CoSchd: Coordinated Scheduling with Channel- and Load-Awareness for Alleviating Cellular Congestion

Huasen Wu, *Member, IEEE*, Xiaojun Lin, *Senior Member, IEEE*, Xin Liu, *Member, IEEE*, Kun Tan, *Member, IEEE*, and Yongguang Zhang, *Fellow, IEEE*

***Abstract*—Although cellular networks can be provisioned according to the peak demand, they usually experience large fluctuations in both channel conditions and traffic load level. Scheduling with both channel- and load-awareness allows us to exploit the delay-tolerance of data traffic to alleviate network congestion, and thus reduce the peak. However, solving the optimal scheduling problem leads to a large-scale Markov Decision Process (MDP) with extremely high complexity. In this paper, we propose a scalable and distributed approach to this problem, called Coordinated Scheduling (CoSchd). CoSchd decomposes the large-scale MDP problem into many individual MDP problems, each of which can be solved independently by each user under a limited amount of coordination signals from the BS. We show that CoSchd is close to optimal when the number of users becomes large. Further, we propose an approximation of CoSchd that iteratively updates the scheduling policy based on online measurements. Simulation results demonstrate that exploiting channel- and load-awareness with CoSchd can effectively alleviate cellular network congestion.**

***Index Terms*—Wireless scheduling, deadline constraint, dual decomposition, large-system asymptotics.**

I. Introduction

A GRAND challenge facing today's mobile service providers is to meet the exponentially increasing demand for mobile broadband services. This problem is particularly severe at the so-called "peak", where the network is heavily loaded at specific times and locations. Currently, wireless providers invest heavily in new spectrum and infrastructure to accommodate the *peak* demand, but such efforts are costly and inefficient: since the network traffic at non-peak times is orders-of-magnitude lower than that at peaks, provisioning network capacity for peak demand will lead to poor utilization of network resources.

An alternative approach is to exploit the delay tolerance of mobile applications to improve the network utilization. Prior work has identified a class of applications that can tolerate some delay, ranging from a few minutes to hours [2–5]. For example, the analysis in [4] shows that more than 55% of multimedia contents in cellular networks are uploaded more than one day after their creation time. More recently, the survey conducted in the TUBE project indicates that users are willing to defer their data transmissions if appropriate incentives are provided, *e.g.*, a discounted price [2]. Motivated by these findings, in this paper we study the scheduling of delay-tolerant traffic to minimize network congestion and improve resource utilization in wireless networks.

There are two directions where delay-tolerance can potentially be exploited to alleviate network congestion: *load-awareness* and *channel-awareness*. On one hand, approaches such as TUBE [2] and CoAST [6] move delay-tolerant traffic to the time and location where the network is less loaded, i.e., *being load-aware,* and thus alleviate network congestion. However, these approaches do not consider users' time-varying wireless channels – hence we classify them as "*load-only*" approaches. On the other hand, noting the temporal variation of channel conditions in wireless networks, a number of channel-aware scheduling schemes have been proposed at the mobile device to improve spectrum efficiency [3, 5, 7]. While this line of work takes advantage of the opportunistic nature of wireless networks, it has been limited to optimizing on a single mobile device. As a result, these schemes are oblivious to traffic-load levels and thus we refer to them as "*channel-only*" approaches. The recent work in [8] proposes mobile-side mechanisms to estimate and react to both channel condition and network load. However, it mainly focuses on reducing the energy consumption of the mobile. To the best of our knowledge, the above two directions have not been investigated jointly for the purpose of reducing network congestion.

In this paper, we study jointly channel- and load-aware scheduling policies for delay-tolerant traffic to reduce network congestion. We consider the scenario of a cellular network serving a sequence of data transfer requests. Each data transfer request has a pre-specified deadline, which is directly tied to the users' overall experience. The network's objective is to schedule these data transfers intelligently to minimize the network congestion cost, subject to their deadline constraints. We define the network congestion cost as the sum of convex functions of the load at each BS/WiFi-hotspot and at each time. With the convexity, the cost function naturally penalizes high peak demand and thus a cost-minimizing solution will tend to smooth out the traffic load across time and location.

The above scheduling problem is a sequential decision problem and can theoretically be cast as a Markov Decision Process (MDP). However, solving such an MDP problem faces challenges of both computational complexity and information collection. First, as the system size increases, the complexity of the MDP problem increases exponentially due to the curse of dimensionality. Compared to the channel-only approach that

This paper is published in IEEE/ACM TRANSACTIONS ON NETWORKING. Published version is available online at http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=7243365.

A preliminary version of this paper has appeared in 2014 Information Theory and Applications Workshop, February 9–14, 2014, San Diego, CA, USA [1].

H. Wu and X. Liu are with Department of Computer Science, University of California, Davis, CA 95616, USA.

X. Lin is with School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47906, USA.

K. Tan and Y. Zhang are with Microsoft Research Asia, Beijing, 100080, China.

only considers one mobile device [3, 5], here the size of the problem is very large, as a typical network may have hundreds of thousands of requests and a large number of BSs and WiFi hotspots. Compared to the load-only approach [2], here the channel uncertainty leads to significant difficulty in determining the amount of load that can be moved under a given policy. Second, if we were to solve the MDP in a centralized manner, the scheduler needs to know all requests and channel evolution statistics for each individual user. Collecting this information may require the BS to track the behaviors of all users, which raises concerns on both signaling overhead and privacy. Thus, decomposition technique and distributed scheduling policies are highly desirable to effectively solve such a large-scale MDP.

In this paper, we propose distributed schemes for solving this type of large-scale MDP problems. We refer to our distributed solution as Coordinated Scheduling (CoSchd). Under CoSchd, the network does not need to know the statistics of all requests, but instead updates a set of congestion signals based on the aggregated network load. At the same time, each user executes an individual decision policy based on the congestion signals and its own channel statistics. The key to this decomposition is to approximate the original problem by exchanging the order of the expectation and the cost function. Specifically, we replace the minimization of the expectation of the cost function with a minimization of the cost of the expected load (see Section IV for details). This approximation allows us to apply duality to decompose the network control, which addresses both the complexity issue and the signaling/privacy issues discussed earlier. Under certain conditions, we show that as the number of users in the system tends to infinity, the proposed CoSchd policy approaches the optimal solution of the original problem. We further propose an approximate version of CoSchd, referred to as CoSchd with Online Update (CoSchd-OU), that iteratively updates the scheduling policy based on online measurements. Finally, we evaluate the performance gains of exploiting channel- and/or load-awareness through simulations. Our simulation results demonstrate the asymptotic optimality of the proposed CoSchd and the benefits of scheduling with channel- and load-awareness. In particular, CoSchd can significantly reduce the network congestion for multi-cell systems with load variations.

In summary, the main contributions of this paper are:

- We study jointly channel- and load-aware scheduling policies for alleviating cellular network congestions. Previous work has only studied channel-aware and load-aware scheduling schemes separately, which are much easier to analyze. To the best of our knowledge, this is the first unified framework that considers both channel- and load-fluctuations to alleviate network congestion and improve resource efficiency.
- We decompose the large scale scheduling problem by dual decomposition and propose a Coordinated Scheduling (CoSchd) policy. CoSchd provides a framework for reducing the computation complexity and signaling overheads in channel- and load-aware scheduling. Under CoSchd, each user solves an individual MDP problem based on its own channel statistics and the congestion signals broadcast by the BS. The BS updates the congestion signals based on the aggregated traffic. We show that CoSchd achieves a near-optimal congestion cost in the many-source regime.
- We propose an approximation of CoSchd, referred to as CoSchd with Online Update (CoSchd-OU), where the congestion signals are updated based on the real-time aggregated traffic. Thus, CoSchd-OU may be even easier to implement.

The remainder of this paper is organized as follows. We first discuss related work in Section II. We define the problem in Section III and present our distributed solution and its near-optimality in Section IV. We present the evaluation results in Section V.

II. Related Work

Opportunistic scheduling has been extensively studied in wireless networks [9–14]. In particular, scheduling with Quality-of-Service guarantees, e.g., deadline constraints, has attracted plenty of attention recently. The Earliest-Deadline-First (EDF) and Least-Laxity-First (LLF) policies have been shown to be optimal for underloaded systems in traditional machine-job scheduling problem without channel-variations [15]. Variants of these policies, e.g., FEDD [16] and L^2HPR [17], have been proposed for deadline-constrained scheduling under wireless channels. Lyapunov-optimization-based policies are studied in [18] and [19] for maximizing the network utility with deadline constraints. Our previous work [20] proposes application-level scheduling policies that are asymptotically optimal in minimizing the deadline-violation-probability in the large-system regime. However, all these studies assume stationary arrival processes without considering the time-dependency of network traffic.

In addition to network-scale scheduling, channel-aware scheduling in the mobile side has also been studied to reduce the energy consumption and improve mobile battery performance. [21, 22] and [23] propose dynamic-programming (DP) based policies for minimizing the energy consumption in a mobile device. Both Wiffler [7] and Bartendr [5] consider the setting of vehicular systems to offload 3G data traffic to either WiFi networks or to time-instants when signal strength is stronger. In [3], Lyapunov-optimization-based algorithm is developed for the access link selection problem to reduce energy consumption of data transfers. These channel-only solutions leverage WiFi availability and signal variability, but do not consider network load fluctuation. The recent work in [8] proposes a LoadSense technique and a Peek-n-Sneak protocol, to estimate and react to both channel condition and network load. However, it mainly focuses on reducing the energy consumption of the mobile.

Load-aware control, e.g., time-dependent pricing, has been proposed to leverage delay tolerance to alleviate cellular network congestion. In particular, TUBE is a theoretical and experimental study that leverages time-dependent pricing to alleviate network congestion [2]. Its pilot trial conducted at Princeton with 50 AT&T data users demonstrates the feasibility of using time-dependent pricing to alleviate network congestion. A more recent work [6] proposes a CoAST approach to reduce the peak by exploiting the small-scale variations in cellular traffic. TUBE and CoAST leverage network load

fluctuations while our work considers not only network load fluctuations, but also user channel variations.

The channel- and load-aware scheduling problem can be viewed as an MDP. One possible way to solve this large scale MDP problem is the mean field approach [24], which approximates a large MDP by a continuous deterministic optimization problem and obtains the optimal policy by solving ordinary differential equations. However, it is not straightforward that this mean-field approach will lead to a decentralized solution as we will propose in this paper. Another possible way is to view this MDP as a factored decentralized MDP (factored Dec-MDP) [25], where a large scale MDP can be divided into independent sub-MDPs. Decomposition techniques have been extensively studied to reduce the complexity of such large scale MDPs [26–28], usually assuming special structure for the global reward function, e.g., a linear-sum of local rewards [28]. However, the congestion cost in our paper is a general convex function of the load contributed by all users, and the scheduler needs to coordinate all users with simple strategies.

III. System Model

We start by considering a single-BS system, where the proposed approach can also be generalized to include multiple BSs and WiFi-hotspots, as discussed in Section IV-E. The problem stated here applies to both the uplink and downlink in cellular networks.

Assume that time is slotted and indexed by $t \in \{0, 1, \ldots, N-1\}$, where N is the number of time-slots in each day. A typical time-slot length ranges from tens of seconds to a few minutes. Because of the large time scale, we assume that a data transfer request will be completed in one time-slot when the request is accepted, as in [2].

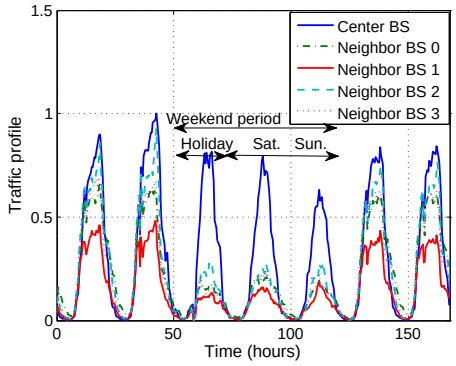


Fig. 1. Normalized cellular load from an anonymous mobile network operator in an urban area, obtained from http://anrg.usc.edu/www/Downloads/.

Data Traffic. In a typical day, a sequence of data transfer requests enter the network with user-specified deadlines. We use the words "user" and "request" interchangeably. The requests depart upon completion or deadline expiration. Mobile users show similar *aggregated* behavior over time (e.g, weekdays), as shown in various measurement studies of cellular traffic [29, 30]. For example, in Fig. 1, real-life load traces of cellular BSs show clear patterns over weekdays and over weekends/holidays.

Consider the scheduling problem in one day, where $t \in \{0, 1, \ldots, N-1\}$. Let $\mathcal{I} = \{1, 2, \ldots, m\}$ be the index set of all users that may request transfers from the BS. For each user $i \in \mathcal{I}$, denote the arrival time and the file size of its request by A_i and B_i, respectively. Assume that A_i's and B_i's are independent across users. A_i follows a distribution that reflects the typical traffic pattern of the day [2, 30]. We assume that the file size B_i is bounded by $B_{\max}$, i.e., $B_i \leq B_{\max}$, and is given as soon as the request arrives.

Each request i is associated with a user- or application-specific deadline D_i, i.e., the maximum delay that a user can tolerate. The deadline ranges from minutes to hours for delay-tolerant traffic [2, 4]. Such a deadline requirement depends on specific applications and can be set in various ways. For example, it could be a default setting in an application, e.g., syncing emails every half an hour; or, it can be learned from user preference. We assume that all transmission tasks should be completed at the end of the day, i.e., $A_i + D_i \leq N - 1$, for simplicity. To guarantee the quality of user experience, we need to constrain the deadline violation probability when scheduling delay-tolerant traffic, as discussed later. Note that in this model we also allow real-time traffic that needs to be transmitted immediately, in which case the deadline is set to be zero.

Channel Dynamics. Each user experiences time-varying channel conditions. We aim at designing scheduling policies that exploit channel variations in the coarse time-scale due to shadowing and user mobility, i.e., slow-fading. The measurements in [31] show that "the channel has a dominant slow-fading component on which the fast-fading component is overlaid", and the slow-fading component remains roughly constant on the order of seconds to minutes depending on the mobility. Thus, we model the channel conditions of user i as a stochastic process $R_i(t)$, where $R_i(t) \geq 0$ denotes the instantaneous rate per unit spectrum resource (e.g., a time-frequency block in LTE) at which the BS can communicate with user i in time-slot t. As suggested in [32], we assume that $R_i(t)$ is a homogeneous Markov chain over a finite set of possible transmission rates, i.e., $R_i(t) \in \{r_1, r_2, \ldots, r_J\}$, where J is the number of possible rates, and $0 = r_1 < r_2 < \ldots < r_J$. We assume that the channel conditions are independent across users and the transition probability matrix for user i is given by

$$\boldsymbol{P}_i = [p^{(i)}_{j_1 j_2}]_{J \times J}, i \in \mathcal{I}, \tag{1}$$

where $p^{(i)}_{j_1 j_2} \in [0, 1]$, $1 \leq j_1, j_2 \leq J$, is the transition probability from state j_1 to state j_2 for user i. We assume that all channel processes achieve the steady state, i.e., following the stationary distribution $\boldsymbol{\pi}^{(i)}$, where $\boldsymbol{\pi}^{(i)} = [\pi^{(i)}_1, \pi^{(i)}_2, \ldots, \pi^{(i)}_J]$ is the stationary distribution for the Markov chain of user i.

When user i in channel condition $R_i(t)$ ($R_i(t) > 0$) is scheduled to transmit a file of size B_i, it consumes $B_i/R_i(t)$ units of spectrum resource. We assume that each user can estimate its current channel condition via measurements of received signal strength and interference levels. Further, the user can learn the transition probability of its channel dynamics based on historical measurements, as in [7, 33].

Scheduling Policy and Base-Station Load. Let Γ denote a general scheduling policy that decides which users to transmit at a given time-slot. We consider the set of all causal policies. Corresponding to each Γ, we let $L_t(\Gamma)$ be the aggregate

amount of spectrum resource consumed by the users transmitting in time-slot t under policy Γ. We express $L_t(\Gamma)$ as

$$L_t(\Gamma) = \sum_{i \in \mathcal{I}} Y_{i,t}(\Gamma), \ t = 0, 1, \ldots, N-1, \tag{2}$$

where $Y_{i,t}(\Gamma)$ is the amount of resource consumed by user i in time-slot t. More precisely, recall that $R_i(t)$ is the instantaneous rate per unit spectrum resource of user i in time-slot t. We then have that for $0 \le t \le N-1$,

$$Y_{i,t}(\Gamma) = \begin{cases} B_i/R_i(t), & \text{if user } i \text{ transmits in slot } t, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Objective. From the network's point of view, the objective is to minimize the total congestion cost in the horizon of N time-slots subject to the deadline violation constraints. Let $f(\cdot)$ be a convex congestion-cost function and $v_i(\Gamma)$ be the deadline violation probability of user i. The scheduling problem is then

$$(\mathcal{P}_0) \qquad \begin{aligned} \underset{\Gamma}{\text{minimize}} \quad & F = \sum_{t=0}^{N-1} \mathbb{E}\big[f\big(L_t(\Gamma)\big)\big], \\ \text{subject to} \quad & v_i(\Gamma) \le \eta_i, \quad \forall i \in \mathcal{I}, \end{aligned} \tag{4}$$

where η_i is the maximum deadline violation probability tolerated by user i. In problem $\mathcal{P}_0$, the convexity of $f(\cdot)$ penalizes peaks and thus favors load that is smoothed over time, which is desirable for network operators. In most of our numerical results, we use the following function $f(l) = \big(l/\bar{C}\big)^{\nu}$, where $\bar{C}$ is a positive constant and $\nu > 1$ is a factor for controlling the penalty.

Table I summarizes the key variables used in this paper.

TABLE I
LIST OF VARIABLES

Variable	Explanation
t	Time-slot index
i	User index
N	Number of time-slots per day
$\mathcal{I}$	Set of users
(A_i, B_i, D_i)	Arrival time, file size, and deadline of user i
$R_i(t)$	Instantaneous rate per unit spectrum resource
$p^{(i)}_{j_1 j_2}$	Transition probability from channel state j_1 to j_2
$Y_{i,t}$	Resource consumed by user i in time-slot t
L_t	Total resource required in time-slot t under policy
l_t	Expectation of L_t
v_i	Deadline violation probability of user i under policy
η_i	Deadline violation probability constraint of user i
$f(\cdot)$	Cost function
F^*	Optimal value of the original problem $\mathcal{P}_0$
$F_\sharp$	Optimal value of the approximate problem $\mathcal{P}_1$

Note that, in principle, $\mathcal{P}_0$ can be viewed as an MDP by taking the waiting time and channel condition of all users as system state. However, solving such an MDP problem in a centralized manner is forbiddingly complex. First, the size of the problem is very large, as a typical network may have hundreds of thousands of users, over a time horizon of a day. In addition, deadline constraint is notoriously difficult to solve in general because of the resource coupling across time and among users. Second, the problem formulation assumes knowledge of all requests and their detailed channel information. In practice, it is not feasible to gather such detailed information in a central entity because of both signaling overhead and privacy concerns.

Next, we will focus on the regime where the number of users is large, and develop a distributed approach for (approximately) solving problem $\mathcal{P}_0$. Our main intuition is the following. In our system, each user can be seen as interacting with the set of all other users. When the number of users is large, the impact of any given user's decision on the overall system should be minimal. Thus, it would be as if each user is interacting with a common entity that includes all users in the system. If we can summarize the effect of all users by some kind of "congestion signal," we may then be able to approximate the original system by another system where each user independently reacts to such a common congestion signal. The challenges are how to design such a common congestion signal and how to establish the (asymptotic) optimality of the decomposition, which will be the focus of the following sections.

IV. Asymptotically Optimal Decomposition

This section studies asymptotically optimal policies for solving the large scale MDP $\mathcal{P}_0$. Note that the objective in (4) is to minimize the *expectation of total cost*. We first propose a lower bound of $\mathcal{P}_0$ by introducing a new problem $\mathcal{P}_1$ that minimizes the *total cost of expectation*. We then propose a distributed policy, referred to as Coordinated Scheduling (CoSchd), and show its asymptotic optimality in the many-source regime.

A. Lower Bound

In the original problem $\mathcal{P}_0$, the cost is a function of the instantaneous load level $L_t(\Gamma)$ and the objective is to minimize the expected total cost. Because the cost function $f(\cdot)$ is convex, the optimal value of $\mathcal{P}_0$ can be lower bounded by exchanging the order of the expectation and the cost function. Specifically, consider the following problem that minimizes the total cost of the expected load level:

$$(\mathcal{P}_1) \qquad \begin{aligned} \underset{\Gamma}{\text{minimize}} \quad & \tilde{F} = \sum_{t=0}^{N-1} f\big(l_t(\Gamma)\big), \\ \text{subject to} \quad & v_i(\Gamma) \le \eta_i, \forall i \in \mathcal{I}, \end{aligned}$$

where $l_t(\Gamma) = \mathbb{E}[L_t(\Gamma)]$ is the expectation of the load level. Let F^* be the optimal value of the original problem $\mathcal{P}_0$ and let $F_\sharp$ be the optimal value of $\mathcal{P}_1$. Because the constraints of $\mathcal{P}_0$ and $\mathcal{P}_1$ are identical and the only difference lies in the objective function, we can easily show the following proposition by the convexity of $f(\cdot)$ and Jensen's inequality [34].

Proposition 1 *The optimal value of problem $\mathcal{P}_1$ provides a lower bound on the value of the original problem $\mathcal{P}_0$, i.e., $F_\sharp \le F^*$.*

As we will see later, thanks to the linearity of expectation operation, the cost of expected load is much easier to deal with than the expectation of cost. Hence, problem $\mathcal{P}_1$ and its lower-bound property are critical in the design and analysis of asymptotically optimal policies. Next, we will study the optimal solution for $\mathcal{P}_1$, and show its asymptotic optimality for the original problem $\mathcal{P}_0$ in the many-source regime.

B. Dual Decomposition

This subsection proposes a decomposition approach for solving problem $\mathcal{P}_1$ based on dual decomposition. Recall that the channel rates satisfy $0 = r_1 < r_2 < \ldots < r_J$ and users could only request to transmit under a positive rate. To use dual decomposition, we first introduce auxiliary variables $h_t \in [0, h_{\max}]$, $(t = 0, 1, \ldots, N-1)$, where $h_{\max} = B_{\max}/r_2$ is the maximum load level in one slot. Let $\boldsymbol{h} = [h_0, h_1, \ldots, h_{N-1}]$. We can rewrite problem $\mathcal{P}_1$ as

$$
\begin{aligned}
(\mathcal{P}_1') \qquad \underset{\Gamma, \boldsymbol{h}}{\text{minimize}} \quad & \tilde{F} = \sum_{t=0}^{N-1} f(mh_t) \\
\text{subject to} \quad & l_t(\Gamma)/m \le h_t,\ 0 \le t \le N-1, \quad (5) \\
& v_i(\Gamma) \le \eta_i, \forall i \in \mathcal{I}, \quad (6) \\
& 0 \le h_t \le h_{\max},\ 0 \le t \le N-1. \quad (7)
\end{aligned}
$$

Let $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_{N-1}]$ be the Lagrange multiplier vector corresponding to the constraints in Eq. (5). It will be clear that $\boldsymbol{\beta}$ serves as the congestion signal provided by the BS over time. Given $\boldsymbol{\beta}$, we formulate and decompose the Lagrangian as follows:

$$
\begin{aligned}
\mathcal{L}(\Gamma, \boldsymbol{h}, \boldsymbol{\beta}) &= \sum_{t=0}^{N-1} f(mh_t) - \sum_{t=0}^{N-1} \beta_t \big[h_t - l_t(\Gamma)/m\big] \\
&= \sum_{t=0}^{N-1} [f(mh_t) - \beta_t h_t] + \frac{1}{m} \sum_{i \in \mathcal{I}} \sum_{t=0}^{N-1} \beta_t y_{i,t}(\Gamma),
\end{aligned} \tag{8}
$$

where $y_{i,t}(\Gamma) = \mathbb{E}[Y_{i,t}(\Gamma)]$ is the expected amount of resource consumed by user i in slot t. Let the objective function of the dual problem be $g(\boldsymbol{\beta})$, i.e.,

$$
g(\boldsymbol{\beta}) = \inf_{\Gamma, \boldsymbol{h}} \mathcal{L}(\Gamma, \boldsymbol{h}, \boldsymbol{\beta}). \tag{9}
$$

Since the Lagrangian has been decomposed, we can use an individual policy Γ_i to minimize the expected consumed resource of user i. Thus, for a given $\boldsymbol{\beta}$, the dual objective function can be obtained by solving the following subproblems:

$$
\begin{aligned}
(\mathcal{SP}_0) \qquad \text{minimize} \quad & \sum_{t=0}^{N-1} [f(mh_t) - \beta_t h_t], \\
\text{subject to} \quad & 0 \le h_t \le h_{\max},\ 0 \le t \le N-1.
\end{aligned}
$$

$$
\begin{aligned}
(\mathcal{SP}_i) \qquad \underset{\Gamma_i}{\text{minimize}} \quad & \sum_{t=0}^{N-1} \beta_t y_{i,t}(\Gamma_i) \\
\text{subject to} \quad & v_i(\Gamma_i) \le \eta_i,\ i \in \mathcal{I}.
\end{aligned}
$$

The master dual-problem is

$$
\begin{aligned}
(\mathcal{D}_1) \qquad \underset{\boldsymbol{\beta}}{\text{maximize}} \quad & g(\boldsymbol{\beta}) \\
\text{subject to} \quad & \boldsymbol{\beta} \ge \mathbf{0}.
\end{aligned}
$$

Since $f(\cdot)$ is convex, subproblem $\mathcal{SP}_0$ can be easily solved by convex optimization algorithms [34]. For subproblem $\mathcal{SP}_i$, we can view it as a constrained sequential decision problem and obtain the optimal policy Γ_i using backward induction [35]. Therefore, the dual problem can be solved efficiently by using (sub-)gradient approach, as will be discussed later.

For a general optimization problem, dual decomposition only guarantees weak duality, i.e., the dual solution only provides a lower bound to the original problem. However, we show below that the duality gap between $\mathcal{P}_1$ and $\mathcal{D}_1$ is zero, and hence there exists an optimal value of $\boldsymbol{\beta}$ such that the algorithms $\mathcal{SP}_0$ and $\mathcal{SP}_i$ combined provide an optimal solution to $\mathcal{P}_1$.

Proposition 2 *Given that the cost function $f(\cdot)$ is convex, the dual problem $\mathcal{D}_1$ have zero duality gap, and thus the dual decomposition approach provides an optimal value to $\mathcal{P}_1$.*

Proof: Strong duality holds for convex optimization problem. To prove the proposition, we convert $\mathcal{P}_1$ to a convex problem with exponentially large number of control variables. Thus, the dual problem $\mathcal{D}_1$ have zero duality gap and generates the same solution as $\mathcal{P}_1$. See Appendix A for details. ■

The proof in Appendix A uses a transformation of policy representations, which will also be useful for design scheduling policies. Hence, we briefly introduce the transformation here. Let Ω_i be the set of possible realizations of channel process $R_i(t)$ for user i. For each realization denoted by $\boldsymbol{r} = [r(0), r(1), \ldots, r(N-1)] \in \Omega_i$, let $\boldsymbol{r}(0:t) = [r(0), r(1), \ldots, r(t)]$ be the first $t+1$ elements of $\boldsymbol{r}$. We only focus on causal policies, and thus the decision is made based on the history of the channel conditions. Let $x_{a_i,w,\boldsymbol{r}(0:a_i+w)} \in [0,1]$ denote the transmission probability of user i when its arrival time is a_i, waiting time is w slots and the channel condition history is $\boldsymbol{r}(0:a_i+w)$. Then, a policy Γ_i for solving subproblem $\mathcal{SP}_i$ can be represented by a decision matrix $\boldsymbol{x}_i = \{\boldsymbol{x}_{a_i}(\boldsymbol{r}) : 0 \le a_i \le N-1, \boldsymbol{r} \in \Omega_i\}$, where each submatrix $\boldsymbol{x}_{a_i}(\boldsymbol{r}) = [x_{a_i,w,\boldsymbol{r}(0:a_i+w)}]$ represents the policy for each pair of arrival time a_i and channel realization $\boldsymbol{r}$. For each decision matrix $\boldsymbol{x}_i$, we define the following transformation, denoted by $\boldsymbol{\varphi}_i = \mathcal{T}(\boldsymbol{x}_i)$, as follows: for each realization $\boldsymbol{r} \in \Omega_i$,

$$
\begin{aligned}
&\varphi_{a_i,w,\boldsymbol{r}(0:a_i+w)} \\
&= \begin{cases}
x_{a_i,w,\boldsymbol{r}(0:a_i+w)}, & \text{if } w = 0, \\
x_{a_i,w,\boldsymbol{r}(0:a_i+w)} \prod_{w'=0}^{w-1} [1 - x_{a_i,w,\boldsymbol{r}(0:a_i+w')}], & \\
\qquad \text{if } w = 1, 2, \ldots, D_i - 1.
\end{cases}
\end{aligned} \tag{10}
$$

Note that $\varphi_{a_i,w,\boldsymbol{r}(0:a_i+w)}$ can be interpreted as the probability that user i under a particular channel realization $\boldsymbol{r}$ transmits at time $a_i + w$. The transformation $\mathcal{T}$ is invertible, where the inverse transformation $\boldsymbol{\varphi}_i = \mathcal{T}^{-1}(\boldsymbol{x}_i)$ can be defined as follows: for a realization $\boldsymbol{r} \in \Omega_i$,

$$
\begin{aligned}
&x_{a_i,w,\boldsymbol{r}(0:a_i+w)} = \\
&\begin{cases}
\varphi_{a_i,w,\boldsymbol{r}(0:a_i+w)}, & \text{if } w = 0, \\
\frac{\varphi_{a_i,w,\boldsymbol{r}(0:a_i+w)}}{\prod_{w'=0}^{w-1}(1 - x_{a_i,w',\boldsymbol{r}(0:a_i+w'))}}, & \\
\quad \text{if } 0 < w \le D_i - 1 \ \&\ \prod_{w'=0}^{w-1}(1 - x_{a_i,w',\boldsymbol{r}(0:a_i+w')}) > 0, & \\
0, \text{if } 0 < w \le D_i - 1 \ \&\ \prod_{w'=0}^{w-1}(1 - x_{a_i,w',\boldsymbol{r}(0:a_i+w')}) = 0. &
\end{cases}
\end{aligned} \tag{11}
$$

C. CoSchd: Coordinated Scheduling

Based on the dual decomposition discussed in the previous subsection, we propose the following distributed algorithm, referred to as Coordinated Scheduling (CoSchd), to solve the

approximate problem $\mathcal{P}_1$. Under CoSchd, each user individually decides its transmission schedule based on the congestion signal from the BS and its own channel characteristics, and the BS updates the congestion signals based on the expected aggregated load, as shown in Algorithm 1.

Algorithm 1 Coordinated Scheduling (CoSchd).

Input: Distributions of A_i, B_i, D_i;
Transition probability matrix $\boldsymbol{P}_i$.
Output: Transmission probability $\bar{\boldsymbol{x}}_i$.
Init: Set $d=1$ and $\beta_t^{(1)}=1$ for all $t=0,1,\ldots,N-1$.
for $d \leftarrow 1,2,\ldots,d_{\max}$
1) Mobile-side: each user $i \in \mathcal{I}$ solves $\mathcal{SP}_i$ and obtains

$$\begin{aligned} \boldsymbol{x}_i^{(d)} &\leftarrow \arg\min \sum_{t=0}^{N-1} \beta_t^{(d)} y_{i,t}(\boldsymbol{x}_i), \\ \boldsymbol{\varphi}_i^{(d)} &\leftarrow \mathcal{T}(\boldsymbol{x}_i^{(d)}); \end{aligned}$$

Each user estimates its expected load $y_{i,t}$ in each slot and reports to the BS;
2) Network-side: the BS collects the load of each user and calculates the aggregated load $l_t^{(d)}$:
The BS solves $\mathcal{SP}_0$ and updates $\beta_t^{(d)}$ using the sub-gradient method:

$$\beta_t^{(d+1)} = \left[\beta_t^{(d)} + \alpha^{(d)}\big(\frac{l_t^{(d)}}{m} - h_t^{(d)}\big)\right]^+, 0 \le t \le N-1, \quad (12)$$

endfor
Averaging: Calculate the average transition probability

$$\bar{\boldsymbol{\varphi}}_i \leftarrow \frac{1}{d_{\max}} \sum_{d=1}^{d_{\max}} \boldsymbol{\varphi}_i^{(d)}; \quad (13)$$

$$\bar{\boldsymbol{x}}_i \leftarrow \mathcal{T}^{-1}(\bar{\boldsymbol{\varphi}}_i). \quad (14)$$

From Algorithm 1, we can see that the network-side operation under CoSchd is simple: the BS solves subproblem $\mathcal{SP}_0$ to obtain the optimal value of $\boldsymbol{h}^{(d)}$, and then updates congestion signals based on load level $\boldsymbol{l}^{(d)}$ and $\boldsymbol{h}^{(d)}$. Next, we focus on the operation on the mobile side.

1) Mobile-side Operation

On the mobile-side, each user operates independently as follows: it generates policies based on its channel characteristics and the congestion signals, and then executes the policy based on the instantaneous channel condition.

For a given congestion signal vector $\boldsymbol{\beta}$, the subproblem $\mathcal{SP}_i$ turns out to be a constrained MDP problem [35]. Then each user only needs to make decisions based on the waiting time and the current channel state. Specifically, each user introduces a cost for deadline violation and minimizes $\mathcal{SP}_i$ plus the deadline violation cost by backward induction. We note that the complexity of the backward induction method for solving the individual MDP is $O(J^2D)$ [36], which can be implemented in most smartphones as in [37, 38].

We now discuss the specifics of the deterministic deadline-constraint case as follows and refer the readers to [35] for the probabilistic deadline constraint case.

For user i arriving at a_i, let $x_{a_i,w,j} \in [0,1]$ ($w = 0,1,\ldots,D_i-1; j=1,2,\ldots,J$) be the probability that user i requests transmission when its waiting time is w and channel state is j. (Thus, the probability $x_{a_i,w,\boldsymbol{r}(0:a_i+w)} = x_{a_i,w,j}$ if $r(a_i+w)=r_j$.) In the deterministic deadline-constraint case, i.e., $\eta_i = 0$, all data must be transmitted before expiration. Therefore, for user i arriving at a_i, it requires that $x_{a_i,a_i+D_i-1,j}=1$. To guarantee a finite transmission cost, we assume that for each user,

$$\mathbb{E}\{B_i/R_i(a_i+D_i-1)|\mathcal{E}_{i,D_i-1}\} < +\infty, \quad i \in \mathcal{I}, \quad (15)$$

where $\mathcal{E}_{i,D_i-1}$ represents the event that user i does not transmit before a_i+D_i-1. Using the principle of optimality and taking the multipliers $\boldsymbol{\beta}$ into account, we can obtain the optimal decision

$$x_{a_i,w,j} = \begin{cases} 1, & \text{if } \frac{\beta_{a_i+w}}{R_i(a_i+w)} \le \mathbb{E}[V^*_{a_i,w+1}|r_j] \\ 0, & \text{otherwise,} \end{cases} \quad (16)$$

where $\mathbb{E}[V^*_{a_i,w+1}|r_j]$ is the expected future cost conditioned on $R_i(a_i+w)=r_j$, which can be calculated by backward induction:

$$\begin{aligned} &\mathbb{E}[V^*_{a_i,w+1}|r_j] \\ &= \begin{cases} \mathbb{E}\big[\frac{\beta_{a_i+D_i-1}}{R_i(a_i+D_i-1)}\big|r_j\big], & \text{for } w = D_i-2, \\ \mathbb{E}\Big[\min\big(\frac{\beta_{a_i+w+1}}{R_i(a_i+w+1)}, V^*_{a_i,w+2}\big)\big|r_j\Big], & \\ & \text{for } 0 \le w \le D_i-3. \end{cases} \end{aligned}$$

After obtaining $\boldsymbol{x}_i$, each user can estimate the amount of required resource as follows:

$$\begin{aligned} y_{i,t} &= \sum_{a=0}^{N-1} \mathbb{P}(A_i=a) y_{i,a,t} \\ &= \sum_{a=0}^{N-1} \mathbb{P}(A_i=a)\mathbb{E}[B_i] \sum_{j=2}^{J} \pi'_{i,a,t,j}/r_j, \end{aligned} \quad (17)$$

where $\pi'_{i,a,t,j}$ is the probability that the user i with arrival time a transmits at slot t under channel condition r_j, i.e.,

$$\pi'_{i,a,t,j} = \begin{cases} \pi_j^{(i)}, & \text{if } t=a, \\ \sum_{j'=1}^{J}(1-x_{a,t-a,j'})\pi'_{i,a,t-1,j'}p_{j'j}, & \text{otherwise.} \end{cases}$$

Finally, the averaging operation given by (13) and (14) is designed to deal with the possible oscillation issues of the subgradient method, as in [39].

2) Asymptotic Optimality of CoSchd

In this section, we show that the proposed CoSchd policy is near-optimal for problem $\mathcal{P}_0$ when the number of users is large.

First, we show that CoSchd provides a near-optimal solution to problem $\mathcal{P}_1$, which is an approximation of $\mathcal{P}_0$. Specifically, we use a constant step-size in (12) and let $\alpha^{(d)} = \alpha$. Let $\tilde{F}_{\mathrm{CoSchd}(\alpha)}$ be the cost value of $\mathcal{P}_1$ under CoSchd with $\alpha^{(d)}=\alpha$. Then, we present the following lemma stating that $\mathrm{CoSchd}(\alpha)$ provides a near-optimal solution of $\mathcal{P}_1$.

Lemma 1 *For CoSchd with a constant step-size $\alpha^{(d)} = \alpha$, the cost of problem $\mathcal{P}_1$ is bounded as follows:*

$$\tilde{F}_{\mathrm{CoSchd}(\alpha)} \leq F_\sharp + \frac{\alpha G^2}{2}, \tag{18}$$

where $F_\sharp$ is the optimal value of $\mathcal{P}_1$, and $G = 2\sqrt{N}h_{\max}$ is an upper bound on the norm of the subgradient in the dual problem.

Proof: We verify that: a) the constraints (5) are linear in $\boldsymbol{\varphi}_i$, and hence $\mathcal{P}_1'$ is convex; b) $\mathcal{P}_1$ satisfies the Slater's condition; c) the value of $|l_t/m - h_t|$ is bounded by $2h_{\max}$, and hence the norm of the subgradient is bounded as $||\boldsymbol{l}/m - \boldsymbol{h}|| \leq 2\sqrt{N}h_{\max}$. Then the conclusion of this lemma can be inferred from Proposition 2 in [39]. ∎

Next, we study the performance of CoSchd for the original problem $\mathcal{P}_0$. Note that CoSchd provides a near-optimal solution to $\mathcal{P}_1$, but $\mathcal{P}_1$ is not equivalent to the original problem $\mathcal{P}_0$. Fortunately, because all users independently solve individual MDPs under CoSchd, the instantaneous load level approaches its expectation as the number of users increases. Using this property, we can show that the proposed approach is asymptotically optimal for $\mathcal{P}_0$ in the many-source regime.

Consider the many-source regime. To study the asymptotic properties of the proposed approach, we consider the following m-scaled system.

Assumption 1 *All users in $\mathcal{I}$ can be divided into K classes. For each class k,*

- *the number of users m_k ($k = 1, 2, \ldots, K$) is proportional to the total number of users m, i.e., $m_k = m\lambda_k$, where $0 < \lambda_k < 1$ is the ratio of class-k users and $\sum_{k=1}^{K} \lambda_k = 1$;*
- *users in class-k have the same deadline requirements, and the same statistics of arrival time and channel dynamics that do not change with m.*

Further, we make the following assumption on the cost function:

Assumption 2 *The cost function $f(\cdot)$ in the m-scaled system is continuous, and is a function of the normalized load, i.e., $f(l) = \tilde{f}(\tilde{l})$, where $\tilde{l} = l/m$.*

For the above m-scaled system, we let $F^{(m)}_{\mathrm{CoSchd}(\alpha)}$ be the cost value of the original problem $\mathcal{P}_0$ under CoSchd with $\alpha^{(d)} = \alpha$, and let $F^{(m)}_\sharp$ be the optimal value of problem $\mathcal{P}_1$. Note that by optimizing on the normalized load-level $\tilde{l}_t$, the constant G in Lemma 1 for the m-scaled system is independent of m. The following proposition then shows the performance of CoSchd in the many-source regime.

Proposition 3 *Under Assumptions 1 and 2, $F^{(m)}_{\mathrm{CoSchd}(\alpha)}$ converges to a value near $F^{(m)}_\sharp$ as m increases, i.e.,*

$$\lim_{m\to\infty} F^{(m)}_{\mathrm{CoSchd}(\alpha)} \leq \lim_{m\to\infty} F^{(m)}_\sharp + \frac{\alpha G^2}{2}, \tag{19}$$

where $G = 2\sqrt{N}h_{\max}$ is an upper bound on the norm of the subgradient in the dual problem, as in Lemma 1.

Proof: Let $\tilde{F}^{(m)}_{\mathrm{CoSchd}(\alpha)}$ be the cost value of $\mathcal{P}_1$ under $\mathrm{CoSchd}(\alpha)$ in the m-scaled system. Because $\tilde{F}^{(m)}_{\mathrm{CoSchd}(\alpha)} \leq F^{(m)}_\sharp + \frac{\alpha G^2}{2}$ according to Lemma 1, we only need to show that $\lim_{m\to\infty} F^{(m)}_{\mathrm{CoSchd}(\alpha)} = \lim_{m\to\infty} \tilde{F}^{(m)}_{\mathrm{CoSchd}(\alpha)}$. Note that $F^{(m)}_{\mathrm{CoSchd}(\alpha)}$ is the sum of the expected costs in each slot, i.e., $F^{(m)}_{\mathrm{CoSchd}(\alpha)} = \sum_{t=0}^{N-1} \mathbb{E}\big[f(L_t)\big]$. Therefore, it suffices to show that under $\mathrm{CoSchd}(\alpha)$, $\lim_{m\to\infty} \mathbb{E}[f(L_t)] = f\big(\mathbb{E}[L_t]\big)$. This can be verified using the fact that, under $\mathrm{CoSchd}(\alpha)$, each user operates independently for given congestion signal $\boldsymbol{\beta}$, and the load level L_t is close to its expectation according to large deviation theories. See Appendix B for details. ∎

According to Proposition 1, $F^{(m)}_\sharp$ provides a lower bound on the optimal value of $\mathcal{P}_0$. The above proposition states that $F^{(m)}_{\mathrm{CoSchd}(\alpha)}$ will be in a neighborhood of $F^{(m)}_\sharp$ as m increases, and thus the decomposition approach is near-optimal for $\mathcal{P}_0$ in the many-source regime.

D. Approximate Implementation of CoSchd

In the previous section, we implement the CoSchd approach to obtain the approximately optimal solution before really running the network. In this section, we propose an approximation version of CoSchd with Online-Update (CoSchd-OU).

As shown in Algorithm 2, in every day, each user solves the individual decision problem $\mathcal{SP}_i$ based on its historic channel and arrival statistics. To reduce the complexity of policy averaging, i.e., Eqs. (13) and (14) in Algorithm 1, here each user directly applies a Exponential-Moving-Averaging policy according to Eq. (20). Then, each user makes decisions based on its channel state and the BS updates the congestion signals based on the actual traffic-load measurements.

Note that under CoSchd-OU, each user requires its channel and arrival statistics to solve $\mathcal{SP}_i$. Once these statistics are known, $\mathcal{SP}_i$ can be solved either online or offline. To avoid delay and energy consumption induced by the MDP computation for $\mathcal{SP}_i$, it is typical for each user to pre-compute the decision policy based on its own channel statistics and the congestion signals broadcast by the BS. Then, the individual policies are carried out online and the congestion signals are also updated in an online manner.

Algorithm 2 CoSchd-OU.

Init:
set $d = 0$ and $\beta_t^{(0)} = 1$ for all $t = 0, 1, \ldots, N-1$.
Iteration: (day d)
1) At time $t = 0$, $\beta_t^{(d)}$ ($t = 0, 1, \ldots, N-1$) is announced to all users;
Each user $i \in \mathcal{I}$ solves $\mathcal{SP}_i$ and calculates the average decision matrix as follows

$$\bar{\boldsymbol{x}}_i^{(d)} = \vartheta \bar{\boldsymbol{x}}_i^{(d-1)} + (1-\vartheta)\boldsymbol{x}_i^{(d)},\ \vartheta \in [0,1]. \tag{20}$$

2) For $t = 0 \to N-1$,
Each user makes decision based on its channel states;
The BS serves requested users and observes the load level $L_t^{(d)}$;
The BS solves $\mathcal{SP}_0$ and updates $\beta_t^{(d)}$ using Eq. (12) with $l_t^{(d)} = L_t^{(d)}$;
3) Set $d \leftarrow d+1$ and go to step 1).

In practice, the arrival pattern may vary gradually over time. Our evaluations show that certain variations of the arrival pattern are acceptable in practice [1]. Specifically, when the arrival pattern varies across different days but still have similarities, the proposed CoSchd-OU policy can provide congestion signals that capture such similarities. Then the peak load can be reduced under these congestion signals.

Another potential issue of CoSchd-OU is that the convergence time may be large because one iteration takes one day. To further reduce convergence time, the BS can use initial congestion signals pre-computed from Algorithm 1. Specifically, each user solves the individual decision problem based on its historic observations and reports the expected load levels to the BS; the BS updates the congestion signals according to the expected load level. The process is repeated until appropriate congestion signals are obtained. Then, CoSchd-OU can be implemented with these initial congestion signals. The averaging step for the primal variables, i.e., Eqs. (13) and (14), in Algorithm 1 can be omitted because we only need the initial congestion signals to start CoSchd-OU. We also note that if the BS can obtain estimates of the statistics for traffic and channel conditions, it can run the CoSchd algorithm in a centralized manner to obtain the initial congestion signals.

E. Multi-cell Networks

For ease of exposition, we have so far focused on the single-cell scenario. Next, we explain how the proposed algorithm can be extended to include multiple BSs and WiFi hotspots.

We note that a cellular BS and a WiFi AP have no conceptual difference in terms of the problem formulation, except that their corresponding congestion cost functions could differ because of the difference in capacity and cost. To extend the results from one BS to multiple BSs, one can expand the objective (i.e., the total congestion cost) to include all BSs' congestion cost at all time slots. Specifically, when there are C BSs, the objective in the multi-cell system becomes

$$\underset{\Gamma}{\text{minimize}} \qquad F = \sum_{c=0}^{C-1}\sum_{t=0}^{N-1} \mathbb{E}\big[f_c\big(L_{c,t}(\Gamma)\big)\big], \qquad (21)$$

where $f_c(\cdot)$ is the cost function of BS c and $L_{c,t}(\Gamma)$ is the load-level in BS c in time-slot t under policy Γ. Similar to Section IV-A, we can approximate the original problem in (21) by exchanging the order of expectation and cost function, i.e.,

$$\underset{\Gamma}{\text{minimize}} \qquad \tilde{F} = \sum_{c=0}^{C-1}\sum_{t=0}^{N-1} f_c\big[\mathbb{E}\big(L_{c,t}(\Gamma)\big)\big], \qquad (22)$$

Then, in the duality-based solution, each BS broadcasts its own congestion signal for each time slot. Instead of a congestion signal vector, we introduce a congestion signal matrix $[\beta_{ct}]_{C\times N}$, where β_{ct} represents the congestion signal broadcast by BS c in time-slot t. Similar to Section IV-B, we can use $[\beta_{ct}]_{C\times N}$ to decompose the primal problem, and rearrange it into the mobile-side and BS-side problems as in $\mathcal{SP}_i$ and $\mathcal{SP}_0$ (except that there are multiple equations similar to $\mathcal{SP}_0$, one for each BS).

[1]The results are omitted here due to space limitations.

On the mobile side, each user maintains a profile of $\big(c_i(t), R_i(t)\big)$, which are the index of the BS that user i connects to in time-slot t and the channel condition with respect to that BS. Upon receiving the congestion signals from all BSs that it may connect to, the user can then compute the decision table regarding when and which BS it may use to complete the data transfer, while meeting the deadline constraints. The load at each BS is thus determined by mobile users' opportunistic decisions. Finally, at the end of the day, after all mobiles perform their data transfer, each BS updates its congestion signal as in Eq. (12).

Using similar techniques as in the setting of the single-cell networks, we can show that CoSchd also achieves near-optimal performance in multi-cell networks. The difference is that the dimensionality of the subgradient in the dual problem becomes NC. Thus, the constant G in Lemma 1 and Proposition 3 becomes $G = 2\sqrt{NC}h_{\max}$.

Using CoSchd can balance the traffic load among multiple BSs. In general, different BSs often have different offered load to begin with, as shown in Fig. 1. With a load-aware scheduling policy, the network would prefer a portion of the data transfers to be moved from heavily-loaded BSs to lightly-loaded BSs. In our CoSchd solution, at a given time a heavily-loaded BS will tend to have a larger value for its congestion signal than a lightly-loaded BS. Therefore, in the mobile-side decision, the threshold to transmit for the heavily-loaded BS will be correspondingly higher, which serves the goal of moving an appropriate amount of traffic to other lightly-loaded BSs. In contrast, under channel-only approaches, mobile devices are only aware of the channel condition at each BS, but not its congestion signal. Thus, it is possible that a mobile device delays its traffic until it connects to a BS with a stronger signal, but only finds that the BS has heavy load. In this case, a channel-only solution may not best alleviate network congestion, while CoSchd performs better, as shown in Section V.

F. From CoSchd to Load-only/Channel-only approaches

We mainly focus on the joint approach in the previous sections. Under the proposed framework, we can also investigate the load-only and channel-only approaches, which are discussed as follows and will be evaluated in Section V.

1) Load-only approach: A load-only approach balances the load without considering the channel variations. To compare with the best performance of this type of policies, we consider an optimal *offline* load-only policy that can be viewed as a modification of TUBE [2]. We assume that the knowledge of the traffic (e.g., distributions of arrival time and deadline) is available by the BS, and the data can be transmitted in any time-slot before the deadline. Then, the corresponding load-balancing problem can be formulated as a convex optimization problem and solved by standard algorithms [34]. We note that the TUBE work focuses on single-cell systems and only *temporal* load-variations are studied in [2]. If one was to also consider *spacial* load-variations, the scheduling problem would be similar to our multi-cell scenarios. However, in that case at least the user-connectivity profile across multiple cells must be taken into account. In other words, the load cannot be considered independently from the connectivity/channel

profiles. Due to this reason, similar to [2], we will not study the load-only approach in the multi-cell setting in Section V.

2) Channel-only approach: When the congestion signals are identical across all time-slots and all BSs, CoSchd degenerates to a channel-only approach. We consider the optimal channel-only policy, where each user applies an individual decision policy to minimize the expectation of the consumed resource based on its own channel condition profile under the deadline constraint. The Bartendr policy proposed in [5] can be viewed as one of the channel-only policies, while a fixed threshold is used for any waiting time. The performance of Bartendr is slightly worse than that of the optimal channel-only policy and will not be evaluated in Section V for more concise presentation.

V. Evaluation

In this section, we evaluate the performance of load-only, channel-only, and CoSchd approaches through simulations. Since the CoSchd policy and its approximation CoSchd-OU have similar performance and CoSchd-OU is more scalable as discussed in Section IV, we only consider CoSchd-OU here and refer it to as CoSchd for simplicity. As a baseline, we also consider *ImTrans*, where all users immediately transfer the data when the requests arrive.

A. Simulation Setup

We use a slot length of 10 minutes and each day is divided into 144 time-slots. For the cost function, we mainly use $f(l) = \left(l/\bar{C}\right)^{\nu}$ that is mentioned in Section III, and set $\nu = 8$ which is large enough for smoothing out the load level according to our experiments. Another type of cost function, i.e., $f(l) = \left([l - \bar{C}]^{+}/\bar{C}\right)^{2}$, will also be considered for comparison purpose. We consider single-cell and multi-cell scenarios, except that the load-only policy will only be evaluated in the single-cell scenario similar to [2]. For multi-cell scenarios, we mainly focus on 2-cell scenarios where we show the impact of system settings on the performance. Further, we also provide simulations for a 7-cell scenario where a center BS is surrounded by 6 neighbor BSs with lower traffic, and reports a higher gain for the proposed CoSchd policy.

1) Traffic Arrival Pattern: We assume that users enter the system according to the profile illustrated in Fig. 1. Specifically, for the single-cell network, the distribution of the arrival time for each user is set based on the weekday traffic profile of the center BS from Fig. 1. For the multi-cell network, we use the weekday traffic profile of the center BS and the neighbor BS 1 (again from Fig. 1) for the 2-cell scenarios. The profile of the neighbor BS 1 is further used for all the neighbor BSs in the 7-cell scenario. To capture the delay-tolerance of traffic, we apply the waiting function proposed in [2], and use the patience indices for the different traffic classes estimated from the U.S. survey in [2]. Specifically, for the delay-tolerant traffic ("Time-Dependent Pricing" traffic in [2]), the probability that user i wants to wait D_i slots is proportional to $\frac{1}{(D_i+1)^{\rho}}$, where the patience index ρ is 2.0 for video traffic and 0.6 for others. For simplicity, we set the maximum deadline violation probability at 0 for all simulations. This is achievable when the data rate is positive with probability 1, as will be discussed later.

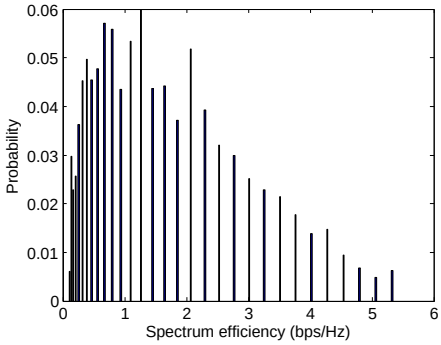


Fig. 2. Distribution of spectrum efficiency.

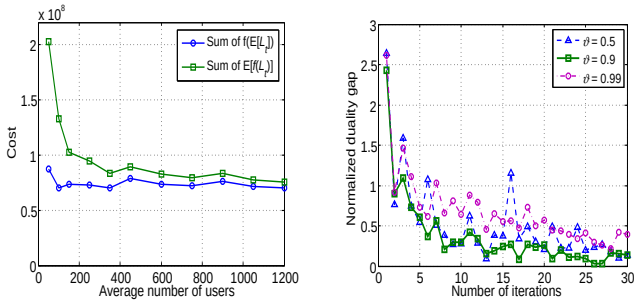
2) Channel and Mobility Profile: We collected a set of Received Signal Strength Indication (RSSI) values from a group of anonymous mobile users to best emulate the spectrum efficiency in cellular networks. RSSI indicates the strength of all received signals, including both desirable signals and interference. We assume that the interference strength is a constant and thus the RSSI value represents the SINR, which determines the spectrum efficiency. We follow the LTE-Advanced standards [40], and map the measured RSSI to the proper modes of Modulation and Coding Scheme (MCS). We use the 5-bit CQI and the distribution of the corresponding spectrum efficiency is shown in Fig. 2. Under this mapping, the data rate is always positive and it is possible to achieve zero deadline violation probability with finite resource. We then model the channel process as a Markov chain, whose transition probabilities are estimated by the empirical transition probabilities obtained from the above measurements. We assume that all users have the same channel statistics.

In 2-cell scenarios, we assume a two-state Markov mobility model, where the probability that a user stays in the c-th cell in the next slot is q_{cc}, and the stationary probability that a user stays in the c-th cell is q_c. We fix the stationary probability at $[q_1, q_2] = [2/3, 1/3]$ for different transition probabilities, so that the traffic arrival pattern is consistent with that in Fig. 1. For the 7-cell scenario, we assume that users only move between the center BS and one of its neighbor BSs. We fix the stationary probability at $[q_1, q_c] = [1/4, 3/4]$ $(2 \leq c \leq 7)$, so that the total traffic at the center BS is two times of that of its neighbors.

B. Convergence of CoSchd

We first demonstrate the asymptotic behavior of the system and the convergence of CoSchd, as shown in Fig. 3. The results in this subsection are obtained by running simulations in the single-cell scenario.

Fig. 3(a) shows the difference between the values of the original problem $\mathcal{P}_0$ and its approximate version $\mathcal{P}_1$. Note that the cost of expected load $f[\mathbb{E}(L_t)]$ is close to the lower bound (Lemma 1) and the expected cost under CoSchd $\mathbb{E}[f(L_t)]$ provides an upper bound on the original problem $\mathcal{P}_0$. As we can see from the figure, the gap between the upper- and lower-bounds becomes smaller as the network scale increases. The two values are close to each other in medium-sized systems, as shown in Fig. 3(a). Hence, minimizing the cost of expected



(a) (b)

Fig. 3. Convergence of CoSchd. (a) The difference between the cost of expected load and the expectation of cost. (b) The evolution of duality gap.

load approximately solves the original problem $\mathcal{P}_0$. Recall that we consider large time-scale scheduling and use a time-slot length of 10 minutes. The measurements in [41, 42] show that in this case, there are usually hundreds of users requesting transmission within each time-slot of 10 minutes. For the rest of the simulation, we set the average number of users in each slot at 400 for single-cell scenarios and 800 for 2-cell scenarios, respectively.

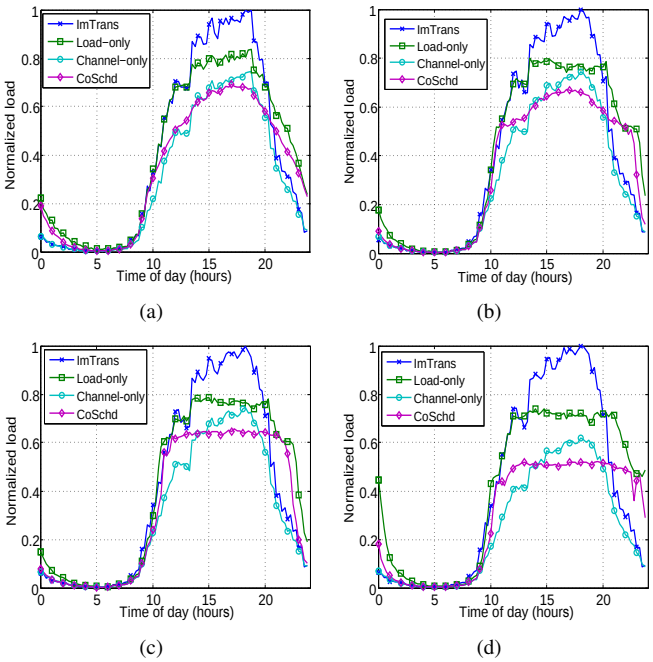
Fig. 3(b) shows the evolution of the duality gap between problem $\mathcal{P}_1$ and its dual problem. The duality gap decreases as the number of iterations increases, and the duality gap is small after several iterations. Comparing the evolutions under different memory factor ϑ (with the same fixed step-size $\alpha = 1$), we can see that with a smaller ϑ, the duality gap decreases faster, but fluctuates more. We set $\vartheta = 0.9$ for the rest of simulations which seems to strike a balance between the convergence speed and fluctuation.

C. Network Load

In this section, we study the network load level under CoSchd for both single-cell and multi-cell scenarios. Because the network load level fluctuates from day to day, we present the one-day load level averaged over the same time-slot of ten days.

1) Single-Cell Scenarios: Fig. 4 shows the network load in single-cell systems. The four subfigures represent different settings. Figs. 4(a) to 4(c) are for the systems with 50% of load being delay-tolerant under different cost functions, while Fig. 4(d) is for the system with 75% of load being delay-tolerant. From Figs. 4(a) to 4(c), we can see that by moving the delay-tolerant traffic into "valleys", the peak load obtained by the load-only policy is about 80% of that under ImTrans. On the other hand, using the channel-only policy, the peak is reduced to about 75% of ImTrans. A similar observation can be made from Fig. 4(d), while the peak load reduction is more significant since there is more delay-tolerant traffic. This finding suggests that channel-awareness can be more effective than load-awareness in wireless systems.

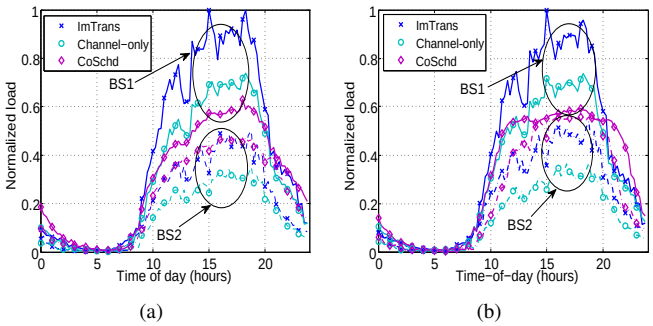
Compared to the load-only and channel-only policies, CoSchd leads to even lower peak consumption by considering both load-awareness and channel-awareness. The additional gain compared to the channel-only policy is about 6% to 12% depending on the cost function. Comparing the load level of CoSchd in Figs. 4(a) to 4(c), we observe that different types



(a) (b) (c) (d)

Fig. 4. Network load level in single-cell systems. (a) 50% delay-tolerant traffic, $f(l) = (l/\bar{C})^\nu$, $\bar{C} = 300$, $\nu = 8$. (b) 50% delay-tolerant traffic, $f(l) = \left([l-\bar{C}]^+/\bar{C}\right)^2, \bar{C} = 250$. (c) 50% delay-tolerant traffic, $f(l) = \left([l-\bar{C}]^+/\bar{C}\right)^2, \bar{C} = 300$. (d) 75% delay-tolerant traffic, $f(l) = \left([l-\bar{C}]^+/\bar{C}\right)^2, \bar{C} = 225$

of cost function result in different shapes of network load. The load at 9pm in Fig. 4(a) is much smaller than that in Fig. 4(c). This is because with the cost function $f(l) = (l/\bar{C})^\nu$ for Fig. 4(a), the scheduler avoids letting users with bad channel conditions transmitting even at periods with medium load levels (e.g. 9pm). In contrast, with the cost function $f(l) = \left([l-\bar{C}]^+/\bar{C}\right)^2$ for Fig. 4(c), more users are allowed to transmit even when the channel conditions are not so good because congestion cost only incurs when the network load exceeds the threshold $\bar{C}$. As a result, the load at periods of medium load levels (e.g., 9am) in Fig. 4(c) can be higher and the peak load is lower. Comparing Fig. 4(b) and Fig. 4(c), we can see that a properly-chosen threshold $\bar{C}$ can result in flat network load and reduced peak load. In practice, the choice of cost function is up to the operators, who can set the cost function based on their capability such as the amount of resource.



(a) (b)

Fig. 5. Load level in multi-cell systems with 50% delay-tolerant traffic (solid-line for BS1 and dash-line for BS2). (a) $f(l) = (l/\bar{C})^\nu$, $\bar{C} = 300$, $\nu = 8$. (b) $f(l) = \left([l-\bar{C}]^+/\bar{C}\right)^2, \bar{C} = 300$

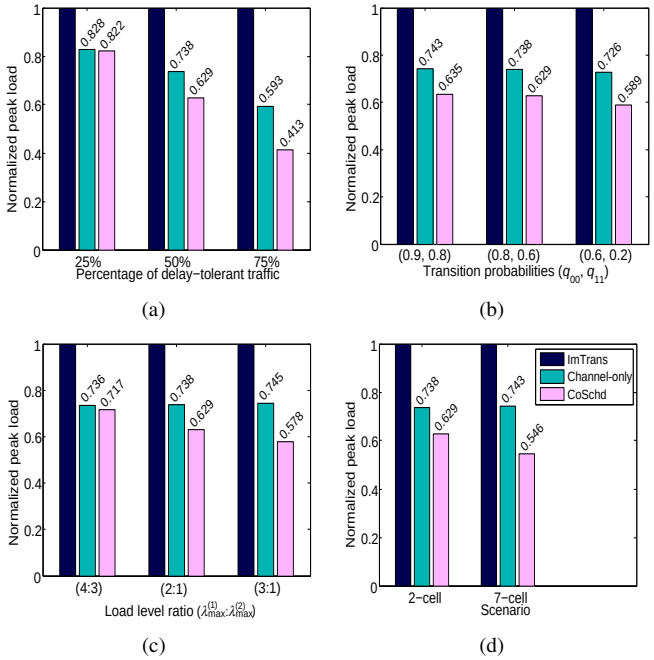


Fig. 6. Peak load level in multi-cell systems (unless expressly stated otherwise we use the following settings: 50% elastic traffic, $(q_{00}, q_{11}) = (0.8, 0.6)$, $\lambda^{(1)}_{\max} : \lambda^{(2)}_{\max} = 2 : 1$. (a) Different elastic percentages. (b) Different transition probabilities. (c) Different load level ratios. (d) Different number of cells.

2) Multi-cell Scenarios: Fig. 5 illustrates the network load level in multi-cell systems with 50% of load being delay tolerant, and with transition probability $(q_{11}, q_{22}) = (0.8, 0.6)$. Figs. 5(a) and 5(b) represent the load level under different cost functions. By moving the delay-tolerant traffic to the neighbor BS (i.e., BS 2), the peak of network load (corresponding to the load in BS 1 at about 6pm) is reduced by about 18% under CoSchd compared to the channel-only policy (Fig. 5(b)). This gain is larger than that in the single-cell settings in Fig. 4. This is because, under the channel-only policy, users defer their transmissions when waiting for good channels. Therefore, a "peak-shedding" effect also occurs under the channel-only approach. Since the temporal fluctuation of traffic is not large in the single-cell system, the room for CoSchd to further move traffic is relatively small. However, when the traffic varies among multiple cells, load-awareness provides higher additional gain.

Next, we evaluate the impact of the different settings on the peak load, i.e., the maximal amount of resource required to serve all the ON users in a BS within a slot, in Fig. 6. Without channel- or load-awareness, the load level is high and imbalance under ImTrans. We normalize the peak loads by the peak-load level of ImTrans. Fig. 6(a) presents the peak-load levels under different percentage of delay-tolerant traffic. From the figure, we can see that as the percentage of delay-tolerant traffic increases, the channel-only and CoSchd policies reduce the peak load more significantly because more data can be deferred to the time/location where the network condition is more favorable. Moreover, as the network load from the real-time traffic becomes less, the benefit of joint channel- and load-awareness becomes more significant and CoSchd achieves even larger performance gain compared to the channel-only policy. For example, the peak load of CoSchd is 14.8% and 31.4% lower than that under the channel-only policy when the percentage of delay-tolerant traffic is 50% and 75%, respectively. Similar trends can be seen from Fig. 6(b), which presents the normalized peak-load level under different transition probability between cells, and Fig. 6(c), which presents the normalized peak-load level under different traffic load levels. From these results, we can see that when the ratio of delay-tolerant traffic is larger, the mobility of users is more frequent, or the traffic load varies more significantly across cells, CoSchd has more opportunities to schedule the traffic from the "peaks" to the "valleys", and results in more balanced load across cells and across time.

Fig. 6(d) shows the peak load levels in the 2-cell and 7-cell scenarios. In the 7-cell scenario, we consider a center BS surrounded by 6 neighbor BSs, where the traffic load in each neighbor BS is half of that in the center BS. From the figure, we can see that the peak load level of the channel-only policy in the 7-cell scenario is similar to that in the 2-cell scenario. In contrast, due to the load-awareness, CoSchd can exploit the additional opportunities to transfer the elastic traffic and achieve much lower peak-load level. Specifically, the peak load level under CoSchd in the 7-cell scenario is only 54.6% of that under ImTrans. Since 50% of traffic is real-time traffic, we can infer that most delay-tolerant traffic has been shifted to the neighbor cells by CoSchd thanks to its load-awareness.

VI. Conclusions

In this paper, we study jointly load- and channel-aware policies for scheduling delay-tolerant traffic for reducing cellular congestion. We present a decomposition technique for solving the large-scale MDP induced from the optimal scheduling problem. Despite the high complexity of the large-scale MDP, we develop a distributed framework, called CoSchd, and show its asymptotic-optimality in the many-source regime. An approximate version of CoSchd is also proposed to reduce the complexity.

The results in this paper are of both practical and theoretical values. Practically, our proposed policy can be implemented in a distributed manner in real systems. Further, our comparative evaluations provide cellular operators with operation guidelines to decide the most appropriate approaches. Specifically, our numerical results suggest that channel-awareness is rather important in wireless networks. For single-cell systems, channel-only may be preferred due to its simplicity and relatively good performance. For multi-cell systems with load variations, CoSchd can attain significant additional gains. Theoretically, the joint approach provides an optimal benchmark for comparing with other solutions. Moreover, the decomposition technique and the proposed CoSchd algorithm can potentially be applied to other large-scale MDP, where multiple agents are weakly coupled through sharing common resources.

Appendix A
Proof of Proposition 2

To prove the optimality of the proposed dual decomposition approach, we show that problem $\mathcal{P}_1$ can be reformulated to a convex optimization problem $\mathcal{P}_2$, albeit with an exponentially

large number of decision variables. Thus, strong duality holds between $\mathcal{P}_2$ and its dual, named $\mathcal{D}_2$.

First, we show that any policy Γ can be represented by a stochastic policy Ψ as follows. Note that each causal policy Γ makes decision based on the history of the arrival sequence and channel processes. To represent the history, for each user $i \in \mathcal{I}$, we introduce $\tilde{A}_i(t)$ to represent its present status in time-slot t. Namely, if the arrival time A_i of user i is equal to a_i (we let $a_i = N$ represent the event that user i does not appear), then $\tilde{A}_i(t) = -1$ if $a_i > t$, and $\tilde{A}_i(t) = a_i$ if $a_i \le t$. Recall that $R_i(t)$ $(i = 0, 1, \ldots, N-1)$ is the channel process of user $i \in \mathcal{I}$, and $m = |\mathcal{I}|$ is the number of users. Hence, the history of the system up to time-slot t is given by

$$\boldsymbol{S}_t = [\tilde{\boldsymbol{A}}_t \ \tilde{\boldsymbol{R}}_t],$$

where $\tilde{\boldsymbol{A}}_t = [\tilde{A}_1(t), \tilde{A}_2(t), \ldots, \tilde{A}_m(t)]^{\mathrm{T}}$ and

$$\tilde{\boldsymbol{R}}_t = \begin{bmatrix} R_1(0) & R_1(1) & \ldots & R_1(t) \\ R_2(0) & R_2(1) & \ldots & R_2(t) \\ \vdots & \vdots & \ddots & \vdots \\ R_m(0) & R_m(1) & \ldots & R_m(t) \end{bmatrix}.$$

Let Ω be the set of possible realizations of arrival sequence and channel processes, i.e., the possible realization of $\boldsymbol{S}_{N-1}$. Then, each policy Γ can be represented by a stochastic policy Ψ, which is a $\Omega \mapsto [0,1]^{m \times N}$ mapping: for each $\boldsymbol{s} \in \Omega$,

$$\Psi(\boldsymbol{s}) = \begin{bmatrix} \psi_1(\boldsymbol{s}_0) & \psi_1(\boldsymbol{s}_1) & \ldots & \psi_1(\boldsymbol{s}_{N-1}) \\ \psi_2(\boldsymbol{s}_0) & \psi_2(\boldsymbol{s}_1) & \ldots & \psi_2(\boldsymbol{s}_{N-1}) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_m(\boldsymbol{s}_0) & \psi_m(\boldsymbol{s}_1) & \ldots & \psi_m(\boldsymbol{s}_{N-1}) \end{bmatrix},$$

where $\boldsymbol{s}_t$ is the history of arrival sequence and channel processes up to time-slot t for the realization $\boldsymbol{s}$, and $\psi_i(\boldsymbol{s}_t) \in [0,1]$ is the transmission probability of user i in time-slot t.

Second, we study the expected resource consumed by user i under $\Psi(\boldsymbol{s})$. For each $\boldsymbol{s} \in \Omega$ where user i arrives in time-slot a_i, we can calculate the probability that user i transmits in slot $a_i + w$ as follows

$$\varphi_i(\boldsymbol{s}_t) = \begin{cases} \psi_i(\boldsymbol{s}_{a_i}), & t = a_i \\ \psi_i(\boldsymbol{s}_{a_i+w}) \prod_{w'=0}^{w-1} [1 - \psi_i(\boldsymbol{s}_{a_i+w'})], & \\ & t = a_i + w, 0 < w \le D_i - 1 \\ 0, & \text{otherwise.} \end{cases}$$

For given $\boldsymbol{s}$, the expected consumed resource of user i in time-slot t is

$$c'_{i,t}(\boldsymbol{s}, \Psi) = \frac{b_i \varphi_i(\boldsymbol{s}_t)}{R_i(t)}.$$

In addition, note that all users should transmit before expiration. Hence,

$$\sum_{w=0}^{D_i-1} \varphi_i(\boldsymbol{s}_{a_i+w}) = 1, \quad \boldsymbol{s} \in \Omega, i \in \mathcal{I}. \tag{23}$$

Moreover, using the relationship between $\varphi_i(\cdot)$ and $\psi_i(\cdot)$, a $\varphi_i(\cdot)$ satisfying (23) can be mapped to a policy Ψ [2].

[2]If $\sum_{w'=0}^{w} \varphi_i(\boldsymbol{s}_{0:a_i+w'}) = 1$ for some $w < D_i - 1$, then for $w' > w$, $\psi_i(\boldsymbol{s}_{i,w})$ can be artificially set to be 0, which will not affect the behavior of Ψ.

Consequently, problem $\mathcal{P}_1$ is equivalent to

$$\begin{aligned} (\mathcal{P}_2) \quad \underset{\Psi, h'}{\text{minimize}} \quad & F = \sum_{t=0}^{N-1} f\big(m h'_t\big), \\ \text{subject to} \quad & \sum_{w=0}^{D_i-1} \varphi_i(\boldsymbol{s}_{a_i+w}) = 1, \ \boldsymbol{s} \in \Omega, i \in \mathcal{I}, \\ & l'_t(\Psi)/m \le h'_t, \quad t = 0, 1, \ldots, N-1, \end{aligned}$$

where

$$l'_t(\Psi) = \sum_{\boldsymbol{s} \in \Omega} \sum_{i \in \mathcal{I}} \pi(\boldsymbol{s}) c'_{i,t}(\boldsymbol{s}, \Psi). \tag{24}$$

We can verify that $\mathcal{P}_2$ is a convex optimization problem because $f(\cdot)$ is a convex function and all the constraints are linear. However, we note that it is impractical to solve $\mathcal{P}_2$ directly because of its large number of variables. Recall that there are $m \times N$ decision variable for each possible state. Assume the channel state of each user can be quantized to J values, then there are $J^{m \times N}$ possible states, and thus $m \times N \times J^{m \times N}$ decision variables, which is clearly intractable. We note that the formulation can be considered as a linear representation of a centralized Markov Decision Policy, which clearly suffers the curve of dimensionality.

Again, we resort to the dual decomposition approach to study $\mathcal{P}_2$. Similar to the approach in Section IV, we can introduce a dual variable for each time slot, and then rearrange the variables that belong to each user. Then, we have a similar format as in $\mathcal{SP}_0$ and $\mathcal{SP}_i$. The dual decomposition approach can also be applied to solve problem $\mathcal{P}_2$ and the strong duality holds.

Appendix B
Proof of Proposition 3

Let $\tilde{F}^{(m)}_{\mathrm{CoSchd}(\alpha)}$ be the cost value of $\mathcal{P}_1$ under $\mathrm{CoSchd}(\alpha)$. Because $\tilde{F}^{(m)}_{\mathrm{CoSchd}(\alpha)} \le F^{(m)}_{\sharp} + \frac{\alpha G^2}{2}$ according to Lemma 1, we only need to show that $\lim_{m\to\infty} F^{(m)}_{\mathrm{CoSchd}(\alpha)} = \lim_{m\to\infty} \tilde{F}^{(m)}_{\mathrm{CoSchd}(\alpha)}$. To achieve this, we first consider the single-class system, i.e., $K = 1$. Since $F^{(m)}_{\mathrm{CoSchd}(\alpha)}$ is the sum of the expected costs in each slot, i.e., $F^{(m)}_{\mathrm{CoSchd}(\alpha)} = \sum_{t=0}^{N-1} \mathbb{E}\big[f(L_t)\big]$, we can prove Proposition 3 if we can show that under $\mathrm{CoSchd}(\alpha)$,

$$\lim_{m\to\infty} \mathbb{E}[f(L_t)] = f\big(\mathbb{E}[L_t]\big), \tag{25}$$

which implies that the "expectation of the cost" approaches the "cost of the expectation" as m increases. As will be seen shortly, this can be verified by the fact that, under $\mathrm{CoSchd}(\alpha)$, each user operates independently when the congestion signal $\boldsymbol{\beta}$ is fixed.

Specifically, fix a time-slot t. Let Y_i $(i = 1, 2, \ldots, m)$ be the amount of resource required by the i-th user in slot t. Since all users in the same class have identical traffic and channel statistics, Y_i's $(i = 1, 2, \ldots, m)$ are i.i.d. random variables. Let $\mathbb{E}[Y_i] = \mu_Y$. The load level is $L_t = \sum_{i=1}^{m} Y_i$ and the normalized load level is $\tilde{L}_t = \frac{1}{m} \sum_{i=1}^{m} Y_i$ with $\mathbb{E}[\tilde{L}_t] = \mu_Y$. Since the file size B_i is bounded, the amount of resource Y_i is bounded and we let $y_{\max} = \max Y_i = \frac{\max B_i}{r_2}$.

Using the Chernoff bound, we have that for a given $\delta > 0$,

$$\mathbb{P}\{\tilde{L}_t \leq \mu_Y - \delta\} \leq e^{-mI_Y(\delta)} \quad (26)$$

$$\mathbb{P}\{\tilde{L}_t \geq \mu_Y + \delta\} \leq e^{-mI_Y(\delta)}, \quad (27)$$

where $I_Y(\delta)$ is a positive number independent of m.

Next, we bound $\mathbb{E}[f(L_t)]$ using the above results and the properties of the cost function. $\mathbb{E}[f(L_t)]$ can be calculated as follows

$$\begin{aligned} & \mathbb{E}[f(L_t)] = \mathbb{E}[\tilde{f}(\tilde{L}_t)] \\ = & \int_0^\infty \tilde{f}(l)\phi_{\tilde{L}_t}(l)\mathrm{d}l \\ = & \left[\int_0^{\mu_Y-\delta} + \int_{\mu_Y-\delta}^{\mu_Y+\delta} + \int_{\mu_Y+\delta}^{\infty}\right]\tilde{f}(l)\phi_{\tilde{L}_t}(l)\mathrm{d}l, \end{aligned}$$

where $\phi_{\tilde{L}_t}(\cdot)$ is the probability density function of $\tilde{L}_t$. Note that $\tilde{f}(l)$ is increasing in l. Thus, from (26) and (27), we have

$$\mathbb{E}[f(L_t)] \geq (1 - 2e^{-mI_Y(\delta)})\tilde{f}(\mu_Y - \delta), \quad (28)$$

and

$$\mathbb{E}[f(L_t)] \leq \tilde{f}(\mu_Y + \delta) + 2e^{-mI_Y(\delta)}\tilde{f}(y_{\max}). \quad (29)$$

For any $\epsilon > 0$, by the continuity of $\tilde{f}(l)$, we can choose a $\delta > 0$ such that $\tilde{f}(\mu_Y - \delta) \geq \tilde{f}(\mu_Y) - \epsilon/2$ and $\tilde{f}(\mu_Y + \delta) \leq \tilde{f}(\mu_Y) + \epsilon/2$. Combining with (28) and (29), we know that there exists an m_1 such that for all $m \geq m_1$, we have

$$|\mathbb{E}[f(L_t)] - f(\mathbb{E}[L_t])| = |\mathbb{E}[f(L_t)] - \tilde{f}(\mu_Y)| \leq \epsilon,$$

and thus (25) holds by taking $\epsilon \to 0$.

For multi-class systems, similar properties can be obtained. Let $\mathcal{I}_k$ ($k = 1, 2, \ldots, K$) be the index set of class-k users. For all $i \in \mathcal{I}_k$, Y_i are i.i.d. random variables because within one class, all users have identical traffic and channel characteristics. For $i \in \mathcal{I}_k$, let $\mu_Y^{(k)} = \mathbb{E}[Y_i]$, and $I_Y^{(k)}(\delta)$ be the value satisfying Eqs. (26) and (27) for class-k users. Then, the load level is $L_t = \sum_{k=1}^K \sum_{i\in\mathcal{I}_k} Y_i$ and the normalized load level is given by

$$\tilde{L}_t = \frac{L_t}{m} = \sum_{k=1}^K \lambda_k \tilde{L}_t^{(k)},$$

where $\tilde{L}_t^{(k)} = \frac{1}{m_k}\sum_{i\in\mathcal{I}_k} Y_i$, and $\mathbb{E}[\tilde{L}_t] = \sum_{k=1}^K \lambda_k \mu_Y^{(k)}$. Because the event $\tilde{L}_t \leq \mathbb{E}[\tilde{L}_t] - \delta$ implies that there exists at least one k satisfying that $\tilde{L}_t^{(k)} \leq \mu_Y^{(k)} - \delta$, we have

$$\begin{aligned} & \mathbb{P}\{\tilde{L}_t \leq \mathbb{E}[\tilde{L}_t] - \delta\} \\ \leq & \sum_{k=1}^K \mathbb{P}\{\tilde{L}_t^{(k)} \leq \mu_Y^{(k)} - \delta\} \leq Ke^{-mI_Y^*(\delta)} \end{aligned} \quad (30)$$

where $I_Y^*(\delta) = \min_{k\in\{1,2,\ldots,K\}} \lambda_k I_Y^{(k)}(\delta)$.

Similarly, when considering the other side of deviation, we have

$$\mathbb{P}\{\tilde{L}_t \geq \mathbb{E}[\tilde{L}_t] + \delta\} \leq Ke^{-mI_Y^*(\delta)}. \quad (31)$$

Therefore, using the same approach in the single-class case, $\mathbb{E}[f(L_t)]$ can be made as close to $\tilde{f}(\mathbb{E}[\tilde{L}_t])$ as desired.

The conclusion then follows by adding the expected costs from time-slot 0 to $N - 1$.

REFERENCES

[1] H. Wu, X. Lin, X. Liu, K. Tan, and Y. Zhang, "Decomposition of large-scale MDPs for wireless scheduling with load- and channel-awareness," in *Information Theory and Applications Workshop (ITA)*, San Diego, CA, Feb. 2014.

[2] S. Ha, S. Sen, C. Joe-Wong, Y. Im, and M. Chiang, "TUBE: Time-dependent pricing for mobile data," in *Proc. ACM SIGCOMM'12*, Helsinki, Finland, Aug. 2012, pp. 247–258.

[3] M. Ra, J. Paek, A. Sharma, R. Govindan, M. Krieger, and M. Neely., "Energy-delay tradeoffs in smartphone applications," in *Proc. ACM MobiSys'10*, San Francisco, CA, June 2010, pp. 255 – 270.

[4] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci, "Taming user-generated content in mobile networks via drop zones," in *Proc. IEEE INFOCOM*, Shanghai, China, Apr. 2011, pp. 2840 – 2848.

[5] A. Schulman, V. Navda, R. Ramjee, N. Spring, P. Deshpandez, C. Grunewald, K. Jain, and V. N. Padmanabhan, "Bartendr: A practical approach to energy-aware cellular data scheduling," in *Proc. ACM MobiCom' 10*, Chicago, Sept. 2010, pp. 85 – 96.

[6] C. Shi, K. Joshi, R. K. Panta, M. H. Ammar, and E. W. Zegura, "CoAST: collaborative application-aware scheduling of last-mile cellular traffic," in *Proc. ACM MobiSys'14*, 2014, pp. 245–258.

[7] A. Balasubramanian, R. Mahajan, and A. Venkataramani, "Augmenting mobile 3G using WiFi," in *Proc. of ACM MobiSys'10*, San Francisco, CA, June 2010, pp. 209–222.

[8] A. Chakraborty, V. Navda, V. N. Padmanabhan, and R. Ramjee, "Coordinating cellular background transfers using LoadSense," in *Proc. ACM MobiCom' 13*, Miami, FL, USA, 2013.

[9] X. Liu, E. K. P. Chong, and N. B. Shroff, "A framework for opportunistic scheduling in wireless networks," *Computer Networks*, vol. 41, no. 4, pp. 451 – 474, Mar. 2003.

[10] X. Lin, N. B. Shroff, and R. Srikant, "A tutorial on cross-layer optimization in wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 81, pp. 1452 – 1463, 2006.

[11] A. Asadi and V. Mancuso, "A survey on opportunistic scheduling in wireless communications," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 4, pp. 1671–1688, 2013.

[12] B. Li and A. Eryilmaz, "Optimal distributed scheduling under time-varying conditions: A Fast-CSMA algorithm with applications," *IEEE Trans. on Wireless Communications*, vol. 12, no. 7, pp. 3278–3288, 2013.

[13] W. Du, M. Li, and J. Lei, "CO-MAP: improving mobile multiple access efficiency with location input," *IEEE Trans. on Wireless Communications*, vol. 13, no. 12, pp. 6643–6654, 2014.

[14] X. Wei and M. J. Neely, "Power aware wireless file downloading: A constrained restless bandit approach," in *Proc. IEEE WiOpt*. IEEE, 2014, pp. 482–489.

[15] A. K.-L. Mok, "Fundamental design problems of distributed systems for the hard-real-time environment," Ph.D. dissertation, MIT, Cambridge, MA, May. 1983.

[16] S. Shakkottai and R. Srikant, "Scheduling real-time traffic with deadlines over a wireless channel," *ACM/Baltzer Wireless Networks*, vol. 8, no. 1, pp. 13 – 26, Jan 2002.

[17] H. Wu, Y. Zhang, and X. Liu, "Laxity-based opportunistic scheduling with flow-level dynamics and deadlines," in *IEEE Wireless Communications and Networking Conference (WCNC)*, Shanghai, China, Apr. 2013.

[18] M. J. Neely, "Opportunistic scheduling with worst case delay guarantees in single and multi-hop networks," in *Proc. IEEE INFOCOM*. IEEE, 2011, pp. 1728–1736.

[19] I.-H. Hou and P.-C. Hsieh, "QoE-optimal scheduling for on-demand video streams over unreliable wireless networks," in *Proc. ACM MobiHoc'15*. Hangzhou, China: ACM, 2015.

[20] H. Wu, X. Lin, X. Liu, and Y. Zhang, "Application-level scheduling with probabilistic deadline constraints," *IEEE/ACM Trans. on Networking*, to appear.

[21] A. Fu, E. Modiano, and J. N. Tsitsiklis, "Optimal transmission scheduling over a fading channel with energy and deadline constraints," *IEEE Transactions on Wireless Communications*, vol. 5, no. 3, pp. 630 – 641, Mar. 2006.

[22] M. Zafer and E. Modiano, "Minimum energy transmission over a wireless channel with deadline and power constraints," *IEEE Transactions on Automatic Control*, vol. 54, no. 12, pp. 2841 – 2852, Dec. 2009.

[23] J. Lee and N. Jindal, "Asymptotically optimal policies for hard-deadline scheduling over fading channels," *IEEE Transactions on Information Theory*, vol. 59, no. 4, pp. 2482 – 2500, Apr. 2013.

[24] N. Gast, B. Gaujal, and J.-Y. Le Boudec, "Mean field for Markov Decision Processes: from discrete to continuous optimization," *Automatic Control, IEEE Trans. on*, vol. 57, no. 9, pp. 2266–2280, 2012.

[25] D. S. Bernstein, S. Zilberstein, and N. Immerman, "The complexity of decentralized control of Markov decision processes," in *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2000, pp. 32–37.

[26] C. Amato, G. Chowdhary, A. Geramifard, N. K. Ure, and M. J. Kochenderfer, "Decentralized control of partially observable Markov decision processes," in *The 52nd IEEE Conference on Decision and Control (CDC)*, 2013.

[27] R. Becker, S. Zilberstein, V. Lesser, and C. V. Goldman, "Solving transition independent decentralized Markov decision processes," *Journal of Artificial Intelligence Research*, vol. 22, no. 1, pp. 423–455, 2004.

[28] N. Meuleau, M. Hauskrecht, K.-E. Kim, L. Peshkin, L. P. Kaelbling, T. L. Dean, and C. Boutilier, "Solving very large weakly coupled Markov decision processes," in *Proc of the 15th National Conference on Artificial Intelligence (AAAI)*, 1998, pp. 165–172.

[29] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz, "Primary user behavior in cellular networks and implications for dynamic spectrum access," *IEEE Comm. Mag.*, vol. 47, no. 3, pp. 88–95, Mar. 2009.

[30] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, "Towards dynamic energy-efficient operation of cellular network infrastructure," *IEEE Communications Magazine*, vol. 49, no. 6, 2011.

[31] R. Margolies, A. Sridharan, V. Aggarwal, R. Jana, N. Shankaranarayanan, V. A. Vaishampayan, and G. Zussman, "Exploiting mobility in proportional fair cellular scheduling: Measurements and algorithms," in *Proc. IEEE INFOCOM*, 2014.

[32] A. Seetharam, K. Kurose, D. Goeckel, and G. Bhanage, "A Markov chain model for coarse timescale channel variation in an 802.16e wireless network," in *Proc. IEEE INFOCOM'12*, Mar. 2012, pp. 1800 – 1801.

[33] A. J. Nicholson and B. D. Noble, "Breadcrumbs: forecasting mobile connectivity," in *Proc of ACM MobiCom'08*, New York, NY, USA, 2008, pp. 46–57.

[34] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[35] M. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, 1st ed. Wiley, 2005.

[36] D. P. Bertsekas, *Dynamic Programming and Optimal Control, Volume I*, 3rd ed. Athena Scientific, 2005.

[37] E. Jung, F. Maker, T. Cheung, X. Liu, and V. Akella, "Markov decision process (MDP) framework for software power optimization using call profiles on mobile phones," *Design Automation for Embedded Systems*, vol. 14, pp. 131–159, 2010.

[38] Y. Chon, E. Talipov, H. Shin, and H. Cha, "SmartDC: Mobility prediction-based adaptive duty cycling for everyday location monitoring," *Mobile Computing, IEEE Transactions on*, vol. 13, no. 3, pp. 512–525, 2014.

[39] A. Nedic and A. Ozdaglar, "Approximate primal solutions and rate analysis for dual subgradient methods," *SIAM Journal on Optimization*, vol. 19, no. 4, pp. 1757–1780, 2009.

[40] 3GPP TS 36.213 V11.0.0, "Physical layer procedures," Sept. 2012.

[41] C. Williamson, E. Halepovic, H. Sun, and Y. Wu, "Characterization of CDMA2000 cellular data network traffic," in *IEEE Conference on Local Computer Networks*. IEEE, 2005, pp. Z000–719.

[42] M. Laner, P. Svoboda, S. Schwarz, and M. Rupp, "Users in cells: a data traffic analysis," in *IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2012, pp. 3063–3068.