u}}, l_{i^\star}^{\mathrm{d}}\}$ according to an *access probability* distribution $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_N, \alpha_{\mathrm{AP}}]$ satisfying: (i) $\alpha_i > 0, \forall i \in \mathcal{N}$, and $\alpha_{\mathrm{AP}} > 0$, and (ii) $\alpha_{\mathrm{AP}} = 1 - \sum_{i=1}^N \alpha_i$. We refer to α_i and α_{AP} as the access probability for user i and the AP, respectively. Therefore,

$$\mathrm{IL}(t) = \begin{cases} l_i^{\mathrm{u}}, & \text{with probability } \alpha_i,\ \forall i \in \mathcal{N}, \\ l_{i^\star}^{\mathrm{d}}, & \text{with probability } \alpha_{\mathrm{AP}}, \end{cases} \tag{4}$$

i.e., $\mathrm{IL}(t)$ is either a UL or the DL with the longest queue. If $\mathbf{X}(t-1)\neq\mathbf{0}$, set $\mathrm{IL}(t) = \mathrm{IL}(t-1)$.

(2) Coordination. In the coordination mini-slot, if the DL of user $i^\star$ is selected as the initiator link ($\mathrm{IL}(t) = l_{i^\star}^{\mathrm{d}}$), the AP sets $X_{i^\star}^{\mathrm{d}}(t)=1$ with probability $p_{i^\star}^{\mathrm{d}}(t)$. Otherwise, it remains silent. If the AP decides to transmit on DL $l_{i^\star}^{\mathrm{d}}$ (i.e., $X_{i^\star}^{\mathrm{d}}(t)=1$), it broadcasts a control packet containing the information of $\mathrm{IL}(t)$ and user $i^\star$ sets $X_{i^\star}^{\mathrm{u}}(t)=1$ if and only if $i^\star \in \mathcal{N}_F$.

If the UL of user i is selected as the initiator link ($\mathrm{IL}(t) = l_i^{\mathrm{u}}$ for some $i\in\mathcal{N}$), the AP broadcasts the information of $\mathrm{IL}(t)$ and user i sets $X_i^{\mathrm{u}}(t)=1$ with probability $p_i^{\mathrm{u}}(t)$. Otherwise, user i remains silent. If user i is FD-capable and decides to transmit (i.e., $X_i^{\mathrm{u}}(t)=1$), it sends a control packet containing this information to the AP and the AP sets $X_i^{\mathrm{d}}(t)=1$.[3]

The transmission probability of the link is selected depending on its queue size $Q_i^j(t)$ at the beginning of slot t. Specifically, similar to [8], [9], link l_i^j chooses logistic form

$$p_i^j(t) = \frac{\exp\left(f(Q_i^j(t))\right)}{1+\exp\left(f(Q_i^j(t))\right)},\ \forall i\in\mathcal{N},\ \forall j\in\{\mathrm{u},\mathrm{d}\}, \tag{5}$$

where $f(\cdot)$ is called the *weight function* which is some positive increasing function to be determined later. Further, if an FD initiator UL (or DL) decides to stop transmitting (after packet transmission in the last slot), it again sends a short coordination message which stops further packet transmissions at the DL (or UL) or the same FD user.

[3]Note that this operation can be done in the same coordination mini-slot since FD user i can simultaneously receive the control packet ($\mathrm{IL}(t)=l_i^{\mathrm{u}}$) from the AP and send its control packet ($X_i^{\mathrm{u}}(t)=1$) back to the AP.

Algorithm 2 H-GMS Algorithm (in slot t)

– If $\mathbf{X}(t-1)=\mathbf{0}$:

1. In the initiation mini-slot, the AP computes $i^\star = \arg\max_{i\in\mathcal{N}} Q_i^{\mathrm{d}}(t)$ (i.e., the index of the user with the longest DL queue). If multiple DL queues have the same length, break ties according to some deterministic rule. The AP chooses an initiator link $\mathrm{IL}(t)$ from $\mathcal{L}(t) = \{l_1^{\mathrm{u}}, \cdots, l_N^{\mathrm{u}}, l_{i^\star}^{\mathrm{d}}\}$ according to an access probability distribution $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_N, \alpha_{\mathrm{AP}}]$.
2. If $\mathrm{IL}(t) = l_{i^\star}^{\mathrm{d}}$, the AP sets:
   - $X_{i^\star}^{\mathrm{d}}(t) = 1$ with probability $p_{i^\star}^{\mathrm{d}}(t)$, or $X_{i^\star}^{\mathrm{d}}(t) = 0$ with probability $\bar{p}_{i^\star}^{\mathrm{d}}(t) = 1 - p_{i^\star}^{\mathrm{d}}(t)$;
   - In the coordination mini-slot, AP broadcasts a control packet containing the information of $\mathrm{IL}(t)$ and user $i^\star$ sets $X_{i^\star}^{\mathrm{u}}(t) = X_{i^\star}^{\mathrm{d}}(t)\cdot \mathbb{1}(i^\star \in \mathcal{N}_F)$;
3. If $\mathrm{IL}(t) = l_i^{\mathrm{u}}$ for some $i\in\mathcal{N}$, in the coordination mini-slot, the AP broadcasts the information of $\mathrm{IL}(t)$ and user i sets:
   - $X_i^{\mathrm{u}}(t) = 1$ with probability $p_i^{\mathrm{u}}(t)$, or $X_i^{\mathrm{u}}(t) = 0$ with probability $\bar{p}_i^{\mathrm{u}}(t) = 1 - p_i^{\mathrm{u}}(t)$;
   - In the same coordination mini-slot, user i sends a control packet containing this information to the AP if $i\in\mathcal{N}_F$, and AP sets $X_i^{\mathrm{d}}(t) = X_i^{\mathrm{u}}(t)$;
4. At the beginning of the data slot,
   - AP activates DL i if $X_i^{\mathrm{d}}(t)=1$;
   - User i activates it UL if $X_i^{\mathrm{u}}(t)=1$;

– If $\mathbf{X}(t-1)\neq\mathbf{0}$, set $\mathrm{IL(t)} = \mathrm{IL(t-1)}$. Repeat Steps 2–4.

(3) Data transmission. After steps (1)–(2), if either a pair of FD UL and DL or an HD link (UL or DL) is activated, a packet is sent on the links in the data slot. The initiator link then starts a new coordination in the subsequent control slot which either leads to more packet transmissions or stops further packet transmissions at the links involved in the schedule.

Remark: The initiation step of the H-GMS algorithm is described as a polling mechanism where the AP draws a link $\mathrm{IL}(t)$ from $\mathcal{L}(t)$ according to the probability distribution $\boldsymbol{\alpha}$. Alternatively, the initiation step can be described in a distributed fashion using an extra mini-slot as follows: user i sends a short initiation message with probability α_i. If AP receives the message, it sends back a clear-to-initiate message and set $\mathrm{IL}(t) = l_i^{\mathrm{u}}$, otherwise (i.e., in case of collision or idleness) $l_{i^\star}^{\mathrm{d}}$ is selected as the initiator link by the AP. This effectively emulates polling user i with probability $\tilde{\alpha}_i = \alpha_i \prod_{i'\neq i}(1-\alpha_{i'})$ and AP with probability $\tilde{\alpha}_{\mathrm{AP}} = 1 - \sum_{i=1}^N \tilde{\alpha}_i$.

C. Main Result (Throughput Optimality)

The system state can be described by a Markov chain $(\mathbf{X}(t), \mathbf{Q}(t))$. Note that $\mathbf{X}(t)$ is determined after the Initiation and Coordination steps. Specifically, let $Y(t)$ indicate whether the initiator link in slot t is activated or not. Under the H-GMS algorithm, given a fixed queue length vector $\mathbf{Q}(t)=\mathbf{Q}$, $Y(t)$ evolves as an irreducible and reversible Markov chain over a finite number of states. Let $i^\star = \arg\max_{i\in\mathcal{N}} Q_i^{\mathrm{d}}(t)$, then the state space of $Y(t)$ can be labeled as $S_Y = \{0, 1, \cdots, N, i^\star\}$, in which 0 means no link is active, $i^\star$ means DL $l_{i^\star}^{\mathrm{d}}$ is active,

and $i \in \{1, \cdots, N\}$ means UL l_i^{u} is active. Let $P(s, s')$ be the transition probability from state $s \in S_Y$ to $s' \in S_Y$ when $\mathbf{Q}(t) = \mathbf{Q}$. Then, under Algorithm 2, we have for $1 \le i \le N$,

$$\begin{aligned} P(0,i) &= \alpha_i p_i^{\mathrm{u}},\ P(i,i) = p_i^{\mathrm{u}},\quad P(i,0) = \bar{p}_i^{\mathrm{u}}, \\ P(0,i^\star) &= \alpha_{\mathrm{AP}} p_{i^\star}^{\mathrm{d}},\ P(i^\star,i^\star) = p_{i^\star}^{\mathrm{d}},\ P(i^\star,0) = \bar{p}_{i^\star}^{\mathrm{d}}, \\ P(0,0) &= 1 - \textstyle\sum_{i=1}^N P(0,i) - P(0,i^\star). \end{aligned} \tag{6}$$

Lemma 1. *Given fixed queue-length vector $\mathbf{Q}(t) = \mathbf{Q}$, the steady-state distribution of Markov chain $Y(t)$ is given by*

$$\begin{aligned} \pi^{\mathbf{Q}}(i) &= \alpha_i \exp(f(Q_i^{\mathrm{u}}))/Z,\ i \in S_Y \setminus \{0, i^\star\}; \\ \pi^{\mathbf{Q}}(i^\star) &= \alpha_{AP} \exp(f(Q_{i^\star}^{\mathrm{d}}))/Z,\ \pi^{\mathbf{Q}}(0) = 1/Z, \end{aligned} \tag{7}$$

where Z is the normalizing constant and $f(\cdot)$ is the weight function from (5).

Proof: Under fixed $\mathbf{Q}$, the Markov chain $Y(t)$ evolves according to the transition probabilities defined by (6). It is easy to check that the steady-state distribution satisfies the detailed balance equations. ∎

The following corollary is immediate as the result of Lemma 1 and the fact that $Y(t)$ uniquely determines $\mathbf{X}(t)$ by (possible) activation of both the UL and DL of an FD user in the coordination step.

Corollary 1. *Let $\mathbf{f}_i = \mathbf{e}_{2i-1} + \mathbf{e}_{2i}$, $i \in \mathcal{N}_F$, be an FD bi-directional transmission schedule, and $\mathbf{h}_i^{\mathrm{u}} = \mathbf{e}_{2i-1}$ ($\mathbf{h}_i^{\mathrm{d}} = \mathbf{e}_{2i}$), $i \in \mathcal{N}_H$, be an HD UL (DL) transmission schedule. Given a fixed queue vector $\mathbf{Q}(t) = \mathbf{Q}$, in steady-state, if $i^\star \in \mathcal{N}_F$,*

$$\begin{aligned} \mathbb{P}\{\mathbf{X} = \mathbf{f}_{i^\star}\} &= [\alpha_{AP} \exp(f(Q_{i^\star}^{\mathrm{d}})) + \alpha_{i^\star} \exp(f(Q_{i^\star}^{\mathrm{u}}))]/Z, \\ \mathbb{P}\{\mathbf{X} = \mathbf{f}_i\} &= \alpha_i \exp(f(Q_i^{\mathrm{u}}))/Z,\ \forall i \in \mathcal{N}_F,\ i \neq i^\star, \\ \mathbb{P}\{\mathbf{X} = \mathbf{h}_i^{\mathrm{u}}\} &= \alpha_i \exp(f(Q_i^{\mathrm{u}}))/Z,\ \forall i \in \mathcal{N}_H. \end{aligned}$$

Otherwise, if $i^\star \in \mathcal{N}_H$,

$$\begin{aligned} \mathbb{P}\{\mathbf{X} = \mathbf{f}_i\} &= \alpha_i \exp(f(Q_i^{\mathrm{u}}))/Z,\ \forall i \in \mathcal{N}_F, \\ \mathbb{P}\{\mathbf{X} = \mathbf{h}_i^{\mathrm{u}}\} &= \alpha_i \exp(f(Q_i^{\mathrm{u}}))/Z,\ \forall i \in \mathcal{N}_H, \\ \mathbb{P}\left\{\mathbf{X} = \mathbf{h}_{i^\star}^{\mathrm{d}}\right\} &= \alpha_{AP} \exp(f(Q_{i^\star}^{\mathrm{d}}))/Z, \end{aligned}$$

where Z and $f(\cdot)$ are as in Lemma 1.

Corollary 1 suggests that, assuming $Y(t)$ is always in steady-state at any time t, when the maximum queue size $\max_{i,j} Q_i^j(t) \to \infty$, the algorithm chooses a GMS schedule with high probability (note that the returned schedule might not be an MWS schedule). However, establishing this intuition is not simple, since the coupling between $\mathbf{X}(t)$ and $\mathbf{Q}(t)$ makes the analysis complicated. On one hand, the dynamics of $\mathbf{X}(t)$ is governed by the Markov chain $Y(t)$, whose transition probability matrix $P^{\mathbf{Q}(t)}$ depends on the queues. On the other hand, the dynamics of $\mathbf{Q}(t)$ depends on the schedule process $\mathbf{X}(t)$. We provide a proof of the stability of the H-GMS algorithm based on the analysis of the fluid limits. The aforementioned coupling gives rise to qualitatively different fluid limits, depending on the time-scale of schedule process convergence compared to the time-scale of the changes in the queue process. Our main result is summarized in the following theorem, whose proof is provided in Section V.

Theorem 1 (Main Result)**.** *For any arrival rate vector $\boldsymbol{\lambda} \in \mathrm{int}(\Lambda_{\mathrm{HD-FD}})$, the system Markov chain $(\mathbf{X}(t), \mathbf{Q}(t))$ is positive recurrent under the H-GMS algorithm, as described in Algorithm 2. The weight function $f(\cdot)$ in* (5) *can be any nonnegative increasing function such that $\lim_{x\to\infty} f(x)/\log(x) < 1$, or $\lim_{x\to\infty} f(x)/\log(x) > 1$ (including $f(x) = x^\beta,\ \beta > 0$).*

D. Variants of the H-GMS Algorithm

In this subsection, we introduce three variants of the H-GMS algorithm which differ only in Step 1 of Algorithm 2. Assuming that $\boldsymbol{\alpha}$ is a uniform distribution over $\mathcal{L}(t) = \{l_1^{\mathrm{u}}, \cdots, l_N^{\mathrm{u}}, l_{i^\star}^{\mathrm{d}}\}$, the first two variants are:

- H-GMS-R: The AP selects a DL *uniformly at random*, i.e., $i^\star \sim \mathrm{Unif}(1, \cdots, N)$;
- H-GMS-L: The AP selects the *longest DL* as described exactly in Algorithm 2.

H-GMS-R and H-GMS-L may not balance FD and HD queues since the initiator link $\mathrm{IL}(t)$ is selected uniformly at random from $\mathcal{L}(t)$. Therefore, we consider a third heuristic variant:

- H-GMS-E: Exactly the same as H-GMS-L except for the access probability $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_N, \alpha_{\mathrm{AP}}]$ being:

$$\begin{aligned} \alpha_i &\propto \max\{\tilde{Q}_i^{\mathrm{u}}/(\textstyle\sum_{i'=1}^N \tilde{Q}_{i'}^{\mathrm{u}} + Q_{i^\star}^{\mathrm{d}}),\ \alpha_{\mathrm{th}}\},\ \forall i \in \mathcal{N}, \\ \alpha_{\mathrm{AP}} &\propto \max\{Q_{i^\star}^{\mathrm{d}}/(\textstyle\sum_{i'=1}^N \tilde{Q}_{i'}^{\mathrm{u}} + Q_{i^\star}^{\mathrm{d}}),\ \alpha_{\mathrm{th}}\}, \end{aligned}$$

  in which $\tilde{Q}_i^{\mathrm{u}}$ is the estimated UL queue length of user i. Specifically, when a user transmits on the UL, it includes its queue length information in the packets and the AP updates $\tilde{Q}_i^{\mathrm{u}}$ using the most recently received UL queue length from user i. We also introduce a minimum access probability, α_{th}, so that the AP selects each UL with probability at least α_{th}.

V. Proof of Theorem 1 via Fluid Limits

We prove Theorem 1 based on the analysis of the fluid limits of the system under Algorithm 2. The proof has three parts: (i) existence of the fluid limits (Lemma 2), (ii) deriving the fluid limit equations (Lemma 3), and (iii) proving the stability of the queues in the fluid limit using a Lyapunov method, which implies the stability of the original stochastic process. The analysis and derivations are similar to the fluid limits of CSMA algorithms [22]–[24] but specialized to the heterogeneous HD-FD network. The specialization allows us to prove throughput optimality for *any* nonnegative increasing weight function f in (5) such that $\lim_{x\to\infty} f(x)/\log(x) < 1$, or $\lim_{x\to\infty} f(x)/\log(x) > 1$ (including $f(x) = x^\beta,\ \beta > 0$).

Define a scaled process $\mathbf{Q}^{(r)}(t)$ where $\mathbf{Q}^{(r)}(t) = \mathbf{Q}(rt)/r$. Note that the queue process $\mathbf{Q}$ is scaled in both time and space by a factor $r > 0$. To avoid technical difficulties, we can simply work with a continuous process by linear interpolation among the values at integer time points. Suppose the scaled process, with $r > 0$, starts from an initial state $\mathbf{Q}^{(r)}(0)$. Any (possibly random) limit of the scaled process $\mathbf{Q}^{(r)}(t)$ as $r \to \infty$ is called a fluid limit. The process $\mathbf{Q}^{(r)}(t)$ can be constructed as follows. At any time $t \ge 0$,

$$\mathbf{Q}^{(r)}(t) = \mathbf{Q}^{(r)}(0) + \bar{\mathbf{A}}^{(r)}(t) - \bar{\mathbf{S}}^{(r)}(t), \tag{8}$$

where for any user $i \in \mathcal{N}$ with UL or DL $j \in \{\mathrm{u}, \mathrm{d}\}$,

$$\bar{A}_i^{j\,(r)}(t) = \frac{1}{r}\sum_{\tau=1}^{rt} A_i^j(\tau),\ \bar{S}_i^{j\,(r)}(t) = \frac{1}{r}\sum_{\tau=1}^{rt} X_i^j(\tau)\mathbb{1}(Q_i^j(\tau) > 0).$$

The following lemma shows that the scaled process converges to the fluid limit in a weak convergence sense, in the metric of uniform norm on compact time intervals. It is possible to show a stronger convergence (i.e., almost sure convergence uniformly over compact time intervals) under certain conditions on the weight functions f (i.e., $\lim_{x\to\infty} f(x)/\log(x) < 1$). Nevertheless, the weak convergence is sufficient for our proofs.

Lemma 2 (Existence of Fluid Limits)**.** *Suppose $\mathbf{Q}^{(r)}(0) \to \mathbf{q}(0)$. Then any sequence r has a subsequence such that $(\mathbf{Q}^{(r)}(t), \bar{\mathbf{A}}^{(r)}(t), \bar{\mathbf{S}}^{(r)}(t)) \Rightarrow (\mathbf{q}(t), \mathbf{a}(t), \mathbf{s}(t))$ along the subsequence. The sample paths $(\mathbf{q}(t), \mathbf{a}(t), \mathbf{s}(t))$ are Lipschitz continuous and thus differentiable almost everywhere with probability one.*

Proof: The proof is standard and follows from Lipschitz continuity of the scaled process, see e.g. [25]. ∎

Next, we show that all the fluid sample paths must satisfy the following equations. The equations do not uniquely describe the fluid limit process but are sufficient to establish stability.

Lemma 3 (Fluid Limit Equations)**.** *Consider any non-negative increasing weight function $f(\cdot)$ in (5), such that $\lim_{x\to\infty} f(x)/\log(x) < 1$, or $\lim_{x\to\infty} f(x)/\log(x) > 1$ (including $f(x) = x^\beta,\ \beta > 0$). Let $\hat{q}_i(t) = \max\{q_i^{\mathrm{u}}(t), q_i^{\mathrm{d}}(t)\}$, for $i \in \mathcal{N}_F$. At any regular point t (i.e., any point where the derivatives of all the functions exist), for any $j \in \{\mathrm{u}, \mathrm{d}\}$,*

$$q_i^j(t) = q_i^j(0) + a_i^j(t) - s_i^j(t),\ i \in \mathcal{N} \tag{9}$$

$$a_i^j(t) = \lambda_i^j t,\ s_i^j(t) = \textstyle\int_0^t \mu_i^j(\tau)\ \mathrm{d}\tau,\ \mu_i^j(t) \in [0,1], \tag{10}$$

$$\mu_i^j(t) \cdot \mathbb{1}(q_i^j(t) = 0, \mathbf{q}(t) \neq \mathbf{0}) = 0,\ i \in \mathcal{N}_H, \tag{11}$$

$$\mu_i^j(t) \cdot \mathbb{1}(\hat{q}_i(t) = 0, \mathbf{q}(t) \neq \mathbf{0}) = 0,\ i \in \mathcal{N}_F, \tag{12}$$

$$\text{if } q_i^j(t) = \hat{q}_i(t),\ \mu_i^j(t) = \max\{\mu_i^{\mathrm{u}}(t), \mu_i^{\mathrm{u}}(t)\},\ i \in \mathcal{N}_F, \tag{13}$$

$$\text{if } \mathbf{q}(t) \neq \mathbf{0},\ \text{then}$$

$$\sum_{i\in\mathcal{N}_F} \max\{\mu_i^{\mathrm{u}}(t), \mu_i^{\mathrm{d}}(t)\} + \sum_{i\in\mathcal{N}_H} (\mu_i^{\mathrm{u}}(t) + \mu_i^{\mathrm{d}}(t)) = 1. \tag{14}$$

The proof of Lemma 3 is deferred to [21]. The following proposition proves the stability of the queues in the fluid limit, which will suffice to complete the proof of Theorem 1.

Proposition 2. *Starting from an initial queue size $\mathbf{q}(0)$, there is a deterministic finite time T by which all the queues at the fluid limit will hit zero.*

Proof: Let $\hat{q}_i(t) = \max\{q_i^{\mathrm{u}}(t), q_i^{\mathrm{d}}(t)\}$, $i \in \mathcal{N}_F$. Consider the Lyapunov function

$$V(\mathbf{q}(t)) = \textstyle\sum_{i\in\mathcal{N}_F} \hat{q}_i(t) + \sum_{i\in\mathcal{N}_H} (q_i^{\mathrm{u}}(t) + q_i^{\mathrm{d}}(t)).$$

Let $\mathcal{U}_H^j(t) := \{i \in \mathcal{N}_H : q_i^j(t) > 0\}$, $j \in \{\mathrm{u}, \mathrm{d}\}$, and $\mathcal{U}_F(t) := \{i \in \mathcal{N}_F : \hat{q}_i(t) > 0\}$. Suppose $V(\mathbf{q}(t)) > 0$ (i.e., $\mathbf{q}(t) \neq \mathbf{0}$). Then based on the fluid limit equations (11)–(14):

(i) The network is draining some subsets $\mathcal{P}_H^{\mathrm{u}}(t) \subseteq \mathcal{U}_H^{\mathrm{u}}(t)$, $\mathcal{P}_H^{\mathrm{d}}(t) \subseteq \mathcal{U}_H^{\mathrm{d}}(t)$, and $\mathcal{P}_F(t) \subseteq \mathcal{U}_F(t)$ of non-zero queues,

(ii) $\hat{q}_i(t)$ for user $i \in \mathcal{P}_F(t)$ is always drained at rate $\max\{\mu_i^{\mathrm{u}}(t), \mu_i^{\mathrm{d}}(t)\}$,

(iii) $\sum_{i\in\mathcal{P}_F(t)} \max\{\mu_i^{\mathrm{u}}(t), \mu_i^{\mathrm{d}}(t)\} + \sum_{i\in\mathcal{P}_H^{\mathrm{u}}(t)} \mu_i^{\mathrm{u}}(t) + \sum_{i\in\mathcal{P}_H^{\mathrm{d}}(t)} \mu_i^{\mathrm{d}}(t) = 1$.

Hence, using (9)–(10), and properties (i)–(iii) above,

$$\begin{aligned}\frac{\mathrm{d}V(\mathbf{q}(t))}{\mathrm{d}t} &\le \textstyle\sum_{i\in\mathcal{N}_F} \max\{\lambda_i^{\mathrm{u}}, \lambda_i^{\mathrm{d}}\} + \sum_{i\in\mathcal{N}_H} (\lambda_i^{\mathrm{u}} + \lambda_i^{\mathrm{d}}) \\ &\quad - \textstyle\sum_{i\in\mathcal{P}_F(t)} \max\{\mu_i^{\mathrm{u}}(t), \mu_i^{\mathrm{d}}(t)\} - \sum_{i\in\mathcal{P}_H^{\mathrm{u}}(t)} \mu_i^{\mathrm{u}}(t) \\ &\quad - \textstyle\sum_{i\in\mathcal{P}_H^{\mathrm{d}}(t)} \mu_i^{\mathrm{d}}(t) \\ &= \textstyle\sum_{i\in\mathcal{N}_F} \max\{\lambda_i^{\mathrm{u}}, \lambda_i^{\mathrm{d}}\} + \sum_{i\in\mathcal{N}_H} (\lambda_i^{\mathrm{u}} + \lambda_i^{\mathrm{d}}) - 1 \le -\delta,\end{aligned}$$

where the last inequality is due to the fact that $\boldsymbol{\lambda} \in \mathrm{int}(\Lambda_{\mathrm{HD-FD}})$, by the assumption of Theorem 1. Thus, there must exist a small $\delta > 0$ such that $\boldsymbol{\lambda}/(1-\delta) \in \Lambda_{\mathrm{HD-FD}}$. Therefore, $V(\mathbf{q}(t))$ will hit zero in finite time $T = V(\mathbf{q}(0))/\delta$, and in fact remains zero afterwards. ∎

Proposition 2 implies the stability (positive recurrence) of the original Markov chain $(\mathbf{Q}(t), \mathbf{X}(t))$ in a similar fashion as [26] (note that the component $\mathbf{X}(t)$ lives in a finite state space). This completes the proof of Theorem 1.

Remark: Recall from Section IV-C that the coupling between the scheduling process, $\mathbf{X}(t)$, and the queue dynamics, $\mathbf{Q}(t)$, makes the proof challenging. In addition, unlike the traditional Q-CSMA algorithm that distributedly approximates MWS, the proposed H-GMS algorithm approximates GMS in a distributed manner.

VI. Benefits of Introducing FD-Capable Nodes

In this section, we illustrate the benefits in terms of throughput (achievable rates) gain obtained from introducing FD-capable nodes into all-HD networks. We define the network (individual) throughput gain as the ratio between the achievable network (individual) throughputs in heterogeneous HD-FD and all-HD networks with the same total number of users.

For simplicity and illustrative purposes, we use a static version of H-GMS-R to demonstrate the throughput gains of both the entire network and individual users. In particular, we assign (constant) access probability $\alpha_i = 1/(2N),\ \forall i \in \mathcal{N}$, and $\alpha_{\mathrm{AP}} = 1/2$ (see Algorithm 2 and Section IV-D). We select (constant) transmission probabilities $p_f^{\mathrm{u}} = p_f^{\mathrm{d}} = p_f,\ p_h^{\mathrm{u}} = p_h^{\mathrm{d}} = p_h \in (0,1)$ for FD and HD users, respectively. By analyzing the CSMA Markov chain (similar to Lemma 1), the network throughput (i.e., sum rates) of the heterogeneous HD-FD network, $S_{\mathrm{HD-FD}}$, is given by

$$S_{\mathrm{HD-FD}} = \frac{\frac{2N_F}{N}\frac{p_f}{1-p_f} + \frac{N_H}{N}\frac{p_h}{1-p_h}}{1 + \frac{N_F}{N}\frac{p_f}{1-p_f} + \frac{N_H}{N}\frac{p_h}{1-p_h}}. \tag{15}$$

Note that the throughput of the benchmark all-HD network is simply $S_{\mathrm{HD}} = p_h$. If $p_f = p_h = p$ (i.e., FD and HD users transmit with the same probability when they capture the channel), (15) becomes $S_{\mathrm{HD-FD}} = (1 + N_F/N) \cdot p$. This implies that under the static H-GMS-R, the network throughput gain achieved by the HD-FD network is $(1 + N_F/N) \in [1, 2]$, which increases with respect to N_F.

Assigning equal transmission probabilities results in FD users having $2\times$ throughput compared to the HD users. We can balance the throughput obtained by FD and HD users by assigning different transmission probabilities. Let $p_h = p$ and $p_f = \chi \cdot p$ for some *transmission probability ratio* χ. In order

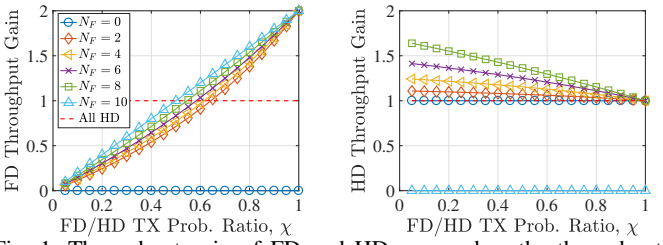

Fig. 1: Throughput gain of FD and HD users when the throughput is compared to the individual throughput of an HD user in the all-HD network under the static H-GMS-R algorithm, with $N = 10$, $N_F \in \{0, 2, \cdots, 10\}$, and $p_h = 0.5$.

to balance the individual throughput of FD and HD users, we lower the priority of FD transmissions by choosing $\chi \in (0, 1]$.

We numerically evaluate the individual user throughput gain. We consider both the benchmark all-HD network (with transmission probability $p_h = p$) and HD-FD networks with $N = 10$ and vary $N_F \in \{0, 2, \cdots, 10\}$ in the latter. We select constant $p_h = 0.5$ and $p_f = \chi p_h$ with varying $\chi \in (0, 1]$. Fig. 1 plots individual throughput gains of an FD or HD user. As Fig. 1 suggests, if FD and HD users are assigned equal transmission probabilities ($\chi = 1$), an FD user gets $2\times$ throughput compared to an HD user. If the transmission probability of the FD users is lowered (by decreasing χ), the throughput of FD and HD users is more balanced. For example, with $\chi = 0.75$, the individual throughput gains of FD and HD users are 43% and 20%, respectively.

The results reveal an interesting phenomenon: when N_F is sufficiently large, at the cost of slightly lowering the priority of FD users, even HD users can experience throughput improvements. This opens up a possibility of designing wireless protocols with different fairness-efficiency tradeoffs incurred by setting different priorities among FD and HD users.

VII. Simulation Results

In this section, we evaluate the performance of different scheduling algorithms in heterogeneous HD-FD networks via simulations. We focus on (i) network-level *delay* performance (represented by the long-term average queue length per link), and (ii) *fairness* between FD and HD users (represented by the relative delay performance between FD and HD users).

Throughout this section, we consider heterogeneous HD-FD networks with one FD AP and 10 users ($N = 10$), with a varying number of FD users, N_F.[4] We choose a rate vector $\mathbf{v} = [v_i^{\mathrm{u}}, v_i^{\mathrm{d}}]_{i=1}^N$ on the boundary of the capacity region $\Lambda_{\mathrm{HD-FD}}$ (see Section III-C) and consider arrival rates of the form $\boldsymbol{\lambda} = \rho\mathbf{v}$, in which $\rho \in (0, 1)$ is the *traffic intensity*. Note that as $\rho \to 1$, $\boldsymbol{\lambda}$ approaches the boundary of $\Lambda_{\mathrm{HD-FD}}$. Since we focus on the fairness between FD and HD users, we assume equal UL and DL arrival rates over all the users. Therefore, for $j \in \{\mathrm{u}, \mathrm{d}\}$, we use $v_f = v_i^j$, $\forall i \in \mathcal{N}_F$, and $v_h = v_i^j$, $\forall i \in \mathcal{N}_H$, to denote the equal UL and DL arrival rates assigned to FD and HD users, respectively. For an *equal arrival rate* model, we have $v_f = v_h = 1/(N_F + 2N_H)$ computed using (1).

[4]The results for heterogeneous HD-FD networks with a different number of users, N, are similar, and thus, omitted.

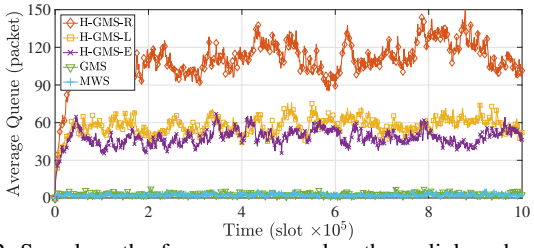

Fig. 2: Sample path of average queue length per link under different scheduling algorithms for a heterogeneous HD-FD network with $N_F = N_H = 5$, and *very high* traffic intensity $\rho = 0.95$.

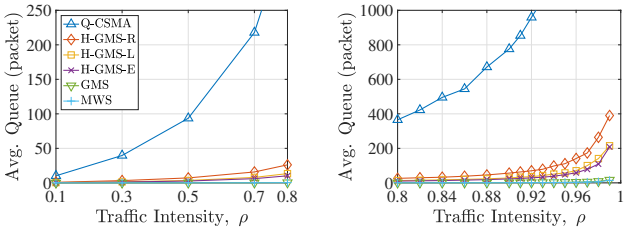

Fig. 3: Long-term average queue length per link in a heterogeneous HD-FD network with $N_F = N_H = 5$ and equal arrival rates, under different scheduling algorithms: (left) low and moderate traffic intensities, and (right) high traffic intensity.

The packet arrivals at each link l_i^j follow an independent Bernoulli process with rate λ_i^j. For each algorithm under a given traffic intensity, ρ, we take the average over 10 independent simulations, each of which lasts for 10^6 slots. For simplicity, we refer to the "queue length of an FD (resp. HD) user" as the sum of its UL and DL queue lengths, and only compare the average queue length between FD and HD users without distinguishing between individual UL and DL.[5] The considered algorithms include:

- Q-CSMA: The standard distributed Q-CSMA algorithm from [8], in which each link (UL or DL) performs channel contention independently and the AP does not leverage the central DL queue information;
- H-GMS-R, H-GMS-L, and H-GMS-E: Three variants of the H-GMS algorithm as described in Section IV-D;
- MWS, GMS: The centralized MWS and GMS algorithms.

In the first four distributed algorithms, the transmission probability of link l in slot t is selected as $p_l(t) = \frac{\exp(f(Q_l(t)))}{1+\exp(f(Q_l(t)))}$ where $f(Q_l(t)) = \log(Q_l(t) + 1)$. We will show that different degrees of centralization at the AP result in performance improvements of H-GMS over the classical Q-CSMA in terms of both delay and fairness.

A. Delay Performance

We first show the queue length dynamics under various scheduling algorithms in an HD-FD network with $N_F = N_H = 5$ and traffic intensity $\rho = 0.95$. This implies that $v_f = v_h = 1/15$, corresponding to a capacity region expansion value of $\gamma = 4/3$ (see Section III-C with $v_h = 1/20$ in the all-HD network). Fig. 2 plots the sample paths of the average queue length of the network (i.e., averaged over all the ULs and DLs) under different algorithms. The result for

[5]We also investigated the fairness between FD and HD UL (or DL) queues and observed similar results.

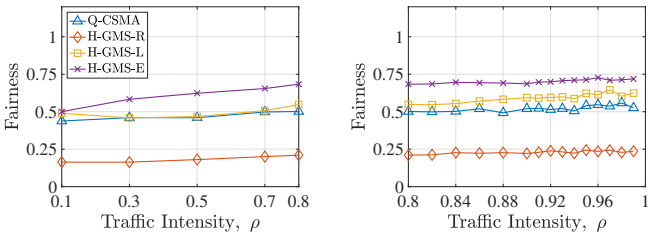

Fig. 4: Long-term average queue length ratio between FD and HD users in a heterogeneous HD-FD network with $N_F = N_H = 5$ and equal arrival rates, under different scheduling algorithms.

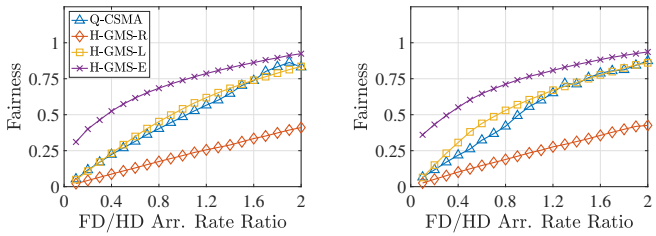

Fig. 5: Long-term average queue length ratio between FD and HD users in a heterogeneous HD-FD network with $N_F = N_H = 5$ and varying ratio between FD and HD arrival rates, with (left) moderate ($\rho = 0.8$), and (right) high ($\rho = 0.95$) traffic intensities.

Q-CSMA algorithm is omitted since, as we will see shortly, its average queue length is at least one order of magnitude larger than that achieved by other algorithms. In addition, Fig. 3 plots the average queue length of the network under low and moderate ($\rho \in [0.1, 0.8)$), as well as high ($\rho \in [0.8, 0.99]$) traffic intensities, respectively.

Figs. 2 and 3 show that, as expected from Theorem 1, all the considered algorithms are throughput-optimal in the sense that they stabilize all network queues. The fully-centralized MWS and GMS have the best delay performance but require high-complexity implementations. Among distributed algorithms, Q-CSMA [8] has the worst delay performance due to the high contention intensity introduced by a total number of $2N$ contending links. By "consolidating" the N DLs into one single DL that participates in channel contention, H-GMS-R, H-GMS-L, and H-GMS-E achieve 5–10$\times$ better delay performance than that of Q-CSMA under all considered ρ.

In particular, H-GMS-L and H-GMS-E have very similar delay performance which is better than that of H-GMS-R, since the AP leverages its central information to always select the DL with the longest queue for channel contention. However, they provide different fairness among FD and HD users due to the choice of access probability distribution $\boldsymbol{\alpha}$ (that is constant for the former and depends on the queue-length estimates for the latter), as we show below.

B. Fairness Between FD and HD Users

We now focus on the fairness between FD and HD. Instead of giving a mathematical representation, we define fairness as the *ratio between the average queue length of FD and HD users*. We use this notion since, intuitively, if an FD user experiences lower average delay (i.e., queue length) than an HD user, then introducing FD capability to the network will imbalance the service rate both users get. Ideally, we would like the proposed algorithms to achieve good fairness performance in the heterogeneous HD-FD networks.

We first evaluate the fairness under different distributed algorithms with equal arrival rates at each link. Fig. 4 plots the fairness between FD and HD users in an HD-FD network with $N_F = N_H = 5$ and varying traffic intensity, ρ. It can be observed that H-GMS-R has the worst fairness performance since the DL participating in the channel contention is selected uniformly at random by the AP. When the traffic intensity is low or moderate, Q-CSMA and H-GMS-L achieve similar fairness of about 0.5. This is because under equal arrival rates, FD queues are about half the length of the HD queues due to the fact that they are being served about twice as often (i.e., an FD bi-directional transmission can be either activated by the FD UL or DL due to the FD PHY capability). When the traffic intensity is high, both H-GMS-L and H-GMS-E have increased fairness performance since the longest DL queue will be served more often due to the central DL queue information at the AP. Furthermore, H-GMS-E outperforms H-GMS-L since, under H-GMS-E, the AP not only has explicit information of all the DL queues, but also has an estimated UL queue lengths that can be used to better assign the access probability distribution $\boldsymbol{\alpha}$.

We also evaluate the fairness under different arrival rates between FD and HD users. Let σ be the ratio between the arrival rates on FD and HD links. It is easy to see that if we assign $v_f = \sigma/(\sigma N_F + 2N_H)$ and $v_h = 1/(\sigma N_F + 2N_H)$, then $\mathbf{v}$ is on the boundary of $\Lambda_{\text{HD}-\text{FD}}$. In this case, we have a capacity region expansion value of $\gamma = 1 + \sigma N_F/(\sigma N_F + 2N_H)$, which depends on both N_F and σ (see Section III-C).

Fig. 5 plots the fairness between FD and HD users with varying σ under moderate ($\rho = 0.8$) and high ($\rho = 0.95$) traffic intensities on the x-axis. It can be observed that as the packet arrival rate at FD users increases, the FD and HD queue lengths are better balanced. When $\sigma = 2$, FD and HD users have almost the same average queue length since the FD queues are served twice as often as the HD queues under Q-CSMA, H-GMS-L, and H-GMS-E. It is interesting to note that the fairness under Q-CSMA and H-GMS-L is almost a linear function with respect to the arrival rate ratio, σ. This is intuitive since, as the FD queues are served about twice as often as the HD queues, increased arrival rates will result in longer queue lengths at the FD users. Moreover, since the FD and HD queues have about the same queue length when σ approaches 2, H-GMS-E does not further improve the fairness since it generates an access probability distribution that is approximately a uniform distribution.

C. Impact of the Number of FD Users, N_F

Lastly, we evaluate the fairness between FD and HD users with varied number of FD (or equivalently, HD) users under the equal arrival rate model. We vary $N_F \in \{1, 2, \cdots, 9\}$. Fig. 6 plots the fairness between FD and HD users under moderate ($\rho = 0.8$) and high ($\rho = 0.95$) traffic intensities.

As Fig. 6 suggests, the fairness depends on the number of FD users, N_F, only under H-GMS-L. This is because under

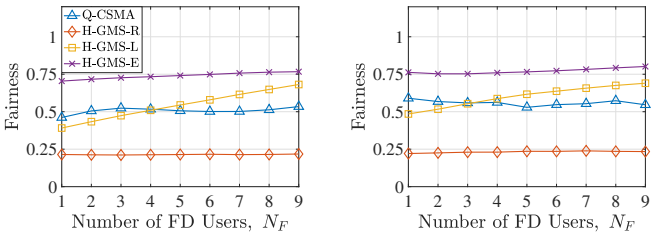

Fig. 6: Long-term average queue length ratio between FD and HD users in a heterogeneous HD-FD network with $N = 10$ and varying $N_F \in \{1, 2, \cdots, N-1\}$, with (left) moderate ($\rho = 0.8$), and (right) high ($\rho = 0.95$) traffic intensities.

equal arrival rate, FD users have about half the queue lengths compared with HD users. As N_F increases, the number of HD DLs at the AP (those with relatively larger queue length) decreases and as a result, the AP is very likely to select an HD DL or UL under the H-GMS-L algorithm, resulting in larger average queue length at the FD users. In addition, H-GMS-E resolves this issue by taking into account the UL queue length estimates. Therefore, the FD users that have smaller queues will be selected with a lower probability so that the longer HD queues will be served at a higher rate. In addition, as N_F increases, H-GMS-L achieves better fairness than that of the classical Q-CSMA by approximating the GMS (instead of MWS as Q-CSMA does) in a distributed manner. Moreover, H-GMS-E has the best fairness performance which is independent of the value of N_F.

VIII. Conclusion

We presented a hybrid scheduling algorithm, H-GMS, for heterogeneous HD-FD infrastructure-based networks. H-GMS is distributed at the users and leverages different degrees of centralization at the AP to achieve good delay performance while being provably throughput-optimal. Its performance was illustrated and compared to the classical Q-CSMA scheduling algorithm through extensive simulations. We also illustrated benefits and fairness-efficiency tradeoffs arising from incorporating FD users into existing HD networks. There are several important directions for future work. We plan to expand the results to multi-channel networks with general topologies and to study the impact of imperfect SIC on the scheduling algorithms and their performance. In addition, a rigorous analysis of the delay performance as in [27] and an experimental evaluation of H-GMS on a real wireless testbed are important steps towards provably-efficient and practical MAC layer for HD-FD networks.

Acknowledgment

This work was supported in part by NSF grant ECCS-1547406 and ARO grant 9W911NF-16-1-0259.

References

[1] A. Sabharwal, P. Schniter, D. Guo, D. W. Bliss, S. Rangarajan, and R. Wichman, "In-band full-duplex wireless: Challenges and opportunities," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 9, pp. 1637–1652, 2014.

[2] M. Duarte, C. Dick, and A. Sabharwal, "Experiment-driven characterization of full-duplex wireless systems," *IEEE Trans. Wireless Commun.*, vol. 11, no. 12, pp. 4296–4307, 2012.

[3] D. Bharadia, E. McMilin, and S. Katti, "Full duplex radios," in *Proc. ACM SIGCOMM'13*, 2013.

[4] J. Zhou, N. Reiskarimian, J. Diakonikolas, T. Dinc, T. Chen, G. Zussman, and H. Krishnaswamy, "Integrated full duplex radios," *IEEE Commun. Mag.*, vol. 55, no. 4, pp. 142–151, 2017.

[5] D. Yang, H. Yüksel, and A. Molnar, "A wideband highly integrated and widely tunable transceiver for in-band full-duplex communication," *IEEE J. Solid-State Circuits*, vol. 50, no. 5, pp. 1189–1202, 2015.

[6] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Trans. Autom. Control*, vol. 37, no. 12, pp. 1936–1948, 1992.

[7] A. Dimakis and J. Walrand, "Sufficient conditions for stability of longest-queue-first scheduling: Second-order properties using fluid limits," *Adv. Appl. Prob.*, vol. 38, no. 2, pp. 505–521, 2006.

[8] J. Ni, B. Tan, and R. Srikant, "Q-CSMA: Queue-length-based CSMA/CA algorithms for achieving maximum throughput and low delay in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 20, no. 3, pp. 825–836, 2012.

[9] J. Ghaderi and R. Srikant, "On the design of efficient CSMA algorithms for wireless networks," in *Proc. IEEE CDC'10*, 2010.

[10] B. Birand, M. Chudnovsky, B. Ries, P. Seymour, G. Zussman, and Y. Zwols, "Analyzing the performance of greedy maximal scheduling via local pooling and graph theory," *IEEE/ACM Trans. Netw.*, vol. 20, no. 1, pp. 163–176, 2012.

[11] M. Chung, M. S. Sim, J. Kim, D. K. Kim, and C.-B. Chae, "Prototyping real-time full duplex radios," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 56–63, 2015.

[12] J. Marašević and G. Zussman, "On the capacity regions of single-channel and multi-channel full-duplex links," in *Proc. ACM MobiHoc'16*, 2016.

[13] J. Marašević, J. Zhou, H. Krishnaswamy, Y. Zhong, and G. Zussman, "Resource allocation and rate gains in practical full-duplex systems," *IEEE/ACM Trans. Netw.*, vol. 25, no. 1, pp. 292–305, 2017.

[14] W. Li, J. Lilleberg, and K. Rikkinen, "On rate region analysis of half-and full-duplex OFDM communication links," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 9, pp. 1688–1698, Sept. 2014.

[15] S. Goyal, P. Liu, O. Gurbuz, E. Erkip, and S. Panwar, "A distributed MAC protocol for full duplex radio," in *Proc. Asilomar'13*, 2013.

[16] S.-Y. Chen, T.-F. Huang, K. C.-J. Lin, Y.-W. P. Hong, and A. Sabharwal, "Probabilistic medium access control for full-duplex networks with half-duplex clients," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2627–2640, 2017.

[17] A. Sahai, G. Patel, and A. Sabharwal, "Pushing the limits of full-duplex: Design and real-time implementation," *arXiv preprint: 1107.0607*, 2011.

[18] X. Xie and X. Zhang, "Does full-duplex double the capacity of wireless networks?" in *Proc. IEEE INFOCOM'14*, 2014.

[19] Y. Yang and N. B. Shroff, "Scheduling in wireless networks with full-duplex cut-through transmission," in *Proc. IEEE INFOCOM'15*, 2015.

[20] M. A. Alim, M. Kobayashi, S. Saruwatari, and T. Watanabe, "In-band full-duplex medium access control design for heterogeneous wireless LAN," *EURASIP J. Wireless Commun. and Netw.*, vol. 2017, no. 1, p. 83, 2017.

[21] T. Chen, J. Diakonikolas, J. Ghaderi, and G. Zussman, "Hybrid scheduling in heterogeneous half-and full-duplex wireless networks," *arXiv preprint: 1801.01108*, 2018.

[22] N. Bouman, S. Borst, J. van Leeuwaarden, and A. Proutiere, "Backlog-based random access in wireless networks: Fluid limits and delay issues," in *Proc. ITC'11*, 2011.

[23] J. Ghaderi, S. Borst, and P. Whiting, "Queue-based random-access algorithms: Fluid limits and stability issues," *Stochastic Systems*, vol. 4, no. 1, pp. 81–156, 2014.

[24] M. Feuillet, A. Proutiere, and P. Robert, "Random capture algorithms fluid limits and stability," in *Proc. IEEE ITA'10*, 2010.

[25] W. Whitt, "Weak convergence of probability measures on the function space $c[0, \infty)$," *Ann. of Math. Stat.*, vol. 41, no. 3, pp. 939–944, 1970.

[26] J. G. Dai, "On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models," *Ann. Appl. Prob.*, pp. 49–77, 1995.

[27] B. Ji, G. R. Gupta, M. Sharma, X. Lin, and N. B. Shroff, "Achieving optimal throughput and near-optimal asymptotic delay performance in multichannel wireless networks with low complexity: a practical greedy scheduling policy," *IEEE/ACM Trans. Netw.*, vol. 23, no. 3, pp. 880–893, 2015.