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

Tingjun Chen, Jelena Diakonikolas, Javad Ghaderi, and Gil Zussman

***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* HD-FD networks (composed of HD and FD users). We first show that *centralized* Greedy Maximal Scheduling (GMS) is throughput-optimal in heterogeneous HD-FD networks. We propose the *Hybrid*-GMS (H-GMS) algorithm, a distributed implementation of GMS that combines GMS and a queue-based random-access mechanism. *We prove that H-GMS is throughput-optimal*. Moreover, we analyze the delay performance of H-GMS by deriving lower bounds on the average queue length. We further demonstrate the benefits of upgrading HD nodes to FD nodes in terms of throughput gains for individual nodes and the whole 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 $16$–$30\times$ 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 [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.

This research was supported in part by NSF grants ECCS-1547406, CNS-1650685, and CNS-1717867, ARO grant W911NF-16-1-0259. A partial and preliminary version of this paper appeared in IEEE INFOCOM'18, Apr. 2018 [1], and some results were presented in the Asilomar Conference on Signals, Systems, and Computers (invited), Oct. 2018.

T. Chen, J. Ghaderi, and G. Zussman are with the Department of Electrical Engineering, Columbia University, New York, NY, USA (email: {tingjun, jghaderi, gil}@ee.columbia.edu).

J. Diakonikolas is with the Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI, USA (e-mail: jelena@cs.wisc.edu).

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.

There are three main approaches to wireless schedulingalgorithms that can guarantee maximum throughput:

**Maximum Weight Scheduling (MWS)** [9], 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)** [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, in general, it is not equivalent to MWS and is not throughput-optimal in general topologies.

**Queue-Length-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 be throughput-optimal. However, they generally experience poor delay performance and suffer from excessive queue lengths.

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. We first show by using the notion of Local Pooling [10], [13] that GMS is throughput-optimal in the considered HD-FD networks. However, since GMS is fully centralized, we leverage ideas from distributed Q-CSMA to develop the *Hybrid*-GMS (H-GMS) algorithm that combines centralized GMS with distributed Q-CSMA. The main feature of the proposed H-GMS algorithm is that instead of approximating MWS (as done in "traditional" Q-CSMA), it approximates GMS.

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 any DL queues or to any 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, 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 (i.e., not MWS). We stress that, unlike MWS and Q-CSMA, H-GMS is built on top of an algorithm (GMS) which is not throughput-optimal in general wireless networks, but for which we are able to show such a result thanks to the special hybrid HD-FD network structure. Due to the critical differences between H-GMS, MWS, and Q-CSMA, to prove the throughput optimality of H-GMS, it is not sufficient to apply any of the existing results in a black-box manner. Thus, to obtain such a result, we need to establish its fluid limits. The results are general, and, unlike most existing work, apply to a wide range of weight functions (see Section IV for more information).

We also present variants of H-GMS with different degrees of centralization. To understand the delay performance of H-GMS, in Section VI, we derive 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 and applies to all its non-adaptive variants. These lower bounds serve as benchmarks when evaluating the delay performance of H-GMS.

Before thoroughly evaluating H-GMS and its variants, we demonstrate 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 double the 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 specific 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 with the number of FD users, and it reaches a gain of 2 when all users are FD-capable. 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*.

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 illustrate 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 16–30$\times$ 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 heterogeneous HD-FD networks.

The rest of the paper is organized as follows. We discuss related work in Section II and introduce the network model and preliminaries in Section III. We developed H-GMS algorithm in Section IV and the proof of its throughput optimality is presented in Section V. The delay analysis of H-GMS and lower bounds on the average queue length are presented in Section VI. We then illustrate the benefits of introducing FD nodes into legacy HD networks in Section VII. We evaluate the performance of different scheduling algorithms via simulations in Section VIII and conclude in Section IX.

## II. Related Work

There has been extensive work dedicated to physical layer FD radio/system design and implementation [3], [4], [6], [8], [14] (see also the review in [2] and references therein). Open-access FD radios based on [8], [15] have been integrated with the ORBIT and COSMOS wireless testbed [16], [17]. Recent research also focused on characterizing and quantifying achievable throughput improvements and rate regions of FD networks in both single- and multi-channel cases with realistic imperfect SIC [18]–[20]. 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 [21]–[26] considering signal-to-noise ratio (SNR) or a specific standard (e.g., IEEE 802.11 where an FD topic interest group was recently formed [27]). For example, [22] 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, [21] considered an all-FD network and proposed a distributed MAC protocol based on the 802.11 DCF. Most relevant to our work are [26] and [28] in terms of the applied techniques and network model, respectively. In particular, [26] proposed a Q-CSMA-based throughput-optimal scheduling algorithm with FD cut-through transmission in all-FD multi-hop networks, where the effect of different classes of users (HD and FD) on the FD transmissions is not studied. On the other hand, [28] proposed a MAC layer algorithm for a heterogeneous HD-FD network and analyzed its throughput based on the IEEE 802.11 distributed coordination function (DCF) model [29]. 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 $[N] = \{1, 2, \cdots, N\}$ 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*.[1] A *schedule* at any time slot $t$ is represented by a vector

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

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 $i^{\text{th}}$ basis vector (i.e., an all-zero vector except the $i^{\text{th}}$ 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[2]:

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

[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 $\max\{\lambda_i^{\text{u}}, \lambda_i^{\text{d}}\}$ with $\lambda_i$ and adding linear inequalities $\lambda_i^{\text{u}} \leq \lambda_i, \lambda_i^{\text{d}} \leq \lambda_i$.

**Algorithm 1** MWS for HD-FD Networks (in slot $t$)

1. Initialize $\mathbf{X}(t) = \mathbf{0}$.
2. Let $\hat{Q}_i(t) = Q_i^{\text{u}}(t) + Q_i^{\text{d}}(t), \ \forall i \in \mathcal{N}_F$. Select user $i^\star = \arg\max \{ \max_{i\in\mathcal{N}_F}\{\hat{Q}_i(t)\}, \max_{i\in\mathcal{N}_H}\{Q_i^{\text{u}}(t), Q_i^{\text{d}}(t)\}\}$. Break ties uniformly at random.
3. • If $i^\star \in \mathcal{N}_F$, set $X_i^{\text{u}}(t) = X_i^{\text{d}}(t) = 1$;
   • If $i^\star \in \mathcal{N}_H$, set $X_i^j(t) = 1$ if $Q_i^j(t) \geq Q_i^{\{\text{u},\text{d}\}\setminus\{j\}}(t)$. Break ties uniformly at random.
4. Use $\mathbf{X}(t)$ as the transmission schedule in slot $t$.

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

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

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 $\Lambda$.

To compare $\Lambda_{\text{HD-FD}}$ with $\Lambda_{\text{HD}}$ and quantify the network throughput gain when a certain number of HD users become FD-capable, similar to [18], we define the *capacity region expansion function* $\gamma(\cdot)$ as follows. Given $\boldsymbol{\lambda}_0$ on the Pareto boundary of $\Lambda_{\text{HD}}$, the capacity region expansion function at point $\boldsymbol{\lambda}_0$, denoted by $\gamma(\boldsymbol{\lambda}_0)$, is defined as

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

$\gamma(\cdot)$ can be interpreted as a function that scales an arrival rate vector on the Pareto boundary of $\Lambda_{\text{HD}}$ 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. Scheduling Algorithms and Main Result

In this section, we develop a hybrid scheduling algorithm tailored for heterogeneous HD-FD networks. We first briefly introduce MWS in the considered networks. We then use Local Pooling [10], [13] to prove that GMS is throughput-optimal in the considered networks, and therefore, MWS [9] is unneeded. Based on that, we present the H-GMS algorithm – a decentralized version of GMS that leverages ideas from distributed Q-CSMA [11], [12]. H-GMS uses information about the DL queues that is available at the AP, but does not require global information about the UL queues. We state the main result (Theorem 4.1) about the throughput optimality of H-GMS and describe its various implementations with different levels of centralization. We later show (in Section VIII) that these variants of H-GMS have different delay performance.

### A. Centralized Max-Weight Scheduling (MWS)

We first describe the throughput-optimal MWS (see Algorithm 1), where in each time slot $t$, the schedule $\mathbf{X}(t) \in \mathcal{S}$ with the maximum sum queue length is selected.

### B. Centralized Greedy Maximal Scheduling (GMS)

We now show that a GMS (see Algorithm 2) is throughput-optimal in *any* collocated heterogeneous HD-FD network, independent of the values of $N_F$ and $N_H$. In both MWS and GMS, 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.

**Algorithm 2** 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 $l_i^{\mathrm{d}}$ 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 $t$.

**Proposition 4.1.** *The Greedy Maximal Scheduling (GMS) algorithm is throughput-optimal in any collocated heterogeneous HD-FD network.*

The proof (see Appendix A) is based on [10, Theorem 1], [13], 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.

### C. Hybrid-GMS (H-GMS) Algorithm

We now present a hybrid scheduling algorithm, H-GMS, which combines the concepts of GMS and Q-CSMA [11], [12]. Instead of approximating MWS [9] 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. Therefore, the number of possible participants under H-GMS in each slot is at most $(N+1)$. Moreover, we show that this hybrid approach yields much better delay performance than Q-CSMA while still achieving throughput optimality.

Algorithm 3 presents the pseudocode for H-GMS, which 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 (independent of the number of users, $N$). 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) $\alpha_{\mathrm{AP}} = 1 - \sum_{i=1}^N \alpha_i$.

**Algorithm 3** H-GMS Algorithm (in slot $t$)

– If $\mathbf{X}(t-1) = \mathbf{0}$:

1. In the initiation mini-slot, the AP computes $i^\star = \arg\max_{i\in\mathcal{N}} Q_i^{\mathrm{d}}(t)$. 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 $\overline{p}_{i^\star}^{\mathrm{d}}(t) = 1 - p_{i^\star}^{\mathrm{d}}(t)$;
   • 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 $\overline{p}_i^{\mathrm{u}}(t) = 1 - p_i^{\mathrm{u}}(t)$;
   • 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.

We refer to $\alpha_i$ and $\alpha_{\mathrm{AP}}$ 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] Note that the real-time information of *all* the UL queue lengths is *not* shared or available at the AP.

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

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

where $f(\cdot)$ is a positive increasing function (to be specified later), 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] 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.

**(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 4.1.** *The initiation step in H-GMS is described as a polling mechanism where the AP draws a link IL$(t)$ from $\mathcal{L}(t)$ according to the access 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 $\alpha_i$. If AP receives the message, it sends back a clear-to-initiate message and sets IL$(t) = l_i^u$, otherwise (i.e., in case of collision or idleness) $l_{i^\star}^d$ is selected as the initiator link by the AP. This effectively emulates polling user $i$ with probability $\widetilde{\alpha}_i = \alpha_i \prod_{i' \neq i}(1-\alpha_{i'})$ and AP with probability $\widetilde{\alpha}_{AP} = 1 - \sum_{i=1}^N \widetilde{\alpha}_i$.*

### D. Main Result: Throughput Optimality of H-GMS

The system state under H-GMS evolves as a Markov chain $(\mathbf{X}(t), \mathbf{Q}(t))$. The following theorem states our main result regarding the positive recurrence of this Markov chain (throughput optimality of H-GMS).

**Theorem 4.1.** *For any arrival rate vector $\boldsymbol{\lambda} \in \text{int}(\Lambda_{HD\text{-}FD})$, the system Markov chain $(\mathbf{X}(t), \mathbf{Q}(t))$ is positive recurrent under H-GMS (Algorithm 3). 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 $f(x) = x^\beta,\ \beta > 0$).*

Establishing Theorem 4.1 is not trivial due to the coupling between $\mathbf{X}(t)$ and $\mathbf{Q}(t)$: The dynamics of the schedule process $\mathbf{X}(t)$ is governed by the queue process $\mathbf{Q}(t)$, while at the same time, the dynamics of $\mathbf{Q}(t)$ depends on $\mathbf{X}(t)$. Depending on the functional shape of the weight function $f(\cdot)$, this coupling gives rise to vastly different behaviors for the Markov chain $(\mathbf{X}(t), \mathbf{Q}(t))$. For functions $f(\cdot)$ that grow slower than $\log(\cdot)$, the convergence of the schedule process $\mathbf{X}(t)$ occurs on a much faster time-scale ("fast mixing") compared to the time-scale of changes in the queue process $\mathbf{Q}(t)$. For more aggressive functions $f(\cdot)$, the convergence of $\mathbf{X}(t)$ occurs on a much slower time-scale ("slow mixing") compared to the time-scale of changes in $\mathbf{Q}(t)$. Nevertheless, Theorem 4.1 states that the system Markov chain is stable (positive recurrent) for a wide range of weight functions. We provide a proof of Theorem 4.1 in Section V based on the analysis of the fluid limits of the system under the H-GMS algorithm.

### E. 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 3.

- H-GMS (Algorithm 3): The AP selects the *longest DL*.
- H-GMS-R: The AP selects a DL *uniformly at random*, i.e., $i^\star \sim \text{Unif}(1, \cdots, N)$ (in step 1 of Algorithm 3).
- 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 $\widetilde{Q}_i^{\text{u}}$ an estimate of UL queue length of user $i$. Specifically, when a user transmits on the UL, it includes its queue length in the packets and the AP updates $\widetilde{Q}_i^{\text{u}}$ using this information contained in the last (i.e., most recently) received packet from user $i$ on the UL. Then, $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_N, \alpha_{\text{AP}}]$ is obtained after normalization, i.e., $\alpha_i = \dfrac{\widetilde{\alpha}_i}{\sum_{i'=1}^N \widetilde{\alpha}_{i'} + \widetilde{\alpha}_{\text{AP}}},\ \forall i \in \mathcal{N},\ \alpha_{\text{AP}} = \dfrac{\widetilde{\alpha}_{\text{AP}}}{\sum_{i'=1}^N \widetilde{\alpha}_{i'} + \widetilde{\alpha}_{\text{AP}}}$. A 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 $l_i^{\text{u}}$ ($\forall i \in \mathcal{N}_H$) 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. The access probability distribution $\boldsymbol{\alpha}$ is *non-adaptive* in H-GMS and H-GMS-R, and is *adaptive* in H-GMS-E. As we will see in Section VIII, the adaptive choice of $\boldsymbol{\alpha}$ helps balance the queue lengths between FD and HD users.

## V. Proof of Theorem 4.1 via Fluid Limits

We prove Theorem 4.1 based on the analysis of the fluid limits of the system under H-GMS (Algorithm 3). The proof has three parts: (i) existence of the fluid limits (Lemma 5.1), (ii) deriving the fluid limit equations for the various choices of $f(\cdot)$ (Lemma 5.3), and (iii) proving the stability of the queues in the fluid limits using a Lyapunov method, which implies the stability of the original stochastic process.

**Part (i): Definition and Existence of Fluid Limits.**

Consider 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 $\mathbf{q}(t)$ 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 $t \geq 0$,

$$\mathbf{Q}^{(r)}(t) = \mathbf{Q}^{(r)}(0) + \overline{\mathbf{A}}^{(r)}(t) - \overline{\mathbf{S}}^{(r)}(t), \tag{6}$$

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

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

Similarly, we denote by $\mathbf{a}(t)$ and $\mathbf{s}(t)$ the limits of the scaled processes $\overline{\mathbf{A}}^{(r)}(t)$ and $\overline{\mathbf{S}}^{(r)}(t)$ as $r \to \infty$, respectively. 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) in the case of $\lim_{x\to\infty} f(x)/\log x < 1$; nevertheless, the weak convergence is sufficient for our proofs.

**Lemma 5.1** (Existence of Fluid Limits)**.** *Suppose $\mathbf{Q}^{(r)}(0) \to \mathbf{q}(0)$. Then any sequence $r$ has a subsequence such that $(\mathbf{Q}^{(r)}(t), \overline{\mathbf{A}}^{(r)}(t), \overline{\mathbf{S}}^{(r)}(t)) \Rightarrow (\mathbf{q}(t), \mathbf{a}(t), \mathbf{s}(t))$ 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., [30]. ∎

**Part (ii): Fluid Limit Equations under H-GMS.**

Recall that the schedule $\mathbf{X}(t)$ at time $t$ is determined after the Initiation and Coordination steps of Algorithm 3. Let $Y(t)$ indicate the initiator link which is activated in slot $t$. 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 $S_Y = \{0, 1, \cdots, N, i^\star\}$, where $Y(t) = 0$ means no link is active, $Y(t) = i^\star$ means DL $l_{i^\star}^{\mathrm{d}}$ is active, and $Y(t) = i$, for $i \in \{1, \cdots, N\}$, means UL $l_i^{\mathrm{u}}$ is active. We further use $\{Y^{\mathbf{Q}}(t)\}_{t\geq t_0}$ to denote the dynamics of $Y(t)$, assuming a fixed queue length vector $\mathbf{Q}(t) = \mathbf{Q}(t_0) = \mathbf{Q}$ for all times $t \geq t_0$. Under the H-GMS algorithm, $\{Y^{\mathbf{Q}}(t)\}_{t\geq t_0}$ evolves as an irreducible and aperiodic Markov chain over the state space $S_Y$. If $Y^{\mathbf{Q}}(t) = i$ for an $i$ which is an initiator UL or DL of an FD user $i \in \mathcal{N}_F$, then the other link of the same FD user will follow the initiator link and become active as well under H-GMS. Due to the activation/deactivation coordination among the initiator link and the follower link, adding the possible follower link does not change the subsequent dynamics of the Markov chain $Y^{\mathbf{Q}}(t)$ under fixed $\mathbf{Q}$.

Let $\mathbf{P}^{\mathbf{Q}} = [P(s, s')]$ be the transition probability matrix of $Y^{\mathbf{Q}}(t)$, where $P(s, s')$ is the transition probability from state $s \in S_Y$ to $s' \in S_Y$. Then, under Algorithm 3, we have

$$\begin{aligned} &P(0,i) = \alpha_i p_i^{\mathrm{u}},\ P(i,i) = p_i^{\mathrm{u}},\quad P(i,0) = \overline{p}_i^{\mathrm{u}},\ \forall i \in \mathcal{N} \\ &P(0,i^\star) = \alpha_{\mathrm{AP}} p_{i^\star}^{\mathrm{d}},\ P(i^\star,i^\star) = p_{i^\star}^{\mathrm{d}},\ P(i^\star,0) = \overline{p}_{i^\star}^{\mathrm{d}}, \\ &P(0,0) = 1 - \textstyle\sum_{i=1}^N P(0,i) - P(0,i^\star). \end{aligned} \tag{7}$$

**Lemma 5.2.** *The steady-state distribution of Markov chain $Y^{\mathbf{Q}}(t)$ is given by*

$$\begin{aligned} \pi^{\mathbf{Q}}(i) &= \alpha_i \exp(f(Q_i^u))/Z,\ i \in S_Y \setminus \{0, i^\star\}; \\ \pi^{\mathbf{Q}}(i^\star) &= \alpha_{AP} \exp(f(Q_{i^\star}^d))/Z,\ \pi^{\mathbf{Q}}(0) = 1/Z, \end{aligned} \tag{8}$$

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

The proof of Lemma 5.2 is in Appendix B. The following corollary is immediate as the result of Lemma 5.2 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 5.1.** *Let $\mathbf{f}_i = \mathbf{e}_{2i-1} + \mathbf{e}_{2i}$, $i \in \mathcal{N}_F$, be an FD bi-directional transmission schedule, and $\mathbf{h}_i^u = \mathbf{e}_{2i-1}$ ($\mathbf{h}_i^d = \mathbf{e}_{2i}$), $i \in \mathcal{N}_H$, be an HD UL (DL) transmission schedule. Given a fixed queue vector $\mathbf{Q}(t) = \mathbf{Q}$, in steady state, if $i^\star \in \mathcal{N}_F$,*

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

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

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

*where $Z$ and $f(\cdot)$ are as in Lemma 5.2.*

Consider a fluid sample path under our H-GMS algorithm. Suppose $\mathbf{q}(t) = \mathbf{q} \neq \mathbf{0}$ at a time $t$. This implies that for $r$ large enough, all the queues with non-zero fluid limit $q_i^j > 0$ are of size $Q_i^j = \mathcal{O}(q_i^j r)$ in the original process, while all the queues with zero fluid limit are of size $Q_i^j = o(r)$ in the original process. Therefore, taking the limit $r \to \infty$ in (8), and noting that the weight function $f(\cdot)$ is a positive increasing function of the queue size, it follows that

$$\begin{aligned} \pi^{\mathbf{Q}}(i) \to 0 \text{ if } q_i^{\mathrm{u}} = 0,\ i \in S_Y \setminus \{0, i^\star\}, \\ \pi^{\mathbf{Q}}(i^\star) \to 0 \text{ if } q_{i^\star}^{\mathrm{d}} = 0,\ \pi^{\mathbf{Q}}(0) = 0. \end{aligned}$$

This shows that a queue with a zero fluid limit *cannot* initiate transmission in steady state. Consequently,

$$\begin{aligned} &\mathbb{P}\{\mathbf{X} = \mathbf{f}_i\} \to 0,\ \text{if } \max\{q_i^{\mathrm{u}}, q_i^{\mathrm{d}}\} = 0,\ i \in \mathcal{N}_F, \\ &\mathbb{P}\{\mathbf{X} = \mathbf{h}_i^{\mathrm{u}}\} \to 0,\ \text{if } q_i^{\mathrm{u}} = 0,\ i \in \mathcal{N}_H, \\ &\mathbb{P}\left\{\mathbf{X} = \mathbf{h}_i^{\mathrm{d}}\right\} \to 0,\ \text{if } q_i^{\mathrm{d}} = 0,\ i \in \mathcal{N}_H. \end{aligned}$$

Hence, in steady state, with high probability, the Markov chain $\mathbf{X}(t)$ never activates an HD link with empty fluid limit queue or an FD link whose both UL and DL queues are empty, i.e., it chooses a *Maximal Schedule* over the non-zero fluid queues (note that the returned schedule might not be a MWS schedule). However, as mentioned in Section IV-D, the Markov chain $\mathbf{X}(t)$ might not always be at its steady state due to coupling between $\mathbf{X}(t)$ and $\mathbf{Q}(t)$. This coupling gives rise to qualitatively different fluid limits, depending on the time-scale of convergence of the schedule process compared to the time-scale of the changes in the queue process. For weight functions $f(\cdot)$, such that $\lim_{r\to\infty} f(r)/\log r < 1$, the schedule process $\mathbf{X}(t)$ is always close to its steady state at the fluid scale; while for functions $f(\cdot)$ with $\lim_{r\to\infty} f(r)/\log r > 1$, this does not happen. Nevertheless, in both cases, the following Lemma establishes a set of equations that the fluid limit sample paths under H-GMS algorithm must satisfy. The equations do not uniquely describe the fluid limit process but are sufficient to establish stability in our setting.

**Lemma 5.3** (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 $\widehat{q}_i(t) = \max\{q_i^u(t), q_i^d(t)\}$, 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 $j \in \{u, d\}$,*

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

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

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

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

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

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

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

The proof of Lemma 5.3 is provided Appendix C. Essentially, (9)–(10) hold for any scheduling algorithm and their proof is standard. $\mu_i^j(t)$ is the rate that queue $q_i^j(t)$ is served at time $t$ in the fluid limit. (11)–(14) imply that at any time, H-GMS chooses a maximal schedule from the queues that are *non-zero* in the fluid limit (i.e., $\mathbf{q}(t) \neq \mathbf{0}$). However, the choice of maximal schedule could be random over the space of such

maximal schedules.

**Part (iii): Stability of the Queues in the Fluid Limit.**

The following proposition proves the stability of the queues in the fluid limit, which completes the proof of Theorem 4.1.

**Proposition 5.2.** *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 reach zero.*

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

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

Let $\mathcal{U}_H^j(t) := \{i \in \mathcal{N}_H : q_i^j(t) > 0\}$, $j \in \{\mathrm{u}, \mathrm{d}\}$, and $\mathcal{U}_F(t) := \{i \in \mathcal{N}_F : \widehat{q}_i(t) > 0\}$. 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) $\widehat{q}_i(t)$ 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,

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

where the last inequality is due to the fact that $\boldsymbol{\lambda} \in \mathrm{int}(\Lambda_{\text{HD-FD}})$, by the assumption of Theorem 4.1. Thus, there must exist a small $\delta > 0$ such that $\boldsymbol{\lambda}/(1-\delta) \in \Lambda_{\text{HD-FD}}$. Therefore, $V(\mathbf{q}(t))$ will hit zero in finite time $T = V(\mathbf{q}(0))/\delta$, and in fact remains zero afterwards. ∎

Proposition 5.2 implies the stability (positive recurrence) of the original Markov chain $(\mathbf{X}(t), \mathbf{Q}(t))$ in a similar fashion as [31] (note that the component $\mathbf{X}(t)$ lives in a finite state space). This completes the proof of Theorem 4.1.

**Remark 5.2.** *We emphasize that, the proposed H-GMS algorithm approximates GMS in a distributed manner, where the fluid limits are largely different from that of the classical Q-CSMA (which approximates MWS in a distributed manner). Further, we establish throughput optimality of H-GMS for a broad family of (almost) any increasing weight function* $f(\cdot)$*.*

## VI. 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 VIII. In particular, we derive two lower bounds: (i) a fundamental lower bound that is independent of the scheduling algorithms, and (ii) an improved lower bound tailored for the developed H-GMS and H-GMS-R.[4] In Section VIII-B, we numerically evaluate these lower bounds and compare them to the average queue length achieved by various scheduling algorithms.

[4]The analysis can possibly be extended to H-GMS-E by incorporating its *time-varying and queue-dependent* access probability. We leave this analysis for future work.

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

$$\overline{Q} = \textstyle\sum_{l\in\mathcal{E}} \mathbb{E}[Q_l]/|\mathcal{E}| = \mathcal{Q}_{\mathcal{E}}/(2N). \tag{15}$$

Therefore, finding a lower bound on $\overline{Q}$ is equivalent to finding a lower bound on $\mathcal{Q}_{\mathcal{E}}$.

### A. A Fundamental Lower Bound

We first derive a fundamental lower bound on $\overline{Q}$ that is independent of the chosen (possibly centralized) scheduling algorithm, based on the following result.

**Proposition 6.3** ([32, Proposition 4.1])**.** *With independent packet arrivals, the expected sum of queue lengths in a clique* $\mathcal{C}$ *under any scheduling policy satisfies*

$$\mathcal{Q}_{\mathcal{C}} = \sum_{l\in\mathcal{C}} \mathbb{E}[Q_l] \geq \sum_{l\in\mathcal{C}} \frac{\lambda_l + \mathrm{Var}\,[A_l] - \lambda_l\lambda_{\mathcal{C}}}{2\,(1-\lambda_{\mathcal{C}})} := \mathcal{Q}_{\mathcal{C}}^{LB}.$$

Note that $\mathcal{Q}_{\mathcal{C}}^{\mathrm{LB}}$ is equivalent to the sum of queue lengths in a standard single-server GI/D/1 queue in clique $\mathcal{C}$. In order to obtain a tight fundamental lower bound in the heterogeneous HD-FD networks, one needs to find the largest clique of links, $\mathcal{E}_{\max}$, with the maximal sum of arrival rates. In particular, we divide $\mathcal{E}$ into two disjoint sets $\mathcal{E} = \mathcal{E}_{\max} \cup \mathcal{E}_{\min}$:

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

where $\{\bar{j}\} = \{\mathrm{u}, \mathrm{d}\} \setminus \{j\}$ and we break ties uniformly at random if $\lambda_i^{\mathrm{u}} = \lambda_i^{\mathrm{d}}$ 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 $\Lambda_{\text{HD-FD}}$ (see (1)). The following proposition gives the fundamental lower bound on the average queue length in the heterogeneous HD-FD networks.

**Proposition 6.4.** *A fundamental lower bound on the average queue length in the considered heterogeneous HD-FD networks, denoted by* $\overline{Q}_{Fund}^{LB}$*, is given by*

$$\overline{Q} \geq \overline{Q}_{Fund}^{LB} := \mathcal{Q}_{\mathcal{E}_{max}}^{LB}/(2N), \tag{16}$$

*where* $\overline{Q}$ *is the average queue length defined in* (15)*, and* $\mathcal{Q}_{\mathcal{E}_{max}}^{LB}$ *is given by Proposition 6.3 for clique* $\mathcal{E}_{max}$*.*

*Proof:* Since a pair of FD UL and DL will always be activated at the same time, it holds that for $\forall i \in \mathcal{N}_F$, $\mathbb{E}[Q_i^j] \geq \mathbb{E}[Q_i^{\bar{j}}]$ if $\lambda_i^j \geq \lambda_i^{\bar{j}}$. By assigning the FD UL/DL with a higher arrival rate to $\mathcal{E}_{\max}$, we construct a maximal clique, $\mathcal{E}_{\max}$, with the maximal possible sum or arrival rates. Although it is possible that two queues from both $\mathcal{E}_{\min}$ and $\mathcal{E}_{\max}$ are served simultaneously, it is still guaranteed that

$$\mathcal{Q}_{\mathcal{E}} \geq \mathcal{Q}_{\mathcal{E}_{\max}} \geq \mathcal{Q}_{\mathcal{E}_{\max}}^{\mathrm{LB}}, \tag{17}$$

and Proposition 6.4 follows directly. ∎

### B. An Improved Lower Bound under H-GMS and H-GMS-R

We now derive an improved lower bound on $\overline{Q}$ for the considered heterogeneous HD-FD networks taking into account the characteristics of the developed H-GMS and H-GMS-R

(e.g., the access probability $\boldsymbol{\alpha}$ and the transmission probability $p(\cdot)$). The result is stated in the following proposition.

**Proposition 6.5.** *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 $\alpha_{max} = \max\{\alpha_1, \cdots, \alpha_N, \alpha_{AP}\}$ be the maximum access probability. The average queue length under H-GMS and H-GMS-R is lower bounded by $\overline{Q}^{LB}_{H\text{-}GMS}$ given by*

$$\overline{Q} \geq \overline{Q}^{LB}_{H\text{-}GMS} := \max\Big\{\overline{Q}^{LB}_{Fund}, \Big(1-\frac{N_F}{2N}\Big)\cdot p^{-1}\Big(\frac{\lambda_{min}/\alpha_{max}}{1-\lambda_{\mathcal{E}_{max}}+\lambda_{min}/\alpha_{max}}\Big)\Big\}, \quad (18)$$

*where $\overline{Q}^{LB}_{Fund}$ is given in Proposition 6.4.*

*Proof:* The proof is based on the workload decomposition rules [33] and can be found in Appendix D. ∎

**Remark 6.3.** *Note that (18) applies to any variant of H-GMS with fixed access probability $\boldsymbol{\alpha}$. The lower bound $\overline{Q}^{\text{LB}}_{H\text{-}GMS}$ depends on: (i) the ratio between the link arrival rate and access probability $\frac{\lambda_{\min}}{\alpha_{\max}}$, and (ii) the weight function $f(\cdot)$ (through $p(\cdot)$). A more aggressive $f(\cdot)$ results in a lower value of $\overline{Q}^{\text{LB}}_{H\text{-}GMS}$. The lower bound can be also applied to H-GMS-E by setting $\alpha_{\max}=1$; however, this will result in a loose lower bound as it ignores the adaptive behavior of $\boldsymbol{\alpha}$.*

## VII. Benefits of Introducing FD-Capable Nodes

In this section, we illustrate the benefits of introducing FD-capable nodes into all-HD networks, in terms of obtained throughput gains. The throughput gains can be expressed for individual users or the network (i.e., the sum rates). We define the network (individual) throughput gain as the ratio between the achievable network (individual) throughput in a heterogeneous HD-FD network and that in an all-HD network with the same total number of users.

For simplicity and illustrative purposes, consider a static version of H-GMS-R, with access probabilities $\boldsymbol{\alpha} = \frac{1}{1+N}\cdot\mathbf{1}$ (see Algorithm 3 and Section IV-E), and fixed transmission probabilities $p_f^{\text{u}} = p_f^{\text{d}} = p_f$, $p_h^{\text{u}} = p_h^{\text{d}} = p_h \in (0,1)$ for FD and HD users in (5), respectively. By analyzing the Markov chain (similar to Lemma 5.2) under fixed $\boldsymbol{\alpha}$, $p_f$, and $p_h$, the network throughput (i.e., sum rates) of the heterogeneous HD-FD network, $S_{\text{HD-FD}}$, is given by

$$S_{\text{HD-FD}} = \frac{\frac{2N_F}{N}\frac{p_f}{1-p_f} + \frac{N_H}{N}\frac{p_h}{1-p_h}}{1+\frac{N_F}{N}\frac{p_f}{1-p_f}+\frac{N_H}{N}\frac{p_h}{1-p_h}}. \quad (19)$$

Note that the throughput of the benchmark all-HD network is simply $S_{\text{HD}} = p_h$. If $p_f = p_h = p$ (i.e., FD and HD users transmit with the same probability when they capture the channel), (19) becomes $S_{\text{HD-FD}} = (1+\frac{N_F}{N})\cdot p$. This implies that under the static H-GMS-R, the network throughput gain achieved by the HD-FD network is $(1+\frac{N_F}{N}) \in [1,2]$, 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* $\chi$. In order to balance the individual throughput of FD and HD users, we lower the priority of FD transmissions by choosing $\chi \in (0,1]$.

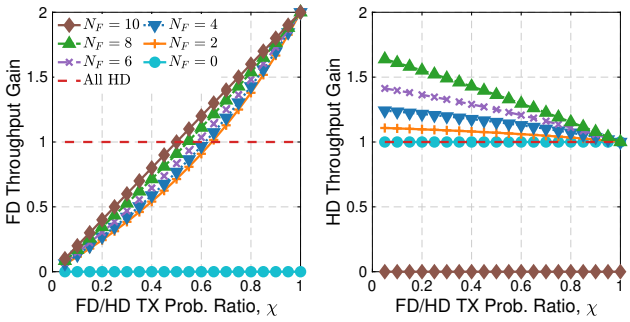



Fig. 1: Throughput gain of FD and HD users when the throughput is compared to the individual throughput of an HD user in the all-HD network under the static H-GMS-R algorithm, with $N = 10$, $N_F \in \{0, 2, \cdots, 10\}$, and $p_h = 0.5$.

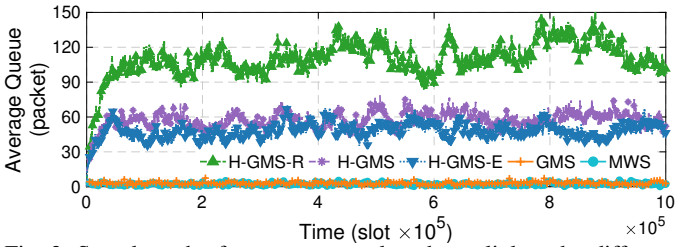



Fig. 2: Sample path of average queue length per link under different scheduling algorithms for a heterogeneous HD-FD network with $N_F = N_H = 5$, and *very high* traffic intensity $\rho = 0.95$.

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 $p_f = \chi \cdot p_h$ 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 $\chi$), 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.

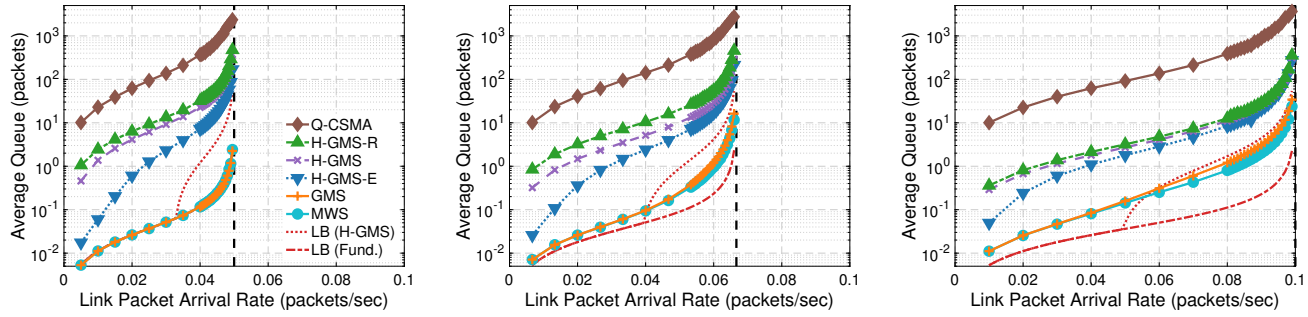
The results reveal an interesting phenomenon: when $N_F$ is sufficiently large, at the cost of slightly lowering the priority of FD users, even HD users can experience throughput improvements. This opens up the possibility of designing wireless protocols with different fairness-efficiency tradeoffs by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

## VIII. 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).

### A. Setup

Throughout this section, we consider heterogeneous HD-FD networks with one FD AP and 10 users ($N = 10$), with




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

Fig. 3: Long-term average queue length per link in a heterogeneous HD-FD network with $N = 10$ and equal arrival rates, under different scheduling algorithms and 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 (16) and (18). The capacity region boundary in each HD-FD network is illustrated by the vertical dashed line.

a varying number of FD users, $N_F$.[5] 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 $\Lambda_{\text{HD-FD}}$ (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 $\Lambda_{\text{HD-FD}}$. 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).

The packet arrivals at each link $l_i^j$ follow an independent Bernoulli process with rate $\lambda_i^j$. For each algorithm under a given traffic intensity, $\rho$, 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: The centralized MWS and GMS algorithms;
- H-GMS, H-GMS-R, and H-GMS-E: Three variants of the H-GMS algorithm as described in Section IV-E;
- Q-CSMA: The standard distributed Q-CSMA algorithm from [12], in which 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 $f(x) = \log(1+x)$ (i.e., $p_l(t) = \frac{1+Q_l(t)}{2+Q_l(t)}$). 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-E). 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. We also consider effects of different weight functions in Section VIII-D.

### B. Delay Performance

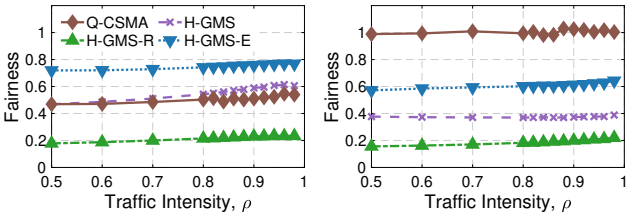
We first consider the queue length dynamics under various scheduling algorithms in an HD-FD network with $N_F = N_H = 5$ and traffic intensity $\rho = 0.95$. This implies that $v_f = v_h = 1/15$, corresponding to a capacity region expansion value of $\gamma = 4/3$ (see Section III-C with $v_h = 1/20$ 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 Q-CSMA algorithm is omitted since, as we will see shortly, its average queue length is at least one order of magnitude larger than those achieved by other algorithms.

Fig. 3 plots the average queue length with varying traffic intensities in HD-FD networks with $N = 10$ and $N_F \in \{0, 5, 10\}$. Recall that in the equal arrival rate model, the relationship between the link packet arrival rate and traffic intensity is $\lambda_i^j = \rho/(N_F + 2N_H)$, $\forall i \in \mathcal{N}$, $\forall j \in \{\mathrm{u}, \mathrm{d}\}$. Fig. 3 shows that the capacity region of the HD-FD networks expands with increased value of $N_F$. Compared with Figs. 3(a), Figs. 3(b) and 3(c) show a capacity region expansion value of $\gamma = 4/3$ for $N_F = 5$, and $\gamma = 2$ for $N_F = 10$, respectively.

Figs. 2 and 3 show that, as expected from Theorem 4.1, 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 $\rho$. 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), as we show below.

Fig. 3 also presents both the fundamental and improved lower bounds on the delay, $\overline{Q}_{\text{Fund}}^{\text{LB}}$ and $\overline{Q}_{\text{H-GMS}}^{\text{LB}}$, given by (16) and (18), respectively. The turning point of $\overline{Q}_{\text{H-GMS}}^{\text{LB}}$ where it starts to deviate from $\overline{Q}_{\text{Fund}}^{\text{LB}}$ is because of the $\max(\cdot)$ operator in (18). As Fig. 3 suggests, the fundamental lower bound, $\overline{Q}_{\text{Fund}}^{\text{LB}}$ is very close to the average queue length obtained by MWS and GMS (they indeed match perfectly in the all-HD network with $N_H = N = 10$). However, in heterogeneous HD-FD networks, $\overline{Q}_{\text{H-GMS}}^{\text{LB}}$ provides a much tighter lower

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




(a) b/w FD and HD users (b) b/w ULs and DLs

Fig. 4: Long-term average queue length ratio between (a) FD and HD users, and (b) ULs and DLs with varying traffic intensity, in an HD-FD network with $N_F = N_H = 5$ and equal arrival rates.

bound on the average queue length achieved by H-GMS, especially with high traffic intensities.

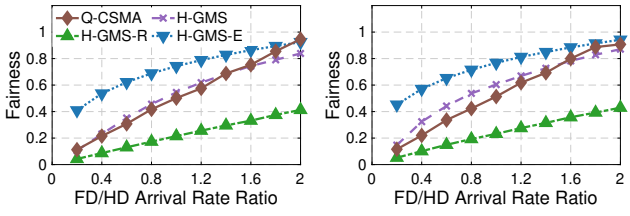
### C. Fairness

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. Ideally, we would like the proposed algorithms to achieve good fairness performance in the considered HD-FD networks. Similarly, we define fairness between ULs and DLs as the *ratio between the average UL and DL queue lengths* to evaluate the effects of different levels of centralization at the AP when operating H-GMS.

*1) Equal Arrival Rates:* We first evaluate the fairness under different distributed algorithms under the *equal arrival rate model*. We focus on traffic intensity regime of $\rho \in [0.5, 1)$ since, as shown in Fig. 3, all links have very small queue lengths with low traffic intensities (e.g., the average queue length is less the 10 packets with $\rho = 0.5$).

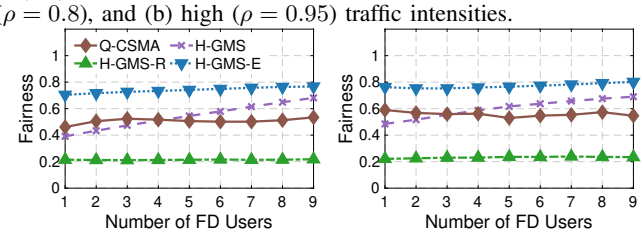
Fig. 4(a) plots the fairness between FD and HD users in an HD-FD network with $N_F = N_H = 5$ and varying traffic intensity, $\rho$. It can be observed that H-GMS-R has the worst fairness performance since the DL participating in the channel contention is selected uniformly at random by the AP. When the traffic intensity is low or moderate, Q-CSMA and H-GMS 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 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 $\boldsymbol{\alpha}$.

Fig. 4(b) presents the fairness between ULs and DLs with the same network setting. It can be seen that Q-CSMA has the best fairness performance of around 1 since all the $2N$ link have equal access probability. The fairness by H-GMS-R between FD and HD users, and between ULs and DLs, are almost identical, and are always the worst among all




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

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




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

Fig. 6: 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 $N_F \in \{1, 2, \cdots, N-1\}$, with (left) moderate ($\rho = 0.8$), and (right) high ($\rho = 0.95$) traffic intensities.

variants of H-GMS. On the other hand, H-GMS-E still has the best fairness performance among all variants of H-GMS by leveraging the information on estimated UL queue lengths.

*2) Different Arrival Rates:* We also evaluate the fairness under different distributed algorithms with *different arrival rates* between FD and HD users. Let $\sigma$ 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 $\Lambda_{\text{HD-FD}}$. 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 $\sigma$ (see Section III-C).

Fig. 5 plots the fairness between FD and HD users with varying $\sigma$ 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, 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, $\sigma$. 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 $\sigma$ approaches 2, H-GMS-E does not further improve the fairness since it generates an access probability distribution that is approximately a uniform distribution.

*3) Impact of the Number of FD Users, $N_F$:* We now 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. This is because under

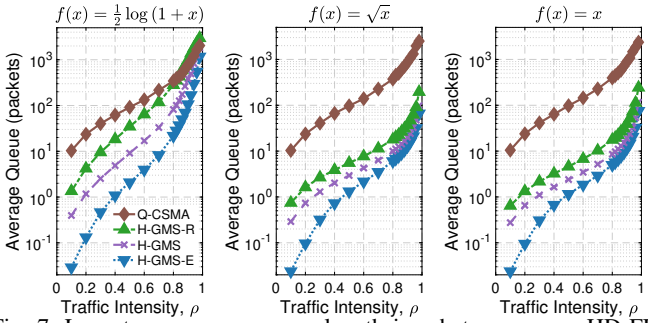



Fig. 7: Long-term average queue length in a heterogeneous HD-FD network with $N_F = N_H = 5$ and equal arrival rates, with different weight functions and $p_l(t) = \frac{\exp(f(Q_l(t)))}{1+\exp(f(Q_l(t)))}$. The results in the case with $f(x) = \log(1+x)$ are shown in Fig. 3(b).

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 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 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 $N_F$.

### *D. Impact of the Weight Function, $f(x)$*

We now evaluate the delay performance of H-GMS under different weight functions and compare it to Q-CSMA. Recall from Theorem 4.1 that H-GMS is throughput-optimal for a broad family of weight functions, $f(x)$, and the relationship between $f(\cdot)$ and the transmission probability $p(\cdot)$ is given by (5). In particular, we consider the following weight functions:

- $f(x) = \frac{1}{2}\log(1+x)$: $\lim_{x\to\infty} \frac{f(x)}{\log x} = \frac{1}{2} < 1$;
- $f(x) = \log(1+x)$: $\lim_{x\to\infty} \frac{f(x)}{\log x} = 1$;
- $f(x) = \sqrt{x}$: $\lim_{x\to\infty} \frac{f(x)}{\log x} = \infty$ ($\beta = \frac{1}{2}$);
- $f(x) = x$: $\lim_{x\to\infty} \frac{f(x)}{\log x} = \infty$ ($\beta = 1$).

Fig. 7 plots the average queue length with varying traffic intensity in an HD-FD network with $N_F = N_H = 5$ and equal arrival rates, and with different weight functions, $f(x)$, as listed above. For each considered $f(\cdot)$, we consider all four distributed algorithms listed in Section VIII-A. The results in the case with $f(x) = \log(1+x)$ are shown in Fig. 3(b). Table I summarizes the improvements in the delay performance achieved by variants of H-GMS compared to Q-CSMA, with the considered weight functions and moderate ($\rho = 0.8$) and extremely high ($\rho = 0.98$) traffic intensities.[6]

The results show that all the scheduling algorithms are throughput-optimal under different choices of $f(x)$ that satisfy the conditions in Theorem 4.1. Overall, the delay performance of Q-CSMA in HD-FD networks has less dependency on $f(x)$ than H-GMS, and variants of H-GMS (especially H-GMS and H-GMS-E) achieve significantly improved delay performance. Moreover, the delay improvement achieved by H-GMS over the classical Q-CSMA becomes more significant with a "more aggressive" weight function. For example, H-GMS with a sublinear/linear weight function ($f(x) = x^\beta$ with $\beta \in \{\frac{1}{2}, 1\}$) achieves 10–20× better delay than with a logarithmic weight function $f(x) = \frac{1}{2}\log(1+x)$. This highlights the importance of the selection of $f(x)$ in the design of H-GMS.

[6]The results in the cases with $f(x) = \sqrt{x}$ and $f(x) = x$ are almost identical (see Fig. 7) and thus omitted in Table I.

TABLE I: Improvements in the delay performance achieved by H-GMS compared with Q-CSMA under three different weight functions with different aggressiveness and, with moderate ($\rho = 0.8$) and extremely high ($\rho = 0.98$) traffic intensities.

| Weight Function, $f(x)$ | $\frac{1}{2}\log(1+x)$ | | $\log(1+x)$ | | $x$ | |
|---|---|---|---|---|---|---|
| Traffic Intensity, $\rho$ | 0.8 | 0.98 | 0.8 | 0.98 | 0.8 | 0.98 |
| $\frac{\overline{Q}_{\text{Q-CSMA}}}{\overline{Q}_{\text{H-GMS-R}}}$ | 1.2 | 0.7 | 14.4 | 8.5 | 22.3 | 9.8 |
| $\frac{\overline{Q}_{\text{Q-CSMA}}}{\overline{Q}_{\text{H-GMS}}}$ | 4.2 | 1.1 | 28.4 | 16.2 | 46.2 | 20.4 |
| $\frac{\overline{Q}_{\text{Q-CSMA}}}{\overline{Q}_{\text{H-GMS-E}}}$ | 15.8 | 1.7 | 52.8 | 25.4 | 79.2 | 31.8 |

## IX. 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. We also derived lower bounds on the average queue length to evaluate the delay performance of H-GMS. We further illustrated various aspects of the performance of H-GMS and compared it to the classical Q-CSMA 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, an experimental evaluation of H-GMS on a real wireless testbed is an important step towards a provably-efficient and practical MAC layer for HD-FD networks.

## 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] 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] 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.

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

[8] 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*, 2019.

[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] 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.

[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] J. Zhou, A. Chakrabarti, P. R. Kinget, and H. Krishnaswamy, "Low-noise active cancellation of transmitter leakage and transmitter noise in broadband wireless receivers for FDD/co-existence," *IEEE J. Solid-State Circuits*, vol. 49, no. 12, pp. 3046–3062, 2014.

[16] T. Chen, M. Baraani Dastjerdi, G. Farkash, J. Zhou, H. Krishnaswamy, and G. Zussman, "Open-access full-duplex wireless in the ORBIT testbed," *arXiv preprint arXiv:1801.03069v2*, 2018.

[17] D. Raychaudhuri, I. Seskar, G. Zussman, T. Korakis, D. Kilper, T. Chen, J. Kolodziejski, M. Sherman, Z. Kostic, X. Gu, H. Krishnaswamy, S. Maheshwari, P. Skrimponis, and C. Gutterman, "Challenge: COSMOS: A city-scale programmable testbed for experimentation with advanced wireless," in *Proc. ACM MobiCom'20 (to appear)*, 2020.

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

[19] 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.

[20] 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.

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

[22] 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.

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

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

[25] A. Tang and X. Wang, "A-duplex: Medium access control for efficient coexistence between full-duplex and half-duplex communications," *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, pp. 5871–5885, 2015.

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

[27] "IEEE 802.11 full duplex topic interest group," https://mentor.ieee.org/802.11/dcn/18/11-18-0191-01-0wng-full-duplex-for-802-11.pptx.

[28] 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.*, no. 1, 2017.

[29] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE J. Sel. Areas Commun.*, vol. 18, 2000.

[30] 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.

[31] 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.

[32] 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.

[33] O. J. Boxma, "Workloads and waiting times in single-server systems with multiple customer classes," *Queueing Systems*, vol. 5, no. 1-3, 1989.

[34] J. Ghaderi and R. Srikant, "The impact of access probabilities on the delay performance of Q-CSMA algorithms in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 21, no. 4, pp. 1063–1075, 2013.

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

[36] 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.

[37] N. Bouman, S. C. Borst, and J. S. van Leeuwaarden, "Delay performance in random-access networks," *Queueing Systems*, vol. 77, no. 2, 2014.


## APPENDIX A
## PROOF OF PROPOSITION 4.1

The proof is based on the structural properties of the interference graph of the heterogeneous HD-FD network. The interference (or conflict) graph is defined as $G_{\rm I} = (\mathcal{V}_{\rm I}, \mathcal{E}_{\rm I})$, where $\mathcal{V}_I$ is the set of network links, and there is an edge between link $l_i$ and link $l_j$ if they interfere with each other. Clearly, the interference graph of a collocated all-HD network is a clique. For the collocated HD-FD network, $\mathcal{V}_{\rm I} = \{v_i^j, i \in \mathcal{V}, j \in \{{\rm u}, {\rm d}\}\}$, where $v_i^j$ corresponds to link $j$ (UL or DL) of user $i$. Since a pair of FD UL and DL can be simultaneously activated, $G_{\rm I}$ is a complete graph with $N_F$ edges missing, each of which has endpoints $(v_i^{\rm u}, v_i^{\rm d})$ for $\forall i \in \mathcal{N}_F$.

It has been shown in [10] that the Greedy Maximal Scheduling (GMS) is throughput-optimal if the interference graph $G_{\rm I}$ satisfies the so called *Overall Local Pooling (OLoP)* condition. We use the following definition and result from [13].

**Definition 1.1** (Co-strongly perfect graph)**.** *A graph $G$ is co-strongly perfect, if and only if $G$ contains a clique that intersects every maximal independent set in $G$.*

**Proposition 1.6** ([13, Definitions 2.2, 2.3, and 5.1])**.** *Every graph that is co-strongly perfect satisfies OLoP.*

We now prove Proposition 4.1. To show that $G_{\rm I}$ is co-strongly perfect, if suffices to find a clique contained in $G_{\rm I}$ that intersects every maximal independent set in $G_{\rm I}$. Recall that $G_{\rm I}$ is a complete graph with $N_F$ edges missing. Let $K = \{v_n^{\rm u}, n \in \mathcal{N}\} \cup \{v_m^{\rm d}, m \in \mathcal{N}_H\} \subseteq \mathcal{V}_{\rm I}$ with $|K| = N_F + 2N_H$. It is easy to see that the induced graph $G(K)$ on $K$ is a clique. In addition, note that the maximal independent set in $G_{\rm I}$ can be (i) $\{v_m^{\rm u}\}$ or $\{v_m^{\rm d}\}$ for some $m \in \mathcal{N}_H$, or (ii) $\{v_n^{\rm u}, v_n^{\rm d}\}$ for some $n \in \mathcal{N}_F$ (there are a total number of $(N_F + 2N_H)$ such maximal independent sets). Thus, $G(K)$ intersects with every maximal independent set in $G_{\rm I}$, which implies that $G_{\rm I}$ is co-strongly perfect and satisfies OLoP. Hence, the centralized GMS algorithm (described in Algorithm 2) is throughput-optimal in *any* collocated heterogeneous HD-FD network.

## APPENDIX B
## PROOF OF LEMMA 5.2

Under fixed $\mathbf{Q}(t) = \mathbf{Q}$, (7) is the transition probability matrix of the discrete time Markov chain $Y^{\mathbf{Q}}$. Recall that the state space of $Y(t)$ is $S_Y = \{0, 1, \cdots, N, i^\star\}$, where $i^\star = \arg\max_{i \in \mathcal{N}} Q_i^{\rm d}$. The detailed balance equations for $Y^{\mathbf{Q}}$ are:

$$\pi^{\mathbf{Q}}(0) \cdot P(0,i) = \pi^{\mathbf{Q}}(i) \cdot P(i,0), \ \forall i \in S_Y \setminus \{0\}. \quad (20)$$

Recall that the transmission probability is given by (5); from (7) and (20), we have

$$\begin{aligned}\pi^{\mathbf{Q}}(i) &= \alpha_i \tfrac{p_i^{\rm u}}{\overline{p}_i^{\rm u}} \cdot \pi^{\mathbf{Q}}(0) \\ &= \alpha_i \exp\left(f(Q_i^{\rm u})\right) \cdot \pi^{\mathbf{Q}}(0), \ \forall i \in S_Y \setminus \{0, i^\star\}, \\ \pi^{\mathbf{Q}}(i^\star) &= \alpha_{\rm AP} \tfrac{p_{i^\star}^{\rm d}}{\overline{p}_{i^\star}^{\rm d}} \cdot \pi^{\mathbf{Q}}(0) = \alpha_{\rm AP} \exp\left(f(Q_{i^\star}^{\rm d})\right) \cdot \pi^{\mathbf{Q}}(0).\end{aligned}$$

Normalizing $\sum_{s\in S_Y}\pi^{\mathbf{Q}}(s)=1$ yields the steady-state distribution (8) in Lemma 5.2, in which

$$Z=1+\sum_{i\in S_Y\setminus\{0,i^\star\}}\alpha_i\exp\left(f(Q_i^{\mathrm{u}})\right)+\alpha_{\mathrm{AP}}\exp\left(f(Q_{i^\star}^{\mathrm{d}})\right).$$

## APPENDIX C
## PROOF OF LEMMA 5.3: FLUID LIMIT EQUATIONS

Equations (9)–(10) hold for any scheduling algorithm and their proof is standard. Equation (9) is obtained by taking the limit in (6). Equation (10) is by applying the Strong Law of Large Numbers to the arrival process. Further, by the Lipschitz continuity of $s$, the derivative of $s_i^j$ (denoted by $\mu_i^j(t)$) exists at any regular point $t$ (almost everywhere) and is bounded by its Lipschitz constant (less than one). Equations (11)–(14) are specific to H-GMS, and we prove them below. We consider two cases depending on the choice of the weight function $f(\cdot)$.
**Case 1**: $\lim_{x\to\infty}f(x)/\log x=b\in(0,1)$.

Recall from Section V that Markov chain $\{Y^{\mathbf{Q}(t)}(s)\}_{s\geq t}$ denotes the dynamics of $Y(s)$, assuming a fixed $\mathbf{Q}(s)=\mathbf{Q}(t)$ for all $s\geq t$. Consider a fluid sample path under the H-GMS algorithm. Suppose $\mathbf{q}(t)\neq\mathbf{0}$ at a regular point $t$. By Lipschitz continuity, we can find a short interval $(t,t+\epsilon)$, such that $\mathbf{q}(\tau)$ is approximately constant ($\approx\mathbf{q}$) for $\forall\tau\in(t,t+\epsilon)$, its actual change being of order $\epsilon$ for non-zero queues. This implies that for $r$ large enough, all the queues with non-zero fluid limit $q_i^j>0$ are of size $\mathcal{O}(q_i^j r)$ in the original process, while all the queues with zero fluid limit are of size $o(r)$ in the original process. Therefore, taking the limit $r\to\infty$ in (8), it follows that for any $\mathbf{Q}(\tau)$, $\tau\in(rt,rt+r\epsilon)$, $\pi^{\mathbf{Q}(\tau)}\to\widetilde{\pi}^{\mathbf{q}}$, where

$$\widetilde{\pi}^{\mathbf{q}}(i)=\alpha_i(q_i^{\mathrm{u}})^b/\widetilde{Z}^{\mathbf{q}},\ i\in S_Y\setminus\{0,i^\star\},$$
$$\widetilde{\pi}^{\mathbf{q}}(i^\star)=\alpha_{\mathrm{AP}}(q_{i^\star}^{\mathrm{d}})^b/\widetilde{Z}^{\mathbf{q}},\ \widetilde{\pi}^{\mathbf{q}}(0)=1/\widetilde{Z}^{\mathbf{q}},$$

$\widetilde{Z}^{\mathbf{q}}=1+\sum_{i'=1}^{N}\alpha_{i'}(q_{i'}^{\mathrm{u}})^b+\alpha_{\mathrm{AP}}(q_{i^\star}^{\mathrm{d}})^b$, and the probabilities are zero for queues which are 0 at the fluid limit. This shows that, with high probability, a queue with a zero fluid limit *cannot* initiate transmission in steady-state. Hence, in equilibrium, the Markov chain never activates an HD link with empty fluid limit queue or an FD link with empty (both) UL and DL queues. Next, we argue that at any $\tau\in(rt,rt+r\epsilon)$, the Markov chain $Y^{\mathbf{Q}(\tau)}$ is at its equilibrium distribution $\pi^{\mathbf{Q}(\tau)}=\widetilde{\pi}^{\mathbf{q}}$ as $r\to\infty$.

**Proposition 3.7** (Mixing time of Markov chain $Y^{\mathbf{Q}}$)**.** *Let $\nu_\tau$ and $\pi$ denote the instantaneous and the equilibrium distribution of Markov chain $\{Y^{\mathbf{Q}}(\tau)\}_{\tau\geq1}$, respectively. Given $0<\zeta<1$, the mixing time is defined as*

$$T_{mix}(\zeta):=\inf\big\{\tau\geq1:\sup_{s\in S_Y}|\nu_\tau(s)-\pi(s)|\leq\zeta\big\}.$$

*Let $\alpha_{min}=\min_i\{\alpha_i\}$ and $Q_{max}=\max_{i,j}\{Q_i^j\}$. Then*

$$T_{mix}(\zeta)\leq\frac{2\exp\left(f(Q_{max})\right)}{\alpha_{min}}\cdot\Big[\log\Big(\frac{2}{\zeta\alpha_{min}}\Big)+f(Q_{max})\Big].$$

*Proof:* The proof follows the application of Raleigh Theorem to characterize the second largest eigenvalue modulus (SLEM) of the transition probability matrix of the Markov chain $Y^{\mathbf{Q}}$. The analysis is similar to [34, Lemma 5] with minor modifications and is omitted. ∎

Hence for the Markov chain $Y^{\mathbf{Q}(rt)}$, the mixing time is $T_{\mathrm{mix}}(1/r)=\mathcal{O}(r^b\log r)$. This shows that for $b<1$, the mixing time is sub-linear in $r$ which completely vanishes when taking the average at the fluid scale, i.e.,

$$\frac{1}{\epsilon}\left(s_i^{\mathrm{u}}(t+\epsilon)-s_i^{\mathrm{u}}(t)\right)\approx\frac{1}{r\epsilon}\sum_{\tau=rt}^{rt+r\epsilon}\mathbb{1}(Y^{\mathbf{Q}(tr)}(\tau)=i)$$
$$\to\widetilde{\pi}^{\mathbf{q}}(i),\ \text{as } r\to\infty$$

where the second convergence is almost surely by the Ergodic Theorem. This indicates that $\mu_i^{\mathrm{u}}(t)=\widetilde{\pi}^{\mathbf{q}}(i)$, $\forall i\in\mathcal{N}$, and similarly, $\mu_{i^\star}^{\mathrm{d}}(t)=\widetilde{\pi}^{\mathbf{q}}(i^\star)$. This implies that $\mu_i^j(t)=0$ for $i\in\mathcal{N}_H$ if $q_i^j(t)=0$, $j\in\{\mathrm{u},\mathrm{d}\}$, which establishes (11). Similarly, considering the coordination among the activation of a pair of FD UL and DL, $\mu_i^j(t)=0$, $j\in\{\mathrm{u},\mathrm{d}\}$, $i\in\mathcal{N}_F$, if $\max(q_i^{\mathrm{u}}(t),q_i^{\mathrm{d}}(t))=0$, giving (12). Also, (13) stems from the fact that once an FD UL (or DL) initiates the transmission at rate $\mu_i^{\mathrm{u}}$ (or $\mu_i^{\mathrm{d}}$), the corresponding DL (or UL), if *non-zero*, can follow and transmit at the same rate. Finally, (14) is due to the fact that if $\mathbf{q}(t)\neq\mathbf{0}$, no queue that is empty in the fluid limit can initiate transmission at a positive rate and thus the non-empty queues transmit at the maximum sum rate of 1.

**Case 2**: $\lim_{x\to\infty}f(x)/\log x=b>1$.

The analysis in this case is similar to the analysis of aggressive CSMA algorithms in [35], [36]. Suppose that $\mu_i^j(t)>0$ and $q_i^j(t)>0$ for some initiator queue. This implies that for some $\epsilon$, $X_i^j(\tau)=1$ for $\forall\tau\in(rt,r(t+\epsilon))$, and $Q_i^j(\tau)\geq(q_i^j(t)-\epsilon)r$. The probability that this queue releases the channel after one packet transmission is less than $(Q_i^j(\tau))^{-b}$ which is $\mathcal{O}(r^{-b})$ for $b>1$. The probability that the link releases the channel during any time $\tau\in(rt,r(t+\epsilon))$ is thus less than $\sum_{\tau=rt}^{r(t+\epsilon)}r^{-b}$ which is $\mathcal{O}(r^{1-b})$ which goes to 0 as $r\to\infty$. This shows that at the fluid limit, if $\mu_i^j(t)>0$ and $q_i^j(t)>0$, then $\mu_i^j(t)=1$. Hence, any positive period of transmission, no matter how short, must be followed by full transmission at rate 1 until the queue has drained on the fluid scale. Furthermore, when the queue hits zero, another non-zero queue will capture the channel without any capture delay (the proof is similar to that of [36, Lemmas 8 and 9]).

This implies that in the heterogeneous HD-FD network, whenever the initiator queue $q_i^j$ belongs to an HD user $i$, the queue $q_i^j$ drains at full rate 1 until it becomes empty at fluid scale. Whenever the initiator queue $q_i^j$ ($j\in\{\mathrm{u},\mathrm{d}\}$) belongs to an FD user $i$, both its UL and DL queues $q_i^{\mathrm{u}}$ and $q_i^{\mathrm{d}}$ can drain at the maximum rate of 1, until the initiator queue hits zero, at which point both queues release the channel (due to the coordination among a pair of FD UL and DL in Algorithm 3). Whenever an HD or FD user releases the channel, another HD or FD user will capture the channel immediately and start transmission at full rate. The choice of which user and which queue captures the channel is randomized over non-zero queues according to access probabilities $\boldsymbol{\alpha}$ and whether $i^\star(t)\in\mathcal{N}_F$ or $i^\star(t)\in\mathcal{N}_H$. Nevertheless, as long as $\mathbf{q}(t)\neq\mathbf{0}$, an HD link with non-zero queue or an FD link with at least one non-zero queue (either UL or DL) will be activated at full rate. This shows that the fluid limits still satisfy (11)–(14).

Hence, the fluid limit equations hold for both cases.

## APPENDIX D
## PROOF OF PROPOSITION 6.5

Recall from Section VI, $\overline{Q} = \frac{\mathcal{Q}_{\mathcal{E}_{\max}}+\mathcal{Q}_{\mathcal{E}_{\min}}}{2N} \geq \frac{\mathcal{Q}_{\mathcal{E}_{\max}}}{2N}$. Since the queueing dynamics in $\mathcal{E}_{\max}$ and $\mathcal{E}_{\max}$ are *not* independent due to the existence of FD users, we wish to find a lower bound on $\mathcal{Q}_{\mathcal{E}_{\max}}$. Denote $Q_{l,\mathcal{E}_{\max}}$ as the queue length of link $l$ at an arbitrary epoch during a *non-serving* interval for the clique $\mathcal{E}_{\max}$. Denote $\mathcal{Q}_{\widetilde{\mathcal{E}_{\max}}} = \sum_{l\in\mathcal{E}_{\max}} \mathbb{E}[Q_{l,\mathcal{E}_{\max}}]$. From the workload decomposition rule [33] applied to a discrete time GI/G/1 system in clique $\mathcal{E}_{\max}$, we have [37]

$$\begin{aligned}\mathcal{Q}_{\mathcal{E}_{\max}} &= \textstyle\sum_{l\in\mathcal{E}_{\max}} \mathbb{E}[Q_l] = \mathcal{Q}^{\text{LB}}_{\mathcal{E}_{\max}} + \sum_{l\in\mathcal{E}_{\max}} \mathbb{E}[Q_{l,\mathcal{E}_{\max}}] \\ &= \mathcal{Q}^{\text{LB}}_{\mathcal{E}_{\max}} + \mathcal{Q}_{\widetilde{\mathcal{E}_{\max}}}. \end{aligned} \tag{21}$$

Note that $\mathcal{Q}^{\text{LB}}_{\mathcal{E}_{\max}}$ and $\mathcal{Q}_{\widetilde{\mathcal{E}_{\max}}}$ are both non-negative, and an immediate lower bound on $\overline{Q}$ is obtained by

$$\overline{Q} = \frac{\sum_{l\in\mathcal{E}} \mathbb{E}[Q_l]}{2N} = \frac{\mathcal{Q}_{\mathcal{E}}}{2N} \geq \frac{\mathcal{Q}_{\mathcal{E}_{\max}}}{2N} \geq \frac{\mathcal{Q}^{\text{LB}}_{\mathcal{E}_{\max}}}{2N}. \tag{22}$$

A key observation to derive the improved lower bound is that assuming the system is stable, in each time slot, the probability that link $l$ transitions from idle state to active state (i.e., *link $l$ is activated*) equals the probability it transitions from active state back to idle state (i.e., *link $l$ is deactivated*). Therefore, for any link $l \in \mathcal{E}_{\max}$,

$$\mathbb{P}\{l \text{ is activated}\} = \mathbb{P}\{l \text{ is deactivated}\}. \tag{23}$$

Let $\partial_l$ denote the set of conflicting links of $l$ including link $l$ itself and recall that the access probability $\boldsymbol{\alpha}$ is *fixed* under H-GMS and H-GMS-R. For $\forall l \in \mathcal{E}_{\max}$,

$$\begin{aligned}&\mathbb{P}\{l \text{ is activated}\} = \mathbb{E}[\alpha_l \cdot p(Q_l) \cdot \mathbb{1}(X_{l'}(t) = 0, \forall l' \in \partial_l)] \\ &\leq \mathbb{E}[\alpha_l \cdot p(Q_l) \cdot \mathbb{1}(X_{l'}(t) = 0, \forall l' \in \mathcal{E}_{\max})] \\ &= \alpha_l \mathbb{E}[p(Q_{l,\mathcal{E}_{\max}})] \cdot \mathbb{P}\{X_{l'}(t) = 0, \forall l' \in \mathcal{E}_{\max}\} \\ &= \alpha_l \mathbb{E}[p(Q_{l,\mathcal{E}_{\max}})] \cdot (1 - \textstyle\sum_{l'\in\mathcal{E}_{\max}} \mathbb{P}\{X_{l'}(t) = 1\}) \\ &= \alpha_l \mathbb{E}[p(Q_{l,\mathcal{E}_{\max}})] \cdot (1 - \textstyle\sum_{l'\in\mathcal{E}_{\max}} \pi_{l'}), \end{aligned} \tag{24}$$

where $\pi_{l'}$ is the steady state probability of link $l'$ being active. Similarly,

$$\begin{aligned}\mathbb{P}\{l \text{ is deactivated}\} &= \mathbb{E}[(1 - p(Q_l)) \cdot \mathbb{1}(X_l(t) = 1] \\ &= (1 - \mathbb{E}[p(Q_l)]) \cdot \pi_l. \end{aligned} \tag{25}$$

Plugging (24) and (25) into (23) yields

$$\begin{aligned}&\alpha_l \mathbb{E}[p(Q_{l,\mathcal{E}_{\max}})] \cdot (1 - \textstyle\sum_{l'\in\mathcal{E}_{\max}} \pi_{l'}) \geq (1 - \mathbb{E}[p(Q_l)]) \cdot \pi_l \\ &\Leftrightarrow \frac{\mathbb{E}[p(Q_{l,\mathcal{E}_{\max}})]}{1 - \mathbb{E}[p(Q_l)]} \geq \frac{\pi_l/\alpha_l}{1 - \sum_{l'\in\mathcal{E}_{\max}} \pi_{l'}} \geq \frac{\lambda_l/\alpha_l}{1 - \lambda_{\mathcal{E}_{\max}}}, \end{aligned} \tag{26}$$

where the last inequality comes from the fact that in steady state, $\lambda_l \leq \pi_l$ for $\forall l \in \mathcal{E}_{\max}$. Recall the definitions of $\lambda_{\min}$ and $\alpha_{\max}$ from Proposition 6.5, it is easy to see that $\min_{l\in\mathcal{E}_{\max}} \lambda_l \geq \lambda_{\min}$ and $\max_{l\in\mathcal{E}_{\max}} \alpha_l \leq \alpha_{\max}$ (under both H-GMS and H-GMS-R. Applying (26) to all $l \in \mathcal{E}_{\max}$, we obtain

$$\begin{aligned}&\sum_{l\in\mathcal{E}_{\max}} \mathbb{E}[p(Q_{l,\mathcal{E}_{\max}})] \geq \frac{1}{1 - \lambda_{\mathcal{E}_{\max}}} \cdot \sum_{l\in\mathcal{E}_{\max}} \frac{\lambda_l}{\alpha_l} (1 - \mathbb{E}[p(Q_l)]) \\ &\geq \frac{1}{1 - \lambda_{\mathcal{E}_{\max}}} \cdot \frac{\lambda_{\min}}{\alpha_{\max}} \cdot \sum_{l\in\mathcal{E}_{\max}} (1 - \mathbb{E}[p(Q_l)]) \\ &\geq \frac{1}{1 - \lambda_{\mathcal{E}_{\max}}} \cdot \frac{\lambda_{\min}}{\alpha_{\max}} \cdot |\mathcal{E}_{\max}| \cdot \Big(1 - \frac{\sum_{l\in\mathcal{E}_{\max}} \mathbb{E}[p(Q_l)]}{|\mathcal{E}_{\max}|}\Big) \\ &\geq \frac{1}{1 - \lambda_{\mathcal{E}_{\max}}} \cdot \frac{\lambda_{\min}}{\alpha_{\max}} \cdot |\mathcal{E}_{\max}| \cdot \Big(1 - p\Big(\frac{\mathcal{Q}_{\mathcal{E}_{\max}}}{|\mathcal{E}_{\max}|}\Big)\Big), \end{aligned} \tag{27}$$

where the last inequality comes from applying Jensen's inequality to the concave increasing function $p(\cdot)$, i.e.,

$$\frac{\sum_{l\in\mathcal{E}_{\max}} \mathbb{E}[p(Q_l)]}{|\mathcal{E}_{\max}|} \leq p\Big(\frac{\mathcal{Q}_{\mathcal{E}_{\max}}}{|\mathcal{E}_{\max}|}\Big).$$

In addition, the left-hand-side of (27) can be upper bounded using Jensen's inequality,

$$\sum_{l\in\mathcal{E}_{\max}} \mathbb{E}[p(Q_{l,\mathcal{E}_{\max}})] \leq |\mathcal{E}_{\max}| \cdot p\Big(\frac{\mathcal{Q}_{\widetilde{\mathcal{E}_{\max}}}}{|\mathcal{E}_{\max}|}\Big) \leq |\mathcal{E}_{\max}| \cdot p\Big(\frac{\mathcal{Q}_{\mathcal{E}_{\max}}}{|\mathcal{E}_{\max}|}\Big), \tag{28}$$

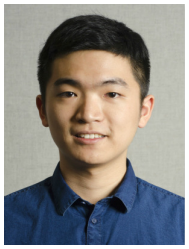
where the last inequality is due to $\mathcal{Q}_{\widetilde{\mathcal{E}_{\max}}} \leq \mathcal{Q}_{\mathcal{E}_{\max}}$ (see (21)). Putting together (27) and (28) yields

$$\mathcal{Q}_{\mathcal{E}_{\max}} \geq |\mathcal{E}_{\max}| \cdot p^{-1}\Big(\frac{\lambda_{\min}/\alpha_{\max}}{1 - \lambda_{\mathcal{E}_{\max}} + \lambda_{\min}/\alpha_{\max}}\Big),$$

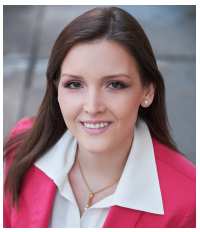
and as a result,

$$\overline{Q} \geq \frac{\mathcal{Q}_{\mathcal{E}_{\max}}}{2N} \geq \Big(1 - \frac{N_F}{2N}\Big) \cdot p^{-1}\Big(\frac{\lambda_{\min}/\alpha_{\max}}{1 - \lambda_{\mathcal{E}_{\max}} + \lambda_{\min}/\alpha_{\max}}\Big). \tag{29}$$

Combining (22) and (29) leads to (18), completing the proof.



**Tingjun Chen** received the B.Eng. degree in electronic engineering from Tsinghua University in 2014. He is currently pursuing the Ph.D. degree in electrical engineering at Columbia University. His research interests are in the areas of wireless networking and systems with focuses on algorithms, optimization, and system design and implementation. He received the Facebook Fellowship, the Wei Family Private Foundation Fellowship, the Columbia Electrical Engineering Armstrong Memorial Award, and the ACM CoNEXT'16 Best Paper Award.



**Jelena Diakonikolas** is an Assistant Professor at the Department of Computer Sciences, University of Wisconsin-Madison. She received her Ph.D. degree from Columbia University in 2016. Her research interests include large-scale optimization and applications in wireless and network systems. She received the 2017 Morton B. Friedman Prize for Excellence at Columbia Engineering and a Qualcomm Innovation Fellowship. In 2016, she was featured on the N$^2$ Women list of "10 Women in Networking/Communications That You Should Watch."

**Javad Ghaderi** is an Associate Professor of Electrical Engineering at Columbia University. His research interests include network algorithms, control and optimization. He received his Ph.D. from the University of Illinois at Urbana-Champaign (UIUC) in 2013. He is the recipient of the Mac Van Valkenburg Graduate Research Award at UIUC, Best Student Paper Finalist at the 2013 American Control Conference, Best Paper Award at ACM CoNEXT 2016, and NSF CAREER Award in 2017.

**Gil Zussman** received the Ph.D. degree in electrical engineering from the Technion in 2004 and was a postdoctoral associate at MIT in 2004–2007. He is a Professor of Electrical Engineering at Columbia University. He is a co-recipient of 7 paper awards including the ACM SIGMETRICS'06 Best Paper Award, the 2011 IEEE Communications Society Award for Advances in Communication, and the ACM CoNEXT'16 Best Paper Award. He received the Fulbright Fellowship, the DTRA Young Investigator Award, and the NSF CAREER Award.