Opportunistic Multicast Scheduling for Unicast Transmission in MIMO-OFDM System

Peng Hui Tan, Jingon Joung, Sumei Sun
Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
Email:{phtan, jgjoung, sunsm}@i2r.a-star.edu.sg

Abstract*—We propose a opportunistic multicast scheduling scheme to exploit content reuse when there is asynchronicity in user requests. A unicast transmission setup is used for content delivery, while multicast transmission is employed opportunistically to reduce wireless resource usage. We then develop a multicast scheduling scheme for the downlink multiple-input multiple-output orthogonal-frequency division multiplexing system in IEEE 802.11 wireless local area network (WLAN). At each time slot, the scheduler serves the users by either unicast or multicast transmission. Out-sequence data received by a user is stored in user's cache for future use. Multicast precoding and user selection for multicast grouping are also considered and compliance with the IEEE 802.11 WLAN transmission protocol. The scheduling scheme is based on the Lyapunov optimization technique, which aims to maximize system rate. The resulting scheme has low complexity and requires no prior statistical information on the channels and queues. Furthermore, in the absence of channel error, the proposed scheme restricts the worst case of frame dropping deadline, which is useful for delivering real-time traffic. Simulation results show that our proposed algorithm outperforms existing techniques by $17\ \%$ to $35\ \%$ in term of user capacity.

***Index Terms*—Multicast scheduling, Lyapunov optimization, multicast precoding, WLAN network**

I. Introduction

To provide satisfactory quality of service (QoS) for multimedia contents, efficient allocation of wireless resource is a necessity. Opportunistic scheduling is one of the most promising techniques. It has been observed that most of user requests are restricted to only a few very popular contents. For such scenario, multicast is an efficient mechanism for one-to-many transmissions over wireless channels [1]–[3]. Contrary to a unicast, in which each user (or STA: station) is supported by an access point (AP) separately at each time slot t_1 or t_2 as illustrated in Fig. 1(a), multicast can support multiple users who request identical content simultaneously as illustrated in Fig. 1(b). Herein, users 1 and 2 belong to a multicast group, which requires message (data chunk, internet protocol packet, or data frame) D1, D2 and D3 from the AP. On the other hand, the user requests usually occur at different times, i.e., asynchronous request. Hence, the AP has to fall back to unicast transmission and loses the exploitation of this content reuse feature. Another approach to deal with the opportunistic demand is harmonic broadcasting and its variants introduced in [4], [5]. These schemes enable each user to start playback within a small delay from its request time. However, the allocation of wireless resource, i.e., scheduling in time slot, was not considered.

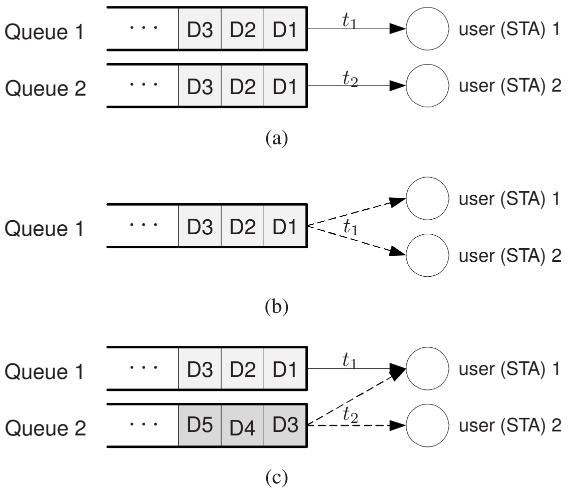

Fig. 1. Illustration of (a) unicast. (b) multicast. (c) opportunistic multicast.

In this work, our goal is to develop efficient transmission and scheduling (i.e., time resource allocation) scheme to exploit the content reuse feature under the opportunistic requests. We refer to this transmission scheme as the opportunistic multicast, and illustrate it in Fig. 1(c). Herein, users 1 and 2 demand for identical content. Since unicast transmission is used, two queues are required at the AP. However, either unicast or opportunistic multicast transmission is performed dynamically at each time slot. For example, if queue 1 is scheduled for transmission, AP sends message D1 to user 1 only, which is the same as the unicast transmission, as user 2 has already received D1. On the other hand, if AP selects queue 2 for transmitting message D3 to (intended) user 2 and if it also knows that user 1 requires D3 in the future, the AP switches from unicast to multicast transmission and sends D3 to both users 1 and 2. Once user 1 receive D3, it stores D3 in its own cache for future use[1]. If D3 appears in queue 1 at the AP later, it will be dropped as it is already cached in user 1. Similarly, user 1 will not request for D3.

The main contribution of this paper is to introduce the opportunistic multicast transmission as an alternative to ac-

[1] Our work is mainly inspired by the caching approach to deliver contents. In [6]–[8], contents are stored in the users' local caches and in dedicated helper nodes distributed in the network. In contrast, our transmission scheme requires no helpers, but relies on multicast transmission to exploit content reuse.

commodate for more users in the network. This transmission scheme can also be used to replace the conventional multicast transmission if retransmission is essential: Instead of a single queue for the multicast group, multiple (virtual) queues can be set up to serve the users in group. Contrast to the study in [1]–[3], we have considered *multiple* multicast groups requesting for different contents. We also propose a multicast scheduling scheme for multiple-input multiple-output (MIMO) orthogonal frequency-division multiplexing (OFDM) systems with deadline constraints for real-time traffic. The proposed scheme also involves multicast precoding and user selection. User selection forms a multicast user group that consists of one intended user and multiple unintended users. Multicast regrouping, which is not considered in [1], is necessary due to the limitation of group size in practical network. We further investigate how to set protocol parameters to maximize the number of users accommodated in the network.

The rest of the paper is organized as follows. We describe system model and formulate a scheduling problem in Sections II and III, respectively. In Section IV, we develop an algorithm based on the Lyapunov optimization. We provide simulation scenario and results in Section V and conclude the paper in Section VI.

II. System Model

We first introduce a transmission procedure in a medium access control (MAC) layer, and then elaborate multicast precoding in a physical (PHY) layer.

A. Transmission Procedure in MAC Layer

The system operates in time slot with duration whose length is a bounded variable. The AP maintains K queues, each of which supports a dedicated user, which implies there are K users in the networks. Multiple data frames are allowed to be transmitted in each time slot as long as the transmission time does not exceed a given bound. Denoting a supported user set by $\mathcal{S}$, we introduce the detailed procedure in MAC layer.

Step 1: At the beginning of each time slot, a user is selected by the scheduler to be served by AP. AP can operate in either unicast or multicast transmission. If the data frames to be transmitted are not required by other users (currently or in the future), unicast transmission (the cardinality of user set $|\mathcal{S}| = 1$) is scheduled and accordingly single user MIMO precoding (beamforming in IEEE 802.11ac [9]) is used. Otherwise, multicast transmission for multiple users ($|\mathcal{S}| > 1$) is selected. The multicast precoding used will be designed in next subsection. For practical reasons, the maximum number of multicast users is limited to four, i.e., $|\mathcal{S}| \leq 4$. To balance the tradeoff between multiuser diversity and multicast gain, multicast regrouping is necessary [2], [3]. To select the multicast users, we use the norm criterion $\mathsf{E}_{n\in\mathcal{N}}\|\mathbf{H}_{n,k}\mathbf{H}_{n,s}^H\|_F^2$, where n is subcarrier index and $\mathcal{N}$ is a subcarrier set; $\mathbf{H}_{n,k}$ and $\mathbf{H}_{n,s}$ is the channel gain matrix of subcarrier n from AP to an intended user k and an unintended user s. These values are sorted in descending order and the first three unintended users are selected to form the multicast group with the intended user k. Note that this multicast user grouping is completely opposite to a multiuser (MU)-MIMO precoding. The link abstraction model for mapping the transmission mode to modulation and coding scheme (MCS) is based on mutual information approach given in [10]. The scheduling priority is based on the head-of-line (HOL) delay, outdated transmission rate (due to outdated channel information), transmission time required, frame length and the number of multicast users. The details on the design of this priority are deferred until the later sections.

Step 2: Next, AP requests for channel state information (CSI) feedback from all users in $\mathcal{S}$. The channel sounding procedure in IEEE 802.11ac [9] is applied. The AP sends a null data packet announcement (NDPA) frame to notify the users to prepare for the channel measurement. The users measure the channel based on the null data packet (NDP) transmitted after the NDPA frame, followed by CSI feedback from one user to the AP. The AP then polls the remaining users for their respective CSI if necessary.

Step 3: Upon receiving the CSI, an MCS is selected for the transmission. The selection approach is the same as in Step 1. For the multicast transmission, the smallest MCS over all users is selected so as to ensure that all users can receive the data frames correctly, which will be used to design multicast precoding later. The number of data frames to be sent is determined by the MCS and the maximum allowable transmission time, which is also known as transmit opportunity (TXOP) in IEEE 802.11ac [9].

Step 4: After transmitting the frames, the AP also expects acknowledgement (ACK) frames from the users. After the first user has sent back the ACK frame, the AP sends ACK requests to the remaining users for their respective ACK if necessary. The ACK procedure is similar to the groupcast with retries (GCR) service in IEEE 802.11aa [11]. After ACKs have arrived, the channel is released for contention.

Step 5: Retransmission of erroneous frames is allowed. Only the erroneous packets for the intended user is retransmitted. In addition, retransmission has a higher priority than scheduling user for new transmission.

Summary of the transmission procedure at MAC layer:

Step 1: Select an intended user for transmission; Select (unintended) user(s) and form unicast or multicast user set; Estimate the MCS for transmission; Scheduling priority among users based on HOL delay, transmission rate and time, and packet length.
Step 2: Request CSI feedback.
Step 3: Determined MCS and number of frames to transmit based on current CSI feedback.
Step 4: Transmit the packets and receive ACK/NACK from users.
Step 5: Retransmission if it is necessary.

Remark 1: The above transmission procedure is not restricted to system with time slot of variable duration. It can be also applied to OFDMA system like LTE.

B. Multicast Precoding in PHY Layer

We consider a downlink *multicast* MIMO-OFDM system with one AP (N_t transmit antennas) and K users (N_r receive antennas each), in which a common message is sent to all the users in $\mathcal{S}$ through N subcarriers. Note that MU-MIMO transmitter sends individual message to each user. For a given time slot t, the scheduler at the AP selects a subset of users $\mathcal{S}$ to serve simultaneously. We assume that channel matrix $\mathbf{H}_{n,k}$ includes large- and small-scale fadings and is static during each transmission. For subcarrier $n \in \mathcal{N} = \{1, \ldots, N\}$, the $N_r \times 1$ received signal of user k is given by

$$\mathbf{r}_{n,k} = \mathbf{H}_{n,k}\mathbf{W}_n\mathbf{x}_n + \mathbf{z}_{n,k},$$

where $\mathbf{W}_n \in \mathbb{C}^{N_t \times N_s}$ denotes the precoding matrix with the Frobenius norm equal to one, i.e., $\|\mathbf{W}_n\|_F^2 = 1$; $\mathbf{x}_n$ denotes the $N_s \times 1$ transmitted symbols with $N_s \leq N_r$ and $\mathsf{E}[\mathbf{x}_n\mathbf{x}_n^H] = \mathbf{I}_{N_s}/N_s$; and $\mathbf{z}_{n,k}$ denotes the additive Gaussian noise (AWGN) with zero mean and $\mathsf{E}[\mathbf{z}_{n,k}\mathbf{z}_{n,k}^H] = N_0\mathbf{I}_{N_r}$, and the superscript H represents the Hermitian transpose.

MU-MIMO precoding is typically designed to mitigate multiuser interferences so maximizing sum rate across all users is justifiable. In contrast, multicast MIMO precoding is designed to maximize the minimum sum rate [12], so that all users can receive the common message $\mathbf{x}$ as pointed out in Step 3 in previous subsection. Remind that the MCS is determined based on the minimum rate user. Thus, the multicast MIMO-OFDM precoding design problem is formulated as follows:

$$\max_{\{\mathbf{W}_n\}} \min_{k\in\mathcal{S}} = \left\{ \sum_{n\in\mathcal{N}} \log_2 \det \left(\boldsymbol{I}_{N_r} + \frac{\mathbf{H}_{n,k}\mathbf{W}_n\mathbf{W}_n^H\mathbf{H}_{n,k}^H}{N_0} \right) \right\} \quad \text{(1a)}$$

$$\text{s.t. } \|\mathbf{W}_n\|_F^2 \leq 1, \ \forall n \in \mathcal{N}, \quad \text{(1b)}$$

where (1b) is for the transmit power constraint. Note that there is no multiuser interference and $\mathbf{W}_n$ is common for all users. This is critical difference between MU-MIMO and multicast. The upper bound of problem (1) can be efficiently obtained by using standard semi-definite progamming techniques.

However, for real-time scheduling of OFDM system, the optimization is computationally complexity-intensive. Instead of the optimal multicast precoding, we consider a near-optimal precoding based on a linear precoding principal, i.e., a precoding matrix $\mathbf{W}_n$ lies in the space spanned by $\{\mathbf{H}_{n,k}^H\}$, as follows [13]:

$$\mathbf{W}_n^* = \alpha \sum_{k\in\mathcal{S}} \left\{ \mathbf{H}_{n,k}^H\mathbf{U}_{n,k}^H \Big/ \left\|\mathbf{H}_{n,k}\mathbf{U}_{n,k}^H\right\|_F^2 \right\}, \ \forall n \in \mathcal{N}, \quad (2)$$

where α is the normalization constant to fulfil $\|\mathbf{W}_n\|_F^2 \leq 1$; $\mathbf{U}_{n,k} = [\mathbf{u}_{n,k,1} \cdots \mathbf{u}_{n,k,N_s}]^H \in \mathbb{C}^{N_s \times N_r}$; and $\mathbf{u}_{n,k,i}$ is a left singular vector corresponding to the ith largest singular value of $\mathbf{H}_{n,k}$. When $N_s = N_r$, without loss of generality, we set $\mathbf{U}_{n,k} = \mathbf{I}_{N_s}$ in (2). The channel gain matrices are replaced by feedback channel estimates in this work.

III. Scheduling Problem Formulation

Consider a finite number of time slot T, whose duration consists of the request for CSI feedback, the transmission of CSI, the transmission of data frames, the transmission of ACK frames, and the backoff period. Note that there is no contention because we do not consider any uplink traffic.

Our goal is to find a scheduling scheme that makes binary transmission decisions $\mu_k[t] \in \{0,1\}$ and frame dropping decisions $\omega_k[t] \in \{0,1\}$ for each time slot of duration $T[t]$. A decision of one implies that positive action is taken. We denote the amount of bit to be transmitted as $b_k[t] = \mu_k[t]\widetilde{b}_k[t]$ and the amount of bit to be dropped as $d_k[t] = \omega_k[t]\widetilde{d}_k[t]$. The $b_k[t]$ values are determined by $\mu_k[t]$ and current CSI $\eta_k[t]$, while the $d_k[t]$ values are determined by $\omega_k[t]$. Since we do not consider MU-MIMO in this work, we have orthogonal channel transmission where $\sum_k \mu_k[t] \leq 1$. In addition, we also constrain $b_k[t]$ and $d_k[t]$ such that no frame fragmentation is required. The dynamics of the queue is modeled as follows:

$$Q_k[t+1] = \max\{0, Q_k[t] - b_k[t] - d_k[t]\} + A_k[t], \quad (3)$$

where $A_k[t]$ is the amount of bits arrived at time slot t. Note that data frames arriving at the current time slot will only be served at the next time slot.

We design our scheduling scheme to maximize the transmission rate and minimize the dropping rate. Hence,the optimization problem is formulated as follows:

$$\max_{\mu_k[t],\omega_k[t]} \frac{\sum_k \overline{b_k} - v_k\overline{d_k}}{\epsilon\overline{T}}, \ \forall k, \quad (4)$$

where v_k and ϵ are the parameters for a maximum deadline constraint and a measuring unit for HOL delay, respectively. Here, we define $\overline{b_k}$ as the time average of transmitted bit $b_k[t]$ as

$$\overline{b_k} = \frac{1}{T}\sum_{t=0}^{T-1} b_k[t], \quad (5)$$

and $\overline{d_k}$ and $\overline{T}$ represent the time average of $d_k[t]$ and $T[t]$, respectively.

We use Lyapunov optimization theory [14], [15] to design scheduling scheme for arbitrary $\eta_k[t]$ and $A_k[t]$. The decision vectors $\boldsymbol{\mu}[t] = [\mu_1[t], \ldots, \mu_K[t]]$ and $\boldsymbol{\omega}[t] = [\omega_1[t], \ldots, \omega_K[t]]$ are chosen by minimizing an upper bound on a drift-plus-penalty ratio [15], which will be defined later. At each time slot, we need to solve a quasiconvex problem. To reduce the complexity, we reformulate the optimization problem as

$$\min_{\mu_k[t],\omega_k[t]} \epsilon\overline{T} - \beta\sum_k \overline{b_k} + \beta\sum_k v_k\overline{d_k}, \ \forall k, \quad (6)$$

where a given parameter β has been added. Note that if $1/\beta$ is the maximum of (4), the problem (6) is equivalent to (4).

Remark 2: In the above formulation, we assume that the time slot has variable duration. If the duration is fixed, $\overline{T}$ is not required in (4) and (6). In addition, the scheduler has the current CSI, which is not true for the transmission procedure described previously. Though outdated CSI is available at the point of making the scheduling decision, the assumption makes the formulation more concise. It is also assumed that the transmission is error-free; therefore, retransmission is unnecessary. If channel error is incurred, we can consider the expectation of $b_k[t]$, $d_k[t]$ and $T[t]$ over the error events. For multiple data frames and retransmission attempts, there are no close-form solutions for these expectation terms. Hence, we devise a heuristic scheduling scheme that behaves properly if the MCS is selected with a low probability of a channel error event.

IV. Proposed Scheduling Scheme

We propose a heuristic scheduling scheme having sequential structure with transmission decision and frame dropping decision. Since we do not consider the overflow of queues, it is a better strategy to serve and then drop the remaining frames. Let $Z_k[t]$ represent the HOL delay at time slot t. The $Z_k[t]$ is updated after $\omega_k[t]$ is made as

$$Z_k[t+1] = \max\{0, \widetilde{Z}_k[t+1] - \phi_k(\omega_k[t])\}. \tag{7}$$

Here, $\widetilde{Z}_k[t+1]$ is an intermediate update on HOL delay after $\mu_k[t]$ has been made as

$$\widetilde{Z}_k[t+1] = \max\{0, Z_k[t] - \psi_k(\mu_k[t])\}. \tag{8}$$

To obtain $Z_k[t+1]$ in (7), $\widetilde{Z}_k[t+1]$ is reduced by the inter-arrival time between the HOL frame and the subsequent frame $M_k[t]$ if the queue is not empty after frames are dropped. If the queue becomes empty, it is reduced by $\widetilde{Z}_k[t+1]$. However, if no frame is dropped, $Z_k[t+1] = \widetilde{Z}_k[t+1]$. Hence, $\phi_k(d_k[t])$ is given by

$$\phi_k[t] = \begin{cases} \min(M_k[t], \widetilde{Z}_k[t+1]), & \text{if } \omega_k[t] = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Similarly, $\psi_k(b_k[t])$ is given by

$$\psi_k[t] = \begin{cases} \min(M_k[t], Z_k[t]), & \text{if } \mu_k[t] = 1, \\ -\epsilon T[t], & \text{otherwise.} \end{cases} \tag{10}$$

In this work, we set $\epsilon = 1,000$, and hence, the HOL delay in (7) and (8) are measured in milliseconds.

Defining the quadratic Lyapunov function

$$L[t] \triangleq \frac{1}{2}\sum_k Z_k[t]^2$$

and the Lyapunov drift on slot t as $\Delta[t] \triangleq L[t+1] - L[t]$, the algorithm is designed to minimize a bound on the following drift-plus-penalty ratio expression [15]:

$$\Delta[t] + V\left\{\epsilon T[t] - \beta\sum_k b_k[t] + \beta\sum_k v_k d_k[t]\right\}, \tag{11}$$

where $V \geq 0$ is a control parameter chosen for performance tradeoff. The Lyapunov drift $\Delta[t]$ is upper bounded as shown in Lemma 1.

Lemma 1: $\Delta[t]$ satisfies

$$\Delta[t] \leq B - \sum_k Z_k[t]\psi_k[t] - \sum_k \widetilde{Z}_k[t+1])\phi_k[t], \tag{12}$$

where B is a finite constant.

Bounding (11) with (12) requires $M_k[t]$, which is a random variable whose value is known only after the scheduling decisions are made. Hence, we approximate $\phi_k[t]$ and $\psi_k[t]$ by $\widetilde{\phi}_k[t]$ and $\widetilde{\psi}_k[t]$, respectively. We define $\widetilde{\phi}_k[t] = \widetilde{Z}_k[t+1]$ if $\omega_k[t] = 1$ and $\widetilde{\psi}_k[t] = Z_k[t]$ if $\mu_k[t] = 1$. To minimize this upper bound, the drift-plus-penalty scheme determines the values of $\mu_k[t]$ and $\omega_k[t]$ decisions every time slot. We label this scheme as a Lyapunov optimization (LO) scheduler and summarize it as follows:

Step 1: Scheduling: For each time slot t, choose $\mu_k[t]$ to

$$\max_{\mu_k[t]} \sum_k Z_k[t]\widetilde{\phi}_k[t] - V\epsilon T[t] + V\beta\sum_k b_k[t] \tag{13}$$

Step 2: Frame Dropping: For each time slot t, choose $d_k[t]$ to

$$\max \;\; \widetilde{Z}_k[t+1]\widetilde{\psi}_k[t] - Vv_k\beta d_k[t] \tag{14}$$

Step 3: Queues Updates: Update the queues $Q_k(t)$, $Z_k(t)$ and $\widetilde{Z}_k[t+1]$ according to (3), (7) and (8), respectively.

If user k' is selected to be served, the objective function in (13) is given by

$$Z_{k'}[t]^2 - \left(\sum_{k \neq k'} Z_k[t] + V\right)\epsilon T_{k'}[t] + V\beta b_{k'}[t], \tag{15}$$

where $b_{k'}[t]$ is the maximum number of bits which can be transmitted while its corresponding transmission time $T[t] = T_{k'}[t]$ is still less than the predetermined threshold $T^{\max}$. As $\sum_k \mu_k[t] \leq 1$ for orthogonal channel transmission, the scheduling problem can be further decomposed into

$$k' = \arg\max_k Z_k[t]^2 - \left(\sum_{j \neq k} Z_j[t] + V\right)\epsilon T_k[t] + V\beta b_k[t].$$

If the AP is multicasting the data frames to the $|\mathcal{S}'|$ users, the value of $b_{k'}[t]$ is increased by $|\mathcal{S}'|$ fold. As usual, the scheduling criterion includes the HOL delay and the transmission rate of the users. In addition, it also includes the transmission time and the number of bits transmitted.

The constraint set for $d_k[t]$ is given by $\{0, L_k[t]\}$, where $L_k[t]$ is the amount of bits (restricted to integer number of frames) that can be dropped from the queue. Solving (14), we have

$$d_k[t] = \begin{cases} L_k[t], & \text{if } \widetilde{Z}_k[t+1]^2 \geq V\beta v_k L_k[t], \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

A. Deterministic Performance Bound

It can be shown that the drift-plus-penalty scheme described comes within $O(1/V)$ of the utility of a genie-aided T'-slot lookahead algorithm with an average delay constraint of $O(V)$

TABLE I
SIMULATION PARAMETERS (FROM IEEE 802.11AC NETWORK).

Parameters	Value
Number of contents	10
Data frame size	1,000 Bytes
Traffic load	0.5, 1, 2 , 5 Mbps
Frame exchange sequence	CSI+Data+ACK+DIFS$^{\diamond}$ + Backoff
Max transmission time	3 msec
Max retransmission attempts	3
$N_t,\ N_r,\ N_s,$	4, 1, 1
SNR	12–45 dB
Bandwidth	20 MHz
MCS	0–7
Deadline $Z^{\max}$	200 msec
Channel model	D [16]
Link abstraction model	Based on mutual information [10]

$^{\diamond}$ DIFS: Distributed Coordination Function (DCF) Interframe Space.

[15]. Furthermore, we can ensure that the frames are dropped with a worst delay given in the following Lemma 2.

Lemma 2: Suppose that $L_k[t] \leq L^{\max}$ is the minimum number of bits to be dropped (restricted to integer number of frames) such that the HOL delay of the subsequent frame has been decreased by more than $T^{\max}$. Then frames are dropped with a maximum value of $Z_k^{\max} = \sqrt{V v_k \beta L^{\max}} \geq Z_k[t]$.

Proof: The proof is shown via induction. By definition, $Z_k[0] = 0 < Z^{\max}$. Hence, it is true for $t = 0$. Now suppose $Z_k[t] \leq Z^{\max} - \epsilon T^{\max}$, this implies $\widetilde{Z}_k[t+1] \leq Z^{\max}$ from (7) and subsequently $Z_k[t+1] \leq Z^{\max}$ from (8). Now suppose $Z^{\max} - \epsilon T^{\max} < Z_k[t] \leq Z^{\max}$, we have $Z^{\max} < \widetilde{Z}_k[t+1] \leq Z^{\max} + \epsilon T^{\max}$ from (7). By design, frame dropping occurs and then $Z_k[t+1] \leq Z^{\max}$. ■

Remark 3: The above scheduling scheme does not consider retransmission. Hence, Lemma 2 no longer holds in this context. An additional mechanism is needed for dropping frames after the deadline. The behavior of the scheduling scheme is explored via simulation in the next section.

V. SIMULATION RESULTS

A. Simulation Framework

Table I lists the simulation parameters used in the IEEE 802.11ac system simulator. The simulator is based on a single cell layout. The link abstraction model is based on the mutual information approach given in [10]. In Fig. 2, the average MAC throughput is plotted against average signal-to-noise ratio (SNR) for the $N_t = 4$, $N_r = 1$ and $N_s = 1$ system in channel D. The dashed lines correspond to the throughput of fixed MCSs. The solid line corresponds the throughput achieved by our link adaptation algorithm [10]. We observe the operating SNR range of this system varies from 18 to 45 dB. We therefore consider the following two deployment scenarios:

- Case 1: 18 dB $\leq$ SNR $\leq$ 45 dB
- Case 2: 30 dB $\leq$ SNR $\leq$ 45 dB

Case 1 attempts to cover the whole operating SNR range, while Case 2 looks at the high SNR region. For both cases, the SNR of the users are uniformly selected from their respective ranges.

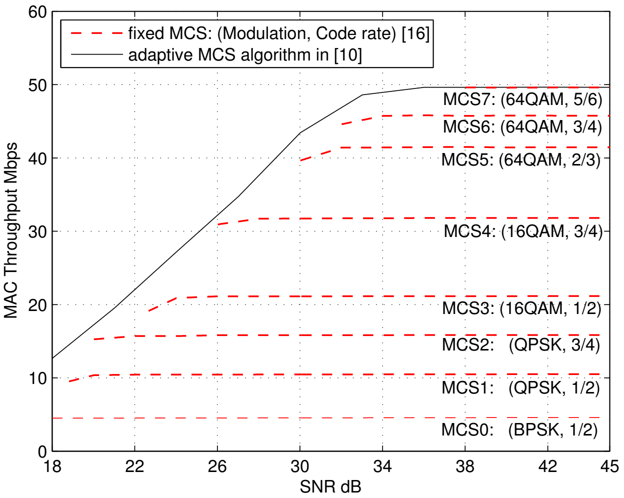

Fig. 2. Evaluation of MAC throughput performance over average SNR.

For traffic model, we consider each user requests for one of the 10 different contents and the selection is done randomly. The start of the frame arrival to the queue is also randomly with the interval of 500 ms. The frames arrive from constant bit rate flow for 2 sec and then pauses for 1 sec. The size of the frame is set to 1,000 Bytes. This cycle is repeated until the simulation is ended. The time duration of each simulation run lasts 30 sec and all simulation results are averaged over 100 sessions.

We now discuss the parameter selection for the LO scheduler. From Fig. 2, the maximum throughput is around 50 Mbps. We select the estimated throughput for LO scheduler to be 25 Mbps and hence $\beta = 4 \times 10^{-5}$. Note that the performance of LO scheduler is not sensitive to the value of estimated throughput as long as the estimated throughput is of the same order as the simulated throughput. The maximum HOL delay of all users is set to $Z^{\max} = 200$ ms and $L^{\max} = 8,000$ bits. This implies that $V v_k = 1.25 \times 10^5$. We vary the value of V from 1 to 10,000 and found that the $V = 1,000$ gives the best performance. Therefore, we set $v_k = 125$.

B. Performance of the Schedulers

Since the traffic load (and hence, the arrival rates) are fixed in the simulation, we consider the number of users that can be supported by the system as our performance metric. This user capacity depends on the the traffic load and the choice of the outage criteria. The outage criteria in our simulation is similar to the evaluation methodology in [17]. We consider a user to be in outage if more than 1% of the frames are either lost or delivered with a delay exceeding the de-jitter buffer delay. The system is considered to be in outage if more than 1% of the users are in outage. Table II lists the user capacities for Case 1 and Case 2 for various schedulers and traffic load.

We first look at the user capacities for the schedulers without multicast transmission. As shown in Table II, the user capacity

TABLE II
CAPACITIES FOR CASE 1 (UNIT: NUMBER OF USERS).

Scheduler	Data Rate: Case 1						Data Rate: Case 2					
	0.5 Mbps		1 Mbps		2 Mbps		1 Mbps		2 Mbps		5 Mbps	
multicast	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes
LO	57	67	29	31	12	12	53	72	27	36	10	13
MLWDF	39	53	21	26	10	11	49	69	26	33	10	12
RR	32	41	17	20	8	9	27	51	19	29	10	11

TABLE III
99 PERCENTILE CACHE SIZE FOR MULTICAST (UNIT: FRAMES).

Scenario	Case 1			Case 2		
Scheduler	Data Rate Mbps					
	0.5	1	2	1	2	5
LO	27	55	80	68	126	292
MLWDF	26	53	77	60	114	253
RR	29	53	80	65	127	267

for traffic load of 0.5 Mbps in Case 1 are 57, 39, and 32 for LO, Maximum-Largest Weighted Delay First (MLWDF) [18], and Round Robin (RR) schedulers, respectively. For Case 1, the LO scheduler has the highest capacity, while the RR scheduler has the lowest capacity. The lack of transmission rate and HOL delay in the calculation for scheduling priority is the reason why RR has the worst capacity. The addition of transmission time in the calculation for scheduling priority allows the LO scheduler to achieve higher capacity than an MLWDF scheduler. The gain is more significant for high density deployment with low data rate. For traffic load of 0.5 Mbps, LO scheduler can support up to 57 users, compared to 39 users for MLWDF scheduler. That implies a gain of 46%. However, the gain vanishes if the SNR of the users are very high as shown in Case 2. From Table II, we see that the LO scheduler has similar user capacity as the MLWDF scheduler, yet it still outperforms the MLWDF scheduler.

Next, let look at the user capacities for the schedulers with multicast transmission. As shown in Table II, the user capacity for Case 1 and traffic load of 0.5 Mbps are 67, 53, and 41 for LO, MLWDF, and RR schedulers, respectively. The gain from multicast is more significant for high density deployment as the opportunity for multicast increases as the number of users increases. For data rate of 0.5 Mbps in Case 1, the LO scheduler can support up to 57 users and 67 users for unicast and multicast mode, respectively. That translates to an increase of 17% in user capacity. Higher gain can be achieved by using multicast mode with MLWDF and RR schedulers, although their user capacities are still lower than that of the proposed LO scheduler. In addition, the gain is more significant if the SNR of the users are very high as shown in Case 2. For data rate of 0.5 Mbps, the user capacity of LO scheduler increases from 53 to 72, which is a gain of 35%.

Finally, we look at the size of cache required at the user's device in Table III. In general, the higher the data rate, the larger the size of cache required. The opportunity for multicasting also determines the size of cache. For a given traffic load, users in Case 2 require larger size of cache than users in Case 1.

VI. CONCLUSION

We have proposed a transmission scheme for exploiting content reuse with opportunistic user requests. The proposed opportunistic multicast transmission is considered in a unicast environment to reduce the wireless resource usage. The Lyapunov optimization approach for the multicast scheduling scheme is designed for real-time traffic. Numerical simulations over WLAN networks have been presented to show the effectiveness of the proposed scheme. It is observed that significant multicast gain (35%) is achievable at higher operating SNR environment with the expense of larger cache memory at user's device.

REFERENCES

[1] H. Won, H. Cai, A. N. I. R. D Young Eun, Katherine Guo, and K. Sabnani, "Multicast scheduling in cellular data networks," *IEEE Trans. Wireless Commun.*, vol. 8, pp. 4540–4549, Sep. 2009.

[2] T.-P. Low, M.-O. Pun, Y.-W. Peter Hong, and C.-C. Jay Kuo, "Optimized opportunistic multicast scheduling (OMS) over wireless cellular networks," *IEEE Trans. Wireless Commun.*, vol. 9, pp. 791–801, Feb. 2010.

[3] J.-T. Tsai and R. L. Cruz, "Opportunistic multicast scheduling for information streaming in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 10, pp. 1776–1785, Jun. 2011.

[4] L.-S. Juhn and L.-M. Tseng, "Harmonic broadcasting for video-on-demand service," *IEEE Trans. on Broadcasting*, vol. 43, pp. 268–271, Mar. 1997.

[5] S. Chand and H. Om, "Geometrico-harmonic data broadcasting and receiving scheme for popular videos," *'IEEE Trans. on Circuits and Systems for Video Technol.'*, vol. 17, pp. 16–25, Jan. 2007.

[6] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," *ArXiv Tech. Report, arXiv:1109.4179v4*, Sep. 2013.

[7] M. Ji, G. Caire, and A. F. Molisch, "Optimal throughput-outage trade-off in wireless one-hop caching networks," in *Proc IEEE Int. Symp. Inform. Theory*, Istanbul, Turkey, Jul. 2013.

[8] D. Bethanabhotla, G. Caire, and M. J. Neely, "Utility optimal scheduling and admission control for adaptive video streaming in small cell networks," in *Proc IEEE Int. Symp. Inform. Theory*, Istanbul, Turkey, Jul. 2013.

[9] IEEE 802.11 Standard, "Wireless LAN medium access control (MAC) and physical layer (PHY) specifications. amendment 4: Enhancement for very high throughput for operation in bands below 6 GHz," *IEEE*, 2013.

[10] P. H. Tan, Y. Wu, and S. Sun, "Link adaptation based on adaptive modulation and coding for multiple-antenna OFDM system," *IEEE J. Select. Areas Commun.*, vol. 26, pp. 1599–1606, Oct. 2008.

[11] IEEE 802.11 Standard, "Wireless LAN medium access control (MAC) and physical layer (PHY) specifications amendment 2: MAC enhancements for robust audio video streaming," *IEEE*, 2012.

[12] N. Jindal and Z.-Q. Luo, "Capacity limits of multiple antenna multicast," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Seattle, WA, Jul. 2006, pp. 1841–1845.

[13] J. Joung, H. D. Nguyen, P. H. Tan, and S. Sun, "Linear precoding for multicast MIMO-OFDM systems," under minor revision, Feb. 2015.

[14] M. Neely, *Stochastic network optimization with application to communication and queueing systems*, ser. Synthesis Lectures on Communication Networks. Morgan and Claypool Publishers, 2010.

[15] ——, "Universal scheduling for networks with arbitrary traffic, channel and mobility," *ArXiv Tech. Report, arXiv:1001.0960v1*, Jan. 2010.

[16] IEEE 802.11 TGn channel model special committee, "TGn channel models for IEEE 802.11 WLANs," *IEEE 802.11-03/940r4*, May 2004.

[17] 3GPP TR 25.814 v7.1.0, "Physical layer aspects for evolved universal terrestrial radio access (UTRA) release 7," Sep. 2006.

[18] A. L. Stolyar and K. Ramanan, "Largest weighted delay first scheduling: Large deviations and optimality," *Ann. Appl. Prob.*, vol. 11, pp. 1–48, Feb. 2001.