# LARGE DEVIATION DELAY ANALYSIS OF QUEUE-AWARE MULTI-USER MIMO SYSTEMS WITH TWO TIMESCALE MOBILE-DRIVEN FEEDBACK

*Junting Chen and Vincent K. N. Lau*

Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology, Hong Kong

## ABSTRACT

Multi-user multi-input-multi-output (MU-MIMO) systems usually require users to feedback the channel state information (CSI) for scheduling. Most of the existing literature on the reduced feedback user scheduling focused on the throughput performance and the queueing delay was usually ignored. As the delay is important for real-time applications, it is desirable to have a low feedback queue-aware user scheduling algorithm for MU-MIMO systems. This paper proposes a two timescale queue-aware user scheduling algorithm, which consists of a queue-aware mobile-driven feedback filtering stage and a SINR-based user scheduling stage. The feedback policy is obtained by solving a queue-weighted optimization problem. In addition, we evaluate the associated queueing delay performance by using the large deviation analysis. The large deviation decay rate for the proposed algorithm is shown to be much larger than the CSI-only scheduling algorithm. Numerical results demonstrate the large performance gain of the proposed algorithm over the CSI-only algorithm, while the proposed one requires only a small amount of feedback.

***Index Terms*—** MU-MIMO, Limited Feedback, Queue-aware, Large Deviation, Random Beamforming

## 1. INTRODUCTION

Multi-user MIMO (MU-MIMO) systems transmit multiple streams of data to a group of users simultaneously by exploiting the spatial degrees of freedom among users. In a cellular network, where the BS equips with $M$ antennas and serves $K$ users, a sum rate of $M \log\log K$ can be achieved at the expense of requiring full *channel state information* (CSI) at the BS [1, 2]. To reduce the CSI feedback overhead, a threshold based feedback control has been proposed in [3–5], where a sum rate capacity $\mathcal{O}(M \log\log K)$ can be achieved requiring only a portion of users feeding back to the BS [3].

While a high throughput is desired in MU-MIMO systems, the delay performance is also crucial for real-time applications. A good user scheduling policy should strike a balance between the throughput and delay. As the CSI indicates *good opportunity to transmit* whereas the *Queue State Information* (QSI) indicates the *urgency* of the data flow, the delay-aware system should incorporate both the CSI and QSI in the user scheduling. Yet, it is challenging to design such a delay-aware MU-MIMO system, as it involves solving *queue-dependent* stochastic optimization problems. Moreover, the delay analysis of the queue *state-dependent* buffer dynamics is very difficult.

In this paper, we consider a MU-MIMO downlink system with a $M$-antenna BS and $K$ multi-antenna mobile users. There are $K$ bursty data flows to each of the $K$ mobiles from the BS. The BS applies random beamforming for MU-MIMO. We propose a two timescale delay-aware user scheduling policy, which consists of a mobile-driven feedback filtering stage (QSI timescale) and a user scheduling stage at the BS (CSI timescale). As a result, *urgent users* are given higher priority to feedback in order to reduce the total feedback overhead. We show that our design can achieve throughput optimality over all the two timescale policies and it only has logarithmic complexity over the number of users $K$. In addition, using the large deviation theory [6], we derive the asymptotic exponential decay rate for the tail probability of the worst case queue in the system. Specifically, we show that the asymptotic decay rate $-\frac{1}{B}\log(\mathrm{P}\{\max_k\ Q_k\} > B)$ scales as $\mathcal{O}(\log K)$ for the proposed delay-aware user scheduling algorithm, which is substantially better than traditional CSI-only MU-MIMO user scheduling scheme.

## 2. RELATION TO PRIOR WORK

Reduced feedback user scheduling for MU-MIMO systems has been studied in [1–5], where the authors only focused on the throughput performance. In the preliminary works for queue-aware designs, the authors in [7] proposed a maximum queue-weighted sum rate policy with a heuristic greedy-based algorithm, but global CSI from all the users is required. On the other hand, the work [8] studied the degradation of the queue stability region for SDMA due to limited feedback. Yet, stability is only a weak form of delay performance. In this work, we propose a novel queue-aware two timescale algorithm to reduce both the feedback overhead and the queueing delay in the system. In addition, we derive the associated asymptotic queue overflow probability for delay performance evaluation.

## 3. SYSTEM MODEL

### 3.1. MU-MIMO System, Bursty Data Source and Queue Model

We consider a downlink MU-MIMO system with a $M$-antenna BS and $K$ geometrically dispersed mobile users ($K \gg M$). Each mobile user has $N$ receive antennas. The BS transmits $M$ data streams to a group of selected users at each time slot, and random beamforming is used. Let $\mathbf{s}(t) = (s_1(t), \ldots, s_M(t))^T$ be the vector of the transmit symbols for the $M$ data streams, where $\mathbb{E}[s_m s_m^*] = 1$. The receive signal $\mathbf{y}_k \in \mathbb{C}^{N\times 1}$ at the $k$-th user is

$$\mathbf{y}_k(t) = \sum_{m=1}^{M} \sqrt{P} H_k \phi_m s_m(t) + \mathbf{n}_k \qquad \forall k \in \mathcal{A}(t)$$

where $\{\phi_1, \ldots, \phi_M\}$ are $M$ (random) orthonormal beamforming vectors, $H_k \in \mathbb{C}^{N\times M}$ is the zero mean, unit-variance circularly symmetric complex Gaussian channel matrix from the transmitter to

the user $k$, $\mathbf{n}_k \in \mathbb{C}^{N\times 1} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_N)$ is the Gaussian additive noise vector, $P$ is the transmit power at the BS, and $\mathcal{A}(t)$ denotes the set of the scheduled users at time slot $t$. The *effective SINR* of the $i$-th beam on the $n$-th receive antenna of the $k$-th user can be calculated as follows,

$$\text{SINR}_{k,n}^{i} = \frac{\left|H_k^{(n)}\phi_i\right|^2}{\sum_{j,j\neq i}\left|H_k^{(n)}\phi_j\right|^2 + 1/P} \quad (1)$$

where $H_k^{(n)}$ denotes the $n$-th row of the channel matrix $H_k$ of user $k$. By selecting the users with the highest SINR on each beam, the transmitter can support near-orthogonal transmission and exploit multi-user diversity without the global CSI $\{H_k\}$ [9].

We assume the channel matrices $\{H_k\}$ to be in *quasi-static* block fading, where each channel realization $H_k$ remains constant during each time slot, but identically and independently distributed (i.i.d.) across different time slots. The mobile users are assumed to have perfect knowledge of their local CSI $H_k$.

Data arrives in packets randomly for different users. Let $A_k(t)$ denote the number of packets that arrive at the BS for user $k$ during time slot $t$, and $\mathbf{A}(t) = (A_1(t), \ldots, A_k(t))$. We assume that the packet arrival $A_k(t)$ are i.i.d. with respect to (w.r.t.) $t$ and independent w.r.t. $k$ according to a general distribution with mean $\mathbb{E}[A_k(t)] = \lambda_k$. The packet has a fixed length of $L$ bits.

Let $D_k(\mathbf{Q}(t), \mathbf{H}(t))$ represents the amount of departure for user $k$ at time slot $t$, where $\mathbf{Q}(t) = (Q_1(t), \ldots, Q_K(t))$ is the vector of queueing backlogs for all the users and $\mathbf{H}(t) = (H_1(t), \ldots, H_K(t))$. The queueing dynamics for user $k$ is given by $Q_k(t+1) = [Q_k(t) - D_k(\mathbf{Q}(t), \mathbf{H}(t))]^+ + A_k(t)$, where the operator $[\mathbf{.}]^+$ represents $[w]^+ = \max\{0, w\}$. According to Little's Law [10], there is no loss of generality to study the queue length $Q_k$ for the purpose of understanding the delay. The goal of the user scheduling controller is to adjust the channel access opportunity for all the users so that their queue lengths (or delay) are minimized while a reasonable system throughput is maintained.

### 3.2. Two Timescale User Scheduling with Reduced Feedback

Conventional throughput optimal (in stability sense) user scheduling policies such as max-weighted-queue (MWQ) algorithms [11] require global CSI and QSI knowledge. However, the CSI is available at the mobile user side while the QSI is available at the BS. Furthermore, the MWQ policy requires solving a queue weighted sum rate combinatorial optimization problem that results in high complexity. To overcome these challenges, we propose a two timescale user scheduling solution as follows.

**Stage I**: *Queue-aware user-driven feedback filtering*. The BS determines and broadcasts the user feedback probability $\{p_k(\mathbf{Q})\}$ based on the user queueing backlogs $\mathbf{Q}(t)$ for every $T$ time slots. Mobile user $k$ attempts to feedback to the BS in the stage II with probability $p_k$. The motivation behind is to save the feedback cost by reducing the lower priority users from feeding back.

**Stage II**: *Dynamic User Scheduling based on SINR feedbacks*. Each user $k$ measures the effective SINRs on each receive antenna $n$ according to (1), finds the strongest beam $i^*(k,n) = \arg\max_{1\leq i\leq M} \text{SINR}_{k,n}^{i}$ and reports it to the BS according to probability $p_k$. At the transmitter side, for each beam $i$, the BS schedules the user who feeds back the highest SINR on that beam. As a result, the stage II user scheduling exploits the multi-user diversity among the set of feedback users.

The two-stage policy can be implemented on a different timescale. The user selection in stage II is done at every time slot $t$, while the user feedback probability $\{p_k(\mathbf{Q})\}$ determined in stage I can be updated once every $T$ time slots. The update period $T$ trades the performance with the control signaling overhead. With a large $T$, there is a smaller signaling overhead in broadcasting $\{p_k(\mathbf{Q})\}$ but the feedback priority results in being driven by outdated QSI.

Note that, as discussed in Section 5, the user scheduling policy that considers only CSI is just a special case of the proposed two timescale policy, while setting $p_k \equiv 1$ for each user $k$ in stage I.

## 4. THE QUEUE-AWARE USER FEEDBACK FILTERING ALGORITHM

In this section, we derive the low complexity *Feedback Filtering Control Algorithm (FFCA)* to determine the probability $\{p_k(\mathbf{Q})\}$ in stage I. We show that, with the FFCA, the two timescale user scheduling achieves the maximum *queue stability region* [11].

Let $\chi_k \in \{0,1\}$ be the feedback indicator of user $k$. According to the stage I policy, $\chi_k$ follows the Bernoulli distribution with probability $p_k$. Let $J_k^i(\mathbf{H}, \boldsymbol{\chi}) \in \{0,1\}$ denotes a mapping from the user feedback CSI to the user scheduling decision for the $k$-th user on the $i$-th beam. The data rate of user $k$ can be written as $R_k(\mathbf{H}, \boldsymbol{\chi}) = \sum_{i=1}^{M} J_k^i \chi_k \log(1+\gamma_k^i)$, where $\gamma_k^i$ is the feedback SINR of user $k$ on the $i$-th beam. In addition, define the conditional feedback cost $S(\mathbf{Q})$ as $S(\mathbf{Q}) \triangleq \mathbb{E}\left[\sum_k \chi_k | \mathbf{Q}\right] = \sum_k p_k(\mathbf{Q})$. We develop the feedback filtering control algorithm as follows.

*Feedback Filtering Control Algorithm (FFCA):* Observing the current queue length $\mathbf{Q}(t)$, users feedback their CSI according to $\{p_k^*(\mathbf{Q}(t))\}$, where $p_k^*(\mathbf{Q}(t))$ is the solution to the following optimization problem,

$$\max_{\{p_k\}} \quad \mathbb{E}\left[\sum_{k=1}^{K} Q_k(t) R_k(\mathbf{H}, \boldsymbol{\chi}) - V S(\mathbf{Q}(t))\right] \quad (2)$$

where $V$ is a positive constant that trades off the performance and the feedback cost. The following theorem justifies the throughput optimality [11] under the FFCA in (2).

**Theorem 1** (Throughput optimality of the FFCA)**.** *The feedback control $\{p_k^*(\mathbf{Q})\}$ given by FFCA achieves the maximum stability region over any two timescale scheduling in the MU-MIMO system.*

*Proof.* Please refer to [12] for the proof. □

We now derive the solution $p_k^*(\mathbf{Q}(t))$ to (2). Define $\eta_k(S) \triangleq \mathbb{E}\left[R_k(\mathbf{H}, \boldsymbol{\chi}) \middle| \chi_k = 1, \sum_i \chi_i = S\right]$ as the average data rate of user $k$ conditioned on $S$ users feedback to the BS (including user $k$). The following lemma characterizes the property of $\eta_k(S)$.

**Lemma 1** (Data rate under a deterministic feedback cost)**.** *Define*

$$\hat{\eta}_k(S) \triangleq M \int_0^{\infty} \log(1+x) N f(x) F(x)^{NS-1} dx$$

*where $F(x) = 1 - \frac{e^{-x/P}}{(1+x)^{M-1}}$ is the cumulative distribution function (CDF) of $\text{SINR}_{k,n}^{i}$ in (1) and $f(x) = \frac{\partial}{\partial x} F(x)$ is the corresponding probability distribution function (PDF). We have*

$$|\eta_k(S) - \hat{\eta}_k(S)| \leq \left(1 - \frac{e^{-1/P}}{2^{M-1}}\right)^{NS}.$$

*Proof.* Please refer to [12] for the proof. □

Since we typically consider a large number of users, it is reasonable to take $\eta_k(S) \approx \hat{\eta}_k(S)$. The following theorem gives the optimal solution to determine the feedback probability $p_k^*(\mathbf{Q}(t))$ in (2).

**Theorem 2** (Global optimal solution to (2)). *Let $\Pi = \{\pi(1), \ldots, \pi(K)\}$ be a permutation of $\{Q_k\}$ such that $Q_{\pi(1)} \geq Q_{\pi(2)} \geq \cdots \geq Q_{\pi(K)}$. The global optimal solution to (2) is given by*

$$p_{\pi(k)} = 1, \qquad 1 \leq k \leq S^* \tag{3}$$

$$p_{\pi(k)} = 0, \qquad \textit{otherwise} \tag{4}$$

*where $S^*$ is an unique integer that satisfies the following condition*

$$U(S^*) \geq U(S^*+1) \textit{ and } U(S^*) \geq U(S^*-1) \tag{5}$$

*where $S^* \in \{1, \ldots, K\}$ and[1] $U(S) \triangleq \sum_{k=1}^{S} Q_{\pi(k)}\eta_{\pi(k)}(S) - VS$.*

*Proof.* Please refer to [12] for the proof. □

Note that the function $U(S)$ can be understood as the queue-weighted sum rate for the first $S$ prioritized users, with a feedback cost regularization (the last term). Theorem 2 implies that, under a two timescale policy, the best choice is to always let the first $S^*$ queue-length-prioritized users feedback (whose queues are large), while keeping the others silent. Meanwhile, the $S^*$ is chosen to maximize the utility $U(S)$. Condition (5) guarantees that we can use a bisection algorithm to search the optimal $S^*$ by evaluating $U(S)$ at most $\log_2(K)$ times. This suggests that only logarithmic complexity is required to solve the FFCA.

## 5. LARGE DEVIATION DELAY ANALYSIS FOR THE WORST CASE USER

In this section, we will study the queueing delay performance of the proposed queue-aware two timescale scheduling policy and illustrate the performance gain over the CSI-only baseline policy. We are interested in the steady state distribution of the worst case queue length, i.e.,

$$\lim_{t\to\infty} \Pr(\max_{1\leq k\leq K} Q_k(t) > B)$$

where $B$ is the buffer size. We denote $Q_{\max}(t) = \max_k Q_k(t)$ as the maximum queue length process and $Q_{\max}(\infty)$ as the steady state of the $Q_{\max}(t)$.

Note that to derive the exact distribution function of $Q_{\max}(\infty)$ is almost impossible due to the complex coupling of the queueing dynamics in the MU-MIMO system. On the other hand, as one is usually concerned about the overflow probability of the queueing system, we can only focus on the asymptotic overflow probability of $Q_{\max}(\infty)$ over an increasing buffer size $B$. Using the large deviation approach [13, 14], this property can be characterized by the large deviation decay rate function $I^*$ defined as follows,

$$I^* \triangleq \lim_{B\to\infty} -\frac{1}{B}\log \Pr\left(Q_{\max}(\infty) > B\right). \tag{6}$$

Using the decay rate function $I^*$ in (6), the queue overflow probability can be written as $\Pr(Q_{\max}(\infty) > B) = e^{-I^*B+o(B)}$, and the exponent $I^*$ characterizes the tail distribution of the worst case queue length $Q_{\max}(\infty)$. It illustrates how quickly the overflow probability drops when the buffer size $B$ grows. A larger decay rate $I^*$ corresponds to a better performance of the scheduling algorithm in the sense of reducing the worst case delay $Q_{\max}$ in the system. In the following, we shall derive the decay rate function $I^*$ for the proposed queue-aware policy and compare it with the CSI-only scheduling policy.

The CSI-only baseline algorithm assumes that each user $k$ feeds back the SINR for the $i^*(k,n)$-th beam on each antenna $n$, where $i^*(k,n) = \arg\max_{1\leq i\leq M} \text{SINR}_{k,n}^i$. Then for each beam $i$, the BS schedules the user who has the highest SINR. The CSI-only baseline scheme corresponds to a special case of the proposed two timescale user scheduling by setting $p_k \equiv 1$ for each user $k$ in stage I.

Consider a special case where the arrivals $A_k$ follow Poisson distributions with parameter $\lambda$. We first characterize the decay rate $I^*$ for the CSI-only algorithm in the following theorem.

**Theorem 3** (Decay rate for the CSI-only algorithm). *Let $\mu_b = \frac{M\log(P\log NK)}{KL}$ and $\lambda_T = \lambda K$. Suppose $\lambda < \mu_b$. The large deviation decay rate of $Q_{\max}(\infty)$ under the CSI-only baseline algorithm is given by*

$$I^*_{\text{baseline}} \approx \log \frac{M\log\left(P\log NK\right)}{\lambda_T L}. \tag{7}$$

*Proof.* Please refer to [12] for the proof. □

We can observe that, under a fixed total arrival rate $\lambda_T$, the CSI-only baseline algorithm has an increasing decay rate $I^* = \mathcal{O}(\log\log\log K)$ with the number of users $K$. This explains the multi-user diversity gain achieved by the CSI-only algorithm.

Similarly, we obtain the following results for the proposed two timescale user scheduling algorithm.

**Theorem 4** (Decay rate for the proposed algorithm). *Let $\mu_0 = \inf_{x\in[0,1]} \mu_p(x)$ and $\lambda_T = \lambda K$, where $\mu_p(x) = \frac{M\log\left(P\log N\hat{S}^*(x)\right)}{L\hat{S}^*(x)}$, $\hat{S}^*(x) = \frac{1}{N}\exp\left(W(\frac{MNx}{V})\right)$ and $W(x)$ is the Lambert W function [15] (defined as $W(x)e^{W(x)} = x$). Suppose $\lambda < \mu_0$. The large deviation decay rate of $Q_{\max}(\infty)$ under the two timescale user scheduling algorithm can be expressed as*

$$I^*_{\text{prop}} \geq (1-\epsilon_s)\log K + \log\frac{M}{\lambda_T L} + \epsilon_s \log R_0 + C \tag{8}$$

*where $\epsilon > 0$ is a small constant, $R_0 = \int_0^1 \log(1+Px)\, dF(x)$, and $C = \int_\epsilon^1 \left\{\log\left[N\log\left(PW\left(\frac{MNx}{V}\right)\right)\right] - W\left(\frac{MNx}{V}\right)\right\} dx$.*

*Proof.* Please refer to [12] for the proof. □

The above result suggests that $I^*_{\text{prop}} = \mathcal{O}(\log K)$ and $I^*_{\text{prop}} \gg I^*_{\text{baseline}}$ for a large number of users $K$. Recall the tail distribution of the maximum queue length $\Pr(Q_{\max}(\infty) > B) = e^{-I^*B+o(B)}$. It implies that the proposed algorithm enjoys a significantly smaller overflow probability, and hence, a smaller worst case delay. This demonstrates the importance of utilizing the queue information for user scheduling, and the queue-aware algorithm benefits more from exploiting the multiuser diversity.

In addition, both of the schemes exploit the gain from the MU-MIMO technology. Note that the large deviation decay rates $I^*_{\text{prop}}$ and $I^*_{\text{baseline}}$ derived here both scale as $\mathcal{O}(\log M \log\log N)$ with the number of data streams $M$ and receive antennas $N$. This scaling property shows the same growth order w.r.t. $M$ and $N$ as that of the MIMO broadcast channel capacity [2].

[1] For mathematical convenience, we define $U(0) = U(K+1) = -\infty$.

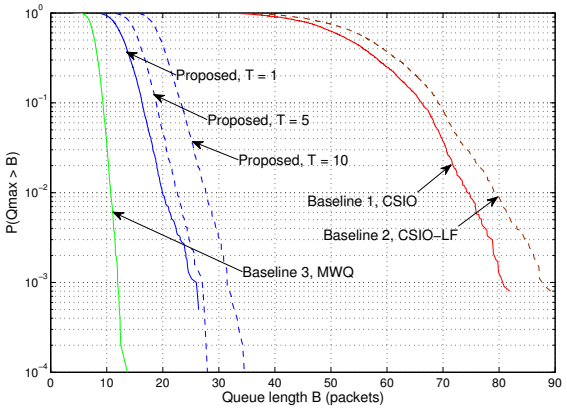


**Fig. 1**. The overflow probability for the worst case queue $\Pr\left(Q_{\max}(\infty) > B\right)$ versus the buffer size $B$.

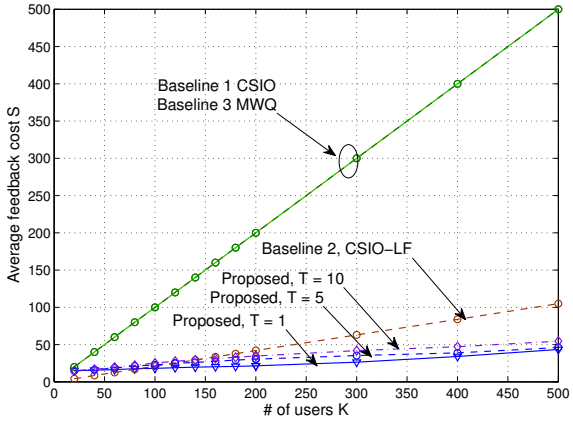


**Fig. 2**. The average feedback cost $\overline{S}$ versus the number of users $K$.

## 6. NUMERICAL RESULTS

In this section, we simulate the queueing delay performance of the proposed two timescale user scheduling algorithm. We consider a MU-MIMO system with $K$ users, and packets arrive to the queue of each user according to a Poisson distribution with rate $\lambda = \lambda_T/K$, where the total arrival rate is $\lambda_T = 7500$ packets/second. Each packet has $L = 8000$ bits. The system bandwidth is 10 MHz and the SNR is 10 dB. The number of transmit and receive antennas are $M = 4$ and $N = 2$, respectively. The scheduling time slot is $\tau = 1$ ms and the simulation is run over $T_{tot} = 100$ seconds.

We compare the performance of proposed algorithm against the following reference baselines. **Baseline 1: CSI-only user scheduling (CSIO)** [4]. At each time slot, all the users feedback the CSI to the BS, and the BS schedules a set of users, each of whom has the highest SINR on the respective beam. **Baseline 2: CSI-only user scheduling with limited feedback (CSIO-LF)** [4]. The scheme is similar to baseline 1 except that the user feeds back to the BS only when its SINR exceeds a threshold $t_{SINR} = 2$ dB. **Baseline 3: Max weighted queue user scheduling (MWQ)** [11]. At each time slot, all the users feedback their CSI to the BS, and the BS selects a set of users so that the instantaneous queue-weighted sum rate $\sum Q_k R_k$ is maximized. Note that the associated user scheduling problem in baseline 3 has much higher complexity for user scheduling and feedback cost. Hence, baseline 3 serves for performance benchmarking purpose only.

Fig. 1 shows the overflow probability for the worst case queue $\Pr\left(Q_{\max}(\infty) > B\right)$ versus the buffer size $B$, under $K = 40$ users. The feedback policy $\{p_k\}$ updates on every $T = 1, 5, 10$ time slots. The proposed scheme significantly outperforms over baselines 1 - 2. It also has similar performance as baseline 3. Fig. 2 demonstrates the average feedback cost $\overline{S}$ (defined as the average number of users feedback to the BS at each time slot) versus the number of users $K$. The feedback cost of the proposed scheme is less than those of all the baselines. Note that although baseline 3 has a smaller worst case queue, it requires all the users feedback to the BS.

Fig. 3 shows the large deviation decay rate over the number of users. The decay rates $I^*$ in (6) are evaluated at buffer size $B_{0.05}$, where the overflow probability $\Pr\left(Q_{\max}(\infty) > B_{0.05}\right) = 0.05$.

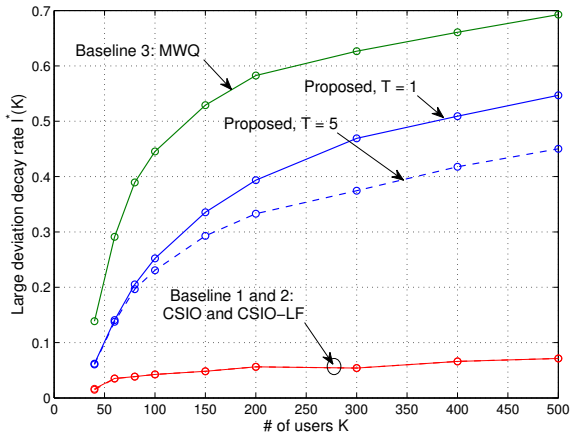


**Fig. 3**. The large deviation decay rate over the number of users.

The decay rate for the proposed scheme grows much faster than those of baselines 1 - 2 with the number of users $K$, demonstrating a better scalability for a large number of users.

## 7. CONCLUSIONS

In this paper, we proposed a novel two timescale delay-aware user scheduling algorithm for the MU-MIMO system. The policy consists of a queue-aware mobile-driven feedback filtering stage and a dynamic SINR-based user scheduling stage. The queue-aware feedback filtering control algorithm in stage I was derived through solving an optimization problem. Under the proposed two timescale user scheduling algorithm, we also evaluated the queueing delay performance for the worst case user using the large deviation analysis. The large deviation decay rate for the proposed algorithm, scaled as $\mathcal{O}\left(\log K\right)$, was shown to be much larger than a CSI-only user scheduling algorithm, which means that the proposed scheme performs better in reducing the worst case delay. The numerical results demonstrated a significant performances gain over the CSI-only algorithm and a huge feedback reduction over the MWQ algorithm.

## 8. REFERENCES


[1] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 528–541, 2006.

[2] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channels with partial side information," *IEEE Transactions on Information Theory*, vol. 51, no. 2, pp. 506–522, 2005.

[3] A. Bayesteh and A. Khandani, "On the user selection for MIMO broadcast channels," *IEEE Transactions on Information Theory*, vol. 54, no. 3, pp. 1086–1107, 2008.

[4] W. Zhang and K. Letaief, "MIMO broadcast scheduling with limited feedback," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 7, pp. 1457–1467, 2007.

[5] S. Sanayei and A. Nosratinia, "Opportunistic downlink transmission with limited feedback," *IEEE Transactions on Information Theory*, vol. 53, no. 11, pp. 4363–4372, 2007.

[6] A. Shwartz, A. Weiss, and R. Vanderbei, *Large deviations for performance analysis*. Citeseer, 1995, vol. 107.

[7] F. She, W. Chen, H. Luo, and D. Yang, "Joint queue control and user scheduling in MIMO broadcast channel under zero-forcing multiplexing," *International Journal of Communication Systems*, vol. 22, no. 12, pp. 1593–1607, 2009.

[8] K. Huang and V. Lau, "Stability and delay of zero-forcing SDMA with limited feedback," *IEEE Transactions on Information Theory*, vol. 58, no. 10, pp. 6499 – 6514, Oct 2012.

[9] J. Chung, C. Hwang, K. Kim, and Y. Kim, "A random beamforming technique in MIMO systems exploiting multiuser diversity," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 5, pp. 848–855, 2003.

[10] J. D. C. Little, "A proof for the queuing formula: L= $\lambda$ w," *Operations Research*, vol. 9, no. 3, pp. 383–387, May 1961.

[11] M. Neely, E. Modiano, and C. Rohrs, "Dynamic power allocation and routing for time-varying wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 1, pp. 89–103, 2005.

[12] J. Chen and V. K. N. Lau, "Large deviation delay analysis of queue-aware multi-user mimo systems with multi-timescale mobile-driven feedback," *submitted to IEEE Transactioins on Signal Processing*, 2012. [Online]. Available: http://arxiv.org/abs/1211.0779

[13] A. Weiss, *Large deviations for performance analysis: queues, communications, and computing*. Chapman & Hall/CRC, 1995.

[14] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*. Springer Verlag, 2009, vol. 38.

[15] R. Corless, G. Gonnet, D. Hare, D. Jeffrey, and D. Knuth, "On the lambertw function," *Advances in Computational mathematics*, vol. 5, no. 1, pp. 329–359, 1996.