resources.

- High Throughput: We focus on a many-user many-channel asymptotic regime. First, we derive an algorithm independent lower bound on the number of channels needed for achieving any non-vanishing throughput as the number of receivers increases. Then, we show that MC-MWC achieves the lower bound in an order sense. Hence, MC-MWC achieves order-optimal throughput in the many-user many-channel asymptotic regime. Furthermore, we prove that the number of channels required by a conventional scheme based on the optimal static channel allocation and capacity-achieving codes is doubly-exponentially larger than that required by MC-MWC.

- Low Delay: Using large deviations theory, we show that the delay of MC-MWC decreases linearly as the number of channels grows, while the delay reduction of conventional channel-allocation based schemes (even incorporating with any coding schemes) is no more than a finite constant. To the best of our knowledge, MC-MWC *is the first wireless multicast scheme in which the delay decreases linearly as the number of channels grows, with no loss in the per-channel-throughput.*

- Low Feedback Overhead: By combining MC-MWC with the recently proposed anonymous feedback technique [6, 12], the total feedback overhead of MC-MWC is a constant, independent of not only the number of receivers in each session but also the number of sessions in the network.

- Trace-driven and numerical results are provided to validate the analytical results and show that 1) MC-MWC achieves significant throughput and delay improvements in the practical scenarios even when the number of users (channels) is not large, 2) the implementation complexity of MC-MWC is low in practice.

The rest of this paper is organized as follows. In Section II, we introduce some related works. In Section III, the system model is described. In Section IV, we introduce Multi-channel Moving Window Codes (MC-MWC). The throughput and delay performance of MC-MWC are analyzed in Section V and Section VI, respectively. In Section VII, we evaluate the performance of our method through simulations. Finally, we conclude the paper in Section VIII.

II. RELATED WORK

In wireless unicast, exploiting the multi-channel resources has been extensively studied. For example, in one line of work, e.g., [13, 14], the multi-channel resources are allocated based on the queue lengths of the users such that throughput optimality is achieved. However, throughput optimality does not necessarily imply good delay performance. Thus, another line of work, e.g., [15, 16] proposed delay-based scheduling policies, which achieve optimal throughput and provably good delay performance for multi-channel wireless unicast.

In wireless multicast, multiple receivers may receive the same transmitted packet, and as a result, the work-conservation principle assumed in wireless unicast is violated [17]. In

addition, rateless codes or network codes, e.g., [2–4, 6–10]. are typically used to achieve good throughput performance. With coding, the receivers need additional time to decode the packets and thus the queueing delay considered in wireless unicast does not fully capture the overall delay. For the above two reasons, the methodologies developed for multi-channel wireless unicast are not applicable to the multicast scenario.

Existing studies about resource allocation in wireless multicast are either channel-statistics based or channel-state based. In the first category, e.g., [18, 19], it is assumed that the transmitter has access to the channel statistics of the receivers. To realize cooperative multicast, the authors in [18] proposed a network-coding based multicast scheme, in which the scheduling policy depends on the channel statistics of the receivers. In [19], to maximize the long-term throughput, the channels are allocated statically according to the channel statistics. The second category, e.g., [9, 17] assumes that the transmitter has perfect knowledge of the channel state information of the receivers and exploits the opportunistic gain in throughput or delay. A throughput-optimal policy was developed in [17] for wireless multicast with dynamic topology. In [9], the authors proposed an instantly decodable network coding scheme which achieves zero decoding delay at the cost of throughput loss. However, collecting channel state information from all receivers incurs a prohibitive large feedback overhead, rendering them impractical when multicast sessions have a large number of receivers. An open question is *how do we obtain a linear delay reduction in the number of channels without loss in the per-channel throughput?*

III. SYSTEM MODEL

We consider a multi-channel, multi-session wireless multicast network with m orthogonal channels and m multicast sessions, as shown in Figure 1. As in [13, 15, 16], for ease of presentation, we assume that the number of channels is equal to the number of sessions. Our throughput and delay analysis can be readily generalized to the scenario when the number of channels scales linearly with the number of sessions.

A. Multicast Sessions

For a multicast session $h \in \{1, \ldots, m\}$, the transmitter needs to send a stream of data packets to a set of receivers, denoted as $\mathcal{S}_h$. For any two different sessions h and h', their sets of receivers $\mathcal{S}_h$ and $\mathcal{S}_{h'}$ are allowed to have arbitrary intersections, i.e., a receiver may be interested in multiple sessions. Let $s_h = |\mathcal{S}_h| \geq 1$ denote the number of receivers in the session h. The number of receivers in different sessions may be different and we assume that $s_1, \ldots, s_m$ are *i.i.d.* across different sessions[3]. Let $\widehat{s}^{(n)}$ denote a random variable which has the same distribution as s_h, where n represents the expected number of receivers in one session, i.e., $n = \mathbb{E}[s_h]$.

In our model, time is slotted. The packets which need to be served in each session arrive stochastically. Let $a_h[t] \in \mathbb{N}$ denote the number of packet arrivals from session h at the

---

[3]This assumption can be relaxed such that $\{s_h\}_h$ have heterogeneous distributions. The details are omitted due to space limitations.
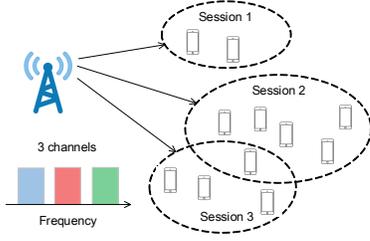
Fig. 1. Multi-channel, multi-session wireless multicast.

Fig. 2. Multi-channel Moving Window Codes.

beginning of time-slot t. We assume that $\{a_h[t]\}_{t \in \mathbb{N}}$ are stationary and bounded random variables, independent across time-slots, independent of the number of receivers $\{s_h\}_h$. We use $\lambda_h \triangleq \mathbb{E}[a_h[t]]$ to denote the expected packet arrival rate in session h.

B. Channel Model

The channel between the transmitter and the receivers is assumed to follow the standard broadcast erasure channel model. Let $c_{i,j}[t]$ denote the channel state of receiver i on channel j in time-slot t, given by

$$c_{i,j}[t] = \begin{cases} 1 & \text{if the packet sent on channel } j \text{ can be} \\ & \text{successfully received by receiver } i \\ & \text{in time-slot } t; \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

As in [2–10], it is assumed that $c_{i,j}[t]$ are *i.i.d.* across different time-slots, and let $\gamma_{i,j} \triangleq \mathbb{P}(C_{i,j}[t] = 1)$ denote the probability of successful packet reception at receiver i on channel j. We make the following assumption on $\{\gamma_{i,j}\}$.

Assumption (Heterogeneous Channel Conditions): The channel statistics $\{\gamma_{i,j}\}$ are random variables that have heterogeneous values for different receivers and different channels. Specifically, $\gamma_{i,j}$ are *i.i.d.*[4] across receivers and across channels. Let $F_\gamma(\cdot)$ and $\bar{\gamma} \triangleq \mathbb{E}[\gamma_{i,j}]$ denote the the CDF and the expectation of $\gamma_{i,j}$, respectively. Moreover, there exists $\kappa > 0$ such that $\lim_{y \to 0^+} \frac{F_\gamma(y)}{y} \geq \kappa$.

The above assumptions are more general than the homogeneous network assumptions in [5, 7, 8, 10], where all $\gamma_{i,j}$ are assumed to be equal and thus fail to reflect the heterogeneous channel conditions in reality. In addition, the existence of κ is fairly weak, since it basically says that there is a non-zero probability that $\gamma_{i,j}$ is close to 0. In Section VII, we will provide simulation results that are obtained using data traces collected from a software defined radio platform. We will see that the analytical results obtained under the above assumptions are also valid in practice.

IV. MULTI-CHANNEL MOVING WINDOW CODES

In this section, we propose a Multi-Channel Moving Window Codes (MC-MWC) approach, which is built on the recent work Moving Window Codes (MWC) [6]. Compared with

---

[4]Note that while $\{\gamma_{i,j}\}_{i,j}$ are *i.i.d.*, they correspond to the probability of successful packet reception of receiver i on channel j. So, these probabilities can be different for different realizations of $\gamma_{i,j}$, hence the channel conditions are heterogeneous.
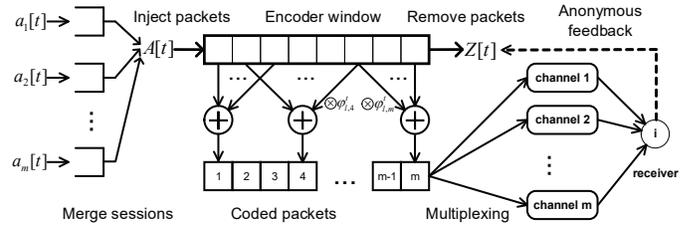
MWC, MC-MWC has two key novelties. First, MWC is designed for the single-channel single-session multicast, and actually we can show that a straightforward generalization of MWC to the multi-channel multi-session multicast setting will lead to poor throughput and poor delay performance. Second, owning to a new Merging strategy, MC-MWC simultaneously achieves good throughput and delay performance by exploiting the multi-channel resources. In the following, we first introduce a Merging strategy, and then propose MC-MWC.

A. A Simple Merging Strategy

The key idea behind MC-MWC is a simple Merging strategy.

Definition 1: (Merging) The m multicast sessions are jointly served by the multi-channel resources as follows:

Tx's side : At the transmitter, the streams of packets from all m sessions are merged to form one large multicast session. Then, the transmitter uses the m channels to send the grouped multicast session to all receivers in $\bigcup_{1 \leq h \leq m} \mathcal{S}_h$.

Rx's side : In each time-slot, a receiver $i \in \bigcup_{1 \leq h \leq m} \mathcal{S}_h$ listens on all the m channels to receive as many packets as it can. Then, the received packets are used to decode the merged multicast session, including the session(s) receiver i is interested in.

Notice that Merging only describes some high-level multicast transmission rules. The coding, transmission, and feedback procedures of a specific example of Merging will be explained next in detail.

In practice, the sessions can be divided into groups and Merging is applied to the sessions within each group. The choice of group size with Merging depends on the performance requirements on throughput, delay, and implementation complexity, as will be systematically analyzed and shown in Sections V, VI, and VII.

B. Multi-Channel Moving Window Codes

As shown in Figure 2, MC-MWC is comprised of three modules, i.e., the encoder at the transmitter side, the decoder at the receiver side, and the feedback modules.

1) Encoder: At the transmitter, all the packets from all sessions $\{1, \ldots, m\}$ are grouped together to form one large multicast session. The packets in the merged multicast session are indexed as $p_1, p_2, \ldots$, where the packets are indexed according to the order of their arrivals and the tie breaking rule for the packets arriving at the same time-slot is arbitrary.

All the newly-arrived packets at the beginning of time-slot t are instantly injected[5] to an encoder window. At time-slot t, the total number of packet arrivals from all the m sessions is denoted as $a[t] \triangleq \sum_{h=1}^{m} a_h[t]$. Then, the number of packets that the encoder has received up to the beginning of time-slot t is $A[t]$, i.e.,

$$A[t] = \sum_{\tau=1}^{t} a[\tau] = \sum_{\tau=1}^{t} \sum_{h=1}^{m} a_h[t]. \tag{2}$$

To prevent the encoder window from growing indefinitely over time, at the beginning of time-slot t, $Z[t]$ packets which are "determinable"[6] at all receivers $\bigcup_{1 \le h \le m} \mathcal{S}_h$ ($Z[t]$ can be determined via a low overhead feedback mechanism described later) are removed from the encoder window.

At time-slot t, the encoder generates m coded packets through linear combinations of the data packets in the encoder window, which are transmitted over the m channels to all receivers $\bigcup_{1 \le h \le m} \mathcal{S}_h$. Specifically, the coded packet $x_j[t]$ which is transmitted on the channel $j \in \{1, \ldots, m\}$ in time-slot t is generated by

$$x_j[t] = \sum_{l=Z[t]+1}^{A[t]} \varphi_{l,j}^t \times p_l, \tag{3}$$

where p_l denotes the lth packet of the merged session, "$\times$" is the product operator on a Galois field $GF(2^q)$, and $\left\{ \varphi_{l,j}^t \right\}_{\forall l,j,t}$ are independently drawn according to a uniform distribution on $\{GF(2^q)\}$. The values of $A[t]$ and $Z[t]$ are embedded in the packet header of $x_j[t]$. Moreover, $\left\{ \varphi_{l,j}^t \right\}_{\forall l,j,t}$ can be known at each receiver by feeding the same seed to the random number generators of the transmitter and all the receivers.

2) Decoder: With Merging, a receiver $i \in \bigcup_{1 \le h \le m} \mathcal{S}_h$ receives all information sent on the m channels and decode all the m sessions including the session(s) it is interested in, i.e., $\{h : i \in \mathcal{S}_h\}$.

To facilitate a clear understanding of the decoding procedure, we restate the definition of *"determinable" packet* that was originally defined in [7].

Definition 2: ("Determinable" packet) A packet p_l is said to be "determinable" at a receiver if the receiver has enough information to express p_l as a linear combination of some packets $p_{l+1}, p_{l+2}, \ldots$ with greater indices, and "indeterminable" otherwise.

Let $S_i[t] \in \mathbb{N}$ be the number of "determinable" packets at receiver i by the end of time-slot t. Define a *virtual decoder queue*

$$Q_i[t] = A[t] - S_i[t] \tag{4}$$

for each receiver $i \in \bigcup_{1 \le h \le m} \mathcal{S}_h$. Then, $Q_i[t]$ is the number of packets which have arrived from one of the m sessions but

---

[5]The injection rate has to be within the capacity region of MC-MWC, which will be discussed and analyzed in Section V.

[6]"Determinable" will be formally defined later in the decoder module.

are "indeterminable" at receiver i at the end of time-slot t.

Suppose at the beginning of time-slot t, the packets $p_1, \ldots, p_{l_1}$ are "determinable" at receiver i and the total number of packets that the encoder received is $A[t] = l_2$. Similar to [6, 7], when the field size 2^q is sufficiently large, for every successfully received packet over the m channels at time-slot t, with a high probability, the next "indeterminable" packet becomes "determinable" until all the packets $p_1, \ldots, p_{l_2}$ are "determinable". To better see this, consider the following simple example. The receiver i receives 3 coded packets from channels $2, 3, 5$ in time-slot 1, i.e., $x_2[1] = p_1 + p_2 + 2p_3$, $x_3[1] = p_1 + 2p_2 + p_3$ and $x_5[1] = 2p_1 + p_2 + p_3$. From the first packet, $p_1 = x_2[1] - p_2 - 2p_3$, thus, by definition, p_1 becomes "determinable". Similarly, from the second and third packets, we have $p_2 = x_3[1] - x_2[1] + p_3$ and $4p_3 = 3x_2[1] - x_3[1] - x_5[1]$, by which p_2 and p_3 are "determinable" in turn. Therefore, the evolution of $S_i[t]$ is given by

$$S_i[t] = S_i[t-1] + \min \left\{ \sum_{j=1}^{m} c_{i,j}[t], A[t] - S_i[t-1] \right\}. \tag{5}$$

Combining Equations (4), (5) and letting $c_i[t] \triangleq \sum_{j=1}^{m} c_{i,j}[t]$, the evolution of $Q_i[t]$ can be expressed as:

$$Q_i[t] = \left\{ Q_i[t-1] + \sum_{h=1}^{m} a_h[t] - \sum_{j=1}^{m} c_{i,j}[t] \right\}^+$$
$$= \left\{ Q_i[t-1] + a[t] - c_i[t] \right\}^+. \tag{6}$$

Similar to [6, 7], decoding occurs at receiver i in time-slot t if and only if all the packets, which have arrived from one of the m sessions are "determinable" at receiver i, i.e.,

$$Q_i[t] = 0, \tag{7}$$

when all packets $p_1, \ldots, p_{A[t]}$ could be decoded.

3) Feedback: Recall that to prevent the encoder window from growing indefinitely over time, $Z[t]$ packets are removed from the encoder window at the beginning of time-slot t. To achieve reliable multicast, these $Z[t]$ packets must be "determinable" at all receivers $\bigcup_{1 \le h \le m} \mathcal{S}_h$ at the beginning of time-slot t, and hence $Z[t]$ must be controlled based on the status feedback from all the receivers. To avoid the prohibitively large overhead caused by traditional per-receiver feedback mechanisms, we modify the anonymous feedback mechanism proposed in the recent work [6, 12] to ensure reliable multicast reception at all receivers $i \in \bigcup_{1 \le h \le m} \mathcal{S}_h$ with a negligible feedback overhead.

Anonymous feedback was devised for a single multicast session [6, 12], which guarantees reliable multicast with constant feedback overhead, regardless of the number of receivers. However, without Merging, applying anonymous feedback to the multi-session multicast would incur an overhead which increases with the number of sessions.

In MC-MWC, anonymous feedback can be applied to the merged session. The key idea is to let the receiver(s), for which

the oldest m packets in the encoder are not all "determinable", send a NAK in a shared feedback channel at the end of each time slot. This feedback is "anonymous" in the sense that the transmitter does not differentiate the actual ID(s) of the receiver(s) sending NAK. As long as the transmitter can detect the existence of NAK signal from feedback, it will keep the oldest m packet in the encoder buffer; otherwise, the oldest m packets are removed from the buffer. This mechanism ensures that

$$Z[t] \le \min_{i \in \bigcup_{1 \le h \le m} \mathcal{S}_h} S_i[t-1], \tag{8}$$

which guarantees reliable multicast. *The anonymous feedback in MC-MWC only requires a short sub-slot to detect the existence of NAK in the shared feedback channel. Thus, the total feedback overhead of MC-MWC is* **a constant**, *not only independent of the number of receivers in each session, but also independent of the number of sessions in the network.* Compared with conventional schemes, e.g., [2–4, 7–10], the feedback overhead of MC-MWC is $\Theta(mn)$ times smaller. Compared with MWC [6], the feedback overhead of MC-MWC is m times smaller. Therefore, MC-MWC effectively reduces the feedback overhead when there are a large number of sessions/receivers. In [12], it is shown that anonymous feedback can be easily implemented and incurs a low overhead in practice.

V. HIGH THROUGHPUT OF MC-MWC

In this section, we analyze the throughput performance of MC-MWC. Noting that multicast is most beneficial when there are a large number of receivers, we focus on a scenario when both the number of receivers and the number of channels scale to infinity. We refer to this setting as the *many-user many-channel asymptotic regime*. We emphasize that although our analysis is in the asymptotic regime, the results provide important insights on the practical scenario when the number of users (channels) is not large (e.g., only 10-20 users per session), as illustrated in Section VII. To facilitate the analysis in this regime, we make the following assumption on the distribution of the number of receivers, i.e., $\{\widehat{s}^{(n)}\}_{n \in \mathcal{N}}$. There is a function $z_s(\cdot)$ such that for any $\alpha \ge 0$ and for any $n \in \mathbb{N}$,

$$\mathbb{P}\left(\frac{\widehat{s}^{(n)}}{\mathbb{E}[\widehat{s}^{(n)}]} \le \alpha\right) \le z_s(\alpha), \tag{9}$$

and

$$\lim_{\alpha \to 0} z_s(\alpha) = 0. \tag{10}$$

Note that this is a mild assumption. For instance, the assumption is satisfied when $\widehat{s}^{(n)}$ follows common discrete distributions such as geometric distribution, discrete uniform distribution, Poisson distribution, degenerate distribution, etc.

A. Algorithm Independent Lower Bound

In the single-channel single-session multicast case, the achievable throughput is bottlenecked by the receiver with the worst channel condition. As a result, with the number

of receivers increasing, it is more and more likely that there exists one receiver whose channel condition happens to be poor, leading to a vanishing multicast throughput. We are interested in understanding the following fundamental question *in a multi-channel, multi-session multicast: how many multi-channel resources are required to achieve a non-vanishing per-session throughput?* We answer the question by deriving an algorithm independent scaling law on the number of channels required to achieve a non-vanishing throughput as the number of receivers increases.

First, given m, n and the channel statistics $\Gamma = \{\gamma_{i,j}\}_{i,j}$, we define the full capacity region as the set of arrival rate vectors that can be stabilized by some multicast scheme that has the instantaneous channel state information (CSI) of all receivers at the transmitter.

$$\Lambda_{\text{Full}}^{(m,n,\Gamma)} = \Big\{ \vec{\lambda}^{(m,n,\Gamma)} = \left(\lambda_1^{(m,n,\Gamma)}, \dots, \lambda_m^{(m,n,\Gamma)}\right) \Big|$$
$$\vec{\lambda}^{(m,n,\Gamma)} \text{ can be stabilized by some}$$
$$\text{multicast scheme with CSI at Tx's side.}\Big\}. \tag{11}$$

Note that it is well known that without coding, $\Lambda_{\text{Full}}^{(m,n,\Gamma)}$ can be expressed as the convex hull of all feasible schedules. Nevertheless, with the possibility of inter-session coding, it is generally very difficult to explicitly characterize $\Lambda_{\text{Full}}^{(m,n,\Gamma)}$ even when the channel statistics $\Gamma = \{\gamma_{i,j}\}_{i,j}$ is given. Instead, we derive a fundamental property of $\Lambda_{\text{Full}}^{(m,n,\Gamma)}$ in the many-user many-channel asymptotic regime.

For any session h, given the channel statistics $\Gamma = \{\gamma_{i,j}\}_{i,j}$, define the maximum achievable throughput of session h in the full capacity region $\Lambda_{\text{Full}}^{(m,n,\Gamma)}$ as

$$\overline{\lambda}_h^{(m,n,\Gamma)} \triangleq \max_{\vec{\lambda}^{(m,n,\Gamma)} \in \Lambda_{\text{Full}}^{(m,n,\Gamma)}} \lambda_h^{(m,n,\Gamma)} \tag{12}$$

Then we have the following theorem regarding $\overline{\lambda}_h^{(m,n,\Gamma)}$.

Theorem 1: If the number of channels scales slower than logarithmically[7] with the expected number of receivers, i.e., $m = O\left((\log n)^\delta\right)$ for some $\delta \in (0,1)$, then the achievable throughput of any multicast scheme vanishes as $n \to \infty$:

$$\overline{\lambda}_h^{(m,n,\Gamma)} \xrightarrow{P} 0, \quad \forall h \in \{1, \dots, m\}, \tag{13}$$

in which $\xrightarrow{P}$ denotes "convergence in probability".

Proof sketch of Theorem 1: The key idea is to define the set of "bottleneck" receivers for a session h as

$$\mathcal{B}_\lambda^{(m,n,\Gamma)} = \left\{ i \in \mathcal{S}_h \Big| \gamma_{i,j} < \frac{\lambda}{m}, \quad \forall 1 \le j \le m \right\}.$$

If $\mathcal{B}_\lambda^{(m,n,\Gamma)} \ne \emptyset$, for any receiver $i \in \mathcal{B}_\lambda^{(m,n,\Gamma)}$, with any possible scheduling and coding scheme, we could upper bound the achievable throughput of receiver i by allocating all m channels to serve receiver i at all time-slots, in which case its throughput is upper bounded by the sum of the capacity of all

---

[7]In this paper, $\log(\cdot)$ denotes the natural logarithm.

m channels,

$$\sum_{j=1}^{m} \gamma_{i,j} < \sum_{j=1}^{m} \frac{\lambda}{m} = \lambda. \tag{14}$$

Thus, when $\mathcal{B}_{\lambda}^{(m,n,\Gamma)} \neq \emptyset$, the achievable throughput of session h is also upper bounded by Equation (14). Then, the focus is to show that for any $\lambda > 0$ and any $p < 1$, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\mathcal{B}_{\lambda}^{(m,n,\Gamma)} \neq \emptyset\right) \geq p.$$

We provide the proof details in Appendix A.

Remark 1.1: Theorem 1 suggests that, to achieve any non-vanishing throughput, the number of channels must scale at least logarithmically with the expected number of receivers, i.e., $m = \Omega\left(\log n\right)$.

B. Order Throughput Optimality of MC-MWC

Now we analyze the achievable throughput of MC-MWC. Given the number of receivers in each session $\{s_h\}_{1\leq h\leq m}$ as well as the channel statistics $\{\gamma_{i,j}\}_{i,j}$, the maximum achievable throughput of MC-MWC can be easily derived as follows. For a receiver i, it receives information on all the m channels, the sum capacity of which is $\sum_{j=1}^{m} \gamma_{i,j}$. Recall the capacity of multicast is limited by the worst receiver, the maximum sum throughput of all sessions is then given by

$$\sum_{h=1}^{m} \lambda_h \leq \min_{i\in\bigcup_{h=1}^{m} \mathcal{S}_h} \sum_{j=1}^{m} \gamma_{i,j}. \tag{15}$$

On the other hand, according to Equation (6), $\{Q_i[t]\}_t$ is a random walk on $[0,\infty)$, and has a steady state distribution if Equation (15) holds. Hence, Equation (15) is achievable by MC-MWC.

Similar to the definition of $\Lambda_{\text{Full}}^{(m,n,\Gamma)}$, we define the capacity region of MC-MWC.

$$\Lambda_{\text{MC-MWC}}^{(m,n,\Gamma)} = \left\{ \vec{\lambda}^{(m,n,\Gamma)} \middle| \vec{\lambda}^{(m,n,\Gamma)} \text{ stabilizable by MC-MWC} \right\}.$$

To characterize $\Lambda_{\text{MC-MWC}}^{(m,n,\Gamma)}$, let us define a convex polytope $\mathcal{P}_{\lambda}^{(m)} \in \mathcal{R}^m$ for a given m and any given $\lambda < \bar{\gamma}$ as

$$\mathcal{P}_{\lambda}^{(m)} = \left\{ \vec{\lambda}^{(m)} = \left(\lambda_1^{(m)}, \ldots, \lambda_m^{(m)}\right) \middle| \right.$$
$$\left. \frac{1}{m}\sum_{h=1}^{m} \lambda_h^{(m)} \leq \lambda, \lambda_h^{(m)} \geq 0, \forall h \right\}. \tag{16}$$

The polytope $\mathcal{P}_{\lambda}^{(m)} \in \mathcal{R}^m$ contains all rate vectors such that the average rate is no greater than λ.

Then we derive the following theorem regarding $\Lambda_{\text{MC-MWC}}^{(m,n,\Gamma)}$.

Theorem 2: If the number of channels scale logarithmically with the expected number of receivers, i.e., $m = \left\lceil \frac{1}{(\bar{\gamma}-\lambda)^2} \log n \right\rceil$ for any $\lambda < \bar{\gamma}$, then the achievable per-session throughput of MC-MWC is lower bounded by λ in the following sense:

$$\lim_{n\to\infty} \mathbb{P}\left(\mathcal{P}_{\lambda}^{(m)} \subseteq \Lambda_{\text{MC-MWC}}^{(m,n,\Gamma)}\right) = 1. \tag{17}$$

Proof sketch of Theorem 2: Let us prove that, with MC-MWC, for any $p \in (0,1)$, if $m = \left\lceil \frac{1}{(\bar{\gamma}-\lambda)^2} \log n \right\rceil$, then $\mathcal{P}_{\lambda}^{(m)} \subseteq \Lambda_{\text{MC-MWC}}^{(m,n,\Gamma)}$ with probability no less than p, as $n \to \infty$.

First, based on Equations (6) and (15), we show that,

$$\mathbb{P}\left(\mathcal{P}_{\lambda}^{(m)} \subseteq \Lambda_{\text{MC-MWC}}^{(m,n,\Gamma)}\right) = \mathbb{P}\left(\frac{1}{m}\min_{i\in\bigcup_{h=1}^{m} \mathcal{S}_h} \sum_{j=1}^{m} \gamma_{i,j} \geq \lambda\right). \tag{18}$$

Note that for any $\tau \geq 1$, it can be shown that

$$\mathbb{P}\left(\left|\bigcup_{h=1}^{m} \mathcal{S}_h\right| \geq \tau mn\right) \leq \frac{1}{\tau}. \tag{19}$$

Then, using Hoeffding's inequality, we derive a lower bound of Equation (18).

$$\mathbb{P}\left(\frac{1}{m}\min_{i\in\bigcup_{h=1}^{m} \mathcal{S}_h} \sum_{j=1}^{m} \gamma_{i,j} \geq \lambda\right)$$
$$\geq \left(1 - \frac{1}{\tau}\right)\left(1 - \lceil\tau mn\rceil e^{-2(\bar{\gamma}-\lambda)^2 m}\right). \tag{20}$$

For any $p \in [0,1)$, let $\tau = \frac{2}{1-p}$ and $m = \left\lceil \frac{1}{(\bar{\gamma}-\lambda)^2} \log n \right\rceil$. Combining with Equations (18) and (20), we have

$$\mathbb{P}\left(\mathcal{P}_{\lambda}^{(m)} \subseteq \Lambda_{\text{MC-MWC}}^{(m,n,\Gamma)}\right) \geq \frac{p+1}{2}\left(1 - o(1)\right),$$

where $o(1)$ converges to 0 as $n \to \infty$.

Noting that $\frac{p+1}{2} \geq p$ and p can be arbitrarily close to 1, the proof is complete.

We provide the proof details in Appendix B.

Remark 2.1: Equation (17) suggests that if the number of channels scales logarithmically with the number of receivers, with high probability MC-MWC can stabilize all the rate vectors such that the average rate is less than $\bar{\gamma}$.

Remark 2.2: Comparing Theorem 2 with Theorem 1, we can see that to achieve a non-vanishing throughput, the number of channels required by MC-MWC achieves the algorithm independent lower bound in an order sense. Hence, MC-MWC achieves order-optimal throughput in the many-user many-channel asymptotic regime.

C. Throughput Gain over a Conventional Scheme

Finally, we compare MC-MWC with a conventional scheme, which can be considered as a straightforward extension of [18, 19] to the multi-channel, multi-session multicast setting. First, let us define capacity-achieving code.

Definition 3: (Capacity-achieving code) A coding scheme is said capacity-achieving if it can achieve the capacity of the broadcast erasure channel.

The conventional scheme is based on the following static channel allocation.

Static Channel Allocation: In the static channel allocation, the transmitter allocates the channels according to the channel statistics $\Gamma = \{\gamma_{i,j}\}_{i,j}$. Let $\vec{g} = (g_1, \ldots, g_m)$:

$\{1, 2, \ldots, m\} \mapsto \{1, 2, \ldots, m\}$ represent a one-to-one mapping from the sessions $\{h\}_{1 \leq h \leq m}$ to the channels $\{j\}_{1 \leq j \leq m}$. The static channel allocation is to find a $\vec{g} = (g_1, \ldots, g_m)$ such that the long-term sum throughput of all sessions is maximized. Note that given a static channel allocation $\vec{g}$, the maximum achievable throughput of any session h is bottlenecked by the receiver with the worst channel condition on the channel g_h, i.e., $\min_{i \in \mathcal{S}_h} \gamma_{i, g_h}$. Then, the optimal static channel allocation is formulated by

$$\max_{\vec{g} = (g_1, \ldots, g_m)} \left(\sum_{h=1}^{m} \min_{i \in \mathcal{S}_h} \gamma_{i, g_h} \right). \qquad (21)$$

Given the optimal channel allocation $\vec{g}^*$ to Equation (21), the transmitter encodes the packets from session h using a capacity-achieving coding scheme, such as rateless codes (e.g., [3, 4]) and some network coding schemes (e.g., [2, 7]), and sends the coded packets over the channel $\vec{g}_h^*$.

With the conventional scheme, given m, n and the channel statistics $\Gamma = \{\gamma_{i,j}\}_{i,j}$, there is a maximum achievable throughput for each session h, denoted as $\overline{\lambda}_h^{(m,n,\Gamma,\text{static})}$. We derive the following theorem regarding $\overline{\lambda}_h^{(m,n,\Gamma,\text{static})}$.

Theorem 3: If the number of channels scales slower than exponentially with the expected number of receivers, i.e., $m = o(v^n)$ for all $v > 1$, then the achievable throughput of the optimal static channel allocation with capacity-achieving codes vanishes as $n \to \infty$, that is,

$$\overline{\lambda}_h^{(m,n,\Gamma,\text{static})} \xrightarrow{P} 0, \quad \forall h \in \{1, \ldots, m\}. \qquad (22)$$

Proof: See Appendix C. ∎

Remark 3.1: Theorem 3 suggests that, to achieve a non-vanishing throughput, the number of channels must scale at least exponentially with the expected number of receivers, i.e., there exists some $v > 1$ such that $m = \Omega(v^n)$.

Remark 3.2: Comparing Theorem 3 with Theorem 2, it can be observed that to achieve a non-vanishing throughput, the multi-channel resource required by the optimal static channel allocation with capacity-achieving codes is doubly-exponentially larger than that required by MC-MWC.

VI. LOW DELAY OF MC-MWC

In this section, we show that MC-MWC achieves significant delay reduction by exploiting the multi-channel resources. We focus on the uniform traffic scenario when the arrival rates for all sessions are equal, i.e., $\lambda_h = \lambda, \forall h$.

Without loss of generality, we focus on analyzing the delay performance of an arbitrary session h. Let p_l^h be the lth packet in session h. The delay of the packet p_l^h with respect to a receiver $i \in \mathcal{S}_h$ is defined as the time between the arrival of the packet at the transmitter to the decoding of p_l^h and all the packets in session h with smaller indices, denoted as $D_{i,l}$.

Then, assuming the system is stationary and ergodic, for receiver $i \in \mathcal{S}_h$, the delay violation probability that the delay

of a packet exceeds a threshold k is given by

$$\mathbb{P}(D_i > k) = \lim_{L \to \infty} \frac{1}{L} \sum_{l=1}^{L} 1_{\{D_{i,l} > k\}}. \qquad (23)$$

We first analyze the delay performance of MC-MWC.

Let the time-slots t_i^d ($d = 1, 2, \ldots$) be the decoding moments of receiver i satisfying (7). Suppose that packet p_l^h arrives at time-slot t, which is between two successive decoding moments $t_i^d < t \leq t_i^{d+1}$. Then, packet p_l^h and all packets with smaller indices in session h will be decoded in time-slot t_i^{d+1}. The delay of packet p_l^h at the receiver i is

$$D_{i,l} = t_i^{d+1} - t. \qquad (24)$$

The following theorem characterizes the delay experienced at a receiver i, which is shown to be independent of the channel statistics of the other receivers and n.

Theorem 4: For a receiver $i \in \mathcal{S}_h$ such that $\sum_{j=1}^{m} \gamma_{i,j} > m\lambda$, the asymptotic decay rate of the delay violation probability of MC-MWC is

$$-\lim_{k \to \infty} \frac{1}{k} \log \mathbb{P}(D_i > k) = \Phi_i^m,$$

where

$$\Phi_i^m = \sup_{\theta \in \mathbb{R}} \left\{ -m \log \mathbb{E}\left(e^{-\theta a_h[t]}\right) - \sum_{j=1}^{m} \log\left(\gamma_{i,j} e^{\theta} + 1 - \gamma_{i,j}\right) \right\}. \qquad (25)$$

In addition, for any $\lambda < \bar{\gamma}$, with probability 1 we have

$$\lim_{m \to \infty} \frac{\Phi_i^m}{m} = \sup_{\theta \in \mathbb{R}} \left\{ -\log \mathbb{E}\left(e^{-\theta a_h[t]}\right) - \mathbb{E}\left[\log\left(\gamma_{i,j} e^{\theta} + 1 - \gamma_{i,j}\right)\right] \right\}. \qquad (26)$$

Proof sketch of Theorem 4: We focus on receiver 1. Define $I_d \triangleq t_1^{d+1} - t_1^d$ as the interval between the dth decoding moment and the $(d+1)^{\text{th}}$ decoding moment, which can be expressed as

$$I_d = \min\left\{ t \geq 1 : \sum_{\tau = t_1^d + 1}^{t_1^d + t} \left(\sum_{h=1}^{m} a_h[\tau] - \sum_{j=1}^{m} c_{1,j}[\tau] \right) \leq 0 \right\},$$

indicating that the decoding process is a renewal process. Noting that analyzing D_1 directly is very difficult, in the first step, we connect $\mathbb{P}(D_1 > k)$ with $\mathbb{P}(I_d > b)$.

Lemma 1: $\mathbb{P}(D_1 > k)$ is upper and lower bounded by

$$\frac{\mathbb{P}(I_d > k)}{m\lambda \mathbb{E}[I_d]} \leq \mathbb{P}(D_1 > k) \leq \frac{k\mathbb{P}(I_d > k) + \sum_{b=k}^{\infty} \mathbb{P}(I_d > b)}{\lambda \mathbb{E}[I_d]}.$$

Then, using large deviations theory, we derive the decay rate for I_d.

Lemma 2: The decay rate of the decoding interval in the steady state is given by

$$-\lim_{b \to \infty} \frac{1}{b} \log \mathbb{P}(I_d > b) = \Phi_1^m,$$

where Φ_1^m is the rate function defined in Equation (25).

By combining Lemma 1 and Lemma 2, the decay rate of D_1 is derived. $\blacksquare$

We provide the proof details in Appendix D.

Remark 4.1: Theorem 4 suggests that with MC-MWC, $\mathbb{P}(D_i > k) \approx e^{-\Phi_i^m k}$. From Equation (26), the delay improvement with MC-MWC is asymptotically linear in the number of channels m. Hence, MC-MWC could dramatically reduce delay by leveraging the multi-channel resources.

Next, we show a lower bound on the delay performance of a general class of conventional schemes defined as follows.

Definition 4: (Channel-allocation based schemes) A multicast scheme is said to be channel-allocation based, if different sessions are allocated to and served by different channels. The channel allocation can be either static over time (as formulated by Equation (21)), or dynamically adapted across time-slots based on the instantaneous channel state information of all the receivers and all time-slots. To reduce the technical difficulties, it is assumed that one session can be allocated with at most one channel in each time-slot, which makes sense considering that there are m channels and m sessions with $\lambda_h = \lambda, \forall h$. Given the channel allocation, the transmitter may serve the multicast sessions by employing any possible coding scheme, including but not limited to rateless codes and network codes, e.g., [2–4, 6–10].

Theorem 5: When $\mathbb{P}(a_h[t] > 1) > 0$, for any receiver $i \in \mathcal{S}_h$, the decay rate of delay in any channel-allocation based scheme is upper bounded by

$$-\lim_{k \to \infty} \frac{1}{k} \log \mathbb{P}(D_i > k) \leq \Phi^{\mathrm{const}}, \qquad (27)$$

where

$$\Phi^{\mathrm{const}} = \sup \left\{ \theta > 0 : \log \mathbb{E}\left(e^{\theta a_h[t]}\right) < \theta \right\}$$

is a constant independent of the number of channels m.

Proof: See Appendix E. $\blacksquare$

Remark 5.1: Theorem 5 suggests that the conventional channel-allocation based schemes (even incorporating any coding schemes), can achieve at most a constant delay improvement (not a function of the number of channels) by exploring the multi-channel resources.

Remark 5.2: The intuition behind Theorem 5 is as follows. For a queueing system, the delay of packets typically originates from two sources: 1) the stochastic and bursty arrival process; 2) the stochastic service process. Note that with a channel-allocation based scheme, a receiver can receive at most one packet from any given session in one time-slot. Even when the receiver could always receive one packet in each time-slot, there is a lower bound of the delay violation probability which originates from the stochastic and bursty arrival process. The lower bound is independent of the number of channels. However, owing to Merging, MC-MWC is not limited to the assumption that a receiver can receive at most one packet from any given session in one time-slot.
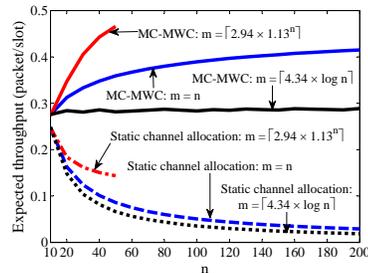

Fig. 3. Throughput behavior of MC-MWC vs. static channel allocation with capacity-achieving codes in a heterogeneous network.

VII. SIMULATIONS

In this section, we provide trace-driven and numerical results to investigate the throughput, delay and implementation complexity performance of MC-MWC.

A. Throughput Performance

To validate our analytical results, we consider a network with heterogeneous channel conditions, where $\gamma_{i,j}$ is uniformly distributed in the interval $[0, 1]$. Each of the m multicast sessions has n receivers. We consider the case which is *most unfavorable* in terms of throughput for Merging: for any two different sessions h and h', the intersection of the sets of receivers $\mathcal{S}_h$ and $\mathcal{S}_{h'}$ is empty.

We compare the achievable throughput of MC-MWC with that of the optimal static channel allocation with capacity-achieving codes. Figure 3 depicts the average achievable throughput of both schemes in 1000 randomly generated network scenarios, under different types of scaling of m with respect to n. It can be observed that the results match well with Theorem 1 and Theorem 3. When $m = \lceil 4.34 \times \log n \rceil$ and thus increases logarithmically with n, MC-MWC achieves a non-diminishing throughput while the throughput of the optimal static channel allocation vanishes as n increases. When $m = \lceil 2.94 \times 1.13^n \rceil$ and therefore scales exponentially with n, the throughput of MC-MWC grows and converges to $\bar{\gamma} = 0.5$, and the throughput of the optimal static channel allocation still decreases with n which verifies the prediction in Theorem 1 that, to achieve a non-diminishing throughput, with the optimal static channel allocation, m has to scale at least exponentially with n. When $m = n$ and thus scales linearly with n, the throughput of MC-MWC increases while the throughput of the optimal static channel allocation decreases with n.

To evaluate the throughput performance of MC-MWC in practice, we collect traces from experiments on a software defined radio platform. We use NI PXIe-1082 platform with NI-5791 RF front end for the experiments. The carrier frequency is set to be 2.5 GHz and the bandwidth used is 40 MHz. Experiments are performed in an indoor lab environment to get the CSI measurements. We note that the effects of multipath is obvious: The SNR difference across this 40 MHz bandwidth can be more than 20 dB. Hence, we collect the CSI traces at 100 different client locations and use these traces to emulate multicast receivers. Groups are formed randomly among these
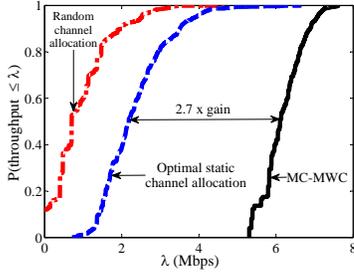
Fig. 4. Multicast throughput in trace-driven simulations with 5 sessions, in which each session has on average 20 receivers.
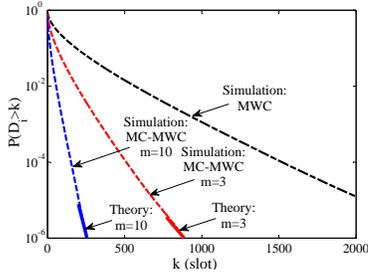

Fig. 6. Average delay vs. the traffic load ρ.

Fig. 5. Delay violation probability.

Fig. 7. Average decoding complexity on CPU/GPU vs. the number of sessions m.

receivers. Packet reception probability is calculated from the subband CSI traces using the method in [20].

Figure 4 shows the CDF of throughput under different schemes, where 5 groups are formed with on average 20 receivers in each group and 1000 random group formations are performed to get the figure. Here we consider the optimal static channel allocation, as well as a random channel allocation strategy, which assigns channels to sessions randomly. Both the optimal static channel allocation and random channel allocation schemes incorporate random linear network codes (RLNC) [2], which is one capacity-achieving code. We can see that MC-MWC shows a 2.7x gain over optimal static channel allocation scheme, which addresses the concern that: *in real world, even when the channel conditions are correlated across receivers and channels, and when the number of receivers/channels is not large, the proposed scheme could still have a significant throughput gain.*

B. Delay Performance

Recall that in the heterogeneous network, MC-MWC achieves a larger throughput than the optimal static channel allocation. To make the delay comparison fair, we consider a homogeneous network scenario, i.e., $\gamma_{i,j} = 0.6$ for all receivers and for all channels, in which the maximum achievable throughput of MC-MWC and the optimal static channel allocation are the same. There are 100 receivers in each session. In each session $h \in \{1, \ldots, m\}$, the packets arrive according to a Bernoulli process with rate $\lambda = 0.54$.

Figure 5 plots the delay violation probability of one receiver $i \in S_h$ with different schemes, i.e., MC-MWC and moving window codes (MWC) [6]. It can be observed that $\mathbb{P}(D_i > k)$ of MC-MWC decays exponentially with k and matches the predicted asymptotic decay rate in Equation (25). Furthermore,
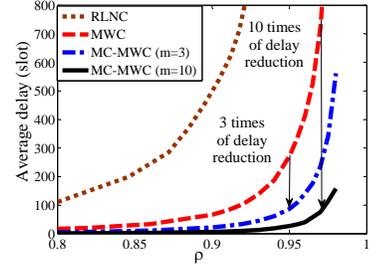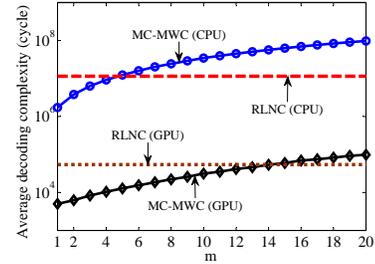
the decay rate is linear in the number of channels m, which shows that MC-MWC achieves a significant delay gain over MWC [6] by exploring the multi-channel resources.

Then, we consider the average delay performance of the same network scenario under different arrival rate λ. Figure 6 shows the average delay of different schemes with respect to the traffic load $\rho \triangleq \frac{\lambda}{\gamma_{i,j}}$. From Figure 6, we have the following observations. First, the average delay of MC-MWC is much lower than that of RLNC [2], which is one of the rateless codes. It is worthy to note that the average delay of other rateless codes, such as LT codes [3] and Raptor codes [4], is close to that of RLNC. Second, compared with MWC [6], MC-MWC achieves a delay reduction which is roughly linear in the number of channels m. Third, the delay gain holds for any load $\rho \in (0, 1)$.

C. Low Implementation Complexity

With MC-MWC, a receiver needs to decode the sessions it may not be interested in. Hence, it is important to understand the decoding complexity of MC-MWC. To this end, we emulate the decoding procedure of MC-MWC on commercial CPU and GPU, and show that the decoding complexity of MC-MWC is affordable in practice for moderate m.

In the emulation, the finite field size in MC-MWC is set to be 256, and we use a simple table loop-up approach [21] to do the operations on the finite field. Each packet has 1024 bytes. We emulate the decoding process at a receiver with $\gamma_{i,j} = 0.6$ for all channels, where packets arrive according to a Bernoulli process with rate $\lambda = 0.48$ such that $\rho = 0.8$. We implement two versions of the decoding algorithm, the serial version on a single core CPU (Intel Core i7-2600 CPU clocked at 3.4 GHz), and the parallel version on a GPU (NVIDIA GeForce GTX 860M). Note that the GPU provides 640 shader units clocked at 1029 MHz, and we use 16 blocks and 64 threads per blocks

in CUDA. For comparison, we also emulate RLNC [2] with a block length of 50 packets, which does not incorporate with Merging.

In Figure 7, we plot the average decoding complexity (in terms of cycles) of different schemes to decode one packet in a session. From Figure 7, we have the following observations. First, as m grows, the decoding complexity of MC-MWC increases on both CPU and GPU, while the decoding complexity of RLNC remains unchanged. This is consistent with our expectation that the decoding complexity with Merging increases with m. Second, for both MC-MWC and RLNC, GPU leverages its ability to decode packets in parallel, and thus dramatically reduce the decoding complexity. Third, despite that MC-MWC requires the receiver to decode m sessions, the decoding complexity of MC-MWC is smaller than that of RLNC when m is small. This is because the moving window coding strategy in general results in a sparser decoding matrix compared with the dense decoding matrix in RLNC, as shown in [6, 22]. Finally, MC-MWC can support 10 sessions, the throughput of each is more than $100Mbps$ on our GPU. Therefore, the decoding complexity of MC-MWC is low in practice for moderate m, in which case MC-MWC already exhibits significant throughput and delay gains as shown previously.

VIII. Conclusion

In this paper, we develop a Multi-Channel Moving Window Codes (MC-MWC) and prove that it achieves high throughput, low delay, and requires very limited feedback. We verify our theoretical results using trace-driven simulations, and show that the complexity of implementing MC-MWC is low in practice. This new approach, which exploits multi-channel capability, moving window coding, and anonymous feedback, has the potential to finally realize in practice the significant promise of wireless multicast.

References

[1] "Cisco visual networking index: Forecast and methodology, 2015-2020." http://www.cisco.com.

[2] T. Ho, M. Médard, R. Koetter, D. R. Karger, M. Effros, J. Shi, and B. Leong, "A random linear network coding approach to multicast," *IEEE Trans. Inf. Theory*, vol. 52, pp. 4413–4430, Oct. 2006.

[3] M. Luby, "LT codes," in *IEEE FOCS 2002*, pp. 271–280, 2002.

[4] A. Shokrollahi, "Raptor codes," *IEEE Trans. Inf. Theory*, vol. 52, pp. 2551–2567, Jun. 2006.

[5] Y. Yang and N. Shroff, "Throughput of rateless codes over broadcast erasure channels," *IEEE/ACM Trans. Netw.*, vol. 23, no. 1, pp. 126–137, 2015.

[6] F. Wu, Y. Sun, Y. Yang, K. Srinivasan, and N. Shroff, "Constant-delay and constant-feedback moving window network coding for wireless multicast: Design and asymptotic analysis," *IEEE J. Select. Areas Commun.*, vol. 33, pp. 127–140, Feb 2015.

[7] J. K. Sundararajan, D. Shah, and M. Médard, "ARQ for network coding," in *IEEE ISIT 2008*, pp. 1651–1655, Jul. 2008.

[8] J. K. Sundararajan, D. Shah, and M. Médard, "Feedback-based online network coding," *CoRR*, vol. abs/0904.1730, 2009.

[9] S. Parastoo, S. Ramtin, and T. Danail, "An optimal adaptive network coding scheme for minimizing decoding delay in broadcast erasure channels," *EURASIP J. Wirel. Commun. Netw.*, vol. 2010, Jan. 2010.

[10] A. Fu, P. Sadeghi, and M. Medard, "Dynamic rate adaptation for improved throughput and delay in wireless network coded broadcast," *IEEE/ACM Trans. Netw.*, in press.

[11] H. Holma and A. Toskala, *LTE for UMTS-OFDMA and SC-FDMA based radio access*. John Wiley & Sons, 2009.

[12] F. Wu, Y. Yang, O. Zhang, K. Srinivasan, and N. B. Shroff, "Anonymous-query based rate control for wireless multicast: Approaching optimality with constant feedback," *ACM MobiHoc*, pp. 191–200, 2016.

[13] S. Bodas, S. Shakkottai, L. Ying, and R. Srikant, "Low-complexity scheduling algorithms for multichannel downlink wireless networks," *IEEE/ACM Trans. Netw.*, vol. 20, no. 5, pp. 1608–1621, 2012.

[14] M. Ouyang and L. Ying, "Approaching throughput optimality with limited feedback in multichannel wireless downlink networks," *IEEE/ACM Trans. Netw.*, vol. 21, pp. 1827–1838, Dec. 2013.

[15] B. Ji, G. R. Gupta, X. Lin, and N. B. Shroff, "Low-complexity scheduling policies for achieving throughput and asymptotic delay optimality in multichannel wireless networks," *IEEE/ACM Trans. Netw.*, vol. 22, no. 6, pp. 1911–1924, 2014.

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

[17] A. Sinha, L. Tassiulas, and E. Modiano, "Throughput-optimal broadcast in wireless networks with dynamic topology," ACM MobiHoc '16, pp. 21–30, 2016.

[18] D. Koutsonikolas, Y. C. Hu, and C.-C. Wang, "Pacifier: High-throughput, reliable multicast without "crying babies" in wireless mesh networks," *IEEE/ACM Trans. Netw.*, vol. 20, pp. 1375–1388, Oct. 2012.

[19] K. C.-J. Lin and D.-N. Yang, "Multicast with intraflow network coding in multirate multichannel wireless mesh networks," *IEEE Trans.on Veh. Technol.*, vol. 62, no. 8, pp. 3913–3927, 2013.

[20] W. Zhou, T. Das, L. Chen, K. Srinivasan, and P. Sinha, "Basic: backbone-assisted successive interference cancellation," *ACM MobiCom*, pp. 149–161, 2016.

[21] M. A. Hasan, "Look-up table-based large finite field multiplication in memory constrained cryptosystems," *IEEE Trans. on Comput.*, vol. 49, pp. 749–758, Jul 2000.

[22] A. Tassi, I. Chatzigeorgiou, and D. E. Lucani, "Analysis and optimization of sparse random linear network coding for reliable multicast services," *IEEE Trans. on Commun.*, vol. 64, no. 1, pp. 285–299, 2016.

[23] S. I. Resnick, *A probability path*. Springer Science & Business Media, 2013.

[24] R. Srikant and L. Ying, *Communication networks: an optimization, control, and stochastic networks perspective*. Cambridge University Press, 2013.

[25] E. Çinlar, "Markov renewal theory: A survey," *INFORMS Management Science*, vol. 21, no. 7, pp. 727–752, 1975.

[26] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*. Springer-Verlag New York, Inc., 2 ed., 2010.

[27] A. Weiss, *Large Deviations for Performance Analysis: Queues, Communications, and Computing*. CRC Press, 1995.

Appendix A
Proof of Theorem 1

Without loss of generality, let us focus on analyzing the achievable throughput of session 1.

We want to show that if the number of channels scales a little bit slower than logarithmically with the expected number of receivers, i.e., $m = O\left((\log n)^{\delta}\right)$ for some $\delta \in (0, 1)$, the achievable throughput of session 1 diminishes to 0 for any possible scheme as $n \to \infty$. It is sufficient for us to show that for any $0 < \lambda \leq 1$ and any $0 < p \leq 1$, the probability that session 1 could achieve throughput λ is less than p as $n \to \infty$.

Define the set of "bottleneck" receivers for session 1 as

$$\mathcal{B}_{\lambda}^{(m,n,\Gamma)} = \left\{ i \in \mathcal{S}_1 \middle| \gamma_{i,j} < \frac{\lambda}{m}, \quad \forall 1 \leq j \leq m \right\}. \quad (28)$$

Notice that if $\mathcal{B}_{\lambda}^{(m,n,\Gamma)} \neq \emptyset$, for any receiver $i \in \mathcal{B}_{\lambda}^{(m,n,\Gamma)}$, with any possible scheme, we could upper bound the achievable throughput of receiver i by allocating all m channels to serve receiver i at all time-slots, in which case its throughput is upper bounded by the sum of the capacity of all m channels,

$$\sum_{j=1}^{m} \gamma_{i,j} < \sum_{j=1}^{m} \frac{\lambda}{m} = \lambda. \quad (29)$$

Recall that the throughput of multicast is bottlenecked by its worst receiver. Thus, when $\mathcal{B}_{\lambda}^{(m,n,\Gamma)} \neq \emptyset$, the achievable throughput of session 1 is also upper bounded by Equation (29).

By the definition in Equation (28), the probability that a receiver $i \in \mathcal{S}_1$ is not in the "bottleneck" set can be given by

$$\mathbb{P}\left(i \notin \mathcal{B}_\lambda^{(m,n,\Gamma)}\right) = 1 - \left(F_\gamma\left(\frac{\lambda}{m}\right)\right)^m, \qquad (30)$$

where $F_\gamma(\cdot)$ is the CDF of $\gamma_{i,j}$.

Subsequently, based on Equation 30, we will show that the probability that $\mathcal{B}_\lambda^{(m,n,\Gamma)} \neq \emptyset$ is high when $m = O\left((\log n)^\delta\right)$ for any $\delta \in (0,1)$.

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{B}_\lambda^{(m,n,\Gamma)} \neq \emptyset\right) \\
=&\mathbb{P}\left(\mathcal{B}_\lambda^{(m,n,\Gamma)} \neq \emptyset \big| s_1 \leq \alpha n\right) \mathbb{P}\left(s_1 \leq \alpha n\right) + \\
&\mathbb{P}\left(\mathcal{B}_\lambda^{(m,n,\Gamma)} \neq \emptyset \big| s_1 > \alpha n\right) \mathbb{P}\left(s_1 > \alpha n\right) \\
\overset{(a)}{\geq}&\left(1 - z_s(\alpha)\right)\left(1 - \mathbb{P}\left(\mathcal{B}_\lambda^{(m,n,\Gamma)} = \emptyset \big| s_1 > \alpha n\right)\right) \\
\overset{(b)}{=}&\left(1 - z_s(\alpha)\right)\left(1 - \left(1 - \left(F_\gamma\left(\frac{\lambda}{m}\right)\right)^m\right)^{s_1}\right) \\
\overset{(c)}{\geq}&\left(1 - z_s(\alpha)\right)\left(1 - \left(1 - \left(F_\gamma\left(\frac{\lambda}{m}\right)\right)^m\right)^{\lfloor \alpha n \rfloor}\right) \\
\overset{(d)}{\geq}&\left(1 - z_s(\alpha)\right)\left(1 - \left(1 - \frac{\lfloor \alpha n \rfloor \left(F_\gamma\left(\frac{\lambda}{m}\right)\right)^m}{1 + (\lfloor \alpha n \rfloor - 1)\left(F_\gamma\left(\frac{\lambda}{m}\right)\right)^m}\right)\right),
\end{aligned}$$
$$(31)$$

where step (a) is based on Equations (9) and (10) and α is picked such that $z_s(\alpha) < p$, in step (b) Equation (30) is applied, step (c) utilizes the condition that $s_1 > \alpha n$, and in step (d), we apply the inequality that $(1+x)^r \leq 1 + \frac{rx}{1-(r-1)x}$ for $x \in [-1,0]$ and $r > 1$.

By the assumption that $\lim_{y \to 0^+} \frac{F_\gamma(y)}{y} \geq \kappa$, when m is large enough, we have $\left(F_\gamma\left(\frac{\lambda}{m}\right)\right)^m \geq \left(\frac{\kappa\lambda}{2m}\right)^m$. Then it is easy to verify that with $m = O\left((\log n)^\delta\right)$ for any given $\delta \in (0,1)$, we have

$$\begin{aligned}
&\lim_{n \to \infty} \lfloor \alpha n \rfloor \left(F_\gamma\left(\frac{\lambda}{m}\right)\right)^m \\
&\geq \lim_{n \to \infty} \lfloor \alpha n \rfloor \left(\frac{\kappa\lambda}{2m}\right)^m = \infty.
\end{aligned} \qquad (32)$$

Inserting Equation (32) into Equation (31), we have

$$\lim_{n \to \infty} \mathbb{P}\left(\mathcal{B}_\lambda^{(m,n,\Gamma)} \neq \emptyset\right) \geq (1 - z_s(\alpha)) \overset{(a)}{\geq} 1 - p, \qquad (33)$$

where step (a) is because α is picked such that $z_s(\alpha) < p$.

Recall that we have shown that when $\mathcal{B}_\lambda^{(m,n,\Gamma)} \neq \emptyset$, the achievable throughput of session 1 is less than λ, according to Equation (33), it is impossible for session 1 to achieve throughput λ with probability p. Since λ and p could be chosen arbitrarily close to 0, by the Cauchy criterion (see Theorem 6.3.1 in [23]), Equation (13) holds, which completes the proof.

APPENDIX B
PROOF OF THEOREM 2

Let us prove that, with MC-MWC, for any $p \in (0,1)$, to guarantee that $\mathcal{P}_\lambda^{(m)} \subseteq \Lambda_{\text{MC-MWC}}^{(m,n,\Gamma)}$ with probability no less than p, it is sufficient to let the number of channels scale logarithmically with the expected number of receivers, i.e., $m = \left\lceil \frac{1}{(\bar{\gamma}-\lambda)^2} \log n \right\rceil$, as $n \to \infty$.

Recall that with MC-MWC, the maximum sum throughput of all sessions is given by Equation (15). On the other hand, according to Equation (6), $\{Q_i[t]\}_t$ is a random walk on $[0,\infty)$, and has a steady state distribution if Equation (15) holds, which suggests that Equation (15) is also achievable by MC-MWC. Hence, we have

$$\mathbb{P}\left(\mathcal{P}_\lambda^{(m)} \subseteq \Lambda_{\text{MC-MWC}}^{(m,n,\Gamma)}\right) = \mathbb{P}\left(\frac{1}{m} \min_{i \in \cup_{h=1}^m \mathcal{S}_h} \sum_{j=1}^m \gamma_{i,j} \geq \lambda\right). \qquad (34)$$

Notice that for any $\tau \geq 1$, we have

$$\begin{aligned}
\mathbb{P}\left(\left|\bigcup_{h=1}^m \mathcal{S}_h\right| \geq \tau mn\right) &\overset{(a)}{\leq} \mathbb{P}\left(\sum_{h=1}^m s_h \geq \tau mn\right) \\
&\overset{(b)}{\leq} \frac{\sum_{h=1}^m \mathbb{E}[s_h]}{\tau mn} = \frac{1}{\tau}, \qquad (35)
\end{aligned}$$

where step (a) uses the fact that $\left|\bigcup_{h=1}^m \mathcal{S}_h\right| \leq \sum_{h=1}^m s_h$ due to possible intersections among the sets $\{\mathcal{S}_h\}_{1 \leq h \leq m}$, and step (b) applies Markov's inequality to the *i.i.d.* random variables $\{s_h\}_{1 \leq h \leq m}$.

With Equation (35), we can derive a lower bound of Equation (34).

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{m} \min_{i \in \cup_{h=1}^m \mathcal{S}_h} \sum_{j=1}^m \gamma_{i,j} \geq \lambda\right) \\
\geq&\mathbb{P}\left(\frac{1}{m} \min_{i \in \cup_{h=1}^m \mathcal{S}_h} \sum_{j=1}^m \gamma_{i,j} \geq \lambda \bigg| \left|\bigcup_{h=1}^m \mathcal{S}_h\right| < \tau mn\right) \times \\
&\mathbb{P}\left(\left|\bigcup_{h=1}^m \mathcal{S}_h\right| < \tau mn\right) \\
\overset{(a)}{\geq}&\left(1 - \frac{1}{\tau}\right) \mathbb{P}\left(\min_{i \in \cup_{h=1}^m \mathcal{S}_h} \sum_{j=1}^m \gamma_{i,j} \geq m\lambda \bigg| \left|\bigcup_{h=1}^m \mathcal{S}_h\right| < \tau mn\right) \\
\geq&\left(1 - \frac{1}{\tau}\right) \mathbb{P}\left[\sum_{j=1}^m \gamma_{i,j} \geq m\lambda\right]^{\lceil \tau mn \rceil} \\
\overset{(b)}{\geq}&\left(1 - \frac{1}{\tau}\right)\left(1 - e^{-2(\bar{\gamma}-\lambda)^2 m}\right)^{\lceil \tau mn \rceil} \\
\overset{(c)}{\geq}&\left(1 - \frac{1}{\tau}\right)\left(1 - \lceil \tau mn \rceil e^{-2(\bar{\gamma}-\lambda)^2 m}\right), \qquad (36)
\end{aligned}$$

where in step (a), Equation (35) is applied, in step (b), Hoeffding's inequality is applied on the *i.i.d.* random variables $\{\gamma_{i,j}\}_{j=1,\dots,m}$ which are bounded by the interval $[0,1]$, and

step (c) utilizes the inequality that $(1 + x)^r \geq 1 + xr$ for $x \geq -1$ and $r > 1$.

Let $\tau = \frac{2}{1-p}$ and $m = \left\lceil \frac{1}{(\bar{\gamma}-\lambda)^2} \log n \right\rceil$. Combining with Equation (36), we have

$$\mathbb{P}\left(\frac{1}{m} \min_{i \in \bigcup_{h=1}^m \mathcal{S}_h} \sum_{j=1}^m \gamma_{i,j} \geq \lambda \right) \geq \frac{p+1}{2}(1 - o(1)), \quad (37)$$

where $o(1)$ converges to 0 as $n \to \infty$.

Noting that $\frac{p+1}{2} \geq p$ and p can be arbitrarily close to 1, by Equation (37) and Equation (34), Equation (17) holds. The proof is complete.

APPENDIX C
PROOF OF THEOREM 3

Without loss of generality, we focus on the analysis of the throughput of session 1.

Now let us prove that if the number of channels scales slower than exponentially with the expected number of receivers, i.e., $m = o(v^n)$ for any $v > 1$, then the achievable throughput of session 1 with the optimal static channel allocation vanishes as $n \to \infty$. Note that it is sufficient for us to prove that with the optimal static channel allocation, to achieve any throughput $\lambda \in (0, \bar{\gamma})$ with any probability $p > 0$ as $n \to \infty$, the number of channels must scale at least exponentially with the expected number of receivers, i.e., there exists some $v > 1$ such that $m = \Omega(v^n)$.

Recall that given $\{\gamma_{i,j}\}_{i \in \mathcal{S}_h, j=1,\dots,m}$, the optimal static channel allocation maximizes Equation (21). By Equation (21), we can upper bound the throughput of session 1 by allowing the session to choose any of the m channels, i.e.,

$$\overline{\lambda}_1^{(m,n,\Gamma,\text{static})} \leq \max_{1 \leq j \leq m} \min_{i \in \mathcal{S}_1} \gamma_{i,j}. \quad (38)$$

Then, the probability that session 1 could achieve throughput λ is upper bounded by

$$\mathbb{P}\left(\overline{\lambda}_1^{(m,n,\Gamma,\text{static})} \geq \lambda \right)$$

$$\leq \mathbb{P}\left(\max_{1 \leq j \leq m} \min_{i \in \mathcal{S}_1} \gamma_{i,j} \geq \lambda \right)$$

$$= \mathbb{P}\left(\max_{1 \leq j \leq m} \min_{i \in \mathcal{S}_1} \gamma_{i,j} \geq \lambda \Big| s_1 \leq \alpha n \right) \mathbb{P}(s_1 \leq \alpha n) +$$

$$\mathbb{P}\left(\max_{1 \leq j \leq m} \min_{i \in \mathcal{S}_1} \gamma_{i,j} \geq \lambda \Big| s_1 > \alpha n \right) \mathbb{P}(s_1 > \alpha n)$$

$$\leq \mathbb{P}(s_1 \leq \alpha n) + \mathbb{P}\left(\max_{1 \leq j \leq m} \min_{i \in \mathcal{S}_1} \gamma_{i,j} \geq \lambda \Big| s_1 > \alpha n \right)$$

$$\overset{(a)}{\leq} z_s(\alpha) + \mathbb{P}\left(\max_{1 \leq j \leq m} \min_{i \in \mathcal{S}_1} \gamma_{i,j} \geq \lambda \Big| s_1 > \alpha n \right), \quad (39)$$

where step (a) uses Equation (9).

Since by our assumption $\{\gamma_{i,j}\}$ are *i.i.d.* across different receivers and different channels, then $\{\min_{i \in \mathcal{S}_1} \gamma_{i,j}\}_{1 \leq j \leq m}$

are also *i.i.d* when s_1 is given. Thus, we have

$$\mathbb{P}\left(\max_{1 \leq j \leq m} \min_{i \in \mathcal{S}_1} \gamma_{i,j} \geq \lambda \Big| s_1 \right) = 1 - \mathbb{P}\left(\min_{i \in \mathcal{S}_1} \gamma_{i,j} < \lambda \Big| s_1 \right)^m$$

$$= 1 - \left(1 - \mathbb{P}\left(\min_{i \in \mathcal{S}_1} \gamma_{i,j} \geq \lambda \Big| s_1 \right) \right)^m$$

$$\overset{(a)}{\leq} m \mathbb{P}(\gamma_{i,j} \geq \lambda)^{s_1}, \quad (40)$$

where step (a) utilizes the inequality that $(1 + x)^r \geq 1 + xr$ for $x \geq -1$ and $r > 1$.

Noticing that the right hand side of Equation (40) decreases as s_1 increases, we could further have

$$\mathbb{P}\left(\max_{1 \leq j \leq m} \min_{i \in \mathcal{S}_1} \gamma_{i,j} \geq \lambda \Big| s_1 > \alpha n \right)$$

$$= \frac{\mathbb{P}\left(\max_{1 \leq j \leq m} \min_{i \in \mathcal{S}_1} \gamma_{i,j} \geq \lambda, s_1 > \alpha n \right)}{\mathbb{P}(s_1 > \alpha n)}$$

$$= \frac{\sum_{s_1 = \lceil \alpha n \rceil}^{\infty} \left\{ \mathbb{P}\left(\max_{1 \leq j \leq m} \min_{i \in \mathcal{S}_1} \gamma_{i,j} \geq \lambda \Big| s_1 \right) \mathbb{P}(s_1) \right\}}{\sum_{s_1 = \lceil \alpha n \rceil}^{\infty} \mathbb{P}(s_1)}$$

$$\leq \max_{s_1 \geq \lceil \alpha n \rceil} \mathbb{P}\left(\max_{1 \leq j \leq m} \min_{i \in \mathcal{S}_1} \gamma_{i,j} \geq \lambda \Big| s_1 \right)$$

$$\overset{(a)}{\leq} m \mathbb{P}(\gamma_{i,j} \geq \lambda)^{\lfloor \alpha n \rfloor}, \quad (41)$$

where step (a) applies Equation (40) for $s_1 \geq \lfloor \alpha n \rfloor$.

Inserting Equation (41) into Equation (39), we have

$$\mathbb{P}\left(\overline{\lambda}_1^{(m,n,\Gamma,\text{static})} \geq \lambda \right) \leq z_s(\alpha) + m \mathbb{P}(\gamma_{i,j} \geq \lambda)^{\lfloor \alpha n \rfloor}. \quad (42)$$

By assumption in Equation (10), we can pick α such that $z_s(\alpha) < p$. Then, to achieve throughput with probability p, we must have

$$\mathbb{P}\left(\max_{1 \leq j \leq m} \min_{i \in \mathcal{S}_1} \gamma_{i,j} \geq \lambda \right) \geq p. \quad (43)$$

Combining Equation (43) with Equation (42), we have

$$m \geq (p - z_s(\alpha)) \left(\frac{1}{\mathbb{P}(\gamma_{i,j} \geq \lambda)} \right)^{\lfloor \alpha n \rfloor}. \quad (44)$$

By the assumption of heterogeneous network that $F_\gamma(\lambda) > 0$ for any $\lambda > 0$, we have $\mathbb{P}(\gamma_{i,j} \geq \lambda) < 1$ for any $\lambda > 0$. Therefore, with static channel allocation, m has to scale at least exponentially with n.

Hence, if m scales slower than exponentially with n, Equation (43) does not hold as $n \to \infty$. Since λ and p can be arbitrarily close to 0, by the Cauchy criterion (see Theorem 6.3.1 in [23]), Equation (22) holds. Thus, the achievable throughput of session 1 vanishes. The proof is complete.

APPENDIX D
PROOF OF THEOREM 4

In this subsection, we analyze the delay violation probability that the delay experienced by a receiver exceeds a given threshold for MC-MWC with general *i.i.d.* packet arrivals. Without loss of generality, we focus on the analysis of the decoding delay of receiver 1.

Notice that the arrival processes of different sessions are symmetric, i.e., $a_h[t]$ is *i.i.d.* across time-slots and sessions, and Merging combines all sessions into one large session and does not differentiate among sessions. The delay violation probabilities of all sessions are the same. Hence, it is sufficient for us to analyze the delay violation probability of the combined session after Merging.

Let us define $I_d \triangleq t_1^{d+1} - t_1^d$. Since $\{t_1^d\}_d$ is set of the decoding moments of receiver 1 that satisfies Equation (7), we know that I_d represents the interval between the dth decoding moment and the $(d+1)^{\text{th}}$ decoding moment. This fact, combined with the evolution of $Q_i[t]$ given by Equation (6), we can expressed I_d as

$$I_d = \min\left\{t \geq 1 : \sum_{\tau=t_1^d+1}^{t_1^d+t} (a[\tau] - c_1[\tau]) \leq 0\right\}$$

$$= \min\left\{t \geq 1 : \sum_{\tau=t_1^d+1}^{t_1^d+t} \left(\sum_{h=1}^m a_h[\tau] - \sum_{j=1}^m c_{1,j}[\tau]\right) \leq 0\right\}.$$

$$(45)$$

The above equation indicates that the decoding process is a renewal process. Let K_d denote the number of packets from all m sessions that arrive at the encoder queue between time-slot t_1^d and time-slot t_1^{d+1}, then it can be expressed as

$$K_d = \sum_{t=t_1^d+1}^{t_1^d+I_d} a[t] = \sum_{t=t_1^d+1}^{t_1^d+I_d} \sum_{h=1}^m a_h[t]. \qquad (46)$$

Notice from Equations (45) and (46) that $\{I_d\}_d$ are *i.i.d.* and $\{K_d\}_d$ are *i.i.d.*, we use $\widehat{I}$ and $\widehat{K}$ to denote random variables which have the same distribution as I_1 and K_1 respectively.

Lemma 3: $\mathbb{P}(D_1 > k)$ is upper and lower bounded by

$$\frac{\mathbb{P}\left(\widehat{I} > k\right)}{m\lambda \mathbb{E}\left[\widehat{I}\right]} \leq \mathbb{P}(D_1 > k) \leq \frac{k\mathbb{P}\left(\widehat{I} > k\right) + \sum_{b=k}^{\infty} \mathbb{P}\left(\widehat{I} > b\right)}{\lambda \mathbb{E}\left[\widehat{I}\right]}.$$

$$(47)$$

Remark 5.3: The proof of Lemma 3 is based on a simple observation. For a given delay threshold $k > 0$, the number of packets decoded after an interval I_d must satisfy the following conditions. 1) If $I_d \leq k$, there is no packets exceeding the threshold k. 2) If $I_d > k$, there are at most mI_d packets which exceed the threshold k. 3) If $I_d > k$, there is at least one packet which exceed the threshold k.

Proof: See Appendix F. ∎

Lemma 3 shows the connection between $\mathbb{P}(D_1 > k)$ and $\mathbb{P}\left(\widehat{I} > b\right)$. Hence, subsequently we study the probability that the decoding interval in the steady state exceeds a certain threshold, i.e., $\mathbb{P}\left(\widehat{I} > b\right), b \in \mathbb{N}$.

Lemma 4: The decay rate of the decoding interval in the steady state is given by

$$-\lim_{b\to\infty} \frac{1}{b} \log \mathbb{P}\left(\widehat{I} > b\right) = \Phi_1^m, \qquad (48)$$

where Φ_1^m is the rate function defined in Equation (25).

Proof: See Appendix G. ∎

Let us pick $\epsilon \in (0, \Phi_1^m)$. By the definition of decay rate, we can find $B_\epsilon \in \mathbb{N}$, such that $\forall b \in \mathbb{N}, b \geq B_\epsilon$, we have

$$e^{-b(\Phi_1^m+\epsilon)} < \mathbb{P}\left(\widehat{I} > b\right) < e^{-b(\Phi_1^m-\epsilon)}. \qquad (49)$$

Combining Equations (47) and (49) yields, for k large enough,

$$\mathbb{P}(D_1 > k) \leq \frac{e^{-k(\Phi_1^m-\epsilon)}}{\lambda \mathbb{E}\left[\widehat{I}\right]} \left(k + \frac{1}{1 - e^{-(\Phi_1^m-\epsilon)}}\right),$$

$$\mathbb{P}(D_1 > k) \geq \frac{e^{-k(\Phi_1^m+\epsilon)}}{m\lambda \mathbb{E}\left[\widehat{I}\right]}. \qquad (50)$$

On account of $\lim_{k\to\infty} \frac{\log k}{k} = 0$, Equation (50) leads to

$$\Phi_1^m - \epsilon \leq -\lim_{k\to\infty} \frac{1}{k} \log \mathbb{P}(D_1 > k) \leq \Phi_1^m + \epsilon.$$

Since ϵ can be arbitrarily close to 0, the decay rate of delay given in Equation (25) is proved.

Lastly, we notice that $\{\gamma_{1,j}\}_{1\leq j\leq m}$ are *i.i.d.* across channels. Hence, by the strong law of large numbers, for any $\lambda < \bar{\gamma}$, with probability 1 we have $\lim_{m\to\infty} \frac{1}{m}\sum_{j=1}^m \gamma_{i,j} = \bar{\gamma} > \lambda$. Thus, Equation (25) is applicable as we take the limit $m \to \infty$. By applying strong law of large numbers to Equation (25), we have Equation (26), which shows that the decay rate of delay increases linearly with the number of channels. The proof is complete.

APPENDIX E
PROOF OF THEOREM 5

Notice that if a session has only one receiver, then the transmitter can dedicatedly minimize the delay of the single receiver. Thus, to derive the lower bound of the delay performance, we assume that there is only one receiver in session h, say $1 \in \mathcal{S}_h$. With either the static channel allocation or the genie-aided dynamic channel allocation, the receiver $1 \in \mathcal{S}_h$ could successfully receive at most one packet in one time-slot. To lower bound the delay performance, we further assume that the receiver successfully receives one packet in each time-slot.

By the definition of delay, a packet is considered as "decoded" only if all packets of smaller indices have been "decoded". Thus, packets which arrive earlier must also be served earlier. Since there is only one receiver in the session, coding across packets at the transmitter is not necessary and may only increase the delay of the receiver. This suggests that the delay of the single receiver is minimized when the transmitter sends the packets of session h in a First-in First-out (FIFO) manner. Hence, we derive a lower bound of the delay by considering the corresponding FIFO queue $\widetilde{Q}_1$, in which $a_h[t]$ packets arrive and one packet gets served at each time-slot.

According to Theorem 10.4.1 in [24], the decay rate of the

queue length for $\widetilde{Q}_1$ is given by Equation (27).

$$-\lim_{k\to\infty}\frac{1}{k}\log\mathbb{P}\left(\widetilde{Q}_1[\infty]>k\right)$$
$$=\sup\left\{\theta>0:\log\mathbb{E}\left(e^{\theta a_h[t]}\right)<\theta\right\}. \tag{51}$$

Now we derive a connection between the decay rate of the queue length and the decay rate of the delay for the FIFO queue $\widetilde{Q}_1$. The key step is to construct a mapping: each time-slot τ at which $\widetilde{Q}[\tau]>k$ is mapped to the $(k+1)^{\text{th}}$ packet in $\widetilde{Q}[\tau]$. Since exactly one packet could be served in $\widetilde{Q}$, different time slots are mapped to different packets. Notice there are k packets ahead of each mapped packet, the delay of each mapped packet is at least k. Based on this fact, we have

$$\mathbb{P}\left(\widetilde{Q}_1[\infty]>k\right)=\lim_{t\to\infty}\frac{1}{t}\sum_{\tau=1}^{t}1_{\{\widetilde{Q}_1[\tau]>k\}}$$
$$\overset{(a)}{\leq}\lim_{t\to\infty}\frac{\sum_{1\leq l\leq\sum_{\tau=1}^{t}a_h[\tau]}1_{\{D_{1,l}\geq k\}}}{t}$$
$$=\lim_{t\to\infty}\frac{\sum_{\tau=1}^{t}a_h[\tau]}{t}\times\frac{\sum_{1\leq l\leq\sum_{\tau=1}^{t}a_h[\tau]}1_{\{D_{1,l}\geq k\}}}{\sum_{\tau=1}^{t}a_h[\tau]}$$
$$\overset{(b)}{=}\lim_{t\to\infty}\lambda\cdot\frac{\sum_{1\leq l\leq\sum_{\tau=1}^{t}a_h[\tau]}1_{\{D_{1,l}\geq k\}}}{\sum_{\tau=1}^{t}a_h[\tau]}, \tag{52}$$

where step (a) is due to the property of the mapping we construct, i.e., each time-slot τ at which $\widetilde{Q}[\tau]>k$ is mapped to a unique packet the delay of which is at least k, and in step (b), strong law of large numbers is applied on the *i.i.d.* random variables $\{a_h[\tau]\}_{\forall\tau}$.

Noticing that $\sum_{1\leq l\leq\sum_{\tau=1}^{t}a_h[\tau]}1_{\{D_{1,l}\geq k\}}$ is the number of packets the delay of which is at least k among the the packets in session h arrived up to time-slot t. Thus, by the definition of delay violation probability (given by Equation (23)), we have

$$\lim_{t\to\infty}\frac{\sum_{1\leq l\leq\sum_{\tau=1}^{t}a_h[\tau]}1_{\{D_{1,l}\geq k\}}}{\sum_{\tau=1}^{t}a_h[\tau]}=\mathbb{P}\left(D_1\geq k\right), \tag{53}$$

which, together with Equation (52), lead to

$$\mathbb{P}\left(\widetilde{Q}_1[\infty]>k\right)\leq\lambda\mathbb{P}\left(D_1\geq k\right). \tag{54}$$

Combining Equations (51) and (54), we can see that for the queue $\widetilde{Q}_1$, the decay rate of the delay is no greater than the decay rate of the queue length. Recall that the delay of $\widetilde{Q}_1$ serves as a lower bound of the delay performance of a receiver with conventional schemes without Merging, Equation (27) is proved.

In addition, it is easy to verify that when $\mathbb{P}\left(a_h[t]>1\right)>0$, the upper bounded given in Equation (27), is a constant independent of m. The proof is complete.

APPENDIX F
PROOF FOR LEMMA 3

Let K_d^k denote the number of packets of the combined session after Merging, with decoding delay greater than the threshold k for the decoding interval $I_d, d\in\mathbb{N}$.

By the definition of delay violation probability (given by Equation (23)), the numerator can be expressed as the sum of the number of packets exceeding the threshold in the decoding intervals,

$$\mathbb{P}(D_1>k)=\lim_{J\to\infty}\frac{\sum_{d=1}^{J}K_d^k}{\sum_{d=1}^{J}K_d}$$
$$=\lim_{J\to\infty}\frac{\sum_{d=1}^{J}K_d^k}{\sum_{d=1}^{J}I_d}\times\lim_{J\to\infty}\frac{\sum_{d=1}^{J}I_d}{\sum_{d=1}^{J}K_d}. \tag{55}$$

Subsequently, we show how to derive the properties for the two limit terms on the right side of Equation (55).

The second limit term is simple. By Equation (7), at a decoding moment t_1^d, all packets up to $A[t_1^d]$ are decoded by receiver 1. If $t=\sum_{d=1}^{J}I_d$, with Equation (2) we have

$$\lim_{J\to\infty}\frac{\sum_{d=1}^{J}I_d}{\sum_{d=1}^{J}K_d}=\lim_{t\to\infty}\frac{t}{\sum_{\tau=1}^{t}\sum_{h=1}^{m}a_h[\tau]}\overset{(a)}{=}\frac{1}{m\lambda}, \tag{56}$$

where in step (a), strong law of large numbers is applied on *i.i.d.* random variables $\{a_h[\tau]\}_{\forall\tau,\forall h}$.

To bound the first limit term in Equation (55), we observe the following facts for the packets decoded after the interval I_d, which can be seen from Equations (45) and the definition of K_d^k.

1) If $I_d\leq k$, there is no packets exceeding the threshold k, i.e., $K_d^k=0$.
2) If $I_d>k$, there are at most mI_d packets which exceed the threshold k, i.e., $K_d^k\leq mI_d$.
3) If $I_d>k$, there is at least one packet which exceed the threshold k, i.e., $K_d^k\geq 1$.

Thus,

$$\lim_{J\to\infty}\frac{\sum_{d=1}^{J}K_d^k}{\sum_{d=1}^{J}I_d}\leq\lim_{J\to\infty}\frac{\sum_{d=1}^{J}1_{\{I_d>k\}}mI_d}{\sum_{j=d}^{J}I_d},$$
$$\lim_{J\to\infty}\frac{\sum_{d=1}^{J}K_d^k}{\sum_{d=1}^{J}I_d}\geq\lim_{J\to\infty}\frac{\sum_{d=1}^{J}1_{\{I_d>k\}}}{\sum_{d=1}^{J}I_d}. \tag{57}$$

Consider $1_{\{I_d>k\}}I_d$ and $1_{\{I_j>k\}}$ as the rewards earned in interval I_d. According to the renewal reward theory (see Theorem 11.4 [25]), we have

$$\lim_{J\to\infty}\frac{\sum_{d=1}^{J}1_{\{I_d>k\}}mI_d}{\sum_{d=1}^{J}I_d}=\frac{m\mathbb{E}\left[1_{\{\widehat{I}>k\}}\widehat{I}\right]}{\mathbb{E}\left[\widehat{I}\right]},$$
$$\lim_{J\to\infty}\frac{\sum_{d=1}^{J}1_{\{I_d>k\}}}{\sum_{d=1}^{J}I_d}=\frac{\mathbb{E}\left[1_{\{\widehat{I}>k\}}\right]}{\mathbb{E}\left[\widehat{I}\right]}. \tag{58}$$

Note that

$$\mathbb{E}\left[1_{\{\widehat{I}>k\}}\widehat{I}\right] = \sum_{b=k+1}^{\infty} b\mathbb{P}\left(\widehat{I}=b\right)$$

$$= k\mathbb{P}\left(\widehat{I}>k\right) + \sum_{b=k}^{\infty}\mathbb{P}\left(\widehat{I}>b\right),$$

$$\mathbb{E}\left[1_{\{\widehat{I}>k\}}\right] = \mathbb{P}\left(\widehat{I}>k\right), \tag{59}$$

which, by combining with Equations (55), (56), (57) and (58), completes the proof of Equation (47).

APPENDIX G
PROOF FOR LEMMA 4

Based on Equation (45), $\mathbb{P}\big(\widehat{I}>b\big), b \in \mathbb{N}$ can be expressed as

$$\mathbb{P}\left(\widehat{I}>b\right) =$$

$$\mathbb{P}\left(\sum_{\tau=1}^{t}(c_1[\tau]-a[\tau]) \le 0, \forall 1 \le t \le b\right)$$

$$\mathbb{P}\left(\sum_{\tau=1}^{t}\left(\sum_{j=1}^{m}c_{1,j}[\tau] - \sum_{h=1}^{m}a_h[\tau]\right) \le 0, \forall 1 \le t \le b\right). \tag{60}$$

Notice that $\{c_1[\tau]-a[\tau]\}_{\tau \in \mathbb{N}}$ are *i.i.d.* random variables and $\mathbb{E}\left[c_1[\tau]-a[\tau]\right] = \sum_{j=1}^{m}\gamma_{1,j} - m\lambda > 0$. According to the Cramer's Theorem (see Theorem 2.1.24 in [26]),

$$\mathbb{P}\left(\sum_{\tau=1}^{t}(\widehat{c}_1[\tau]-\widehat{a}[\tau]) \le 0\right) = e^{-t\Phi_1^m+o(t)}, \tag{61}$$

where Φ_1^m is the rate function defined in Equation (25). According to the Ballot's Theorem (see Theorem 3.3 in [27]),

$$\mathbb{P}\left(\sum_{\tau=1}^{t}(\widehat{c}_1[\tau]-\widehat{a}[\tau]) \le 0, \forall 1 \le t \le b\right)$$

$$= e^{-b\Phi_1^m+o(b)},$$

if and only if Equation (61) holds. Hence, we get the decay rate regarding the decoding interval I, Equation (48) is proved.