Balancing Queueing and Retransmission: Latency-Optimal Massive MIMO Design

Xu Du, *Member, IEEE*, Yin Sun, Ness B. Shroff, and Ashutosh Sabharwal, *Fellow, IEEE*

***Abstract*— One fundamental challenge in 5G URLLC is how to optimize massive MIMO systems for achieving low latency and high reliability. A natural design choice to maximize reliability and minimize retransmission is to select the lowest allowed target error rate. However, the overall latency is the sum of queueing latency and retransmission latency, hence choosing the lowest target error rate does not always minimize the overall latency. In this paper, we minimize the overall latency by jointly designing the target error rate and transmission rate adaptation, which leads to a fundamental tradeoff point between queueing and retransmission latency. This design problem can be formulated as a Markov decision process, which is theoretically optimal, but its complexity is prohibitively high for real-system deployments. We managed to develop a low-complexity closed-form policy named Large-arraY Reliability and Rate Control (LYRRC), which is proven to be asymptotically latency-optimal as the number of antennas increases. In LYRRC, the transmission rate is twice of the arrival rate, and the target error rate is a function of the antenna number, arrival rate, and channel estimation error. With simulated and measured channels, our evaluations find LYRRC satisfies the latency and reliability requirements of URLLC in all the tested scenarios.**

***Index Terms*— 5G mobile communication, mobile communication, multiuser channels, queueing analysis, cross layer design, channel rate control, time-varying channels, precoding, OFDM, channel estimation.**

I. Introduction

NEXT-GENERATION cellular systems, labeled as 5G, are targeting low latency and ultra-high reliability to support new forms of applications, e.g. mission critical communications. One of the key technologies for 5G will be massive MIMO, where the base-stations will be equipped with tens to hundreds of antennas [1]–[4]. In this paper, we explore how to leverage the large number of spatial degrees of freedom to minimize latency while ensuring high reliability.

Current cellular system design follows a layered approach. The queueing latency[1] is managed at MAC and higher layers, while the target (block) error rate[2] is managed separately by the physical layer to maximize the physical layer throughput. For example, the transmission rate (usually referred to as modulation and coding scheme [5]) is often adapted to meet a fixed target error rate of around 10%. This decoupled design is shown to be nearly throughput optimal [6] for single-antenna systems. However, such a decoupled design may not achieve low latency.

As 5G pushes to low latency (10-100$\times$ lower than the LTE system [7]) and ultra-high reliability, it is of paramount importance to control the latency and service unreliability caused by retransmissions. The Ultra-Reliable Low-Latency Communication (URLLC) has a reliability requirement of 99.9999% [8], i.e., the probability of packet successful delivery within 4 round of transmissions (0.25 ms/5G frame) should be higher than 99.9999%. To satisfy such reliability requirement, the target error rate cannot exceed 3.16%. For a given set of possible target error rates, it might be natural to choose the lowest one, which leads to the highest link reliability and shortest retransmission latency. However, since the overall latency is the sum of latency due to queueing and due to retransmissions, a very small target error rate might result in long queueing latency and does not always minimize the overall latency. In this paper, we achieve reliability guaranteed latency minimization by finding the target error rate and the transmission rate adaptation that jointly minimize the overall latency.

While it is widely known that the target error rate reduces with a higher transmission power or a lower transmission rate, the relationship between the target error rate and overall latency is more complex. There is a tradeoff between retransmission latency and queueing latency, both of which are impacted by the target error rate: On the one hand, the retransmission latency reduces as the target error rate reduces. On the other hand, if the system is fixed to an extremely low target error rate, few packets can be transmitted in each frame, i.e., the transmission time to send the same amount of packets increases, and packets have to wait for a longer time in the

Manuscript received February 20, 2019; revised August 4, 2019 and November 8, 2019; accepted December 21, 2019. Date of publication January 10, 2020; date of current version April 9, 2020. This work was supported in part by the National Science Foundation under Grant CCF-1813078, Grant CNS-1518916, Grant CNS-1314822, Grant CNS-1618566, Grant CNS-1719371, and Grant CNS-1409336 and in part by the Office of Naval Research under Grant N00014-17-1-2417. The associate editor coordinating the review of this article and approving it for publication was S. Buzzi. *(Corresponding author: Xu Du.)*

Xu Du was with the Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005 USA. He is now with the Facebook Inc., Menlo Park, CA 94025 USA (e-mail: xdurice@gmail.com).

Yin Sun is with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849 USA (e-mail: yzs0078@auburn.edu).

Ness B. Shroff is with the Department of Electronics and Communication Engineering, The Ohio State University, Columbus, OH 43210 USA, and also with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: shroff@ece.rice.edu).

Ashutosh Sabharwal is with the Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005 USA (e-mail: ashu@rice.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2019.2963830

[1]In this paper, we use queueing latency to represent the waiting time that packets spend in the MAC-layer queue. And overall latency denotes the total latency caused by retransmission and waiting at the MAC-layer queue.

[2]In this paper, we use the target error rate when emphasizing the design of transmission control. And we use block error rate when emphasizing the probability of decoding error under a given transmission control.

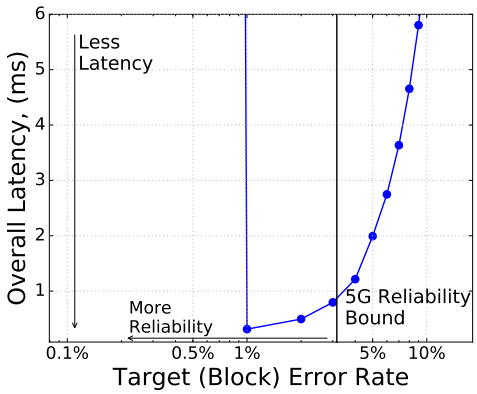


Fig. 1. An example illustrating the overall latency for different target error rates, where the transmission rate has been optimized for each given target error rate. A massive MIMO uplink system with 4 single-antenna users and 32 base-station antennas is considered. The channel traces are measured in an over-the-air channel on the Rice Argos platform and the base-station estimates the channel based on 8 pilot symbols per user. Please find the evaluation details in Section VI.

queue. Therefore, under a given arrival process, the queueing latency increases as the target error rate reduces. The situation is further complicated by the fact that current mobile users adapt their transmission power, which makes the feasible (transmission rate, target error rate) tuple time-varying. Fig. 1 depicts an example of the minimum overall latency achieved at different target error rates where the transmission rate is optimized for given target error rate; the details on how to optimize the transmission rate will be discussed later in Section III. For the specific example in Fig. 1, a target error rate (1%) smaller than both the LTE target error rate (10%) and the URLLC reliability requirement (target error rate of 3.16%) results in the minimum overall latency. It demonstrates a need for finding an appropriate target error rate that minimizes the overall latency by balancing the queueing latency with the retransmission latency.

In this paper, we model practical massive MIMO systems with retransmissions. To minimize the overall latency from both queueing and retransmission, we optimize the target error rate and transmission rate adaptation. The main contributions of this paper are the following:

- We formulate a latency minimization problem for massive MIMO systems, in which the target error rate and transmission rate are jointly optimized for minimizing the overall latency, subject to the reliability constraint of URLLC. The arrival process is a discrete random process that is memoryless. This optimization problem is cast as a constrained Markov decision process and solved by value iteration.
- Because Markov decision process does not provide much insight on the optimal control, we develop a deterministic control policy for massive MIMO with a large number of antennas and a constant arrival rate. We note that there exists an important 5G URLLC type data traffic, e.g., time-sensitive and throughput-hungry virtual reality (VR) service [9], which has a constant data arrival rate. This deterministic control policy is named as Large-arraY Reliability and Rate Control (LYRRC), which has a low complexity and is in a closed form: If the packet arrival rate is λ, the transmission rate of LYRRC is 2λ. In addition, the target error rate of LYRRC is $F_\eta\left[\frac{1}{M^{1-\rho}}\left(1+\frac{K}{\tau}+p_I\right)\right]$, where F_η is the CDF of the effective channel gain (defined later), M is the number of base-station antennas, K is the number of users, ρ is the traffic arrival load over link capacity, p_I is the power of the interference from neighboring cells, and τ is the number of pilots. LYRRC is proven to be asymptotically optimal as the number of antennas grows to infinity. Furthermore, the total latency achieved by LYRRC can be expressed as a closed-form function of the number of base-station antennas M, the number of pilots τ, the number of served users K, and ρ. In particular, for $\rho \in [0,1)$, we show that the average waiting time diminishes to zero as M increases to infinity.
- To verify LYRRC's performance in the real world, we measure massive MIMO channels on the 2.4 GHz with Rice Argos platform [2], which consists of a 64-antenna base-station and four mobile users. The numerical experiments based on the measured and simulated channels show that LYRRC with 5G self-contained frame [5], [10] can simultaneously meet the 1 ms latency and 99.9999% reliability criterion. In the same scenario, the best latency of transmission rate control policies with a fixed target error rate of 10% is more than 5 ms. The evaluations demonstrate that LYRRC can provide $400\times$ latency reduction compared to current LTE transmission control, which has a target error rate of 10% and fixed per-frame transmission power control. Compared to the best queue-length based rate adaptation policy with a fixed target error rate of 10%, LYRRC achieves a $20\times$ latency reduction.

Related Work: The majority of the massive MIMO literature focuses on the achievable rate maximization, which assumes full-buffer and does not model the upper layer latency from queueing. Massive MIMO was shown to provide higher spectral efficiency [11], [12], wider coverage [11], [12] and easier network interference management [11], [13], [14] than traditional MIMO. This work differs from previous massive MIMO physical layer work in that we provide reliability guaranteed latency-optimal transmission control. Prior work also optimized the retransmission process, either for throughput [6] or energy efficiency [15] maximization. Additionally, cross-layer optimization [16]–[19] have been proposed for latency reduction. For a point-to-point system, past studies [20]–[23] showed that using the queue-length information for transmission rate control can reduce queueing latency. Finally, stochastic network calculus [24] is used to capture the latency violation probability of multi-input single-output systems with perfect rate adaptation. Thus, the perfect rate adaptation of past work implies no decoding error or retransmission latency.

The remainder of this paper is structured as follows. In Section II, we provide a physical layer abstraction and network model for a single user latency minimization problem. Section III provides an algorithm to solve the formulated latency minimization problem. A simple and yet latency-optimal transmission control policy, LYRRC, is inves-

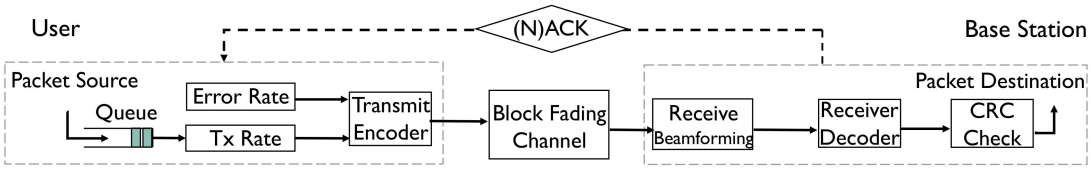


Fig. 2. Single-user uplink system consisting of a single antenna user and an M-antenna base-station.

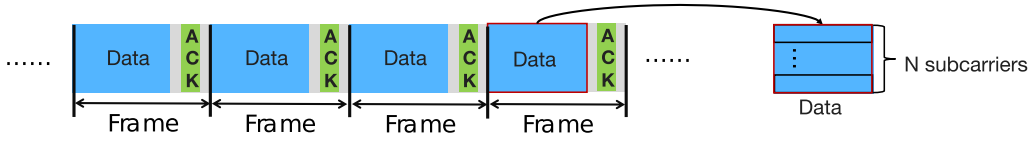


Fig. 3. Structure of the self-contained frames. Each self-contained frame consists of uplink data resource blocks (blue), downlink feedback signals (green) and the guard periods (gray). The transmitted data is encoded over N subcarriers with a single code-block.

tigated in the large-array regime in Section IV. In Section V, we extend our single-user analytical results to multiuser massive MIMO systems. We provide numerical results in Section VI and conclude in Section VII.

Notations: We use boldface to denote vectors/matrices. We use $|\cdot|$ to denote the magnitude of a complex number. And the l_2 norm of a complex vector is $\|\cdot\|$. The complex space is $\mathbb{C}$. The space of real value is $\mathbb{R}$ whose positive half is denoted as $\mathbb{R}^+$. The following notations are used to compare two non-negative real-valued sequences $\{a_n\}$, $\{b_n\}$: $a_n = O(b_n)$ if $\lim_{n\to\infty} \frac{a_n}{b_n} \leq \infty$; $a_n = o(b_n)$ if $\lim_{n\to\infty} \frac{a_n}{b_n} = 0$. And $f_1(M) \cong f_2(M)$ denotes that $\lim_{M\to\infty} \frac{f_1(M)}{f_2(M)} = 1$.

II. System Model and Problem Formulation

A. System Model

We consider a massive MIMO uplink system. The single-user case is considered first in Sections II-IV, and is depicted in Fig. 2. The extension to multi-user systems will be presented later in Section V. Each user is equipped with a single antenna and the base station has M antennas. Based on the physical layer procedures defined in the first 5G release [5], we consider that the system operates in self-contained frames, as shown in Fig. 3. A self-contained frame consists of both data transmission and an immediate ACK/NACK. Without loss of generality, the duration of each frame is of 1 unit and Frame t spans the time interval $[t, t+1)$, $t \geq 0$. In each frame, the user first transmits encoded data packets to the base-station. The base-station then feeds back an ACK or NACK to signal whether a decoding error occurred. The feedback is assumed to be error free.

1) Physical Layer Model: During the uplink data transmission, the received signal by the base-station over the wideband channel is

$$\mathbf{y}_n = \sqrt{\gamma}\mathbf{h}_n x_n + \mathbf{z}_n, \quad n = 1, ..., N, \tag{1}$$

where n is the subcarrier index, N is the total number of subcarriers, x_n is the transmitted signal, $\mathbf{z}_n \in \mathbb{C}^M$ is a zero-mean circularly symmetric complex Gaussian noise vector, and $0 < \gamma \leq 1$ is the large-scale channel gain. We model the channel fading processes as block Rayleigh fading, where the small-scale fading vector $\mathbf{h}_{t,n}$ maintain the same during each frame and varies independently across frames and subcarriers. In this paper, we may omit the frame index t in $\mathbf{h}_{t,n}$ when the frame index is clear from the context. During each frame, the user transmits τ uplink pilots, each with power p_τ. Let $\hat{\mathbf{h}}_n$ be the estimated channel vector by the base-station via the MMSE estimator. The estimated channel satisfies that [11], [12]

$$\mathbf{h}_n = \hat{\mathbf{h}}_n + \mathbf{e}_n, \tag{2}$$

where $\mathbf{e}_n \in \mathbb{C}^M$ is a zero-mean, circularly symmetric complex Gaussian noise vector with variance of $\frac{1}{1+\gamma p_\tau \tau}$. After applying conjugate beamforming, the obtained signal is

$$\begin{aligned}\hat{x}_n &= \hat{\mathbf{h}}_n^H \mathbf{y}_n = \hat{\mathbf{h}}_n^H \left[\sqrt{\gamma}\left(\hat{\mathbf{h}}_n + \mathbf{e}_n\right) x_n + \mathbf{z}_n\right] \\ &= \sqrt{\gamma}\hat{\mathbf{h}}_n^H \hat{\mathbf{h}}_n x_n + \sqrt{\gamma}\hat{\mathbf{h}}_n^H \mathbf{e}_n x_n + \hat{\mathbf{h}}_n^H \mathbf{z}_n,\end{aligned} \tag{3}$$

where the three terms on the right hand side represent the desired signal, signal loss from imperfect channel knowledge, and noise, respectively. The receive SINR on Subcarrier n is [14], [25]

$$\mathrm{SINR}_n = \frac{\gamma p}{\frac{\gamma p}{1+\gamma p_\tau \tau} + 1} \left\|\hat{\mathbf{h}}_n\right\|^2, \tag{4}$$

where $p = |x_n|^2$ is the power of uplink data transmission.

The user is aware of the large-scale channel gain γ and the distribution of the small-scale channel fading via the estimation of a periodic indication signal broadcast by the base-station [5]. During each frame, all uplink packets to be transmitted are encoded in a single code block that spans all N subcarriers. The block error rate of the uplink transmission ϵ is a function of the transmission power. A closed-form characterization of the block error rate appears to be intractable when the code-block length is finite [26]. Hence, we employ the following block error rate approximation that was developed in [6], [26]–[29]. Let L be the number of information bits in each packet, and r_t is the number of transmitted packets in Frame t. We refer to r_t as the *transmission rate*. The block error rate of a code block with a code-block length L_{code} can

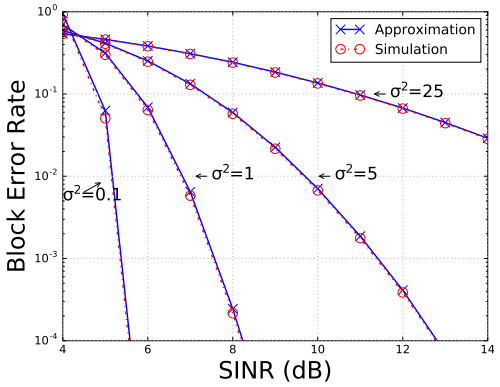


Fig. 4. Block error rate of a coded system as a function of SINR mean with $N = 1$. In simulation, the channel gain follows the normal distribution with labeled variance. The approximations are obtained by (6). And the simulation is done with LDPC code [31] and sparse parity-check matrix comes from the DVB-S.2 standard. The transmission is at a rate of 1.5 bits per symbol (8-QAM, 0.5 code rate).

be approximated as

$$\epsilon \approx \text{Prob}\left[\sum_{n=1}^{N}\log\left(1+\text{SINR}_n\right)-\frac{\nu}{\sqrt{L_{\text{code}}}}\leq rL\right] \quad (5)$$

$$\approx \text{Prob}\left[\sum_{n=1}^{N}\log\left(\text{SINR}_n\right)\leq rL\right], \quad (6)$$

where ν is the channel dispersion [26], [28] due to finite block length and is upper bounded by $\log_2(e)$. For a systems with strong channel coding, [26] shows that (5) closely captures the block error rate when $L_{\text{code}} > 100$. The approximation in (6) is derived by considering sufficiently large code-block length [6], [27], [29] and high SINR regime [6], [27]. Fig. 4 provides an illustration of the approximated block error rate in (6), in which an LDPC-based massive MIMO system is considered and the code-block length is chosen according to DVB-S.2 standard. Our simulations confirm the conclusions drawn from past works [6], [27], [29]. We hence adopt[3] (6) as the block error rate model.

2) Buffer Dynamics With Retransmission: We assume that there is no packet in the buffer at time 0. During each frame, λ new packets arrive in the queue[4] and each packet contains L-bits. In each frame, the user receives downlink ACK/NACK feedback from the base-station. Upon ACK, the transmitted packets are removed from the buffer. Upon NACK, the transmitted packets remain at the buffer queue head[5]. We use the indicator function 1_t to represent decoding success, $1_t = 1$ means success and $1_t = 0$ otherwise. The distribution of the 1_t is determined by the chosen target error rate ϵ as $P[1_t = 1] = 1-\epsilon$ and $P[1_t = 1] = \epsilon$.

[3]One can also use the block error rate approximation (5) which is more accurate in the low SINR and short code-block length regime. In this case, the effective channel gain in (12) and power mapping in (13) should be modified accordingly.

[4]Our model and analysis can be directly generalized to the case where the number of new arrival packets across frames follow an independent and identically distribution.

[5]It is possible to reduce the power of retransmissions via the joint decoding of failed packets and retransmissions as in HARQ. For mathematical tractability, we consider that the receiver discards undecoded packets.

At time t, let q_t be the queue-length of the buffer, and r_t be the number of packets to be transmitted at Frame t as per the control decision. The queue-length evolves according to

$$q_{t+1} = \min\left[\max\left(q_t+\lambda-1_t r_t,\lambda\right),B\right], \quad (7)$$

where B is the size of the buffer and r_t is the number of transmitted packets in Frame t. If the buffer cannot store all the packets waiting to be transmitted, an overflow event occurs. The number of dropped packets due to the buffer overflow is given by

$$b_t = \max\left(q_t+\lambda-1_t r_t-B,\lambda-B\right). \quad (8)$$

The average number of dropped packets due to overflow, measured in packets per frame, is $\lambda_{\text{drop}} = \lim_{T\to\infty}\sum_{t=0}^{T-1} b_t/T$. When packet overflow happens, the dropped packets induce significant latency to time-sensitive applications. We assume that each overflowed packet introduces a large latency penalty D_{drop}. We are interested in minimizing the overall latency (from arrival to successfully delivery). We consider the stationary policies are complete, i.e., the minimum latency can be achieved by a stationary policy. Under a stationary policy, the queueing latency of successfully served packets are $\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\frac{q_t}{\lambda-\lambda_{\text{drop}}}$, which is derived by using Little's Law [30]. To summarize, if a packet is dropped, its latency is D_{drop} and if a packet is successfully served (not dropped), its latency is $\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\frac{q_t}{\lambda-\lambda_{\text{drop}}}$. The average latency is then

$$D = \frac{\lambda-\lambda_{\text{drop}}}{\lambda}\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\frac{q_t}{\lambda-\lambda_{\text{drop}}}+\frac{\lambda_{\text{drop}}}{\lambda}D_{\text{drop}}$$
$$= \frac{\bar{q}}{\lambda}+\frac{\lambda_{\text{drop}}}{\lambda}D_{\text{drop}}, \quad (9)$$

where $\frac{\lambda-\lambda_{\text{drop}}}{\lambda}$ is the proportion of successfully served packets and $\bar{q}$ is the average queue-length, i.e., $\lim_{T\to\infty}\sum_{t=0}^{T-1}\frac{q_t}{T}$.

3) Transmission Power Adaptation: We consider the transmission power of the user to satisfy a long-term power constraint of P. In Frame t, the transmission power is adapted, based on the transmission rate r_t, and the number of pilots τ, to achieve the target error rate ϵ. The transmission power is quantified in the sequel: Substituting (4) into (6), the block error rate is approximated as

$$\epsilon \approx \text{Prob}\left[\left(\prod_{n=1}^{N}\kappa_n\right)^{1/N} \leq \frac{\exp\left(rL/N\right)}{M}\left(\frac{1}{1+\gamma p_\tau\tau}+\frac{1}{\gamma p}\right)\right]. \quad (10)$$

where κ_n is the the per-antenna gain of small-scale channel fading, given by

$$\kappa_n \triangleq \left\|\hat{\mathbf{h}}_n\right\|^2/M. \quad (11)$$

The per-antenna gain κ_n is the arithmetic mean of the small-scale channel gain across the M antennas because the received signals with different antennas are combined during the linear beamforming. The left-hand-side of the

inequality of (10) is determined by the small-scale fading, and the right-hand-side of (10) is a constant independent of small-scale fading. For the ease of subsequent presentation, we define

$$\eta \triangleq \left(\prod_{n=1}^{N} \kappa_n\right)^{1/N}, \tag{12}$$

which is called *effective channel gain*. The effective channel gain (12) is the geometric mean across the N subcarriers because the maximum outage-free rate [26] can be approximated by the logarithmic of the product of the per-subcarrier SINR_n. Let $F_\eta(x) \triangleq \mathrm{Prob}(\eta \le x)$ denote the cumulative distribution function (CDF) of the effective channel η. And the inverse CDF of η is $F_\eta^{-1}(\epsilon) \triangleq \inf\{x \in \mathbb{R}^+ : \epsilon \le F_\eta(x)\}$. Recall that the transmission power is adapted to achieve the target error rate, from (10), we have

$$p(r, \epsilon, \tau) = \left[\frac{M\gamma F_\eta^{-1}(\epsilon)}{\exp(rL/N)} - \frac{\gamma}{1+\gamma p_\tau \tau}\right]^{-1}, \tag{13}$$

where F_η^{-1} is the inverse CDF of the effective channel gain η in (12). When τ increases, the base-station has a more accurate channel estimation and the needed transmission power (at the same rate with the same reliability) reduces. One can observe that the required transmission power increases with the transmission rate r and the packet size L, and decreases with the number of base-station antennas M, the number of subcarriers N, and the number of pilots τ.

B. Single-User Latency Minimization Problem

We now formulate the single-user latency minimization problem. The objective of the joint target error rate and transmission rate control is to minimize the average packet latency under a long-term average power constraint. The *system state* is the queue-length q_t, whose state space is $\mathcal{Q} = \{0, 1, ..., B\}$. The transmission controller determines the number of transmitted packets r_t at the beginning of each frame based on the queue-length q_t, as well as the target error rate ϵ that remains constant in all frames over time. Recall that the the transmission rate is the number of transmitted packets r_t. We consider the set of stationary policies such that $r_t = \mu(q_t)$, where $\mu : \mathcal{Q} \to \mathbb{R}^+$ is a function. And the target error rate ϵ is chosen from a finite set $\mathcal{E}$. Finally, the transmission power p_t is adapted based on the designed rate r_t, target error rate ϵ, and number of pilot τ as in (13). Both the transmission rate function μ and the resulting transmission power are independent of the exact small-scale fading $\mathbf{h}_n$ as it is unknown to the user.

For any target error rate ϵ and transmission rate function μ, we assume that the resulted Markov chain of the system states is ergodic, i.e., the unichain condition is satisfied. The associated unique steady state of the system is denoted as π. The latency minimization problem is formulated as:

$$\min_{\substack{\epsilon\in\mathcal{E},\\ r_t=\mu(q_t),\\ \mu:\mathcal{Q}\to\mathbb{R}^+}} \quad D = E\left[\frac{\bar{q}}{\lambda} + \frac{\lambda_{\mathrm{drop}}}{\lambda} D_{\mathrm{drop}}\right] \tag{14a}$$

$$\text{s.t.} \quad E\left[\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} p(r_t, \epsilon, \tau)\right] \le P, \tag{14b}$$

$$\epsilon \le \epsilon_{\max}, \tag{14c}$$

$$\text{State Transition Model (4)-(8)}, \tag{14d}$$

where $\epsilon_{\max}$ it the maximum allowed target error rate due to reliability requirement. For 5G URLLC, $\epsilon_{\max} = (1 - 99.9999\%)^{1/4} = 3.16\%$. The optimal objective value of (14) is denoted as D^*, or $D^*(M)$ when we need to emphasize the dependence on the number of antennas M. Hence, $D^*(M)$ captures the minimum overall latency D^* as a function of the number of base-station antennas M.

III. Latency-Optimal Single-User Transmission Control

In this section, we first formulate the latency minimization problem (14) as a constrained average cost Markov Decision Process (MDP) and solve it by a proposed algorithm. The proposed algorithm can also solve the latency-optimal control for general point-to-point MIMO systems by replacing the per-subcarrier SINR in (4) with the SINR of the MIMO system. The effective channel gain in (12) and power mapping in (13) also should be modified accordingly.

A. Lagrange Duality of the MDP

For a target error rate $\epsilon \in \mathcal{E}$, and a stationary transmission rate adaptation $\mathcal{Q} \to \mathbb{R}^+$, based on the definition of average latency (9), we define the induced latency cost mapping d on each state action pair as

$$d(q_t, r_t, \epsilon) = \frac{q_t}{\lambda} + \frac{b_t}{\lambda} D_{\mathrm{drop}},$$

where b is the number of the dropped packet due to buffer overflow as shown in (8). In Frame t, a latency cost and a transmission power cost are incurred. The average overall latency of the problem in infinite horizon equals

$$D_\pi = E_\pi\left[\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} d(q_t, r_t, \epsilon)\right].$$

Similarly, utilizing the transmission power characterization in (13), the average power is

$$P_\pi = E_\pi\left[\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} p(r_t, \epsilon, \tau)\right].$$

Given an average power constraint P, the objective of the joint target error rate selection and transmission rate control is restated as a constrained MDP as

$$\begin{aligned} &\text{Minimize} && D_\pi \\ &\text{subject to} && P_\pi \le P, \epsilon \le \epsilon_{\max}, \\ &&& \text{State Transition Model (4)-(8).} \end{aligned} \tag{15}$$

The constrained MDP (15) is converted to an unconstrained MDP via Lagrange's relaxation as

$$\begin{aligned} &\text{Minimize} && D_\pi + \beta P_\pi \\ &\text{subject to} && \epsilon \leq \epsilon_{\max}. \end{aligned} \tag{16}$$

For ergodic MDP, [22], [32] provide a sufficient condition under which the unconstrained MDP is also optimal for the original constrained problem (14). For all policies such that $P_\pi = P$, the sufficient condition provided by [22], [32] is satisfied. Thus, when the constraint is binding, there exists zero-duality gap between original problem (14) and the unconstrained MDP (16), i.e., their optimal solution is the same.

We now present the algorithm to solve (16) in Section III-B. The closed-form solution of (16) and the characterization of the array-latency tradeoff $D^*(M)$ are presented in Section IV.

B. A Value Iteration Based Algorithm

Problem (16) is an MDP with an average cost criterion in infinite horizon. To find the optimal target error rate, we need to find the optimal transmission rate adaptation and the corresponding achievable latency for each $\epsilon \in \mathcal{E}$ that is smaller than $\epsilon_{\max}$. Furthermore, for each target error rate ϵ, we can use binary search method to find the smallest β that satisfies the long-term power constraint P in (16). Such β corresponds to the latency-optimal solution for (15) because that, for each ϵ, the average power is monotonically non-decreasing on $\beta > 0$. Finally, for each ϵ and β, we thus find the optimal transmission rate adaptation μ^* by considering α-discounted problem [33] of (16). We now present a solution to each of the discounted problem. For each system state q, define value cost function as

$$V_\alpha(q) \triangleq \min_\mu E_\pi \left\{ \sum_{t=0}^{\infty} \alpha^t \left[d(r_t, q_t, \epsilon) + \beta p(r_t, \epsilon, \tau) \right] \right\},$$

where $\alpha \in (0,1)$ is the discount factor. For each ϵ and β, we need to find a stationary transmission rate adaptation for all α-discounted problem with $\alpha \in (0,1)$, i.e., the Blackwell optimal policy. For the considered *finite* state MDP, the Blackwell optimal policy [33] exists and is also optimal for the average cost problem (16). The Bellman's equation of the above α-discounted problem is then

$$\begin{aligned} V_\alpha^*(q) = \min_\mu \Big\{ & d(r, q, \epsilon) + \beta p(r, \epsilon, \tau) \\ & + \left[(1-\epsilon) V_\alpha^*(\min(q + \lambda - r, B)) \right. \\ & \left. + \epsilon V_\alpha^*(\min(q + \lambda, B)) \right] \Big\}, \end{aligned} \tag{17}$$

whose state transition is described by (6), (7), and (8). Using dynamic programming with value iteration [33] over (17), we can solve the α-discounted problem. Since the discounted cost V_α is bounded, [33] shows that solving (17) generates the optimal transmission rate control μ^*.

We summarize the above steps in Algorithm 1, which solves (15) to find the optimal target error rate and transmission rate adaptation. To provide insights on the structure of optimal transmission controls, we now resent a closed-form characterizations when $M \to \infty$ in Section IV.

Algorithm 1: Latency-Optimal Joint Target Error Rate and Transmission Rate Control

Input : Average power constraint P, number of antennas M, number of subcarriers N, distribution of packet arrival a, large-scale channel gain γ, CDF of effective channel gain η, number of pilots τ, pilots power p_τ.

Output: Optimal target error rate ϵ^*, optimal transmission rate adaptation μ^*, minimum achievable latency D^*.

For $\epsilon \in \mathcal{E}$ *that* $\epsilon \leq \epsilon_{\max}$ **do** *// Find minimum latency for each* $\epsilon \in \mathcal{E}$

- $\beta_{min} = 0$, $\beta_{max} = z$; *// z is a very large but finite number*
- **while** $\beta_{\min}/\beta_{\max} < 1 - \delta$ **do** *// Find smallest β that satisfies the average power constraint, δ is a small constant that controls the algorithm output accuracy*
  - $\beta \leftarrow (\beta_{\max} + \beta_{\min})/2$;
  - Initialize $V_\alpha^0(q)$ for every system state in $\mathcal{Q}$ and $n = 1$;
  - Solve for V_α^1 from V_α^0 via value iteration as (17);
  - **while** $V_\alpha^n \neq V_\alpha^{n-1}$ **do** *// Find optimal μ for each β and ϵ*
    - Update V_α^n from V_α^{n-1} via value iteration as (17);
  - Compute the corresponding power P_{tmp};
  - **if** $P_{\text{tmp}} > P$ **then**
    - $\beta_{\min} = \beta$;
  - **else**
    - $\beta_{\max} = \beta$;
- Denote the solved transmission rate function as $\mu_\epsilon(q_t)$ and the resulted latency as D_ϵ.

Optimal policy extraction: $\epsilon^* = \arg\min_{\epsilon \in \mathcal{E}, \epsilon \leq \epsilon_{\max}} D_\epsilon$, $\mu^*(q_t) = \mu_{\epsilon^*}(q_t)$, and $D^* = D_{\epsilon^*}$.

IV. Large-Array Latency-Optimal Control

In this section, we derive the latency-optimal control for the single-user problem in (14) when the number of base-station antennas $M \to \infty$. For the single-user system in Rayleigh fading, the per-antenna gain κ_n in (11) satisfies the following [11, A.2.4], [12], [14].

- *Mean*: The per-antenna gain mean is a constant that is independent of M, i.e.,

$$E[\kappa_n] = \frac{\tau p_\tau \gamma}{\tau p_\tau \gamma + 1}, \tag{18}$$

- *Variance*: The per-antenna gain variance is inversely proportional to M, i.e.,

$$\text{Var}[\kappa_n] = \frac{1}{M} \left(\frac{\tau p_\tau \gamma}{\tau p_\tau \gamma + 1} \right)^2. \tag{19}$$

In Section V, we will show that a multiuser massive MIMO channel can be decoupled into parallel single-user channels. For each of the decoupled channels, the per-antenna gain is also of variance that is inversely proportional to M.

Based on condition (18), the achievable SINR grows with the number of base-station antennas M linearly. As the focus of the current section is on the asymptotic analysis with $M \to \infty$, we can view $\log M$ as the link "capacity". In the same

spirit, we define the system utilization factor to be a constant $\rho \in [0,1)$ as

$$\rho \triangleq \frac{\lambda L}{N \log M}, \tag{20}$$

where λ is the packet arrival rate, L is the number of bits in each packet, and N is the number of subcarriers. By (20), the packet arrival rate λ increases with M and equals $\frac{N \log M}{L\rho}$. Conceptually, the term $N \log M$ can be viewed as the total "capacity" of the wideband link and λL can be viewed as the data load. Thus, the utilization factor ρ can be interpreted as the ratio between the offered data load and the total link "capacity".

We also make the following assumptions for mathematical tractability. We consider an infinite buffer (i.e., $B \to \infty$), thus no buffer overflow or overflow latency occurs. And the target error rate ϵ can be chosen from a continuous set $(0,1)$.

A. Array-Latency Scaling Lower Bound

Notice that a trivial lower bound of $D^*(M)$ is 1 frame, which is the first transmission attempt of a packet. This 1 frame latency lower bound can only be achieved if the target error rate is exactly zero. We now provide a tighter lower bound of the array-latency curve $D^*(M)$.

Theorem 1 Latency Scaling Lower Bound: The optimum array-latency curve $D^*(M)$ satisfies

$$D^*(M) - 1 \geq \frac{\epsilon_o}{1-\epsilon_o}, \tag{21}$$

where ϵ_o is given by

$$\epsilon_o = F_\eta \left[\frac{1}{M^{(1-\rho)}} \left(\frac{1}{\gamma P} + \frac{1}{\gamma p_\tau \tau} \right) \right], \tag{22}$$

where $F_\eta(\cdot)$ is the CDF of the effective channel gain η in (12)*, $\rho \in [0,1)$ is the utilization factor in* (20)*, and τ is the number of pilots.*

Proof: The main idea is to lower bound the overall latency by the packet retransmission latency, which monotonically increases with the target error rate. To complete the proof, we use Jensen's inequality to show that there exists a minimum target error rate ϵ_o such that for any $\epsilon < \epsilon_o$ the long-term throughput is smaller than λ. Appendix A provides the proof details. □

Theorem 1 presents a latency lower bound. For any transmission rate adaptation, ϵ_o is the minimum target error rate that leads to a long-term throughput no smaller than λ. And if the target error rate is smaller than ϵ_o, the queue-length process will not stable. By the definition of η (12), the per-antenna mean (18), and the per-antenna variance (19), Chebyshev's inequality can be used to show that ϵ_o converges (in probability) to 0 as the number of base-station antenna M increases to infinity. The channel hardening effect can explain such convergence. The latency lower bound (21) hence converges to 0 as $M \to \infty$.

If τp_τ is small, the channel estimation error is large. As a result, both ϵ_o and the latency lower bound are large. In this case, neither high reliability nor low latency can be met. Hence, sufficiently good channel estimation is necessary for achieving high reliability and low latency.

B. Large-Array Optimal Target Error Rate and Transmission Rate Control

In this subsection, we present a simple transmission control policy that meets with the latency lower bound in (20) asymptotically as $M \to \infty$.

Definition: We define the Large-arraY Reliability and Rate Control (LYRRC) as

$$\begin{cases} \epsilon^* = \epsilon_o \\ \mu^* : r_t(q_t) = \min(q_t,\ 2\lambda) \end{cases}, \tag{23}$$

where ϵ_o is given by (22)*.*

The LYRRC policy contains two parts: a target error rate of ϵ_o and an transmission rate control policy μ^*. The transmission rate adaptation μ^* describes a simple thresholding rule: If there are more than 2λ packets in the buffer queue, i.e., $q \geq 2\lambda$, 2λ packets will be transmitted. If less than 2λ packets are currently in the buffer, all packet in the queue will be scheduled for transmission in the frame. In each frame, based on the transmission rate of $\min(q_t, 2\lambda)$, the user utilizes power adaptation (13) to achieve the target error rate target ϵ_o.

To evaluate LYRRC, we now first derive the latency with arbitrary target error rate $\epsilon < \frac{1}{2}$ and transmission rate policy μ^*. We next prove the asymptotic optimality of LYRRC (23) by comparing the achieved latency to the minimum latency lower bound in Theorem 1.

1) Latency Performance of Transmission Rate Adaptation μ^*:

Lemma 1: Under any target error rate $\epsilon < \frac{1}{2}$ and transmission rate adaptation $r_t(q_t) = \min(q_t,\ 2\lambda)$, the overall latency is $1 + \frac{\epsilon}{1-2\epsilon}$.

Proof: The main idea is to compute the steady state distribution of the queue-length, which is a Markov chain with infinite countable states. Appendix B provides the complete proof. □

Lemma 1 provides a closed-form characterization of the transmission rate adaptation μ^* when the maximum buffer-length is infinite. To provide insights on the proof of Lemma 1, we consider the associated Markov chain of the buffer-length. The buffer-length state transition under any target error rate $\epsilon \in (0,1)$, which is not necessarily equal to ϵ_o, and the transmission rate adaptation μ^* is depicted in Fig. 5. By Little's Law, the overall latency equals to the ratio between the average queue-length and the arrival rate λ. Notice that λ is the difference between the adjacent states in Fig. 5. Hence, the average queue-length is in proportional with λ (see Appendix B for a rigorous proof). As a result, the overall latency depends only on the target error rate ϵ, but not on λ.

To summarize, the transmission rate control policy μ^* applies a negative drift $-\lambda$ with probability $(1-2\epsilon)$ towards the minimum queue-length λ. To minimize the latency as $M \to \infty$, the queue-length needs to be regulated towards the minimum queue-length λ. This regulation is achieved by selecting a smaller target error rate.

By using Lemma 1, we have that the achieved latency of LYRRC is $D_{\text{LYRRC}}(M) = 1 + \frac{\epsilon_o}{1-2\epsilon_o}$. As mentioned above, the target error rate ϵ_o of LYRRC (23) reduces as

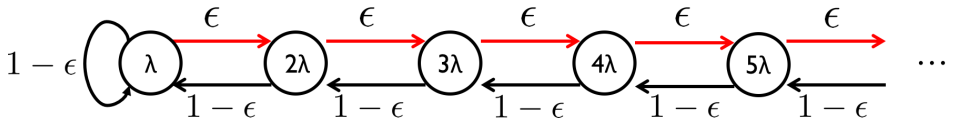


Fig. 5. Evolution of the queue-length q_t under any target error rate $\epsilon \in (0,1)$ and the transmission rate adaptation μ^* as a Markov chain. If $\epsilon > 0.5$, the average queue-length hence queueing latency is infinite.

the number of base-station antennas increases. The achieved latency D_{LYRRC} reduces with more base-station antennas. We now prove the asymptotic optimality of LYRRC.

2) Asymptotic Optimality of LYRRC:

Theorem 2 Optimal Large-Array Control: For any $\rho \in [0,1)$ and positive τ, as $M \to \infty$, LYRRC (23) guarantees that the overall latency is within a vanishing gap from optimal as

$$D_{\text{LYRRC}}(M) - D^*(M) \cong (\epsilon_o)^2, \ M \to \infty, \tag{24}$$

where $D_{\text{LYRRC}}(M) = 1 + \frac{\epsilon_o}{1-2\epsilon_o}$ is the overall latency by LYRRC, and ϵ_o is given by (22).

Proof: We first characterize the gap between latency under LYRRC and minimum latency by combining Lemma 1 and Theorem 1. The proof is complete by using the large deviation theory to show that the power constraint is satisfied. Please see Appendix C for details. □

Recall that $f_1(M) \cong f_2(M)$ denotes that $\lim_{M\to\infty} \frac{f_1(M)}{f_2(M)} = 1$. Theorem 2 establishes the asymptotic optimality of LYRRC. In addition, the latency gap between the lower bound and LYRRC increases as the channel estimation error increases (τ reduces). Furthermore, Lemma 1 and Theorem 2 suggest that the latency-optimal target error rate increases for systems with fewer base-station antennas. Hence, the reliability and low-latency design objectives of 5G URLLC does not always matches with each other for practical massive MIMO system with finite M. Finally, we note that LYRRC can achieve optimal-latency for any $\rho \in [0,1)$, which seems to contradict the transmission rate of $\min(q_t, 2\lambda)$. This can be explained by the fact that we are considering a wireless link with power adaptation and the probability of transmit at 2λ reduces as $M \to \infty$. Therefore, using larger transmission power (over a few frames) can increase the peak transmission rate beyond the long-term average rate. We next combine Theorem 2 and Theorem 1 to characterize the scaling of the array-latency curve $D^*(M)$ in closed-form.

Theorem 3 Large-Array Latency Scaling: As $M \to \infty$, for any positive τ and $\rho \in [0,1)$, the optimum latency converges to 1 frame as

$$D^*(M) - 1 \cong \epsilon_o, \ M \to \infty \tag{25}$$

where $F_\eta(\cdot)$ is the CDF function of the effective channel gain η, and ϵ_o is given by (22).

Proof: Theorem 1 provides a latency lower bound. The optimal joint control in Theorem 2 serves as an achievability proof and provides an upper bound. The proof is complete by showing that the ratio of the upper bound and the lower bound converges to 1 as $M \to \infty$. □

Theorem 3 provides a closed-form characterization of the large-array latency. In *closed-form*, it describes the minimum latency D^* as a function of the utilization factor ρ, the channel estimation error, and the number of base-station antennas M. As $M \to \infty$, $\epsilon_o \to 0$. Thus, both the retransmission and queueing latency converges to 0 frame. Finally, we comment on the impact of imperfect channel state information. For any $\tau > 0$, the latency convergence to the 1 frame as $M \to \infty$. For a practical system with finite M, more accurate channel leads to smaller latency.

V. Multi-User Extension

In this section, we now consider the K-user latency minimization problem over the lossy channel. In this section, suffix $[k],\ k = 1, 2, \cdots, K$ denotes the user index. The long-term power constraint of User k is $P[k]$. The multiuser controller decides the target error rate $\epsilon[k]$ and the transmission rate $r_t[k]$ of User k. The buffer dynamic of each user is identical to that of the single user counterpart that is described in Section II-A.2.

To minimize the system latency of the K users at the same time, we associate positive weights $\omega_k,\ k = 1, \ldots, K$ to users. The multiuser latency minimization problem is then

$$\begin{aligned} \min_{\substack{\epsilon[k],\ r_t[k] \\ \forall k}} \ & \sum_{k=1}^{K} \omega_k D[k] \\ \text{s.t. } & E\left[\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} p_t[k] \le P[k]\right], \ \forall k, \\ & \epsilon[k] = \text{Prob}\left[\sum_{n}^{N} \log\left(\text{SINR}_{t,n}[k]\right) \le r_t[k] L\right], \ \forall k, \\ & \epsilon[k] \le \epsilon_{\max}[k], \ \forall k, \end{aligned} \tag{26}$$

where $\epsilon_{\max}[k]$ is the maximum allowed target error rate (minimum reliability) of User k. And $\text{SINR}_{t,n}[k]$ is the receiver SINR of the n-th subcarrier in Frame t for User k. Here, the buffer length $q_t[k]$ and buffer overflow $b_t[k]$ of User k is given by (7) and (8), respectively.

To detect signals from the K users, the base-station applies receive beamforming. Let matrix $\mathbf{H}_n \in \mathbb{C}^{M\times K}$ denotes the uplink small-scale channel fading between the M-antenna base-station and the K users. Throughout this section, we consider user channels follow i.i.d. Rayleigh fading. Finally, the base-station receives an inter-cell interference that is modeled by an additive white Gaussian noise of power p_I, which is independent of the estimated channel.

Let the estimated channel and estimation error be $\hat{\mathbf{H}}_n$ and $\tilde{\mathbf{H}}_n$, respectively. With the MMSE estimator, the estimation

error between each base-station antenna and User k is an complex Gaussian random variable with zero mean and variance of $\frac{1}{\tau p_\tau[k]\gamma[k]+1}$. Here, τ and $p_\tau[k]$ are the number of uplink pilots and the pilot power, respectively. The base-station use the estimated channel to generate zero-forcing receive beamformers to detect the uplink signal of each user. The receive beamforming matrix is $\mathbf{V}_n \triangleq \left(\hat{\mathbf{H}}_n^H\hat{\mathbf{H}}_n\right)^{-1}\hat{\mathbf{H}}_n^H$. On Subcarrier n, the received signal of User k is [11], [12]

$$\hat{x}_k = \sqrt{p[k]\,\gamma[k]}x_k + \left[\left(\mathbf{H}^H\mathbf{H}\right)^{-1}\hat{\mathbf{H}}_n^H\left(\mathbf{z}+\mathbf{z}_I-\tilde{\mathbf{H}}\mathbf{x}\right)\right]_K, \tag{27}$$

where $\mathbf{z}$ and $\mathbf{z}_I$ are the receiver noise and inter-cell interference, respectively. Similarly to past work [28], [29] on retransmission, we compute the SINR by treating the interference as the worst case Gaussian noise. And the effective SINR for User k on Subcarrier n is

$$\mathrm{SINR}_n[k] = \frac{p_k\gamma_k}{\left(1+p_I+\sum_{i=1}^K\frac{p[i]\gamma[i]}{\tau p_\tau[i]\gamma[i]+1}\right)\left[\left(\hat{\mathbf{H}}_n^H\hat{\mathbf{H}}_n\right)^{-1}\right]_{kk}}, \tag{28}$$

where $[\cdot]_{kk}$ denotes the k-th diagonal element of a matrix. A crucial property of the SINR_n term (28) is that the randomness of both the channel variation and the interference is concisely described by the inverse of the estimated channel, which is a random matrix.

For a practical uplink system where each user is unaware of other users' channel or queue information, the joint target error rate and transmission rate adaptation design appears intractable. To see the difficulty of the joint policy design, we consider the following example. For each user, the inter-beam interference in (28) depends on *other* users' large-scale fading and transmission power. Recall that each user's transmission power changes in each frame based on its current queue-length. Thus, it is extremely difficult for each user with only local knowledge (queue-length and large-scale fading) to infer the exact value of $\sum_{i=1}^K\frac{p[i]\gamma[i]}{\tau p_\tau[i]\gamma[i]+1}$ and hence the proper transmission power. As a result, the target error rate and transmission rate policy cannot be designed distributedly by each user, which is undesirable for a practical uplink system.

Here, we proceed with the observation that, in real-world systems, the pilot power is usually required to be higher than the data signal power [5]. Hence, the $\sum_{i=1}^K\frac{p[i]\gamma[i]}{\tau p_\tau[i]\gamma[i]+1}$ term is upper bounded by $\frac{K}{\tau}$, which can be viewed as a *worst* cast interference penalty. Each user then adjusts its power based on the SINR loss upper bound. Substituting the SINR expression (28) of the multiuser system into (6), we then have that the target error rate as

$$\epsilon \approx \mathrm{Prob}\left[\left(\prod_{n=1}^N\kappa_n\right)^{1/N} \le \left(1+\frac{K}{\tau}+p_I\right)\frac{\exp(rL/N)}{Mp\gamma}\right], \tag{29}$$

where the per-antenna gain κ_n is

$$\kappa_n = \left\{M\left[\left(\hat{\mathbf{H}}_n^H\hat{\mathbf{H}}_n\right)^{-1}\right]_{kk}\right\}^{-1}. \tag{30}$$

Similarly to the single-user case, we also compute the per-frame transmission power as

$$p(r,\epsilon,\tau) = \left(1+\frac{K}{\tau}+p_I\right)\frac{\exp(rL/N)}{F_\eta^{-1}(\epsilon)\,M\gamma}, \tag{31}$$

where ϵ is the scheduled reliability target (target error rate) and r is the transmission rate (in unit of packet). Here, $\approx$ in (29) is because that each user considers the upper bound of inter-beam interference.

The per-antenna gain (30) is independent of the large-scale channel, transmission power, and hence queue-length of the other $K-1$ users. For each user, the distribution of the effective channel η in (12) then becomes independent of the channel, queue-length, and power of the other users. Therefore, we can decouple the multiuser problem. By adopting a new distribution of the effective channel gain η (generated by (30)) and the new power mapping (31), the multiuser problem is decoupled to K independent single user problems (14). Each of the single-user problems can be solved by Algorithm 1. We now further demonstrate that the large-array analytical results in Section IV also apply to the considered multiuser systems.

Theorem 4: For multiuser uplink systems, LYRRC becomes

$$\begin{cases}\epsilon^*[k] = F_\eta\left[\frac{1}{M^{1-\rho[k]}}\left(1+\frac{K}{\tau[k]}+p_I\right)\frac{1}{\gamma P}\right]\\ \mu^*[k] : r_t[k] = \min\left(q_t[k],\ 2\lambda[k]\right).\end{cases} \tag{32}$$

As $M\to\infty$, for positive $\tau[k]$ and $\rho[k]\in[0,1)$, each user operates under LYRRC achieves the minimum latency of

$$D^*[k]-1 \cong \epsilon^*[k],\ k=1,2,\ldots,K,\ M\to\infty. \tag{33}$$

Proof: With random matrix theory, we prove by adopting similar steps as in the single-user case. The key is step is to compute the mean and variance of (30). Please find the proof in Appendix D. □

Recall that $f_1(M)\cong f_2(M)$ denotes that $\lim_{M\to\infty}\frac{f_1(M)}{f_2(M)}=1$. LYRRC, therefore, indeed provides the latency-optimal target error rate and transmission rate policies to the multiuser massive MIMO system. And Theorem 4 also captures the minimum latency of each user.

In conclusion, for any non-negative weights ω_k, we can convert the K user optimization problem into K parallel single user problems. For finite M, Algorithm 1 solves each of the single user problems and provides the optimal target error rate and transmission rate policy. Furthermore, each user operates using LYRRC distributedly is asymptotically latency-optimal.

We end this section by discussing some possible extensions of the multiuser system analysis.

The first extension is the general multiuser MIMO systems with user correlation. For massive MIMO, the user channels are expected to become mutually orthogonal as M increases, which is usually referred to as "favorable propagation" [11], [12]. The favorable proportion is expected to hold in massive MIMO systems [11], [12] and is verified

(a) Argos Array (b) Over-the-air Measurement Setup (c) User Locations

Fig. 6. Argos [2] Massive MIMO base-station and the over-the-air measurements setup. The background map of Fig. Fig. 6c is generated by Google Maps [37]. The black single antennas denotes the locations of the mobile users.

by recent massive MIMO measurements [34], [35]. However, for small scale multiuser systems, user channels might be significantly correlated, and the multiuser scheduling problem cannot be fully decoupled. While spatial multiplexing correlated user leads to smaller SINR, spatial multiplexing only non-correlated users can lead to longer queueing latency. Hence, we expect a latency-minimizing scheduler should balance a tradeoff between longer queueing time and smaller SINR.

The second extension is to model the pilot contamination and base-station array correlation, which both can reduce the SINR. The pilot contamination [11], [12] is caused by pilot reuse and leads to both non-coherent and coherent interference. In particular, without proper pilot decontamination, coherent interference can grow linearly with the number of base-station antennas. Recent research [12], [36] demonstrates that via multicell joint transmission, the massive MIMO system can reject the coherent interference if the covariance matrix of pilot sharing users is asymptotically linearly independent. Under the same condition, [12], [36] shows that the effective SINR can grow linearly with M without bound with pilot contamination and base-station array correlation. Therefore, it is reasonable to use a finite p_I to model the power of the residual inter-cell interference after pilot decontamination.

Finally, we consider the latency-minimum transmission control of multicell systems with pilot contamination and base-station array correlation as an important future work. Note that [12], [36] shows that the SINR can also grow linearly with M, which implies that the mean of the per-antenna gain would be lower bounded by a positive constant. Computing the variance condition and finding the optimal transmission control for this generalized setup is beyond the scope of this paper. To evaluate the impact of the spatial correlation, we utilize over-the-air measured channels in Section VI.

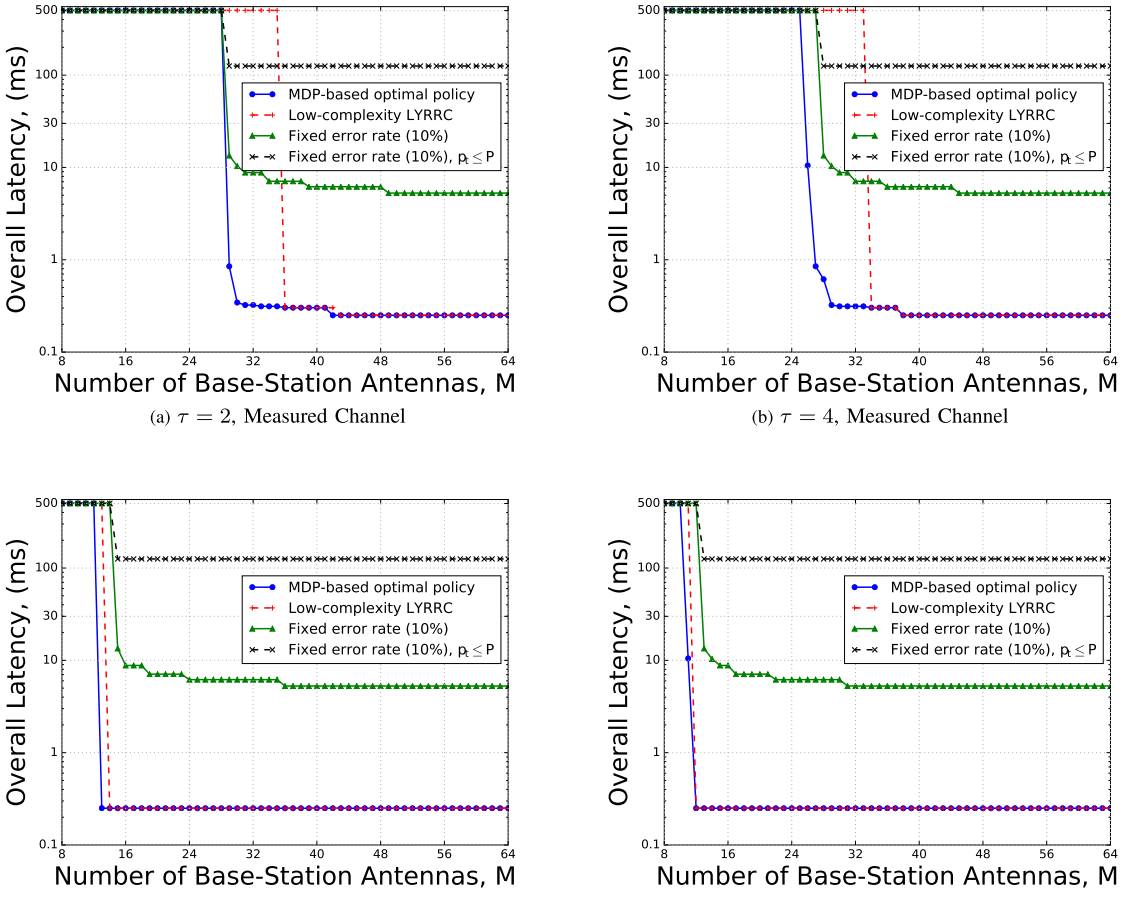
VI. Numerical Results

In this section, we utilize measured channels and simulated channels to confirm our previous analysis in Section III and Section V. During the numerical evaluation, the latency duration is captured in the unit of second, which is obtained by multiplying frame duration to latency measured in the unit of frame. We measure the over-the-air channels between mobile clients and a 64-antenna massive MIMO base-station with Argos system [2] on the campus of Rice University. Figure 6a and 6b describes the Argos array and the over-the-air measurement setup. We measured the 2.4 GHz Wi-Fi channel (20 MHz, 52 non-empty data subcarriers) for four pedestrian users in non-line-of-sight environments, which are denoted by Fig. 6c. For each user, we take channel measurements over 7900 frames of all subcarriers. The effective measured SNR between each mobile user and each base-station antenna is higher than 15 dB. In simulations, we consider measured over-the-air channel traces as the perfect channel.

The base-station adopts MMSE estimator to estimate τ uplink pilots, each of power 20 dBm, from the users. Using the estimated channel, the base-station generates zero-forcing receive beamformers to decode the signal of each user. The users are assumed to follow average power constraint of 20 dBm with large-scale fading of -10 dB. The maximum buffer length B is 10. The packet arrival rate is uniform over the time at the rate of 5 packets per frame. And the packet size L is 52 bits per OFDM symbol. The latency penalty of dropped packets from buffer overflow is 0.5 s. And each self-contained frame is considered of duration 0.25 ms. The state space of the target error rate is $[1\%,\ 2\%,\ \ldots,\ 20\%]$, $[0.1\%,\ 0.2\%,\ \ldots,\ 0.9\%]$, and $[0.01\%,\ 0.02\%,\ \ldots,\ 0.09\%]$. Each user is under a maximum target error rate constraint of 3.16%, which is equivalent to the 5G URLLC reliability constraint of 99.9999% (over 1 ms). And the power of the inter-cell interference equals the receiver noise floor.

Fig. 7 provides the latency performance comparison of four different policies over the measured channels and simulated i.i.d. Rayleigh fading channels. The blue lines are the optimal array-latency curves under the proposed joint reliability and transmission rate adaptation, which is obtained by Algorithm 1. The red lines are the proposed low-complexity LYRRC (23), which was discussed in Section IV. The green colored lines capture the latency under optimal transmission rate adaptation but fixed reliability (target error rate of 10%). And the black lines are the latencies of fixed reliability (10% target error rate) and transmission rate adaptation under a peak power constraint, which is currently deployed in LTE and Wi-Fi systems.

Over measured and simulated channels, the proposed joint control (blue and red lines) clearly provides better latency performance than the two fixed-reliability counterparts. Allowing target error rate to be adaptive on the number of antennas M turned out to reduce the latency significantly. Compared to



(a) $\tau = 2$, Measured Channel

(b) $\tau = 4$, Measured Channel

(c) $\tau = 2$, i.i.d. Rayleigh Fading

(d) $\tau = 4$, i.i.d. Rayleigh Fading

Fig. 7. The solved latency under four different policies over measured and simulated 4-user channels. Algorithm 1 generated policy and LYRRC (23) are labeled by blue and red,respectively. Green lines is the policy of fixed reliability (target error rate) and power adaptation based on queue-length. The peak power constrained policy with fixed reliability is in black. The traffic arrival rate is 5 packets per frame, each of size 52-bits per OFDM symbol. The pilots power is 20 dBm. The average power constraint is 20 dBm with large-scale fading of -10 dBm.

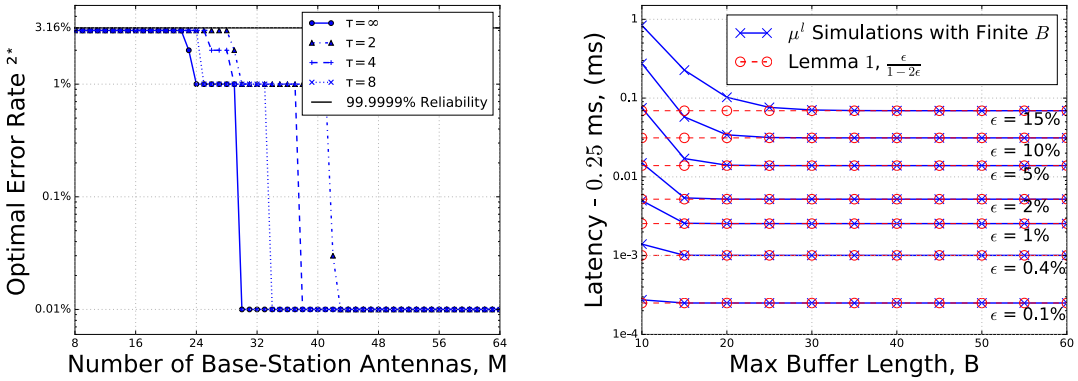
the fixed target error rate with peak power control, a $400\times$ latency reduction is observed when $M > 30$. Additionally, when M is larger than 30, we find that the proposed joint control can provide a $20\times$ latency reduction compared to the state-of-the-art control that fixes target error rate and adapts transmission rate [20]–[23] (based on the number of antennas and queue length). The large-array asymptotic latency-optimal control, LYRRC, turned out to be near latency-optimal when M is larger than 30. Finally, we find policies that fixed target error rate at 10% leads to at least 5 ms latency and cannot satisfy the URLLC latency requirement.

Fig. 7 captures the influence of imperfect channel state information on latency. For a multiuser uplink system, the inter-beam interference (30) reduces with the number of pilots τ. And achieving the same target error rate becomes more power expensive with larger inter-beam interference. Therefore, over measured and simulated channels, the latency increases as τ reduces.

Fig. 7 also demonstrates that the spatial correlation of the base-station antennas reduces the minimum achievable latency. With the same number of pilots τ, a lower latency is observed in i.i.d. Rayleigh fading channels than that in measured channels. The increased latency can be explained by the reduced system capacity from spatial correlation [11], [12]. We further remark that LYRRC achieves near optimal latency performance over both measured and simulated channels when $M > 36$.

We now comment on the optimal target error rate that minimizes the latency. Fig. 8a describes the latency-optimal target error rate obtained during solving the latency minimization problems in Fig. 7. The latency-optimal target error rate increases as τ reduces due to less accurate channel estimation, which agrees with LYRRC. Additionally, due to the reliability constraint, the solved latency-optimal target error rates satisfy the 5G reliability requirement (target error rate of 3.16%).

Finally, we use simulations to verify our structural analysis in Section IV. Fig. 7 confirms that LYRRC (23) is near latency-optimal for M larger than a finite number of 38. One technical contribution independent of the massive MIMO system is a simple transmission rate adaptation μ^l as $\min(q,\ 2\lambda)$, which is referred to as "rule of double" and is part of LYRRC. Lemma 1 captures that, when buffer size $B \to \infty$, the resulted latency by using μ^l and a target error rate $\epsilon < 0.5$ is $1 + \frac{\epsilon}{1-2\epsilon}$. Fig. 8b shows the resulted latency by using μ^l with a finite buffer size. The (large-buffer) asymptotic latency turned out to accurately approximate the system latency when B is larger than 30. And as the target reliability increases (target error rate reduces), buffer overflow is less likely to happen and the latency approximation in Lemma 1 becomes increasingly accurate.



(a) Latency-optimal Target Error Rate and Number of Base-Station Antennas M

(b) Latency under μ^l (23) with finite buffer length.

Fig. 8. Fig. 8a shows the computed error rate that provides minimum latency in the measured channels. And the resulted minimum latencies are shown in Fig. 7 (in blue). Fig. 8b verifies the latency characterization under "rule of double" in Lemma 1.

VII. Conclusion

In this work, we study the latency-optimal cross-layer control over wideband massive MIMO channels. By identifying a tradeoff between queueing and retransmission latency, we find that a lower physical layer target error rate does not always guarantee lower latency. We present algorithms that generate the optimal target error rate and transmission rate policies. We show that to achieve the minimum latency, the target error rate can no longer be considered fixed and needs to be adapted based on the number of base-station antennas, channel estimation accuracy, and the traffic arrival rate. Our results also demonstrate that massive MIMO systems have the potential to achieve both high reliability and low latency and are a promising candidates of 5G URLLC.

Appendix A
Proof of Theorem 1

We use a per packet argument. Since infinite buffer is assumed in this section, no packet is dropped and *all* packets will be successfully received with a variable number of transmissions due to the potential channel-induced error. For any target error rate ϵ, let r be the average number of retransmissions. The sum of the retransmission latency and transmission time equals

$$1+\sum_{r=0}^{\infty}\text{Prob}\left(r\right)r=1+\sum_{r=0}^{\infty}r\left(1-\epsilon\right)\epsilon^{r}=1+\frac{\epsilon}{1-\epsilon}, \tag{34}$$

which is a lower bound of the total latency because the queueing latency is ignored. To finish the proof, we now lower bound ϵ under the long-term power constraint P. Under the steady state, the average transmission rate equals to the packet arrival rate, i.e.,

$$\lambda=E_{\pi}\left[r\left(1-\epsilon\right)\right]=E_{\pi}\left[r\right]\left(1-\epsilon\right). \tag{35}$$

The power function (13) is convex on r. We can apply Jensen's inequality and (20) to obtain a lower bound for the average transmission power as

$$\begin{aligned}P&=E_{\pi}\left[p\left(r,\epsilon,\gamma\right)\right]\\&\geq\left\{\frac{\gamma F_{\eta}^{-1}\left(\epsilon\right)}{\exp\left[\left(\frac{\rho}{1-\epsilon}-1\right)\log M\right]}-\frac{\gamma}{1+\gamma p_{\tau}\tau}\right\}^{-1},\end{aligned} \tag{36}$$

Function F_{η}^{-1} is an inverse CDF and is non-decreasing. From (36), the ϵ is lower bounded as

$$F_{\eta}^{-1}\left(\epsilon\right)\geq M^{-\left[1-\rho/\left(1-\epsilon\right)\right]}\left(\frac{1}{\gamma P}+\frac{1}{\gamma p_{\tau}\tau}\right).$$

Using the monotonicity of the CDF, a lower bound on the target error rate ϵ is then

$$\epsilon\geq F_{\eta}\left[\frac{1}{M^{\left(1-\rho\right)}}\left(\frac{1}{\gamma P}+\frac{1}{\gamma p_{\tau}\tau}\right)\right]. \tag{37}$$

We finish the proof by combining (37) and (34).

Appendix B
Proof of Lemma 1

We compute the queueing latency by considering the steady state. Under transmission rate adaptation μ^l, the buffer length process (7) is rewritten as $q_{t+1}=\max\left[q_t+\left(1-2\ 1_t\right)\lambda,\lambda\right]$. The buffer length process under μ^l thus constitutes a Markov chain with *countably infinite states* [38]. The distribution of 1_t is determined by target error rate ϵ as $\text{Prob}\left(1_t=1\right)=\epsilon$ and $\text{Prob}\left(1_t=0\right)=1-\epsilon$. The state transition is shown in Fig. 5. Denote the steady state distribution of the buffer length as π_q. We then have that

$$\begin{aligned}\pi_{\lambda}&=\left(1-\epsilon\right)\pi_{\lambda}+\left(1-\epsilon\right)\pi_{2\lambda}\\\pi_{i\lambda}&=\epsilon\pi_{(i-1)\lambda}+\left(1-\epsilon\right)\pi_{(i+1)\lambda},\quad i\geq 2,\end{aligned}$$

where $\sum_{i=0}^{N}\pi_{i\lambda}=1$. The steady state distribution is then computed as

$$\pi_{i\lambda}=\left(1-\frac{\epsilon}{1-\epsilon}\right)\left(\frac{\epsilon}{1-\epsilon}\right)^{i-1},\quad i=1,2,\ldots. \tag{38}$$

Using (38), the average latency is then computed as

$$\begin{aligned}\frac{1}{\lambda}E_{\pi_q}[q] &= \frac{1}{\lambda}\left(\sum_{i=1}^{\infty}\pi_{i\lambda}i\lambda\right)\\ &= \sum_{i=1}^{\infty}\left(\frac{\epsilon}{1-\epsilon}\right)^{i-1}i-\sum_{i=1}^{\infty}\left(\frac{\epsilon}{1-\epsilon}\right)^{i}i\\ &= 1+\frac{\epsilon}{1-2\epsilon},\end{aligned}$$

which completes the proof.

Appendix C
Proof of Theorem 2

We characterize the gap between latency under LYRRC as

$$\begin{aligned}D_{\mathrm{LYRRC}}-D^*(M) &= (D_{\mathrm{LYRRC}}-1)-(D^*-1)\\ &\le \frac{\epsilon_o}{1-2\epsilon_o}-\frac{\epsilon_o}{1-\epsilon_o}\\ &= \frac{(\epsilon_o)^2}{(1-2\epsilon_o)(1-\epsilon_o)},\end{aligned}\tag{39}$$

where the last step is obtained via applying Theorem 1 and (37). Equ. (39) provides the characterization of the latency gap. To finish the proof, it is sufficient to show that the average power constraint P is satisfied under the large-array simple control.

With utilization factor ρ (20), the packet arrival rate scales as $\lambda=(\rho N\log M)/L$. Using the per-frame power (13) and the definition of ϵ_o (23), the transmission power with rate r is

$$P^{\epsilon_o}(r)=\left[\left(\frac{1}{P}+\frac{\gamma}{\tau\gamma p_\tau+1}\right)\frac{1}{M^{\rho(r/\lambda-1)}}-\frac{\gamma}{\tau\gamma p_\tau+1}\right]^{-1}.\tag{40}$$

Since we assume empty buffer at time 0 and constant arrival rate of λ, the transmission rates under policy μ^l is either λ or 2λ. Based on the queue length steady state characterization (38), we have that $\mathrm{Prob}(u=\lambda)=\pi_\lambda=1-\frac{\epsilon_o}{1-\epsilon_o}$ and $\mathrm{Prob}(r=2\lambda)=\sum_{i=2}^{\infty}\pi_{i\lambda}=\frac{\epsilon_o}{1-\epsilon_o}$. Conditioning on the rate expression in (40), the average power under LYRRC is

$$\begin{aligned}P^{\epsilon_o,\mu^l} &= \frac{1-2\epsilon_o}{1-\epsilon_o}P^{\epsilon_o}(\lambda)+\frac{\epsilon_o}{1-\epsilon_o}P^{\epsilon_o}(2\lambda)\\ &= \frac{1-2\epsilon_o}{1-\epsilon_o}P+\frac{\epsilon_o}{1-\epsilon_o}P^{\epsilon_o}(2\lambda).\end{aligned}\tag{41}$$

We want to show that the power constraint is satisfied, i.e., $P^{\epsilon_o,\mu^l}\le P$. Using (40), the second power consumption term of (41) is upper bounded as

$$\frac{\epsilon_o}{1-\epsilon_o}P^{\epsilon_o}(2\lambda)\le\epsilon_o P^{\epsilon_o}(2\lambda)\le\epsilon_o M^{\rho}.\tag{42}$$

Therefore, the sufficient condition (41) is equivalent to

$$\lim_{M\to\infty}\epsilon_o\exp(\rho\log M)=\lim_{M\to\infty}\epsilon_o M^{\rho}=0.\tag{43}$$

Before proving (43), we first present an upper bound of ϵ_o. The effective channel gain η (12) is the average of N i.i.d. random variables $\log\kappa$. For $x<0$, we thus have an upper bound as

$$\begin{aligned}F_\eta(x) &= F_{\sum_{n=1}^{N}\log\kappa_n}(Nx)\le F_{\log\kappa}(Nx)\\ &= Pr(\kappa\le\exp(Nx)),\end{aligned}\tag{44}$$

where the last step is by the definition of CDF. We now upper-bound (44) as the follows.

$$\begin{aligned}F_\eta(x) &\le Pr(\kappa-E[\kappa]\le\exp(Nx)-E[\kappa])\\ &\le Pr(|\kappa-E[\kappa]|\ge E[\kappa]-\exp(Nx)).\end{aligned}$$

Here, the last term denotes the probability that κ has a larger deviation (to its mean) than $E[\kappa]-\exp(Nx)$. Using Chebyshev's Inequality, a new upper-bound is obtained as

$$\begin{aligned}F_\eta(x) &\le \frac{\mathrm{Var}[\kappa]}{(E[\kappa]-\exp(Nx))^2}\\ &= \frac{1}{M}\frac{1}{(\frac{\tau p_\tau\gamma}{1+\tau p_\tau\gamma}-\exp(Nx))^2}\left(\frac{\tau p_\tau\gamma}{1+\tau p_\tau\gamma}\right)^2\\ &= O\left(\frac{1}{M}\right),\end{aligned}\tag{45}$$

where the last step is by conditions (18) and (19). By the definition of ϵ_o, using the above upper bound proves (43) and completes the proof.

Appendix D
Proof of Theorem 4

The multi-user mapping (31) can be viewed as a scaled version of (13) when $\tau=\infty$. Recall that the proof of Theorem 1 and Lemma 1 is independent of the distribution of the per-antenna gain κ_n. To complete the proof, we only need to prove the mulituser version of Theorem 2 by following the same derivations as in Appendix C. As the proof from (39) to (44) is also independent to the distribution of κ_n, we finish the proof by proving that (30) satisfies (45). By the multiuser setup in Section V, κ_n (30) equals $\frac{\tau p_\tau[k]\gamma[k]}{\tau p_\tau[k]\gamma[k]+1}\frac{1}{M[\mathbf{W}^{-1}]_{kk}}$, where $\mathbf{W}$ is a $K\times K$ central complex Wishart matrix with M degrees of freedom and covariance matrix of $\mathbf{I}$. Since $\frac{\tau p_\tau[k]\gamma[k]}{\tau p_\tau[k]\gamma[k]+1}$ is a positive constant, we only need to capture the mean and variance of $\frac{1}{M[\mathbf{W}^{-1}]_{kk}}$to verify (45). We first check the mean condition by Jensen's inequality as $E\left[\frac{1}{M[\mathbf{W}^{-1}]_{kk}}\right]\ge\frac{1}{E\left[M[\mathbf{W}^{-1}]_{kk}\right]}$. Using the first moments of inverse Wishart [39] gives that

$$E\left[M\left[\mathbf{W}^{-1}\right]_{kk}\right]=\frac{1}{K}E\left[M\,\mathrm{tr}\left(\mathbf{W}^{-1}\right)\right]=\frac{M}{M-K}.\tag{46}$$

Therefore, the per-antenna gain is lower bounded by a constant as $M\to\infty$. Recall that the κ_n in systems with perfect channel case serves as an upper bound. In the upper bound case, the per-antenna gain expectation is 1 as $M\to\infty$. By random matrix theory [39], the variance of the trace of inverse Wishart satisfies

$$\begin{aligned}\mathrm{Var}\left[\mathrm{tr}\left(\mathbf{W}^{-1}\right)\right] &= E\left\{\left[\mathrm{tr}\left(\mathbf{W}^{-1}\right)\right]^2\right\}-E\left[\mathrm{tr}\left(\mathbf{W}^{-1}\right)\right]^2\\ &= \frac{MK}{\left((M-K)^2-1\right)(M-K)^2}.\end{aligned}$$

Using Taylor's expansion, we complete the proof by checking the variance as

$$\mathrm{Var}\left[\frac{1}{M\left[\mathbf{W}^{-1}\right]_{kk}}\right] = \frac{\mathrm{Var}\left[M\,\mathrm{tr}\left(\mathbf{W}^{-1}\right)\right]_{kk}}{E\left[\frac{1}{M\left[\mathbf{W}^{-1}\right]_{kk}}\right]^4} + o\left(\frac{1}{M}\right) = O\left(\frac{1}{M}\right),\ M \to \infty.$$

REFERENCES

[1] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[2] C. Shepard *et al.*, "Argos: Practical many-antenna base stations," in *Proc. 18th Annu. Int. Conf. Mobile Comput. Netw. (Mobicom)*. New York, NY, USA: ACM, 2012, pp. 53–64, .

[3] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[4] G. Caire, "On the ergodic rate lower bounds with applications to massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3258–3268, May 2018.

[5] *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures*, document TS 36.213, 3GPP, 2019.

[6] P. Wu and N. Jindal, "Coding versus ARQ in fading channels: How reliable should the PHY Be?" *IEEE Trans. Commun.*, vol. 59, no. 12, pp. 3363–3374, Dec. 2011.

[7] M. Lauridsen, L. C. Gimenez, I. Rodriguez, T. B. Sorensen, and P. Mogensen, "From LTE to 5G for connected mobility," *IEEE Commun. Mag.*, vol. 55, no. 3, pp. 156–162, Mar. 2017.

[8] C.-P. Li, J. Jiang, W. Chen, T. Ji, and J. Smee, "5G ultra-reliable and low-latency systems design," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Jun. 2017, pp. 1–5.

[9] E. Bastug, M. Bennis, M. Medard, and M. Debbah, "Toward interconnected virtual reality: Opportunities, challenges, and enablers," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 110–117, Jun. 2017.

[10] K. K. Mukkavilli *et al.*, "Self-contained time division duplex (TDD) subframe structure for wireless communications," U.S. Patent 20 160 270 115, Sep. 15. 2016, .

[11] T. L. Marzetta and H. Yang, *Fundamentals of Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[12] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO Networks: Spectral, energy, and hardware efficiency," *FNT Signal Process.*, vol. 11, nos. 3–4, pp. 154–655, 2017.

[13] V. Jungnickel *et al.*, "The role of small cells, coordinated multipoint, and massive MIMO in 5G," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 44–51, May 2014.

[14] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, 2013.

[15] W. Su, S. Lee, D. A. Pados, and J. D. Matyjas, "Optimal power assignment for minimizing the average total transmission power in hybrid–ARQ Rayleigh fading links," *IEEE Trans. Commun.*, vol. 59, no. 7, pp. 1867–1877, Jul. 2011.

[16] X. Lin, N. Shroff, and R. Srikant, "A tutorial on cross-layer optimization in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1452–1463, Aug. 2006.

[17] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks-part I: Theoretical framework," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 614–624, Mar. 2005.

[18] S. Shakkottai, T. S. Rappaport, and P. C. Karlsson, "Cross-layer design for wireless networks," *IEEE Commun. Mag.*, vol. 41, no. 10, pp. 74–80, Oct. 2003.

[19] B. Makki, A. G. I. Amat, and T. Eriksson, "On noisy ARQ in block-fading channels," *IEEE Trans. Veh. Technol.*, vol. 63, no. 2, pp. 731–746, Feb. 2014.

[20] R. Berry and R. Gallager, "Communication over fading channels with delay constraints," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1135–1149, May 2002.

[21] M. J. Neely, "Optimal energy and delay tradeoffs for multiuser wireless downlinks," *IEEE Trans. Inf. Theory*, vol. 53, no. 9, pp. 3095–3113, Sep. 2007.

[22] M. Goyal, A. Kumar, and V. Sharma, "Optimal cross–layer scheduling of transmissions over a fading multiaccess channel," *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3518–3537, Aug. 2008.

[23] J. Cao and E. M. Yeh, "Power-delay tradeoff analysis for communication over fading channels with feedback," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2008, pp. 614–618.

[24] J. Arnau and M. Kountouris, "Delay performance of MISO wireless communications," in *Proc. 16th Int. Symp. Modeling Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)*, May 2018, pp. 1–8.

[25] J. Hoydis, S. Ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. Select. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.

[26] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[27] P. Wu and N. Jindal, "Performance of hybrid-ARQ in block-fading channels: A fixed outage probability analysis," *IEEE Trans. Commun.*, vol. 58, no. 4, pp. 1129–1141, Apr. 2010.

[28] S. Schiessl, J. Gross, M. Skoglund, and G. Caire, "Delay performance of the multiuser MISO downlink under imperfect CSI and finite–length coding," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 765–779, Apr. 2019.

[29] R. Jurdi, S. R. Khosravirad, H. Viswanathan, J. G. Andrews, and R. W. Heath, Jr., "Outage of periodic downlink wireless networks with hard deadlines," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1238–1253, Feb. 2019.

[30] D. P. Bertsekas, R. G. Gallager, and P. Humblet, *Data Networks*, vol. 2. Upper Saddle River, NJ, USA: Prentice-Hall, 1992.

[31] *MATLAB Communications Toolbox Release 2015b*, MathWorks, Natick, MA, USA, 2015.

[32] D.-J. Ma, A. Makowski, and A. Shwartz, "Estimation and optimal control for constrained Markov chains," in *Proc. 25th IEEE Conf. Decision Control*, Dec. 1986, pp. 994–999.

[33] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 1. Belmont, MA, USA: Athena Scientific, 2007.

[34] X. Gao, O. Edfors, F. Rusek, and F. Tufvesson, "Massive MIMO performance evaluation based on measured propagation data," *IEEE Trans. Wireless Commun.*, vol. 14, no. 7, pp. 3899–3911, Jul. 2015.

[35] X. Du and A. Sabharwal, "Shared angles-of-departure in massive MIMO channels: Correlation analysis and performance impact," *IEEE Trans. Wireless Commun.*, to be published.

[36] E. Bjornson, J. Hoydis, and L. Sanguinetti, "Massive MIMO has unlimited capacity," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 574–590, Jan. 2018.

[37] *Google Map*. Accessed: Nov. 27, 2018. [Online]. Available: https://www.google.com/maps/@29.7201813,-95.3994342,88m/data=!3m1!1e3

[38] R. G. Gallager, *Discrete Stochastic Processes*, vol. 321. Berlin, Germany: Springer, 2012.

[39] A. M. Tulino and S. Verdú, "Random matrix theory and wireless communications," *FNT Commun. Inf. Theory*, vol. 1, no. 1, pp. 1–182, 2004.

Xu Du (Member, IEEE) received the B.E. degree in information technology from Zhejiang University, China, in 2013. He received the M.S. and Ph.D. degrees in electrical and computer engineering from Rice University, Houston, TX, USA, in 2015 and 2019, respectively. He is currently a Research Scientist with Facebook Inc., Menlo Park, CA, USA. His research interests include computer system optimization, information theory, and algorithm design for quality of service enhancement.

Yin Sun received the B.Eng. and Ph.D. degrees in electronic engineering from Tsinghua University, in 2006 and 2011, respectively. He was a Post-Doctoral Scholar and a Research Associate with The Ohio State University from 2011 to 2017. He is an Assistant Professor with the Department of Electrical and Computer Engineering, Auburn University, Alabama. His research interests include information freshness, networking, information theory, and machine learning. He is the Co-Founder of the Age of Information Workshop. He has coauthored a book *Age of Information: A New Metric for Information Freshness* and *Synthesis Lectures on Communication Networks* (Morgan & Claypool Publishers, 2019). He has coauthored articles that received the Best Student Paper Award from the IEEE WiOpt 2013 and the Best Paper Award from the IEEE WiOpt 2019. He is a Guest Editor of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS for the special issue on age of information in real-time systems and networks.

Ness B. Shroff received the Ph.D. degree in electrical engineering from Columbia University in 1994. He joined Purdue University immediately thereafter as an Assistant Professor with the School of Electrical and Computer Engineering. At Purdue University, he became a Full Professor of electronics and communication engineering and the Director of the University-Wide Center on Wireless Systems and Applications in 2004. In 2007, he joined The Ohio State University, where he is the Ohio Eminent Scholar Endowed Chair of networking and communications with the Department of Electronics and Communication engineering and the Department of Computer Science and Engineering. He holds or has held visiting (chaired) professor positions at Tsinghua University, Beijing, China, Shanghai Jiaotong University, Shanghai, China, and IIT Bombay, Mumbai, India. He currently serves as the Steering Committee Chair for ACM Mobihoc. He serves as the Editor-at-Large for the IEEE/ACM TRANSACTIONS ON NETWORKING. He has received numerous best paper awards for his research and is listed in Thomson Reuters' on The World's Most Influential Scientific Minds. He is noted as a Highly-Cited Researcher by Thomson Reuters. He also received the IEEE INFOCOM Achievement Award for seminal contributions to scheduling and resource allocation in wireless networks.

Ashutosh Sabharwal (Fellow, IEEE) research interests include wireless theory, protocols, and open-source platforms. He is the Founder of the WARP Project (warp.rice.edu), an open-source project that was used at more than 125 research groups worldwide and used by more than 500 research articles. He received the 2017 Jack Neubauer Memorial Award, the 2018 Advances in Communications Award, the 2019 ACM Sigmobile Test-of-Time Award, and the 2019 Mobicom Best Community Contribution Paper Award.