Fairness and Delay in Heterogeneous Half- and Full-Duplex Wireless Networks

Tingjun Chen*, Jelena Diakonikolas†, Javad Ghaderi*, Gil Zussman*
*Department of Electrical Engineering, Columbia University, New York, NY
†Department of Statistics, UC Berkeley, Berkeley, CA, USA
Email: {tingjun, jghaderi, gil}@ee.columbia.edu, jelena.d@berkeley.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. In [1], we focused on the design and performance evaluation of scheduling algorithms for *heterogeneous* HD-FD networks and presented the distributed *Hybrid*-Greedy Maximal Scheduling (H-GMS) algorithm. H-GMS combines the centralized Greedy Maximal Scheduling (GMS) and a distributed queue-based random-access mechanism, and is *throughput-optimal*. In this paper, we analyze the delay performance of H-GMS by deriving two lower bounds on the average queue length. We also evaluate the fairness and delay performance of H-GMS via extensive simulations. We show that in heterogeneous HD-FD networks, H-GMS achieves 16–$30\times$ better delay performance and improves fairness between FD and HD 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 [2]. Recent work has demonstrated physical layer FD operation [3]–[6], 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., [5]–[8]), 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. Traditionally, three approaches have been used for the design of wireless scheduling algorithms that can guarantee maximum throughput:

- **Maximum Weight Scheduling** (MWS) [9], which is a centralized policy that schedules non-conflicting links with the maximum total queue length;
- **Greedy Maximal Scheduling** (GMS) [10], 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, it is generally not equivalent to MWS and is not throughput-optimal in general topologies.
- **Queue-based Random Access Algorithms** (e.g., Q-CSMA) [11], [12], 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 [1], we developed the Hybrid-GMS (H-GMS) algorithm, which is a distributed scheduling algorithm that combines the concepts of GMS and Q-CSMA. Essentially, instead of approximating MWS in a decentralized manner (as in traditional Q-CSMA), H-GMS approximates GMS, which is easier to decentralize in the considered HD-FD networks. H-GMS leverages 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. As shown in [1], [13], H-GMS yields much better delay performance than Q-CSMA while still achieving throughput optimality.

In this paper, we analyze the delay performance of H-GMS in heterogeneous HD-FD networks by deriving two lower bounds on the average queue length: (i) a fundamental lower bound that is independent of the scheduling algorithm, and (ii) a stronger lower bound that takes into account the characteristics of the developed H-GMS. We also evaluate the performance of H-GMS in terms of delay and the fairness between FD and HD users through extensive simulations.

II. Related Work

There has been extensive work dedicated to physical layer FD radio/system design and implementation [2]–[6], [14], and an open-access FD radio design has been integrated with the ORBIT wireless testbed [15]. Recent research also focused

on characterizing and quantifying achievable throughput improvements and rate regions of FD networks with realistic imperfect SIC [16], [17]. 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 [18]–[20] considering signal-to-noise ratio (SNR) or a specific standard (e.g., IEEE 802.11). Most relevant to our work are [20] and [21] in terms of the applied techniques and network model, respectively. In particular, [20] proposed a Q-CSMA-based throughput-optimal scheduling algorithm with FD cut-through transmission in all-FD multi-hop networks, where how HD/FD users are affected by FD transmissions is not studied. [21] proposed a MAC protocol for a heterogeneous HD-FD network and analyzed its throughput based on the IEEE 802.11 distributed coordination function model. 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, with 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 $[N] = \{1, 2, \ldots, N\}$ where 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 independent 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*. A *schedule* at any time slot t is represented by a vector

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

where $X_i^{\text{u}}(t)$ (resp. $X_i^{\text{d}}(t)$) is equal to 1 if the UL (resp. DL) of user i is scheduled to transmit a packet in time slot t and $X_i^{\text{u}} = 0$ (resp. $X_i^{\text{d}} = 0$), 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:

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

where $[\cdot]^+ = \max(0, \cdot)$. $\mathbf{Q}(t) = [Q_i^{\text{u}}(t), Q_i^{\text{d}}(t)]_{i=1}^N$ denotes the queue vector, and $\mathbb{1}(\cdot)$ denotes 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 [9]. Therefore, the capacity region of the heterogeneous HD-FD network is given by $\Lambda_{\text{HD-FD}} = \text{Co}(\mathcal{S})$, where $\text{Co}(\cdot)$ 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 :

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

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 $\Lambda_{\text{HD}} = \text{Co}(\mathbf{e}_1, \ldots, \mathbf{e}_{2N})$, or equivalently

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

A scheduling algorithm is called *throughput-optimal* if it can keep the network queues stable for all arrival rate vectors $\boldsymbol{\lambda} \in \text{int}(\Lambda)$, where $\text{int}(\Lambda)$ denotes the interior of Λ. To compare $\Lambda_{\text{HD-FD}}$ with Λ_{HD} and quantify the network throughput gain when a certain number of HD users become FD-capable, 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 $\boldsymbol{\lambda}_0$, $\gamma(\boldsymbol{\lambda}_0)$, is defined as

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

$\gamma(\cdot)$ 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 $\Lambda_{\text{HD-FD}}$, as N_F users become FD-capable. It is not hard to see that $\gamma : \Lambda_{\text{HD}} \to [1, 2]$.

IV. The Hybrid-GMS (H-GMS) Algorithms

In this section, we briefly describe the H-GMS algorithm presented in [1]. Each slot t is divided into a short control slot and a data slot, where the control slot contains only two control mini-slots. 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 the following three steps:

(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), 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., $i^\star(t) = \arg\max_{i\in\mathcal{N}} Q_i^{\text{d}}(t)$. If multiple DLs have equal (largest) queue length, it breaks ties according to some deterministic rule. Then, the AP randomly selects an initiator link $\text{IL}(t)$ from the set $\mathcal{L}(t) = \{l_1^{\text{u}}, \ldots, l_N^{\text{u}}, l_{i^\star}^{\text{d}}\}$ according to an *access probability* distribution

$\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_N, \alpha_{\text{AP}}]$ satisfying: (i) $\alpha_i > 0, \forall i \in \mathcal{N}$, and $\alpha_{\text{AP}} > 0$, and (ii) $\alpha_{\text{AP}} = 1 - \sum_{i=1}^{N} \alpha_i$. We refer to α_i and α_{AP} as the access probability for user i and the AP. Therefore,

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

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

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

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

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 [11], [12], link l_i^j chooses logistic form

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

where $f(\cdot)$ is a positive increasing function called the *weight function*. 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) 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.

The following theorem on the positive recurrence of the system Markov chain $(\mathbf{X}(t), \mathbf{Q}(t))$ under H-GMS (throughput optimality of H-GMS) was established and proved in [1], [13].

Theorem 4.1. *For any arrival rate vector $\boldsymbol{\lambda} \in \text{int}(\Lambda_{\text{HD-FD}})$, the system Markov chain $(\mathbf{X}(t), \mathbf{Q}(t))$ is positive recurrent under H-GMS. The weight function $f(\cdot)$ in* (5) *can be any non-negative increasing function such that $\lim_{x\to\infty} f(x)/\log x < 1$, or $\lim_{x\to\infty} f(x)/\log x > 1$ (including $f(x) = x^\beta, \ \beta > 0$).*

We also introduce three variants of the H-GMS algorithm, which differ only in Step 1 of H-GMS.

- **H-GMS**: The AP selects the *longest DL*.
- **H-GMS-R**: The AP selects a DL *uniformly at random*, i.e., $i^\star \sim \text{Unif}(1, \ldots, N)$ (in step 1 of H-GMS).
- **H-GMS-E**: Exactly the same as H-GMS except for the access probability being set according to:

$$\widetilde{\alpha}_i \propto \max\{\widetilde{Q}_i^{\text{u}}/(\textstyle\sum_{i'=1}^{N} \widetilde{Q}_{i'}^{\text{u}} + Q_{i^\star}^{\text{d}}), \ \alpha_{\text{th}}\}, \ \forall i \in \mathcal{N},$$

$$\widetilde{\alpha}_{\text{AP}} \propto \max\{Q_{i^\star}^{\text{d}}/(\textstyle\sum_{i'=1}^{N} \widetilde{Q}_{i'}^{\text{u}} + Q_{i^\star}^{\text{d}}), \ \alpha_{\text{th}}\},$$

where $\widetilde{Q}_i^{\text{u}}$ is an estimate of UL queue length of user i. Specifically, when a user transmits on the UL, it includes its queue length in the packet and the AP updates $\widetilde{Q}_i^{\text{u}}$ using the most recently received packet. The minimum access probability $\alpha_{\text{th}} > 0$ has been introduced to ensure that each link is selected with a non-zero probability. Otherwise, an HD UL l_i^{u} ($\forall i \in \mathcal{N}_H$) with a zero queue-length estimate would never be selected by the AP (i.e., $\widetilde{Q}_i^{\text{u}} = 0$ and thus $\widetilde{\alpha}_i = 0$), and the AP would never receive any updated information of $\widetilde{Q}_i^{\text{u}}$ since $\widetilde{\alpha}_i$ would remain zero. Then, $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_N, \alpha_{\text{AP}}]$ is obtained after normalization, i.e.,

$$\alpha_i = \frac{\widetilde{\alpha}_i}{\sum_{i'=1}^{N} \widetilde{\alpha}_{i'} + \widetilde{\alpha}_{\text{AP}}}, \ \forall i \in \mathcal{N}, \ \alpha_{\text{AP}} = \frac{\widetilde{\alpha}_{\text{AP}}}{\sum_{i'=1}^{N} \widetilde{\alpha}_{i'} + \widetilde{\alpha}_{\text{AP}}}.$$

V. Lower Bounds on the Average Queue Length

In this section, we analyze the delay performance of H-GMS in terms of the average queue length in order to provide a benchmark for the performance evaluation in Section VI. We adopt the following notation. Given a set of links $\mathcal{L}$, we use $\lambda_{\mathcal{L}} = \sum_{l\in\mathcal{L}} \lambda_l$ to denote the sum of arrival rates, and use $\mathcal{Q}_{\mathcal{L}} = \sum_{l\in\mathcal{L}} \mathbb{E}[Q_l]$ to denote the expected sum of queue lengths of $\mathcal{L}$ in steady state. The average queue length in a given heterogeneous HD-FD network, $(\mathcal{N}, \mathcal{E})$, is defined by

$$\overline{Q} = \textstyle\sum_{l\in\mathcal{E}} \mathbb{E}[Q_l]/|\mathcal{E}| = \mathcal{Q}_{\mathcal{E}}/(2N). \tag{6}$$

We divide $\mathcal{E}$ into two disjoint sets $\mathcal{E} = \mathcal{E}_{\max} \cup \mathcal{E}_{\min}$:

$$\begin{cases} \mathcal{E}_{\max} = \{l_i^j : \forall i \in \mathcal{N}_F \text{ if } \lambda_i^j \geq \lambda_i^{\bar{j}}\} \cup \{l_i^{\text{u}}, l_i^{\text{d}} : \forall i \in \mathcal{N}_H\}, \\ \mathcal{E}_{\min} = \{l_i^j : \forall i \in \mathcal{N}_F \text{ if } \lambda_i^j < \lambda_i^{\bar{j}}\}, \end{cases}$$

where $\{\bar{j}\} = \{\text{u}, \text{d}\} \setminus \{j\}$ and we break ties uniformly at random if $\lambda_i^{\text{u}} = \lambda_i^{\text{d}}$ for $\forall i \in \mathcal{N}_F$. Essentially, $\mathcal{E}_{\max}$ includes the UL *and* DL of each HD user, and the higher arrival rate link (UL *or* DL) of each FD user. As a result, $\lambda_{\mathcal{E}_{\max}}$ approaches 1 as $\boldsymbol{\lambda}$ approaches the boundary of $\Lambda_{\text{HD-FD}}$. Our main results on the lower bounds on the average queue length are summarized in the following two propositions, whose proofs are in [13].

Proposition 5.1 (A fundamental lower bound)**.** *The average queue length in the considered heterogeneous HD-FD networks is lowered bounded by $\overline{Q}_{Fund}^{LB}$, which is given by*

$$\overline{Q} \geq \overline{Q}_{Fund}^{LB} := \mathcal{Q}_{\mathcal{E}_{max}}^{LB}/(2N), \tag{7}$$

where $\mathcal{Q}_{\mathcal{E}_{max}}^{LB}$ is given by [22, Proposition 4.1]

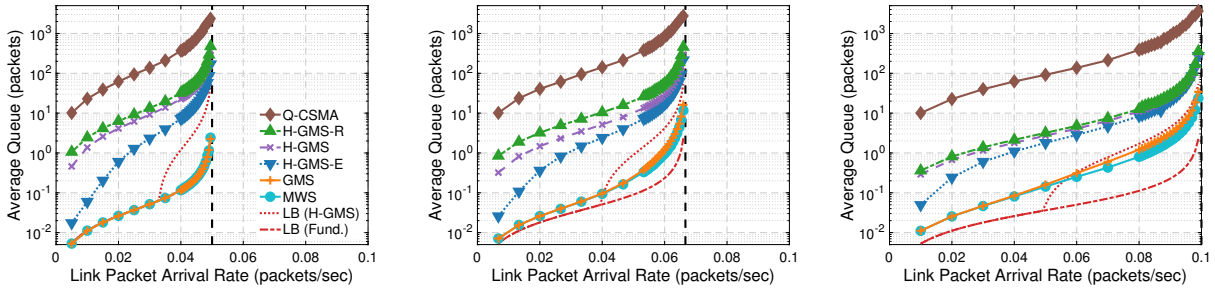
$$\mathcal{Q}_{\mathcal{E}_{max}}^{LB} := \sum_{l\in\mathcal{E}_{max}} \frac{\lambda_l + \text{Var}\left[A_l\right] - \lambda_l \lambda_{\mathcal{E}_{max}}}{2\left(1 - \lambda_{\mathcal{E}_{max}}\right)}.$$

Proposition 5.2 (An improved lower bound)**.** *Let $p^{-1}(\cdot)$ be the inverse of the transmission probability $p(\cdot)$ given by* (5)*. Let $\lambda_{min} = \min_{i\in\mathcal{N}}\{\lambda_i^u, \lambda_i^d\}$ be the minimum link arrival rate and $\alpha_{max} = \max\{\alpha_1, \ldots, \alpha_N, \alpha_{AP}\}$ be the maximum access probability. The average queue length under H-GMS and H-GMS-R is lower bounded by $\overline{Q}_{H\text{-}GMS}^{LB}$ given by*

$$\overline{Q} \geq \overline{Q}_{H\text{-}GMS}^{LB} := \max\Big\{\overline{Q}_{Fund}^{LB}, \Big(1 - \frac{N_F}{2N}\Big) \cdot p^{-1}\Big(\frac{\lambda_{min}/\alpha_{max}}{1 - \lambda_{\mathcal{E}_{max}} + \lambda_{min}/\alpha_{max}}\Big)\Big\}. \tag{8}$$

VI. Simulation Results

In this section, we evaluate the performance of different scheduling algorithms in heterogeneous HD-FD networks via



(a) $N_F = 0$, $N_H = 10$ (b) $N_F = 5$, $N_H = 5$ (c) $N_F = 10$, $N_H = 0$

Fig. 1: Long-term average queue length per link in a heterogeneous HD-FD network with $N = 10$ and equal arrival rates, under varying number of FD users, N_F: (a) $N_F = 0$, (b) $N_F = 5$, and (c) $N_F = 10$. Both the fundamental and improved lower bounds on the delay are also plotted according to (7) and (8). The capacity region boundary in each HD-FD network is illustrated by the vertical dashed line.

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).

• **Setup**. We consider heterogeneous HD-FD networks with one FD AP and 10 users ($N = 10$), with a varying number of FD users, N_F. 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 $\Lambda_{\text{HD-FD}}$ 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 $\Lambda_{\text{HD-FD}}$. 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 UL and DL arrival rates assigned to FD and HD users, respectively. Since we focus on the fairness between FD and HD users, we assume equal UL and DL arrival rates over all the users. For this *equal arrival rate* model, we have $v_f = v_h = 1/(N_F + 2N_H)$ (see (1)).

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.

The considered algorithms include: MWS, GMS, H-GMS, H-GMS-R, H-GMS-E, and Q-CSMA. We adopt the Q-CSMA algorithm from [12] where each link (UL or DL) performs channel contention independently and the AP does not leverage the central DL queue information. In the last 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 the weight function $f(x) = \log(1 + x)$. We set $\boldsymbol{\alpha} = \frac{1}{1+N} \cdot \mathbf{1}$ for H-GMS and H-GMS-R, and $\alpha_{\text{th}} = 0.01$ for H-GMS-E (see Section IV). 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.

• **Delay Performance**. We first consider the delay performance of various scheduling algorithms. Fig. 1 plots the average queue length with varying traffic intensities in HD-FD networks with $N = 10$ and $N_F \in \{0, 5, 10\}$. Fig. 1 shows that the capacity region of the HD-FD networks expands with increased value of N_F. Compared with Fig. 1(a), Figs. 1(b) and 1(c) show a capacity region expansion value of $\gamma = 4/3$ for $N_F = 5$, and $\gamma = 2$ for $N_F = 10$, respectively.

Fig. 1 shows that all the considered algorithms are throughput-optimal – 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 [12] has the worst delay performance due to the high contention intensity introduced by a total of $2N$ contending links. By "consolidating" the N DLs into one DL that participates in channel contention, H-GMS-R, H-GMS, and H-GMS-E achieve at least 9–16×, 16–30×, and 25–50× better delay performance than Q-CSMA, respectively, under different traffic intensities ρ. In particular, H-GMS and H-GMS-E have similar delay performance which is better than for H-GMS-R, since the AP leverages its central information to always select the longest queue DL for channel contention. However, H-GMS and H-GMS-E 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).

Fig. 1 also plots both the fundamental and improved lower bounds on the average queue length, $\overline{Q}^{\text{LB}}_{\text{Fund}}$ and $\overline{Q}^{\text{LB}}_{\text{H-GMS}}$, given by (7) and (8), respectively. As Fig. 1 suggests, the average queue lengths obtained by MWS and GMS are very close to $\overline{Q}^{\text{LB}}_{\text{Fund}}$ (they indeed match perfectly in the all-HD network). However, in heterogeneous HD-FD networks, $\overline{Q}^{\text{LB}}_{\text{H-GMS}}$ provides a much tighter lower bound on the average queue length achieved by H-GMS, especially with high traffic intensities.

• **Fairness between FD and HD users**. Our next focus is on the fairness performance of H-GMS. Here, we define fairness between FD and HD users 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.

We first evaluate the fairness under different distributed algorithms with *equal arrival rates* at each link. Fig. 2 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. With low or moderate traffic intensity, Q-CSMA and H-GMS 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 since they are being served about twice as often (i.e., an FD bi-directional

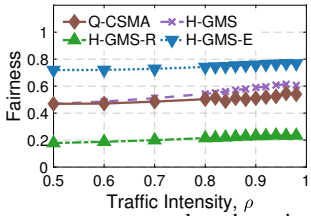
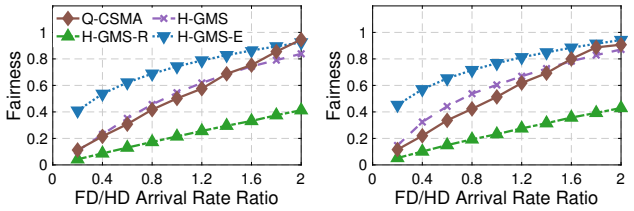


Fig. 2: Long-term average queue length ratio between FD and HD users with varying traffic intensity in a heterogeneous HD-FD network with $N_F = N_H = 5$ and equal arrival rates.



(a) Traffic intensity $\rho = 0.8$ (b) Traffic intensity $\rho = 0.95$

Fig. 3: 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 (a) moderate ($\rho = 0.8$), and (b) high ($\rho = 0.95$) traffic intensities.

transmission can be activated by either the FD UL or DL due to the FD PHY capability). With high traffic intensity, both H-GMS 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 since, under H-GMS-E, the AP not only has explicit information of all the DL queues, but also has estimated UL queue lengths that can be used to better assign the access probability distribution $\boldsymbol{\alpha}$.

We also evaluate the fairness with *different arrival rates* between FD and HD users. Denote by σ the ratio between FD and HD link arrival rates, resulting in $v_f = \sigma/(\sigma N_F + 2N_H)$ and $v_h = 1/(\sigma N_F + 2N_H)$. Fig. 3 plots the fairness between FD and HD users with varying σ under moderate ($\rho = 0.8$) and high ($\rho = 0.95$) traffic intensities. 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, and H-GMS-E. It is interesting to note that the fairness under Q-CSMA and H-GMS 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 the generated $\boldsymbol{\alpha}$ is approximately a uniform distribution.

VII. Conclusion

In [1], we presented H-GMS, a distributed scheduling algorithm designed for heterogeneous HD-FD networks while achieving throughput optimality. In this paper, we analyzed the delay performance of H-GMS by deriving two lower bounds on the average queue length. Then, via extensive simulations, we evaluated the delay and fairness performance of H-GMS in HD-FD networks with various settings.

Acknowledgment

This work was supported in part by ARO grant 9W911NF-16-1-0259, NSF grants ECCS-1547406, CNS-1650685, and CNS-1717867.

References

[1] T. Chen, J. Diakonikolas, J. Ghaderi, and G. Zussman, "Hybrid scheduling in heterogeneous half-and full-duplex wireless networks," in *Proc. IEEE INFOCOM'18*, 2018.

[2] 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.

[3] 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.

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

[5] 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.

[6] T. Chen, M. B. Dastjerdi, J. Zhou, H. Krishnaswamy, and G. Zussman, "Wideband full-duplex wireless via frequency-domain equalization: Design and experimentation," in *Proc. ACM MobiCom'19 (to appear)*, 2019.

[7] 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.

[8] H. Krishnaswamy and G. Zussman, "1 Chip 2x the bandwidth," *IEEE Spectrum*, vol. 53, no. 7, pp. 38–54, 2016.

[9] 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.

[10] 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.

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

[12] 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, 2012.

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

[14] 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, 2015.

[15] T. Chen, M. Baraani Dastjerdi, J. Zhou, H. Krishnaswamy, and G. Zussman, "Open-access full-duplex wireless in the ORBIT testbed," *arXiv preprint arXiv:1801.03069*, 2018.

[16] 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.

[17] 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.

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

[19] 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, 2017.

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

[21] 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.

[22] G. R. Gupta and N. B. Shroff, "Delay analysis for wireless networks with single hop traffic and general interference constraints," *IEEE/ACM Trans. Netw.*, vol. 18, no. 2, pp. 393–405, 2010.