Qos-Aware Predictive Rate Allocation over Heterogeneous Wireless Interfaces

Sherif ElAzzouni*, Eylem Ekici* and Ness B. Shroff†
*Department of Electrical and Computer Engineering
†Department of Electrical and Computer Engineering & Department of Computer Science and Engineering
Ohio State University, USA
Email: elazzouni.1@osu.edu, ekici@ece.osu.edu, shroff@ece.osu.edu

***Abstract*—The rapid growth of mobile data traffic is straining cellular networks. A natural approach to alleviate cellular networks congestion is to use, in addition to the cellular interface, secondary interfaces such as WiFi, Dynamic spectrum and mmWave to aid cellular networks in handling mobile traffic. The fundamental question now becomes: How should traffic be distributed over different interfaces, taking into account different application QoS requirements and the diverse nature of radio interfaces. To this end, we propose the Discounted Rate Utility Maximization (DRUM) framework with interface costs as a means to quantify application preferences in terms of throughput, delay, and cost. The flow rate allocation problem can be formulated as a convex optimization problem. However, solving this problem requires non-causal knowledge of the time-varying capacities of all radio interfaces. To this end, we propose an online predictive algorithm that exploits the predictability of wireless connectivity for a small look-ahead window w. We show that, under some mild conditions, the proposed algorithm achieves a constant competitive ratio independent of the time horizon T. Furthermore, the competitive ratio approaches 1 as the prediction window increases. We also propose another predictive algorithm based on the "Receding Horizon Control" principle from control theory that performs very well in practice. Numerical simulations serve to validate our formulation, by showing that under the DRUM framework: the more delay-tolerant the flow, the less it uses the cellular network, preferring to transmit in high rate bursts over the secondary interfaces. Conversely, delay-sensitive flows consistently transmit irrespective of different interfaces' availability. Simulations also show that the proposed online predictive algorithms have a near-optimal performance compared to the offline prescient solution under all considered scenarios.**

I. Introduction

Cellular networks are witnessing unprecedented growth of demand on mobile traffic data. This growth is straining cellular networks, as it is becoming clear that operators cannot increase capacity to meet the demand by deploying more base stations. Thus, alternative approaches to capacity increase must be undertaken to provide users with the bandwidth needed to support their applications. One such approach is exploiting alternative Radio Access Technologies (RATs) that may be available to smart phones such as WiFi, Bluetooth, mmWave, etc., to aid the cellular network in data transmission [1]. Furthermore, the Dynamic Spectrum Access (DSA) technology [2] could also be used to aid the cellular network in data transfer. DSA is a technology that enables mobile users to use spectrum that belongs to some other entity, as long as the spectrum owner does not experience any interference caused by DSA. For example, users could limit their use of that spectrum to times when the spectrum owner is absent. We collectively refer to those alternative RATs (WiFi, Bluetooth, DSA, etc.) as Secondary Interfaces (SI). An interesting question arises from this proposal: **How can we best distribute mobile traffic over heterogeneous RATs**, taking into account the inherent differences between interfaces? Cellular networks are ubiquitous but have high cost on the operator in terms of congesting the cellular network, as well as having high energy consumption that may drain the phone battery. WiFi networks, if accessible, are usually free, but WiFi coverage is not always present. Furthermore, it is typical for public places to throttle WiFi rates. DSA is usually free-of-charge and has high rates. However, the connection is intermittent as the user is only allowed to access this spectrum when the spectrum owner is absent. These heterogeneous properties necessitate a framework for mobile users to take all these factors into account and make a decision on traffic allocation that is optimal in terms of throughput, Quality of Service (QoS) constraints satisfaction and cost.

In general, two approaches have been employed to address this question. The first approach is **Intelligent Network Selection** by choosing the suitable RAT for each application. This approach assumes that the user is allowed to use only one RAT at any given time. The current default policy employed by Android phones falls into this category: Android default policy is to choose WiFi over LTE whenever possible with the option of setting some applications to only use WiFi. This is known as delayed offloading, where delay tolerant applications are only allowed to use WiFi. If WiFi is not immediately available, then these applications will delay their transmissions up to a deadline or until a WiFi connection is established. This policy was proposed and analyzed in [3] and [4, 5], respectively. However, this solution is more suitable for 3G deployments, where WiFi consistently offers significantly higher rates than the cellular network. More recently, however, [6] has shown that this is not the case in LTE deployments. In particular, it was shown that LTE outperforms WiFi in terms of rate 40%

This work has been funded in part by NSF grants CNS-1618566, CNS-1421576, CNS-1731698, and Office of Naval Research grant N00014-17-1-2417.

of the time. One feature that has been used in the literature is the predictability of wireless connectivity. In particular, it was shown in [7] that short-term future wireless connectivity can be forecast accurately. Thus, a slightly delay-tolerant application can delay a transmission if the connectivity forecast indicates that preferable network conditions will be available in the near future. Using this predictive ability, [8] modeled the problem as a Finite-Horizon Markov Decision Process where the user follows a network-selection policy that minimizes the expected energy when uploading a single file before a deadline. In [9], a Lyapunov drift-plus-penalty approach was taken for network selection with the aim of minimizing power subject to queue stability. In [10], the different applications' QoS needs in terms of throughput, delay and cellular cost were quantified as a utility function. Then, each user solves an open-loop planning problem to choose the best transmission time for each application according to its QoS and the availability of WiFi at any given time. Aside from client-controlled solutions, [11, 12] proposed network-controlled centralized solutions to the problem, where a single entity has a perfect view of all RAT states, of all users, and can assign users to RATs accordingly. The centralized control, however, is not realistic in today's settings as different networks are often controlled by different entities.

The second possible approach to solve the problem is **Simultaneously utilizing all RATs at any given time**. This is sometimes referred to as multihoming. This approach is more flexible as it gives the user the opportunity to utilize the entire bandwidth from all available RATs at any given time. Our problem formulation takes this approach. The current de-facto solution for utilizing different RATs is the MPTCP protocol [13]. However, MPTCP suffers from a variety of problems such as high energy consumption [14] and over-utilization of the cellular link [15], which might cause an increased monetary cost to the user. Several other approaches have been proposed to exploit different RATs: [16] models the scheduling over different RATs as a Mixed-Linear-Integer-Program and propose a greedy heuristic to defer delay tolerant flows to later times. In [17], the problem is considered with flow-interface assignment constraints. A minimum deficit round-robin policy is proposed and it is shown that this policy is max-min fair. However, the result strongly depends on the policy being work-conserving, which might cause over-utilization of a metered cellular link. In [15], managing several RATs is studied for the case of transmission of video chunks. The problem is formulated as a 0-1 min-knapsack problem that takes into account different bandwidths, usage costs, and deadlines of video chunks. A practical online heuristic is proposed and good performance is established via simulations. In [18], the authors proposed an integrated transport layer that modifies SCTP to allow the exploitation of heterogeneous RATs in vehicular networks. The paper uses a Network Utility Maximization formulation with link costs. However, the issues of QoS differentiation per application, application-level fairness and temporal variation in secondary capacity are not addressed as implementation details are emphasized.

There are many challenges in solving the problem of rate-allocation over heterogeneous RATS: 1) Cellular and Secondary interfaces have different costs. While the cellular network is usually metered, secondary RATs such as public WiFi are often free to use. This may cause the optimal policy in terms of throughput, cost, and QoS to be **non-work conserving** which complicates the problem. 2) Secondary interfaces are inherently intermittent and unreliable. WiFi have limited coverage and DSA is only allowed to access spectrum in absence of the spectrum owner. 3) Different applications have different requirements in terms of delay and throughput. Thus, a good rate-allocation policy has to incorporate individual application requirements when allocating rates. However, all these applications share the same RATs that have limited capacity. Thus, the allocations of all applications are coupled. Our contributions can be summarized as follows:

1) We apply the DRUM framework, proposed in [19], in the context of allocating rates to different cellular users to exploit temporal diversity, to the problem of application rate allocation over different RATs. We demonstrate that using the DRUM framework with a discount tied to application delay sensitivity results in a fair allocation with desirable characteristics in terms of balancing the per-application trade-off between throughput, delay, and cost.

2) We propose two online low complexity algorithms that exploit limited look-head predictions of future connectivity.

3) We analyze one of those online algorithms and show that: in the presence of a prediction window of length w time slots, under some mild conditions, the online algorithm achieves a reward that is no less than $(1 - \frac{c}{w+1})$Reward(OPT), where c is a constant and OPT is the prescient offline solution. Thus, the proposed algorithm is constant-competitive independent of the time horizon T, and approaches the optimal reward as the prediction window increases. Simulations show that, in practice, these proposed algorithms perform much better than the theoretical bound under all considered scenarios.

Our work relates to [9] by being predictive and QoS-aware. The difference is that [9] does not offer differentiated service to flows. The formulation in [18] relates to our formulation of utility with link costs. However, [18] does not differentiate between flows, nor does it consider time variations of secondary RAT, and instead, attempts to solve a static optimization problem every time slot. Perhaps the closest work to our problem are [10, 16] which considered all three factors of throughput, delay and cost. [16] considered inelastic traffic that needs to be served over T time slots, whereas [10] considered a mixture of inelastic and fixed-time elastic traffic. However, both of these papers assumed full knowledge (or a good estimate) on all future connectivity, and used heuristics to find good solutions for the hard traffic assignment problem.

II. System Model

We consider a mobile user running N applications, where each application creates a flow i. At each time slot, the mobile user can use the cellular network, the secondary network or both to transmit traffic belonging to flow i. We denote the rate

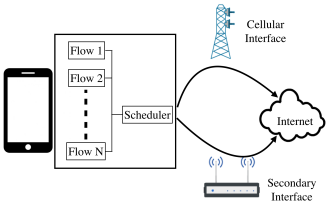

Fig. 1. System Model

received by flow i over the cellular network at time t as $y_i[t]$, and the rate received by flow i over the secondary network as $x_i[t]$.

A. Channel Model: We consider a smart-phone with two interfaces as shown in Fig. 1: a cellular interface and a secondary interface. The extension of the formulation and the online algorithms to the case of multiple secondary interfaces is straightforward. The secondary interface has a time varying capacity $c[t]$ every time slot to capture the effects of intermittence, unreliability, and possible user mobility. We do not have any statistical assumptions on $c[t]$. We assume the user can accurately predict the secondary capacity up to a future window of w time slots. The predictability assumption has been used extensively for similar problems [3, 9, 20], and the feasibility of WiFi prediction was shown in [7, 21]. We assume that the cellular interface has a constant normalized capacity equal to 1 every time slot, i.e., we assume that the cellular operator offers a constant rate to the user throughout the time-horizon. This captures the effect of ubiquity of the cellular network in contrast to intermittence of the secondary network. Although the cellular network is affected by fading, we assume that the cellular operator can employ scheduling, resource-block allocation, MIMO, etc., to guarantee that the client gets a constant rate every time slot over the problem horizon. The "time-slot" in the system is in the order of a few seconds, a sufficient time for the state of secondary interface connectivity to change. In our model, the rates $x_i[t]$ and $y_i[t]$ take continuous non-negative values. Finally, we assume that all queues carrying different flows are infinitely backlogged, i.e., we assume that the flows are elastic.

B. Flow Utility: We use the Discounted Rate Utility framework introduced in [19] to capture the utility of flow i

Definition 1. *(β-Discounted Rate [19]): For a given $\beta \in [0,1]$, we define the β-discounted rate of flow i at time $t \geq 0$ as*

$$R_i^{(\beta_i)}[t] \triangleq \frac{\sum_{\tau=0}^{t} \beta^{t-\tau}(x_i[\tau]+y_i[\tau])}{\sum_{\tau=0}^{t}\beta^{t-\tau}} \tag{1}$$

As an illustrative example, we write down the β-discounted rate for $\beta = 0$, $\beta \in (0,1)$ and $\beta = 1$ as follows:

$$R_i^{(\beta_i)}[t] = \begin{cases} x_i[t]+y_i[t] & \text{if } \beta = 0, \\ \dfrac{\sum_{\tau=0}^{t} \beta^{t-\tau}(x_i[\tau]+y_i[\tau])}{\sum_{\tau=0}^{t}\beta^{t-\tau}} & \text{if } \beta \in (0,1), \\ \dfrac{1}{t}\displaystyle\sum_{\tau=0}^{t}(x_i[\tau]+y_i[\tau]) & \text{if } \beta = 1. \end{cases} \tag{2}$$

The β-discounted rate ties the utility of a certain flow to both the throughput and average delay by adding a weight β to the history of allocated rates. Closer inspection of (2) shows when $\beta = 0$, the β-discounted rate is equal to the instantaneous rate, representing maximum delay sensitivity as no weight is given to the rate allocation history. When $\beta = 1$, the β-discounted rate represents the time-average of allocated rates since time 0. This is suitable for modeling a flow with no delay sensitivity. To summarize, an increase in β models less delay sensitivity, thus, an application with a high value of β can afford to wait for "favorable" transmission opportunities, whereas lower β represents a flow that emphasizes importance of delay over possible cost. Finally, we model the cost of using the cellular network as a linear coefficient p_c. Thus, every flow i has to pay a cost of $p_c y_i[t]$ every time slot, in order to transmit at rate $y_i[t]$ on the cellular network. This cost corresponds to a cellular operator that meters usage of the cellular network. Furthermore, cellular cost helps as a factor discouraging delay tolerant applications from using the cellular interface if they can afford to wait and transmit on the secondary interface. This encapsulates the idea of delayed offloading. However, while most existing literature of delayed offloading considers inelastic traffic that should be transmitted in full, our model considers elastic traffic that balances the trade-off between throughput and delay by using β as a control knob.

III. Problem Formulation

We formulate the Finite-Horizon Discounted-Rate Utility Maximization (DRUM) problem. For a horizon of T slots and N flows, each having its own discount factor $\beta^{(i)}$ we can write the problem as:

$$\textbf{P1}: \max_{\mathbf{x}[1],\mathbf{y}[1],\ldots,\mathbf{x}[T],\mathbf{y}[T]} \sum_{t=1}^{T}\sum_{i=1}^{N} w_i U(R_i^{(\beta_i)}[t]) - p_c y_i[t] \tag{3}$$

subject to (4)

$$\sum_{i=1}^{N} x_i[t] \leq c[t], \qquad t = 1,\ldots,T \tag{5}$$

$$\sum_{i=1}^{N} y_i[t] \leq 1, \qquad t = 1,\ldots,T \tag{6}$$

$$x_i[t], y_i[t] \geq 0, \qquad i = 1,\ldots,N, \text{and } t = 1,\ldots,T, \tag{7}$$

where the bold notation in $\mathbf{x}[t], \mathbf{y}[t]$ refers to the allocation of all flows $(x_1[t], x_2[t], \ldots, x_N[t])$ and $(y_1[t], y_2[t], \ldots, y_N[t])$ at time t, respectively. Also, w_i is a positive weight and $U()$ is a suitable concave non-decreasing utility function that aims to achieve fairness between different flows. Examples of utility functions that provide fairness are the α-fairness functions of the form $U(r) = \frac{r^{(1-\alpha)}}{1-\alpha}$ that were introduced in [22]. However, the difference between that formulation and the standard Network Utility Maximization (NUM) framework is that the utility function is taken over a β-discounted rate that puts a weight β on the history of rate allocated. Every flow i is parameterized with the pair $(w_i, \beta^{(i)})$ where w_i indicates a higher priority in rate allocation and a lower β_i indicates higher sensitivity to delay. The constraint (5) ensures that the

sum of the rates allocated on the secondary interface does not exceed the instantaneous capacity $c[t]$. Similarly, the constraint (6) ensures that the sum of rates allocated on the cellular interface does not exceed the constant normalized cellular capacity.

The problem **P1** is a standard constrained convex optimization problem with $2NT$ decision variables (rate per flow per time-slot per interface) and $2T + 2NT$ constraints. Solving this problem requires non-causal knowledge of secondary capacities $(c[1], c[2], \ldots, c[T])$. In the next section, we provide two predictive online solutions that depend on the knowledge of capacities up to a window w and have theoretical bounds on worst-case performance as well as good practical performance.

IV. Online Predictive Rate Allocation

A. Receding Horizon Control (RHC): RHC, also referred to in control literature as Model Predictive Control (MPC) [23, 24], is a feedback control technique that provides an online solution to the original problem by approximating the original problem as a sequence of open-loop optimization problems over the prediction horizon $[t, t + w]$. After solving the open-loop problem and obtaining the solution, the algorithm implements the first step of the solution only, i.e., $(\mathbf{x}[t], \mathbf{y}[t])$, updates the state, finds the new prediction at time $t + w + 1$ and repeats the procedure at time $t + 1$. We now give a detailed description of the algorithm.

System State: Since the optimization is over the β-discounted rate, which is a function of the rates allocated in the past, the system has to keep a "memory" of past allocations. Since the equivalent rates in (2) are updated as a discounted sum, it is sufficient to save a vector $\mathbf{R}[t-1] = (R_1^{(\beta_1)}[t-1], R_2^{(\beta_2)}[t-1], \ldots, R_N^{(\beta_N)}[t-1])$ of the equivalent rates at time $t-1$. Define the control input $\theta_i[t] = (x_i[t], y_i[t])^T$ where $(\,)^T$ is the vector transpose notation. It can be shown from (2), that the β-discounted rate of flow i is updated over time as follows

$$\begin{aligned} R_i^{(\beta_i)}[t] &= \beta_i R_i^{(\beta_i)}[t-1]\Big(1 - \frac{\beta_i^{t-1}(1-\beta_i)}{1-\beta_i^t}\Big) + \frac{1-\beta_i}{1-\beta_i^t} 1^T \theta_i[t] \\ &\approx \beta_i R_i^{(\beta_i)}[t-1] + (1-\beta_i) 1^T \theta_i[t] \end{aligned} \tag{8}$$

To obtain the online rate allocation, we first solve the following open-loop RHC optimization problem. Let $A = [1, 0]^T$ and $B = [0, 1]^T$. Also define the vector $\mathbf{\Theta}(\mathbf{R}[t-1]) = (\theta[t|\mathbf{R}[t-1]], \theta[t+1|\mathbf{R}[t-1]], \ldots, \theta[t+w|\mathbf{R}[t-1]])$ as the $2 \times (w+1)$ vector that solves the following open-loop optimization problem:

$$\textbf{P2:} \max_{\theta[t],\ldots,\theta[t+w]} \sum_{\tau=t}^{t+w} \sum_{i=1}^{N} w_i U(R_i^{(\beta_i)}[\tau]) - p_c B^T \theta_i[\tau] \tag{9}$$

subject to

$$R_i^{(\beta_i)}[\tau] = \beta_i R_i^{(\beta_i)}[\tau - 1] + (1-\beta_i) 1^T \theta_i[\tau], \tag{10}$$
$$\tau = t, t+1, \ldots, t+w, \text{ and } i = 1, \ldots, N$$

$$\sum_{i=1}^{N} A^T \theta_i[\tau] \le c[\tau], \qquad \tau = t, t+1, \ldots, t+w \tag{11}$$

$$\sum_{i=1}^{N} B^T \theta_i[\tau] \le 1, \qquad \tau = t, t+1, \ldots, t+w \tag{12}$$

$$\theta_i[\tau] \ge 0, \qquad i = 1, \ldots .N, \text{ and } \tau = t, t+1, \ldots, t+w \tag{13}$$

After solving the RHC optimization problem, the scheduler implements only the first step of the solution, i.e.,

$$\theta_{\text{RHC}}[t] = \mathbf{\Theta}(\mathbf{R}[t-1])[t] \tag{14}$$

The state $\mathbf{R}[t]$ is then updated according to (8), the new prediction $c[t + w + 1]$ is obtained from the predictor, and the procedure is repeated to obtain the updated solution.

B. Average Fixed Horizon control (AFHC): The AFHC algorithm was proposed in [25] and analyzed in [26] for online convex optimization (where the objective function is unknown every time slot) with switching costs. AFHC is more amenable to theoretical analysis and sometimes outperforms RHC. Similar to RHC, AFHC approximates the offline problem with a series of open-loop optimization problems. However unlike RHC, AFHC does not only implement the first step of the solution and discards the rest of the solution. Instead, AFHC saves all solutions from all open-loop approximations and averages them out. Thus, both algorithms have the same time complexity. However, AFHC needs $2N$ space whereas RHC only needs N space. We next give the algorithm formally.

First, we define the Fixed Horizon Control parametrized by k where $k = 0, 1, \ldots, w$. FHC$^{(k)}$ only solves the problem at time slots $\Omega_k = \{z : z \equiv k \bmod (w+1)\}$, i.e., FHC$^{(k)}$ solves the problem P2 every $w+1$ slots at times $k, k+(w+1), k+2(w+1), \ldots$ etc., and implements the solutions for the entire horizon. Let $\theta^{(k)}[t]$ be the solution obtained by FHC$^{(k)}$ for time t, we have

$$[\theta^{(k)}[t], \theta^{(k)}[t+1], \ldots, \theta^{(k)}[t+w]] = \mathbf{\Theta}(\mathbf{R}[t-1]), \forall t \in \Omega_k \tag{15}$$

For example FHC$^{(0)}$ will implement the solution by solving the problem at $0, w+1, 2(w+1), \ldots$ etc., FHC$^{(1)}$ will implement the solution by solving the problem at $1, 1+(w+1), 1+2(w+1), \ldots$ etc., and so on up to $k = w$.

To complete the AFHC solution, at time slot $t \in \Omega_k$, the scheduler will first solve FHC$^{(k)}$ to obtain the allocation $[\theta^{(k)}[t], \theta^{(k)}[t+1], \ldots, \theta^{(k)}[t+w]]$ and then sets

$$\theta_{i,\text{AFHC}}[t] = \frac{\sum_{k=0}^{w} \theta_i^{(k)}[t]}{w+1}, \forall i = 1, \ldots, N \tag{16}$$

V. Competitive Ratio of AFHC

A natural question to ask is how well does the proposed algorithm perform under different conditions: number of flows, (w_i, β_i) of each flow, length of prediction window, etc. While it is known that deriving competitive ratios for general online convex optimization problems is hard [26], under some mild conditions, we are able to derive a lower bound on the competitive ratio (the competitive ratio here is w.r.t to reward

rather than cost, so we are looking for lower bounds to the ratio between rewards achieved by the online algorithm and offline algorithm, respectively).

Definition 2. *(Competitive Ratio): An algorithm **ALG** is said to be γ-competitive if*

$$\inf_{c[1],c[2],\ldots,c[T]} \frac{\textit{Reward(ALG)}}{\textit{Reward(OPT)}} \geq \gamma \tag{17}$$

where Reward( ) is the function that computes the objective function according to (3)*, and OPT is the offline prescient solution of P1.*

Note that the definition we use here is slightly different from the definition used in most online algorithms' literature. Conventionally, an algorithm is called c-competitive if Reward(ALG) $\geq \frac{1}{c}$ Reward(OPT). We choose to use $\gamma = \frac{1}{c}$ in Definition 1, instead of the conventional c-competitive notation, since it makes our results more intuitive and understandable.

Theorem 1. *Given N flows, all with weights $w_i = 1$ and $\beta \in [0, 1)$. Under the following assumptions:*

- *A1* $U(\mathbf{0}) = 0$ *and* $U(r) - p_c r > 0$ *for all* $r > 0$.
- *A2 The (sub)-gradient of $U(\,)$ is uniformly bounded by G over the feasible domain.*
- *A3* $c[t] \leq c_{\max}, \forall t \in 1, 2, \ldots, T$*, we take the cellular capacity to be* $y[t] = y_c, \forall t$ *(instead of the normalized value* 1 *in the RHS of* (6) *to derive a more general result.)*

then, under AFHC:

$$\textit{Competitive Ratio} \geq 1 - \frac{1}{w+1} \frac{G\beta_{max}}{D(1-\beta_{max})} \tag{18}$$

where

$$D = \min\left(\frac{U(y_c) - p_c y_c}{y_c}, \frac{U(y_c + c_{\max}) - p_c y_c}{y_c + c_{\max}} \right) \tag{19}$$

Before proving the theorem, we discuss the implications of this result. First, it is clear that the bound improves with increased prediction window w. This is expected since better foresight enables the scheduler to make better instantaneous decisions. Second, the factor $\frac{1}{1-\beta}$ implies that an increase in β worsens the bound. This is also expected since a delay tolerant flow has more flexibility on when it should be allocated optimally. Thus, prediction window discards important information about future opportunities to defer delay tolerant transmissions. Since the competitive ratio bound is valid for all sequences of secondary capacities $(c[1], c[2], \ldots, c[T])$, even cases where an adversary can observe the system state and controls, then generate future secondary capacities accordingly. The bound is expectedly loose for practical cases, and the empirical competitive ratio does not increase sharply with β as suggested by the bound.

In order to prove Theorem 1, three Lemmas are needed. The proof generally follows the same approach as [25] by bounding the difference between the reward of OPT and the reward of FHC$^{(k)}$ over a short horizon, and then using Jensen Inequality to bound the competitive ratio. The difference between our proof and [25] is that: 1. The formulation in [25] minimizes a convex function of the current control action only plus a switching cost that penalizes difference between current and previous control actions, whereas in our formulation, the reward is a function of a state that depends on both control action and previous state. Thus, we need extra steps of using first-order conditions to bound difference between rewards (Lemma 2). 2. The way we define the Competitive Ratio in (17) requires finding a linear underestimator of the reward function (Lemma 1).

Lemma 1. *Given a vector of β-discounted rate vectors $(\mathbf{R}[1], \mathbf{R}[2], \ldots, \mathbf{R}[T])$, the total reward achieved by this vector according to* (3) *satisfies the following linear bound*

$$\sum_{t=1}^{T} \sum_{i=1}^{N} U(R_i^{(\beta_i)}[t]) - p_c y_i[t] \geq \sum_{t=1}^{T} \sum_{i=1}^{N} D R_i^{(\beta_i)}[t] \tag{20}$$

where D is given by (19)*.*

Proof. Taking $\mathbf{1}^T \theta_i[t] = x_i[t] + y_i[t]$ in (8) we get the following bound:

$$y_i[t] \leq \frac{1}{1-\beta_i}(R_i^{(\beta_i)}[t] - \beta R_i^{(\beta_i)}[t-1]). \tag{21}$$

The reward per-flow every time slot can be bounded as follows

$$U(R_i^{(\beta_i)}[t]) - p_c y_i[t] \geq U(R_i^{(\beta_i)}[t]) - \frac{p_c}{1-\beta_i}(\, R_i[t] - \beta R_i[t-1]). \tag{22}$$

Setting $\mathbf{R}[0] = \mathbf{0}$ and summing over all flows and over all time slots, we have the following inequality:

$$\sum_{t=1}^{T} \sum_{i=1}^{N} U(R_i^{(\beta_i)}[t]) - p_c y_i[t] \geq \sum_{t=1}^{T} \sum_{i=1}^{N} U(R_i^{(\beta_i)}[t]) - p_c R_i^{(\beta_i)}[t]. \tag{23}$$

By noting that the linear cost at the RHS cannot exceed $p_c y_c$ (since the cellular allocation cannot exceed the cellular capacity), we can refine the bound on the LHS of (23)

$$\begin{aligned}
&\geq \begin{cases} \sum_{t=1}^{T} \sum_{i=1}^{N} U(R_i^{(\beta_i)}[t]) - p_c R_i^{(\beta_i)}[t] & \text{for } \sum_{i=1}^{N} R_i[t] \leq y_c \\ \sum_{t=1}^{T} \sum_{i=1}^{N} U(R_i^{(\beta_i)}[t]) - p_c y_c & \text{for } \sum_{i=1}^{N} R_i[t] \geq y_c \end{cases} \\
&\overset{(a)}{\geq} \begin{cases} \sum_{t=1}^{T} \sum_{i=1}^{N} \frac{U(y_c) - p_c y_c}{y_c} R_i^{(\beta_i)}[t] & \text{for } \sum_{i=1}^{N} R_i[t] \leq y_c \\ \sum_{t=1}^{T} \sum_{i=1}^{N} \frac{U(y_c + c_{\max}) - p_c y_c}{y_c + c_{\max}} R_i^{(\beta_i)}[t] & \text{for } \sum_{i=1}^{N} R_i[t] \geq y_c \end{cases} \\
&\geq \sum_{t=1}^{T} \sum_{i=1}^{N} D R_i^{(\beta_i)}[t]
\end{aligned}$$

where inequality (a) comes from the fact that the two summand functions on the LHS are concave in $R_i[t]$, thus, each of those two one-dimensional functions can be lower bounded by a straight line connecting points $(0, U(y_c) - p_c y_c), (U(y_c) - p_c y_c, U(y_c + c_{\max}) - p_c y_c)$, respectively. Those two straight lines in turn lie between lines $\frac{U(y_c) - p_c y_c}{y_c} R_i[t]$ and $\frac{U(y_c + c_{\max}) - p_c y_c}{y_c + c_{\max}} R_i[t]$, thus taking the minimum will give us a lower bound everywhere in the domain. □

The next Lemma provides a bound on the difference between the reward achieved by the offline solution and the reward achieved by the approximation P2. We first write the reward achieved in the interval $[t, t+w]$ with a control decision vector $(\theta[t], \theta[t+1], \ldots, \theta[t+w])$ as a function of the initial state at time $t-1$ as follows:

$$g(\mathbf{R}[t-1];\theta[t],\ldots,\theta[t+w]) = \sum_{\tau=t}^{t+w}\sum_{i=1}^{N} U(R_i^{(\beta_i)}[\tau]) - p_c B^T\theta_i[\tau]$$
$$= \sum_{i=1}^{N}\sum_{\tau=t}^{t+w} U(\beta^{\tau-t+1}R_i^{(\beta_i)}[t-1] + \sum_{\eta=t}^{\tau}\beta^{\eta-t}(1-\beta)1^T\theta[\eta])$$
$$-p_c\sum_{i=1}^{N}\sum_{\tau=t}^{t+w} B^T\theta_n[\tau]. \quad (24)$$

Lemma 2. *Denote the offline solution (OPT) of problem P1 and the resulting states as $(\theta^*[1],\theta^*[2],\ldots,\theta^*[T])$ and $(\mathbf{R}^*[1],\mathbf{R}^*[2],\ldots,\mathbf{R}^*[T])$, respectively. Suppose we were running the FHC$^{(k)}$ algorithm from time 0 up to time $t\in\Omega_k$. Let the system state at time $t-1$ be $\mathbf{R}^{(k)}[t-1]$ and denote the online solution at time t of problem P2 given $\mathbf{R}^{(k)}[t-1]$ as $\mathbf{\Theta}^{(k)}(\mathbf{R}^{(k)}[t-1]) = (\theta^{(k)}[t|\mathbf{R}^{(k)}[t-1]],\theta^{(k)}[t+1|\mathbf{R}^{(k)}[t-1]],\ldots,\theta^{(k)}[t+w|\mathbf{R}^{(k)}[t-1]])$. Under assumptions A1-A3 of Theorem 1, the following inequality holds:*

$$g(\mathbf{R}^{(k)}[t-1];\mathbf{\Theta}^{(k)}(\mathbf{R}^{(k)}[t-1])) \geq g(\mathbf{R}^*[t-1];\theta^*[t],\ldots,\theta^*[t+w])$$
$$-\sum_{i=1}^{N}\frac{G\beta_i}{1-\beta_i}(R_i^{(\beta_i)^*}[t-1] - R_i^{(\beta_i)^{(k)}}[t-1]) \quad (25)$$

where G is the uniform bound of the (sub)-gradient of the function $U(\,)$ over the domain as stated by A2 in Theorem 1.

Proof. By concavity of the function $U(\,)$, it is straight-forward to see that the function $g(\mathbf{R}[t-1];\mathbf{\Theta}(\mathbf{R}[t-1])$ is concave in the variable $\mathbf{R}[t-1]$. By first order conditions of concavity in the variable $\mathbf{R}[t-1]$ only (where $\mathbf{\Theta}(\mathbf{R}[t-1])$ are treated as parameters):

$$g(\mathbf{R}^*[t-1];\theta^*[t],..,\theta^*[t+w]) \leq g(\mathbf{R}^{(k)}[t-1];\theta^*[t],..,\theta^*[t+w])$$
$$+\nabla g(\mathbf{R}^{(k)}[t-1];\theta^*[t],\ldots,\theta^*[t+w]))^T(\mathbf{R}^*[t-1]-\mathbf{R}^{(k)}[t-1]) \quad (26)$$

Where ∇ is the gradient operator w.r.t. $\mathbf{R}[t-1]$. The first term in the RHS of (26) can be bounded as follows:

$$g(\mathbf{R}^{(k)}[t-1];\theta^*[t],..,\theta^*[t+w]) \leq g(\mathbf{R}^{(k)}[t-1];\mathbf{\Theta}^{(k)}(\mathbf{R}^{(k)}[t-1])) \quad (27)$$

This is because, given the initial state $\mathbf{R}^{(k)}[t-1]$, $\mathbf{\Theta}^{(k)}(\mathbf{R}^{(k)}[t-1])$ is the maximizing vector of $g(\mathbf{R}^{(k)}[t-1];\theta[t],..,\theta[t+w])$ according to the formulation in P2.
To bound the second term in the RHS, we can use the expression in (24) to explicitly derive the gradient w.r.t. the vector $\mathbf{R}[t-1]$. The i^{th} term of the gradient vector can be bounded as follows:

$$\frac{\partial g(\,)}{\partial R_i[t-1]} = \sum_{\tau=t}^{t+w}\beta_i^{\tau-t+1}U'\Big(\beta_i^{\tau-t+1}R_n[t-1])$$
$$+\sum_{\eta=t}^{\tau}\beta_i^{\eta-t}(1-\beta_i)1^T\theta_i[\eta])\Big) \overset{(b)}{\leq} G\sum_{\tau=t}^{t+w}\beta_i^{\tau-t+1} \leq \frac{G\beta_i}{(1-\beta_i)} \quad (28)$$

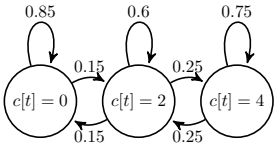

Fig. 2. Secondary Interface capacity process

Where (b) comes from assumption *A2* in Theorem 1. Taking the inner product of the gradient in (28) and $(\mathbf{R}^*[t-1] - \mathbf{R}^{(k)}[t-1])$ and combining the bounds gives the desired result. □

The next Lemma bounds the total reward achieved by the FHC$^{(k)}$ algorithm as a function of the total reward achieved by OPT. With some abuse of the notation, we denote the reward gained over the horizon of AFHC, FHC$^{(k)}$, and OPT as $g_{1:T}(\theta^{\text{AFHC}})$, $g_{1:T}(\theta^{(k)})$, and $g_{1:T}(\theta^*)$, respectively.

Lemma 3. *Given any $k=0,1,\ldots,w$, the following holds*

$$g_{1:T}(\theta^{(k)}) \geq g_{1:T}(\theta^*)$$
$$-\sum_{\tau\in\Omega_k}\sum_{i=1}^{N}\frac{G\beta_i}{1-\beta_i}(R_i^{(\beta_i)^*}[t-1] - R_i^{(\beta_i)^{(k)}}[t-1]) \quad (29)$$

The proof of this Lemma is straight-forward by summing the expression in (25) over the set Ω_k.

Proof of Theorem 1. The reward obtained over the horizon by the AFHC control algorithm can be lower bounded as follows

$$g_{1:T}(\theta^{(\text{AFHC})}) \overset{(c)}{\geq} \frac{1}{w+1}\sum_{k=1}^{w+1} g_{1:T}(\theta^{(k)})$$
$$\overset{(d)}{\geq} g_{1:T}(\theta^*)$$
$$-\frac{1}{w+1}\sum_{k=1}^{w+1}\sum_{\tau\in\Omega_k}\sum_{i=1}^{N}\frac{G\beta_i}{1-\beta_i}(R_i^{(\beta_i)^*}[t-1] - R_i^{(\beta_i)^{(k)}}[t-1])$$
$$\geq g_{1:T}(\theta^*) - \frac{1}{w+1}\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{G\beta_i}{1-\beta_i}R_i^{(\beta_i)^*}[t-1]$$
$$\geq g_{1:T}(\theta^*) - \frac{1}{w+1}\frac{G\beta_{\max}}{1-\beta_{\max}}\sum_{i=1}^{N}\sum_{t=1}^{T}R_i^{(\beta_i)^*}[t-1]$$

Where (c) is by Jensen Inequality (averaging property of AFHC) and (d) is a result of Lemma 3. Dividing both sides by $g_{1:T}(\theta^*)$ results in the Competitive Ratio (CR) lower bound

$$\text{CR} \geq 1 - \frac{1}{w+1}\frac{G\beta_{\max}}{1-\beta_{\max}}\frac{\sum_{t=1}^{T}\sum_{i=1}^{N}R_i^{(\beta_i)^*}[t-1]}{g_{1:T}(\theta^*)} \quad (30)$$

Using Lemma 1 to bound $g_{1:T}(\theta^*)$ linearly and noticing that $\mathbf{R}[0]=\mathbf{0}$ gives the desired result. □

VI. Numerical Results

To validate our formulation, we first assume that the secondary capacity generation process follows the Markov chain in Fig. 2. We simulate the case when two flows exist, a delay sensitive flow with $\beta = 0$ and a delay tolerant flow

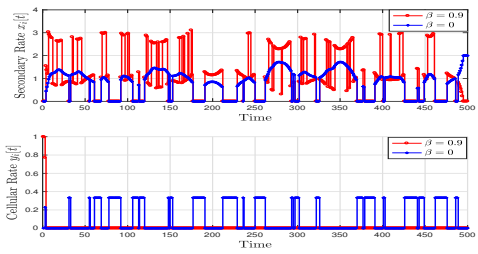

Fig. 3. Optimal Rate Allocations of heterogeneous flows, $U(r) = log(1 + r)$, $p_c = 0.75$.

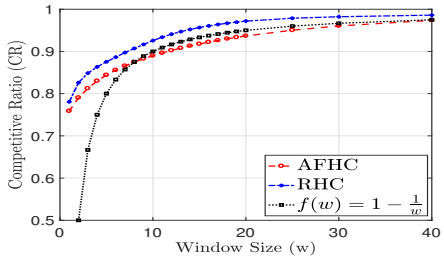

Fig. 4. Competitive Ratios of RHC and AFHC compared to the function $f(w) = 1 - \frac{1}{w+1}$. $U(r) = \frac{r^{(1-\alpha)}}{1-\alpha}$, $p_c = 0.55$, $\beta = \{0, 0.7, 0.95\}, \alpha = 0.5$.

with $\beta = 0.9$. Fig. 3 shows the allocation of OPT when the utility function $U(r) = \log(1 + r)$ and the cellular price, $p_c = 0.75$. We note the following: The delay tolerant flow cellular usage is almost non-existent (less than 1% of the total flow rate) whereas the delay sensitive flow uses the cellular interface whenever the secondary connectivity is weak or absent. In Fig. 3, the delay sensitive flow transmits 15% of its traffic over the cellular interface. The delay tolerant flow transmits in bursts whenever a secondary network is available. Furthermore, the delay tolerant flow tends to increase its rate whenever it predicts a period with no secondary connectivity at the expense of the delay sensitive flow. The burstiness effect can be captured by noticing the following: the means of total rate allocated to the delay-tolerant and the delay sensitive flows are 1.01 and 0.87 respectively, whereas the variance of the total rate allocated to the delay-tolerant flow is 1.11, more than 4 times that of the delay sensitive flow 0.27.

In Fig. 4, we simulate three flows with β values equal to $\{0, 0.7, 0.95\}$, representing different delay sensitivities for a horizon $T = 500$. The secondary capacity evolves as the

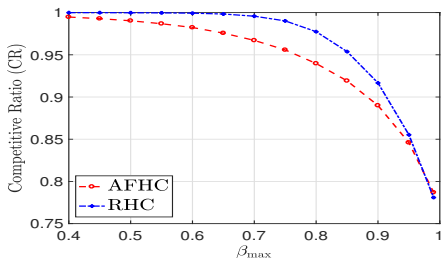

Fig. 5. Competitive Ratios of RHC and AFHC as a function of $\beta_{\max}$. $U(r) = \frac{r^{(1-\alpha)}}{1-\alpha}$, $p_c = 0.55$, $\beta = \{0, 0.3, \beta_{max}\}, \alpha = 0.5$, $w = 3$.

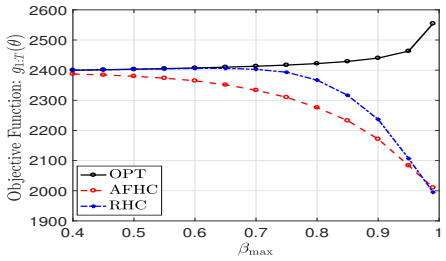

Fig. 6. Reward obtained by OPT, RHC and AFHC as a function of $\beta_{\max}$. $U(r) = \frac{r^{(1-\alpha)}}{1-\alpha}$, $p_c = 0.55$, $\beta = \{0, 0.3, \beta_{max}\}, \alpha = 0.5$, $w = 3$.

Markov Chain shown in Fig. 2. We take the utility function as the α-fairness function with $\alpha = 0.5$. We set p_c to be equal to 0.55. We plot the empirical competitive ratio as a function of w. We see that the RHC performance is slightly superior to AFHC for all values of prediction window w. Naturally, we see that the competitive ratio increases with the increase of w. Interestingly, the empirical rate of increase is very similar to the growth of the function $1 - \frac{1}{w+1}$, which suggests that our theoretical lower bound matches the empirical results order-wise up to a constant factor. However our findings confirm that the CR converges to 1 as w increases. It is worth noting that the α-fairness functions do not satisfy the assumption A2 in Theorem 1, since the gradient is unbounded. However, this can be fixed by modifying the utility functions to be $U(r) = \frac{(\epsilon+r)^{(1-\alpha)}}{(1-\alpha)} - \frac{\epsilon}{1-\alpha}$. This will cause the gradient to be bounded by ϵ. A small ϵ approximates α-fairness functions efficiently at the expense of a loose lower bound on the competitive ratio. Thus, to get a fairly tight bound we have to either use a larger ϵ or refine the bound in Lemma 2 by using special properties of the utility function such as strong concavity.

In Fig. 5, we use the same setup as the previous case to simulate the empirical performance of three flows with parameters $\{0, 0.3, \beta_{\max}\}$, where $\beta_{\max}$ is varied between 0.4 and 0.99. We use a small prediction window with $w = 3$. Our lower bound have suggested that there might be some performance degradation of online algorithms as flows become more delay tolerant. However, while our lower bound suggests very fast degradation in performance (as $\frac{1}{1-\beta_{\max}}$), simulations show that degradation happens at a much slower rate, and that a small window, $w = 3$ achieves over 85% of the utility of OPT, even as $\beta_{\max}$ is increased to 0.95. In Fig. 6, we plot the objective function of OPT, AFHC, and RHC under the same setup. Fig. 6 shows that while Reward(OPT) is guaranteed to be non-decreasing with increased $\beta_{\max}$, since more delay tolerance enables flows to defer transmissions until favorable conditions appear. On the other hand, Reward(AFHC) and Reward(RHC) are not guaranteed to increase with $\beta_{\max}$.

In Fig. 7, we compare the lower bound derived from Theorem 1 to the empirical lower bound. For this figure we assume that $c[t]$ is an i.i.d random variable that takes a value uniformly distributed between 0 and 5. We use a modified α-fairness function as our utility value function. In particular, we use

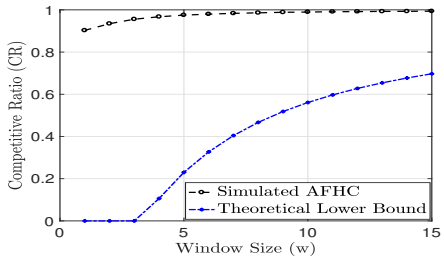

Fig. 7. Comparison between theoretical lower bound and simulated competitive ratio, $\beta = \{0.2, 0.4, 0.6\}$, $U(r) = \frac{(1+r)^{(1-\alpha)}}{1-\alpha} - \frac{1}{1-\alpha}$, $\alpha = 0.5$.

$U(r) = \frac{(1+r)^{(1-\alpha)}}{1-\alpha} - \frac{1}{1-\alpha}$, which is concave increasing and satisfies assumptions A1-A3. The system for this figure has 3 flows with $\beta = \{0.2, 0.4, 0.6\}$. We can see that the lower bound is loose. This is due to two reasons. First, the lower bound covers all possible sequences of secondary capacities including adversarial cases, where an adversary can view the scheduler's decision and generate future capacities to minimize the scheduler's reward. In Fig. 7, the capacities are sampled as an i.i.d sequence, which naturally performs much better than the adversarial worst-case. Second, The approximation in Lemma 1 is very coarse since it must apply to all possible utility functions. We can see that up to $w = 2$, there is no theoretical guarantee for performance whereas practically, AFHC achieves over 90% of the utility achieved by OPT. However, the bound gets tighter as the prediction window increases.

VII. Conclusion

In this paper, we have studied the problem of application rate allocation over different radio interfaces. We have addressed the issue of different delay requirements of applications using the discounted-rate framework. We have proposed two online predictive algorithms to handle the intermittence of secondary interface(s). We have shown that for the AFHC algorithm, the competitive ratio is $1 - \Omega(\frac{1}{w+1})$ when using a prediction window of length w. We have tested our algorithms for a number of practical scenarios using different utility functions. The empirical performance of the proposed online algorithms are consistently near-optimal using small prediction windows.

We intend to extend our work in two directions: 1. We plan to consider systems with arrivals, whereby flows of different classes arrive randomly. The flows are then served at a rate determined by the proposed algorithms and exit the system once they receive a total rate equal to their random size. The interesting questions include a characterization of the stable region of the proposed algorithms (in the sense of [27]), as well as the effect of (w_i, β_i) on the mean flow response time. 2. We plan on testing our algorithm in realistic scenarios using real world traces of user mobility and mobile flows that belong to real applications, which will enable us to compare the performance of the proposed algorithms to other solutions in the literature.

References

[1] M. Wang, J. Chen, E. Aryafar, and M. Chiang, "A survey of client-controlled hetnets for 5g," *IEEE Access*, 2017.

[2] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, "Next generation/dynamic spectrum access/cognitive radio wireless networks: A survey," *Computer networks*, 2006.

[3] A. Balasubramanian, R. Mahajan, and A. Venkataramani, "Augmenting mobile 3g using wifi," in *Proceedings of MobiSys*. ACM, 2010.

[4] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, "Mobile data offloading: How much can wifi deliver?" *IEEE/ACM Transactions on Networking (TON)*, 2013.

[5] F. Mehmeti and T. Spyropoulos, "Is it worth to be patient? analysis and optimization of delayed mobile data offloading," in *INFOCOM*. IEEE, 2014.

[6] S. Deng, R. Netravali, A. Sivaraman, and H. Balakrishnan, "Wifi, lte, or both?: Measuring multi-homed wireless internet performance," in *Proceedings of the 2014 Conference on Internet Measurement Conference*. ACM, 2014.

[7] A. J. Nicholson and B. D. Noble, "Breadcrumbs: forecasting mobile connectivity," in *MobiCom*. ACM, 2008.

[8] M. H. Cheung and J. Huang, "Optimal delayed wi-fi offloading," in *WiOpt*. IEEE, 2013.

[9] H. Yu, M. H. Cheung, L. Huang, and J. Huang, "Predictive delay-aware network selection in data offloading," in *GLOBECOM*. IEEE, 2014.

[10] Y. Im, C. Joe-Wong, S. Ha, S. Sen, M. Chiang *et al.*, "Amuse: Empowering users for cost-aware offloading with throughput-delay tradeoffs," *IEEE Transactions on Mobile Computing*, 2016.

[11] R. Mahindra, H. Viswanathan, K. Sundaresan, M. Y. Arslan, and S. Rangarajan, "A practical traffic management system for integrated lte-wifi networks," in *MobiCom*. ACM, 2014.

[12] H. Deng and I.-H. Hou, "Online scheduling for delayed mobile offloading," in *INFOCOM*. IEEE, 2015.

[13] A. Ford, C. Raiciu, M. Handley, S. Barre, and J. Iyengar, "Architectural guidelines for multipath tcp development," Tech. Rep., 2011.

[14] A. Nikravesh, Y. Guo, F. Qian, Z. M. Mao, and S. Sen, "An in-depth understanding of multipath tcp on mobile devices: Measurement and system design," in *MobiCom*. ACM, 2016.

[15] B. Han, F. Qian, L. Ji, V. Gopalakrishnan, and N. Bedminster, "Mp-dash: Adaptive video streaming over preference-aware multipath." in *CoNEXT*, 2016.

[16] O. B. Yetim and M. Martonosi, "Adaptive delay-tolerant scheduling for efficient cellular and wifi usage," in *WoWMoM*. IEEE, 2014.

[17] K.-K. Yap, T.-Y. Huang, Y. Yiakoumis, S. Chinchali, N. McKeown, and S. Katti, "Scheduling packets over multiple interfaces while respecting user preferences," in *Proceedings of the ninth ACM conference on Emerging networking experiments and technologies*. ACM, 2013.

[18] X. Hou, P. Deshpande, and S. R. Das, "Moving bits from 3g to metro-scale wifi for vehicular network access: An integrated transport layer solution," in *ICNP*. IEEE, 2011.

[19] A. Eryilmaz and I. Koprulu, "Discounted-rate utility maximization (drum): A framework for delay-sensitive fair resource allocation," in *WiOpt*. IEEE, 2017.

[20] S. Deng, A. Sivaraman, and H. Balakrishnan, "Delphi: A software controller for mobile network selection," 2016.

[21] J. Pang, B. Greenstein, M. Kaminsky, D. McCoy, and S. Seshan, "Wifi-reports: Improving wireless network selection with collaboration," *IEEE Transactions on Mobile Computing*, 2010.

[22] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Transactions on Networking (ToN)*, 2000.

[23] J. Mattingley, Y. Wang, and S. Boyd, "Receding horizon control," *IEEE Control Systems*, 2011.

[24] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. Scokaert, "Constrained model predictive control: Stability and optimality," *Automatica*, 2000.

[25] M. Lin, Z. Liu, A. Wierman, and L. L. Andrew, "Online algorithms for geographical load balancing," in *Green Computing Conference (IGCC)*. IEEE, 2012.

[26] N. Chen, A. Agarwal, A. Wierman, S. Barman, and L. L. Andrew, "Online convex optimization using predictions," in *SIGMETRICS Performance Evaluation Review*. ACM, 2015.

[27] J. Liu, A. Proutière, Y. Yi, M. Chiang, and H. V. Poor, "Stability, fairness, and performance: A flow-level study on nonconvex and time-varying rate regions," *IEEE Transactions on Information Theory*, 2009.