Delay Estimation and Fast Iterative Scheduling Policies for LTE Uplink

Akash Baid, Ritesh Madan, Ashwin Sampath

Delay Estimation and Fast Iterative Scheduling Policies for LTE Uplink

Akash Baid (WINLAB, Rutgers University), Ritesh Madan (Accelera MB), and Ashwin Sampath (Qualcomm)

***Abstract*—We design fast iterative policies for resource allocation in the uplink of LTE. We generalize recent works on iterative delay and queue based scheduling policies to more general system settings. We model all constraints due to contiguous bandwidth allocation, peak transmit power and fractional power control. We design a novel mechanism for inferring the packet delays approximately from the buffer status reports (BSR) and construct a new non-differentiable objective function which enables delay based scheduling. For frequency flat fading, we construct an $O(N \log L)$ optimal resource allocation algorithm for N users and L points of non-differentiability in the objective function. For a frequency diversity scheduler with M sub-bands, the corresponding complexity is essentially $O(N(M^2+L^2))$. Through detailed system simulations (based on NGMN and 3GPP evaluation methodology) which model H-ARQ, finite resource grants per sub-frame, realistic traffic, power limitations, interference, and channel fading, we demonstrate the effectiveness of our schemes for LTE.**

I. Introduction

Wideband cellular systems such as LTE allow for resource allocation with high granularity of a resource block (RB) of 1 ms by 180 KHz [1]. While control signalling and the general framework for the physical and medium access control (MAC) layers is specified to enable efficient use of spectral resources, the exact resource allocation algorithms for power and frequency allocation can be designed by an implementor. Moreover, each cell can serve on the order of a thousand active connections over a bandwidth of 20 MHz. Hence, in order to take advantage of the flexibility allowed in resource allocation, the resource allocation algorithms have to be computationally simple. Many schedulers in the literature entail maximizing the weighted sum of rates in each subframe. For example, the weights could be based on utility functions of average rate [2], [3], the queue length [4], [5], or head-of-line delay [6], [7]. In the uplink, the resource allocation problem must consider the maximum transmission power of a mobile and the constraints on the transmission power imposed by fractional power control to limit inter-cell interference [1], [8]. When contiguous bandwidth allocation is considered, the problem of maximizing the weighted sum rate in each subframe on the UL can be posed as a constrained convex optimization problem. For N users and M sub-bands general purpose methods can solve the problem in $O((NM)^3)$. With peak UE power constraints, a $O(NM)$ per iteration subgradient algorithm was obtained in [9]. Interior point methods (which have faster convergence) with an $O(NM^2)$ (if $N >> M$) Newton iteration were obtained in [10] for uplink resource allocation with additional fractional power control constraints. However non-differentiable objective functions are not considered under the framework in [10].

Also relevant to our paper are recent results on low complexity iterative scheduling algorithms. Many papers prior to these results had considered scheduling to maximize the sum of weighted rates in subframe n, where the weights were based on the arrivals and departures in the queue of a user until subframe $n-1$. The iterative policies in [11], [12] take into account how the weights change in subframe n to determine the resource allocation in *that* subframe. The results in these papers shed a remarkable insight that when the rate grows linearly with bandwidth (no peak power constraints at the transmitter), as the number of users in the system grow, these rules lead to much smaller per-user queues and delays, respectively, compared with previous approaches. However, the complexity of these algorithms grow with the resource granularity even if the coherence bandwidth does not grow. In this paper, we construct a continuous but non-differentiable concave reward function based on packet delays. It can be shown that the matching algorithm in [11] is an approximate algorithm to maximize this reward function in every subframe.

Motivated by the above observation, we consider resource allocation to maximize a continuous (possibly non-differentiable) concave reward function. We first consider a channel model where the channel gain in the frequency domain is flat and formulate the resource allocation problem as a non-differentiable convex optimization problem. Note that in typical cellular environments, the channel gains can be fairly correlated even for frequencies 2 to 5 MHz apart [13] – hence, the assumption of frequency flat fading is a reasonable one when the total bandwidth is up to 5 MHz (28 RBs) or lower, or if the UEs are allocated to sub-bands ($<$ 5 MHz) over a slower time-scale based on interference and channel statistics. We use subgradient analysis to design algorithms with $O(N \log L)$ cost per iteration (with small number of iterations) for N users and L points of non-differentiability in the objective function. We also design a novel mechanism to estimate head-of-line delays of queues at UEs with low complexity via only queue length information contained in the buffer status reports (BSR). We note our techniques are equally applicable for enabling delay based scheduling in the PCF and HCF modes in WiFi [14]. We demonstrate the improvement in performance due to our techniques through numerical results obtained via comprehensive numerical simulations based on 3GPP evaluation methodology [15]. Finally, when frequency selective fading is considered, we show how interior point methods with complexity of $O(NM^2 + NL^2)$ per Newton iteration can be obtained; note that in practice $N >> L, M$. A

978-3-901882-45-6/2012 - Copyright is with IFIP

longer version of the paper containing more detailed analysis and additional results is available at [16] for reference.

II. System Model

A. Channel Model, Power, Rate

We focus on the uplink of a single cell in LTE with N UEs and the total bandwidth divided into M sub-bands of equal bandwidth B, with B less than the coherence bandwidth of each user. The maximum transmit power of each UE is P. The channel gain for UE i on sub-band j is G_{ij}; we focus on the scheduler computation in a subframe, and don't explicitly show the dependence of quantities on time t. The base-station can measure the G_{ij}s via decoding the sounding reference signal (SRS) [1]. Fractional power control in LTE limits the amount of interference a UE causes at base-stations in neighboring cells. A UE which is closer to the cell edge inverts a smaller fraction of the path loss to the serving base-station than a UE which is closer to the serving base-station [8]. Thus the transmit powers of a UE on different sub-bands satisfy [10]:

$$p_{ij} \leq \gamma_{ij} b_{ij}, \forall i, j, \quad \sum_{j=1}^{M} p_{ij} \leq P,$$

where b_{ij} is the bandwidth allocated to UE i on sub-band j and γ_{ij} is a sub-band specific constant.

The interference PSD at the serving base-station on sub-band j (denoted as I_j) can be measured by the base-station periodically over unassigned frequency resources. The value depends on the interference coordination algorithm used [17]. When a UE transmits with power p_{ij} over bandwidth b_{ij} on sub-band j, it achieves a rate given by (treating interference as noise) $b_{ij}\psi\left(G_{ij}p_{ij}/b_{ij}I_j\right)$ where $\psi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is an increasing concave and differentiable function which maps the SINR to spectral efficiency.

B. Control Signaling

Single carrier frequency division multiple access (SC-FDMA) is used in the LTE uplink [1] and so a UE can be granted a number of 180 kHz resource blocks in a contiguous manner in frequency. The resource allocation to the UEs is computed by the base-station every subframe (1 ms) and signalled to the UEs via resource grants which include the contiguous set of RBs allocated to the UE and the modulation and coding scheme (MCS). We assume a constant number of maximum allowable re-transmissions for all UEs and do not adapt the re-transmission power and resource assignment through additional control signalling available in LTE.

Buffer status report (BSR) and scheduling request (SR) are transmitted by the UEs to inform the base-station about new packet arrivals at the UE. SR is one bit of information used to indicate the arrival of packets in an empty buffer at the UE and is used by the scheduler to start allocating resources to the UE. BSRs contain a quantized value of the number of bytes pending transmission at the UE[1], and are generated either periodically or when the queue goes from an empty to non-empty state. BSR reports are transmitted only when resources are allocated to the UE and thus provides only a coarse grain information about the queue length at the UE.

[1]We ignore the effect of quantization in BSR, but the methods in this paper extend easily to quantized BSR.

III. Reward Functions

In this section, we define the reward functions that we use for the optimization problem and relate it to the schemes used in earlier works. We assume each UE to have one active LC which supports either best effort or delay QoS traffic.

A. Best Effort

A flow, i, which is best-effort is associated with an average rate $x_i(t) \in \mathbb{R}_+$ in subframe t which is updated as follows:

$$x_i(t+1) = (1-\alpha_i)x_i(t) + \alpha_i r_i, \quad \forall t \geq 0, \tag{1}$$

where r_i is the rate at which UE i is served in the current subframe, and $0 < \alpha_i < 1$ is a user specific constant. The user experience in subframe t is modeled as a strictly concave increasing function $U_i : \mathbb{R}_+ \mapsto \mathbb{R}$ of the average rate $x_i(t)$. We greedily maximize the total utility at each time-step, i.e., the reward function for UE i with best effort traffic, at time t is [18]

$$f_i(r_i) = \frac{1}{\alpha_i} U_i((1-\alpha_i)x_i(t) + \alpha_i r_i). \tag{2}$$

If we set $f_i(r_i) = U'(x_i(t))r_i$, and let $\alpha_i \to 0$ in equation (1), the resulting scheduler is identical to that in [3]. Thus, our analysis offers a computationally efficient method to implement the scheduling policy in [3] for the LTE uplink with fractional power control.

B. Delay QoS Traffic

Here the user experience is a function of the packet delays. User experience is acceptable when the packet delays are lower than a certain tolerable value. The packet arrival process is assumed to be independent of the times at which the packets are served. Traffic for applications such as voice calls and live video chatting fall in this category.

At time t, let $\pi_i(t)$ be the number of packets in the queue of UE i. Denote the sizes and the delays of these $\pi_i(t)$ packets by $\{s_i(1), \ldots, s_i(\pi_i(t))\}$ and $\{d_i(1), \ldots, d_i(\pi_i(t))\}$. Then for a UE i with delay QoS traffic, we define the reward function as:

$$\begin{aligned} f_i(r_i) = & \sum_{j=1}^{n_i^{\text{serv}}(r_i)} s_i(j) d_i(j) \\ & + \left(r_i \Delta - \sum_{j=1}^{n_i^{\text{serv}}(r_i)} s_i(j) \right) d_i(n_i^{\text{serv}}(r_i)+1) \end{aligned} \tag{3}$$

where Δ is the length of a subframe (1 ms) and $n_i^{\text{serv}}(r_i)$ is the number of packets from UE i served fully if UE i is scheduled at rate r_i, i.e.,
$n_i^{\text{serv}}(r_i) = \max\left\{k : \sum_{j=1}^{k} s_i(j) \leq r_i\Delta\right\}$.

Lemma 3.1: $f_i(r_i)$ is a continuous concave function.

Proof: Concavity follows from the observation that $d_i(1) > \ldots > d_i(\pi_i(t))$ and continuity is immediate from definition. ■

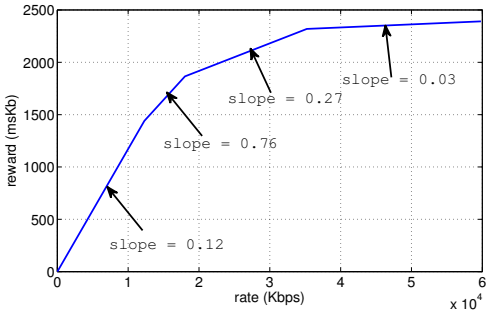

Fig. 1. Example reward function for delay QoS flow.

Example: Consider a UE with delay QoS traffic and four packets in the queue with delays (in ms) at time t given by $d_1 = 120, d_2 = 76, d_3 = 27, d_4 = 3$, and packet sizes (in KB) are $s_1 = 1.5, s_2 = 0.7, s_3 = 2.1, s_4 = 3$. Then the corresponding reward function f_i is shown in Fig. 1.

C. Iterative Queue and Delay Based Policies

If we restrict the model in [11] to frequency flat fading, i.e., a user is either connected to no server or all servers at any time, the algorithm in that paper can be interpreted as one which approximately maximizes the reward function in equation (3). In this work, we consider the maximization of the reward function in (3) for a much more general model with multiple rate options, peak power constraints, and different transmit PSD constraints on different sub-bands. We also note that the complexity of the algorithm in [11] is $O(NR^2)$ for N users and R RBs – when there are multiple RBs in each sub-band of bandwidth B; the complexity of our algorithms is lower. Finally, similar connections can be drawn between the scheme in [12] for frequency flat fading and using an objective function based on sums of squares of queues as in [19]; we believe similar connections can be drawn for the frequency selective fading model through further analysis.

IV. Estimation of Packet Delays

We now describe a method to infer approximate packet delays at the eNB via the mechanisms available in LTE. The main intuition is as follows: if the base-station estimates the queue length at time t to be say, 1000 bytes, but later decodes a BSR which was created at time t and has value 1300 bytes, the base-station can deduce that 300 bytes arrived between time t and the time at which the previous BSR was created. This information about the time interval during which the 300 bytes arrived can be used for making resource allocation decisions – specifically, scheduling policies based on packet delays can be implemented. The main complexity is due to re-transmissions which can lead to the BSR report arriving out of order at the base-station. A similar approach has been independently proposed in [20] recently, however it ignores the effect of retransmission failures in the analysis.

Let T^{retx} be the maximum amount of time between the first transmission of a MAC packet and the latest time when it can be re-transmitted for H-ARQ (for example, if we configure 6 as the maximum number of re-transmissions, $T^{\text{retx}} = 48$ subframes). We estimate the number of bytes that arrived, $A_i(t)$ in each subframe t. The buffer status reports are denoted by a sequence of random three tuples:

$$\{B_i(1), \tau_i(1), \delta_i(1)\}, \{B_i(2), \tau_i(2), \delta_i(2)\}, \ldots$$

where $B_i(1)$ is the buffer size reported in first BSR, $\tau_i(1)$ is the time at which first BSR was received, and $(\tau_i(1) - \delta_i(1))$ is the time at which the first BSR was generated, and so on. $C_i(t)$ denotes the number of bytes scheduled for transmission from UE i, $\hat{C}_i(t)$ the number of bytes which were successfully received from UE i, and $F_i(t)$ the number of bytes that failed the final re-transmission for UE i, at time t.

We maintain the history of estimated queue length for each UE i for duration T^{retx}, denoted by $Q_i(t - T^{\text{retx}} : t)$. Then, we update the Q matrix and the arrival vector A, at each t as follows:

For every t, i

1) *Scheduled Bytes:* $Q_i(t) = Q_i(t-1) - C_i(t)$.
2) *Failed Bytes:* $Q_i(t) = Q_i(t) + F_i(t)$.
3) *BSR report:* If a BSR report is received at time t, i.e., there is n such that $\tau_i(n) = t$, then update queue state as follows: If the base-station has not received any BSR report created after time $t - \delta_i(n)$, then

$$Q_i(t - \delta_i(n) : t) = Q_i(t - \delta_i(n) : t) + A_i(t - \delta_i(n))$$

where arrival $A_i(t - \delta_i(n)) = B_i(t) - Q_i(t - \delta_i(n))$ otherwise for

$$\underset{\{m:\ \tau_i(m) < t\}}{\arg\min} \ [\tau_i(m) - \delta_i(m) - (\tau_i(n) - \delta_i(n))]$$

update

$$A_i(t - \delta_i(n)) = B_i(t) - Q_i(t - \delta_i(n))$$
$$A_i(\tau_i(m) - \delta_i(m)) = A_i(t - \delta_i(m)) - A_i(t - \delta_i(n))$$

$$Q_i(t - \delta_i(n) : \tau_i(m) - \delta_i(m) - 1)$$
$$= Q_i(t - \delta_i(n) : \tau_i(m) - \delta_i(m) - 1) + A_i(t - \delta_i(n))$$

Note that Q_i can have negative entries.

V. Frequency Flat Fading

Here, we consider the resource allocation to N UEs over a single sub-band with bandwidth B and frequency flat fading. We drop the dependence of quantities in the general model on the sub-band j – for example, we denote channel gain from UE i to the eNB as G_i. We allow for contiguous allocation – this is a reasonable approximation when B is larger than a few RBs. Rounding techniques in, for example, [9] can be used to obtain integral solutions. The optimization problem to maximize the sum of rewards for all UEs over the bandwidth allocation vector $b \in \mathbb{R}_+^N$ in a subframe is:

$$\begin{aligned} \max. \quad & \sum_{i=1}^{N} f_i \left(b_i \psi \left(\frac{G_i \min(\gamma_i b_i, P)}{I b_i} \right) \right) \\ \text{s.t.} \quad & 0 \le b_i \le b_i^{\max}, \ \forall i, \quad \sum_{i=1}^{N} b_i \le B \end{aligned} \tag{4}$$

where $b_i^{\max}$ is the maximum bandwidth that UE i can use based on the estimated queue length, $Q_i(t)$, for UE i, and satisfies:

$$b_i^{\max}\psi\left(\frac{G_i\min(\gamma_i b_i^{\max},P)}{Ib_i^{\max}}\right)=Q_i(t)/\Delta$$

where we recall that Δ is the length of a subframe (1 ms). Since, the function on the left is an increasing function of $b_i^{\max}$, we can compute $b_i^{\max}$ efficiently via a bisection search. Problem (4) is a convex optimization problem (with non-differentiable objective function) due to the lemma which follows.

Lemma 5.1: The objective function in optimization problem (4) is concave in the b_is for $b_i \geq 0$, for all i.

Proof: Consider the function $g:\mathbb{R}_+\mapsto\mathbb{R}_+$ defined by $g(x)=x\psi(c/x),\ \forall x>0, c\in\mathbb{R}_+$ is constant. Since, ψ is assumed to be concave, it is easy to verify (via showing that the second derivative is always negative) that g is concave as well. Since, (i) the sum of concave functions is concave, and (ii) the composition of one concave function with another is concave, to show that the objective function is concave, it is sufficient to show that the following function is concave

$$h(x)=x\psi\left(\frac{\min(c_1x,c_2)}{x}\right),\ \forall x\geq 0, c_1,c_2\in\mathbb{R}_+ \text{ are constant}$$

Note that the above function is well defined for $x\geq 0$. Since, ψ is an increasing function, we can write $h(x)=\min\{x\psi(c_1),x\psi(c_2/x)\}$, which is the minimum of two concave functions, and hence, concave. ■

A. Characterization of Optimal Solution

We define a function which maps the bandwidth allocation b_i to achievable rate for user i:

$$h_i(b_i)=b_i\psi\left(\frac{G_i\min(\gamma_i b_i,P)}{Ib_i}\right)$$

We denote the sub-differential of a function $g:\mathbb{R}\mapsto\mathbb{R}$ at x by $\partial g(x)$. For continuous concave functions over the set of reals, the subdifferential at x is the set of slopes of lines tangent to f at x.

Let $b^\star\in\mathbb{R}_+$ denote the solution to the resource allocation problem (4). The following lemma shows that an optimal allocation in a given subframe is one for which the following quantities are equal for all users with non-zero bandwidth allocation: for best effort user, the marginal utility times the incremental rate when more bandwidth is allocated to it, and for delay QoS user, the delay of the oldest packet which is not served completely times the incremental rate when more bandwidth is allocated to it.

Lemma 5.2: There exists a $\lambda^\star>0$ such that if i is best effort, then

$$\lambda^\star\in U'((1-\alpha)x_i(t)+\alpha_i r_i^\star)\partial h_i(b_i^\star),\quad \text{if } b_i^\star>0$$
$$\lambda^\star<U'((1-\alpha)x_i(t))\min\partial h_i(0),\quad \text{if } b_i^\star=0$$

else, if i is delay QoS and $b_i^\star>0$,

- if $\sum_{j=1}^{n_i^{\text{serv}}(r_i^\star)}s_i(j)<r_i^\star\Delta$, $\lambda^\star\in d_i(n_i^{\text{serv}}(r_i^\star)+1)\partial h_i(b_i^\star)$
- else if $\sum_{j=1}^{n^{\text{serv}}(r_i^\star)}s_i(j)=r_i^\star\Delta$

$$\lambda^\star\in[d_i(n_i^{\text{serv}})\min\partial h_i(b_i^\star), d_i(n_i^{\text{serv}}+1)\max\partial h_i(b_i^\star)]$$

else, if i is delay QoS and $b_i^\star=0$,

$$\lambda^\star<d_i(1)\min\partial h_i(0)$$

where $r_i^\star=h_i(b_i^\star)$.

Proof: The lemma follows from standard arguments in, for example [21], the definitions of f_i's, and that the subdifferential of f_i for delay QoS user i is given by

$$\partial f_i(r_i)=\begin{cases} d_i(n_i^{\text{serv}}+1), & \sum_{j=1}^{n_i^{\text{serv}}(r_i^\star)}s_i(j)<r_i^\star\Delta\\ [d_i(n_i^{\text{serv}}),d_i(n_i^{\text{serv}}+1)], & \sum_{j=1}^{n^{\text{serv}}(r_i^\star)}s_i(j)=r_i^\star\Delta\end{cases}$$

■

We now evaluate the sub-differential of h_i for $x\geq 0$, which is bounded because γ_i is assumed to be bounded.

$$\partial h_i(x)=\begin{cases}\left\{\psi\left(\frac{G_i(t)\gamma_i}{I}\right)\right\}, & \text{if } x<P/\gamma_i\\ \left\{\psi\left(\frac{G_i(t)P}{Ix}\right)-\frac{G_i(t)P}{x}\psi'\left(\frac{G_i(t)P}{Ix}\right)\right\}, & \text{if } x>P/\gamma_i\\ \left[\psi\left(\frac{G_i(t)\gamma_i}{I}\right)-\frac{G_i(t)P}{x}\psi'\left(\frac{G_i(t)\gamma_i}{I}\right),\ \psi\left(\frac{G_i(t)\gamma_i}{I}\right)\right], & \text{if } x=P/\gamma_i\end{cases}$$

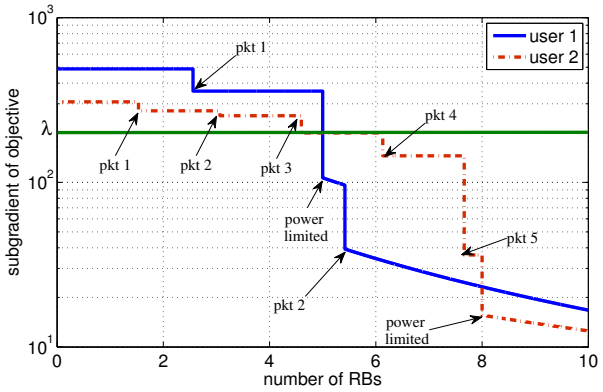

Fig. 2. Optimality condition

We illustrate the optimality condition via a two user example. The total bandwidth to be shared is 10 RBs, or 1800 KHz. All packets are of size 500 bits. The packet delays of the two users in the given subframe are

User 1: $[450, 330, 135, 80, 20]$
User 2: $[170, 150, 140, 110, 80, 20]$

The rate at which the users can be served as a function of the RBs are given by:

$$h_1(b_1)=\begin{cases}b_1\log_2\left(1+10^{0.05}\right) & b_1\leq 5*180\text{khz}\\ b_1\log_2\left(1+10^{0.05\frac{5*180}{b_1}}\right) & b_1>5*180\text{khz}\end{cases}$$

$$h_2(b_2)=\begin{cases}b_2\log_2\left(1+10^{0.4}\right) & b_2\leq 8*180\text{khz}\\ b_2\log_2\left(1+10^{0.4\frac{8*180}{b_2}}\right) & b_2>8*180\text{khz}\end{cases}$$

where the 5 and 8 RB thresholds (and corresponding SINRs of 0.5 dB and 4 dB) are derived from fractional power control constraints in Section II-A. The subgradient of the rewards for both the users as a function of bandwidth allocation, and

the optimal bandwidth allocation are shown in Fig 2 – the optimal resource allocation is 5 RBs to each user, and the optimal dual variable $\lambda^\star$ is shown in the figure. For each user, the figure also shows the number of RBs required to fully serve a given number of packets and the number of RBs at which the user becomes power limited, i.e., the maximum peak power constraint limits the transmission power rather than the fractional power control which limits the transmit PSD.

B. Computation of Optimal Solution

The optimization problem (4) entails the maximization of the sum of concave functions subject to a linear inequality constraint. While, in principle, the optimal resource allocation scheme can be computed via a bisection search on the dual variable λ, two difficulties arise: (i) There may be multiple values of b_i for which the subgradient of $f_i \circ h_i$ is equal to λ. See, for example, the first packet for user 1 in Fig. 2. As a result the dual function is non-differentiable and the bisection search may not converge [22]. (ii) If λ belongs to the sub-differential at a point b_i of non-differentiability of either f_i or h_i, the values of the gradient of $f_i \circ h_i$ may be arbitrarily different at $(b_i+\epsilon)$ and $(b_i-\epsilon)$ for an arbitrarily small ϵ. This can also be seen in Fig 2. We use Algorithm 1 to compute the optimal solution of problem (4). The convergence analysis is almost identical to that in Sec. 6 in [22]. An accurate solution can typically be computed in about 10 iterations.

Algorithm 1: Bisection search for optimal λ

Given starting value of $\underline{\lambda}$, $\overline{\lambda}$, $\underline{b}$, $\overline{b}$ and tolerance ϵ.
repeat
  Bisect: $\lambda = (\underline{\lambda} + \overline{\lambda})/2$.
  Allocate bandwidth for all i:
  if $\lambda > \max \partial f_i(0) \max \partial h_i(0)$ **then**
    set $b_i = 0$.
  else
    b_i is such that

$$\begin{aligned}\lambda \in &[\min \partial f_i(r_i) \times \min \partial h_i(b_i), \\ &\max \partial f_i(r_i) \times \max \partial h_i(b_i)]\end{aligned} \tag{5}$$

    where

$$r_i = \left(b_i \psi \left(\frac{G_i(t) \min(\gamma_i b_i, P)}{I b_i} \right) \right)$$

  end
  Update: if $\sum_{i=1}^N b_i - B > 0$, $\underline{\lambda} = \lambda$, $\underline{b} = b$, else $\overline{\lambda} = \lambda$, $\overline{b} = b$.
until $|\underline{\lambda} - \overline{\lambda}| < \epsilon$
Feasible Solution: if $\sum_i \underline{b_i} - \sum_i \overline{b_i} > 0$ **then**
  set $\alpha = \frac{B - \sum_i \overline{b_i}}{\sum_i \underline{b_i} - \sum_i \overline{b_i}}$.
else
  set $\alpha = 0$.
end
$b = \alpha \underline{b} + (1-\alpha)\overline{b}$

The starting values of $\overline{\lambda}$ and $\underline{\lambda}$ can be generated using the following simple lemma (proof is straightforward and omitted); the values of $\overline{b}$ and $\underline{b}$ are obtained by repeating the *Allocate Bandwidth* step in Algorithm 1 for dual variables $\overline{\lambda}$ and $\underline{\lambda}$, respectively.

Lemma 5.3: The optimal dual variable $\lambda^\star$ satisfies $\underline{\lambda} \le \lambda^\star \le \overline{\lambda}$ where

$$\overline{\lambda} = \max_{i=1,\ldots,N} \left[\psi \left(\frac{G_i(t)\gamma_i}{I} \right) \max \partial f_i(0) \right]$$

$$\begin{aligned}\underline{\lambda} = &\left[\psi \left(\frac{G_i(t)P}{IB} \right) - \frac{G_i(t)P}{B} \psi' \left(\frac{G_i(t)P}{IB} \right) \right] \\ &\times \max \partial f_i \left(B \psi \left(\frac{G_i(t)P}{IB} \right) \right), \text{ for some } i\end{aligned}$$

The main computational step in each iteration of Algorithm 1 entails solving (5) N times – we now show this can be done in $O(\log L)$ time when the reward function f_i for user i is non-differentiable at at most L points. The composition of function f_i with h_i is a concave function as shown in Lemma (5.1). Hence, to compute the bandwidth allocation for UE i as given in equation (5), we can use a bisection on b_i. First we obtain how many packets should be served fully such that the corresponding bandwidth required, b_i, satisfies equation (5) in $O(\log L)$ time. Then, we compute b_i.

We compute the range of subgradients for packet η as

$$\begin{aligned}\underline{b} = h_i^{-1} \left(\frac{\sum_{k=1}^{\eta-1} s_i}{\Delta} \right), \quad \overline{b} = h_i^{-1} \left(\frac{\sum_{k=1}^{\eta} s_i}{\Delta} \right) \\ SG(\eta) = d_i(\eta)[\min \partial h_i(\overline{b}), \min \partial h_i(\underline{b})]\end{aligned} \tag{6}$$

where we recall $d_i(\eta)$ and $s_i(\eta)$ are the delay and size for ηth packet queued at UE i. Note that the inverse of h_i is simple when $b_i < P/\gamma_i$; otherwise it can be computed via bisection.

The number of packets to be served completely is $\eta = \underline{\eta} - 1$. Note that h_i has at most one point of discontinuity, say $\hat{b}_i$. If $\underline{b} \le \hat{b}_i \le \overline{b}$ for $\eta = \underline{\eta} - 1$ in (6), then $b_i = \hat{b}$ if $\lambda / d_i(\eta) \in \partial h_i(\hat{b_i})$; else update $\underline{b}$ or $\overline{b}$ appropriately. A similar method can be used for best effort traffic and the analysis is omitted here due to lack of space.

VI. Simulation Results

A. Simulation Framework

The algorithms in the previous section were simulated using a detailed system simulator where the MAC layer signalling was modeled faithfully, and the PHY layer performance was abstracted via modeling of fading channels, transmission power, and capacity computations as in [15]. A hexagonal regular cell layout with three sectors per site was simulated with the parameters as noted in Table I. For fractional power control parameter values ($P_0 = -60$ dBm, $\alpha = 0.6$) similar to those in [8], a 19 cell (57 sector) simulation with wrap around was first performed to determine the interference over thermal (IoT) at the base-station of a cell to be 6 dB on an average. In subsequent simulations, only one cell was simulated with the IoT assumed to be constant in time and frequency. This drastically reduces the simulation time while still accounting for the inter-cell interference.

Parameter	Value
Channel Profile	ITU-T PedA
Mobile Speed	3 km/hr
Log-Normal Shadowing	σ =8.9 dBm
Intra-site Shadowing Correlation	1.0
Inter-site Shadowing Correlation	0.5
Cell Radius	1 km
No. of UEs/cell	20
No. of RBs	110
Max UE Tx Power	23 dBm
No. of Tx & Rx Antenna	1
eNB & UE Antenna Gains	0 dBi
Thermal Noise Density	-174 dBm/Hz
BSR periodicity	5 ms
max. number of retransmission	6

TABLE I
SIMULATION PARAMETERS

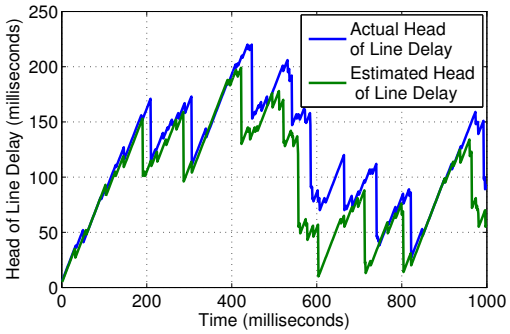

Fig. 3. HoL delay estimation performance

The time varying channel gains, G_i's, were assumed to be measured perfectly at the base-station in each subframe. The MCS was picked on the basis of the channel gain from the UE and a rate adaptation algorithm to target an average of two H-ARQ transmissions for successful decoding was used. We use the mutual information effective SINR metric (MIESM) [23]; we first obtain the effective SINR according to the modulation alphabet size and then use that value to simulate an event of packet loss according to the packet error rate for the effective SINR. We model the timelines for Scheduling Request (SR), resource grants, Hybrid-ARQ, ACK/NACKs, and BSR as described in Sec. II. We assume error free transmission of control messages in our simulations. Two types of traffic, *live video* and *streaming video* were modelled as per the description in [19].

B. Results

We consider two topologies for simulation: a *macro-cell* with the path loss between the base station and UEs randomly selected between 100 dB and 135 dB [24], *micro-cell* with path loss in the range 107 dB to 115 dB. We simulate three scheduling algorithms: (i) *Iterative Delay* which maximizes the reward function in Sec. III-B, (ii) *Iterative Queue* which minimizes sum-of-squares of queue lengths as in [19] and similar to [12], (iii) *non-iterative maximum weight* where a UE with the highest queue length times spectral efficiency for first RB is allocated bandwidth until the queue is drained or the UE becomes power limited before allocation to the next UE. We note that the computational algorithms in this paper are applicable to computing resource allocation for scheduling policies (i) and (ii), and that policies similar to (iii) do not consider the *change* in reward function of the UE in a given subframe.

1) **Macro cell Topology:** We consider 20 UEs with a mix of live video and streaming video traffic. Since live video has a tighter requirement for packet delays, we bias the scheduler to assign live video users 5x priority compared to streaming video users for same packet delay. Simulations were performed for low load and high load cases:
(1) *High Load:* 5 UEs have live video traffic, each with a mean rate of 300 kbps. For the other 15 UEs with streaming video traffic, we mimic an adaptive-rate streaming mechanism in which the data rate for each user depends on the quality of its channel to the base-station, i.e. a user close to the base-station transmits a better quality video compared to a cell-edge user. For simulating high-load, the traffic parameters are varied for each UE such that they generate traffic at 80% of the average data rate they received with full buffer traffic.
(2) *Low Load:* 5 UEs have live video traffic with a mean rate of 200 kbps. The UEs with streaming video traffic are now set to operate at 40% of their full buffer average data rate.

We first study the performance of the delay estimation mechanism described in Section IV. Figure 3 shows the estimated head of line (HoL) delay and the actual HoL delay at a UE over a period of 1 second. The estimated values can be seen to follow the actual delays but the accuracy is limited by the granularity of BSR messages, i.e., if there are multiple arriving packets between two successive BSR messages, the packets are bundled as one in our mechanism resulting in relatively small errors in HoL estimation.

Next we show the performance of the head of line delay based scheduling scheme computed as the solution to the optimization problem in (4) with the reward function in (3). Figure 4 shows the median and 95th percentile delays of the live video UEs for the two baseline and the head of line delay based schedulers for low and high loads. The delays experienced by the live video users are consistently less in the case of HoL delay based scheduling with the non-iterative scheme resulting in an average 95th percentile delay 1.6x higher than with the HoL delay scheduling. The queue based scheme also results in slightly higher delays, on an average 1.1x compared to 95th percentile delays for HoL scheduling. A more pronounced improvement is observed for the streaming video users, as shown in the delay plots in Figure 5. In this case, the non-iterative and queue based schemes result in 6.2x and 5x more delays compared to HoL delay scheduling in terms of 95th percentile latencies. Finally, Figure 6 shows the combined delay numbers for uplink packets from all the UEs in the high load simulation. As can be seen from the figure, the iterative queue based and delay based schemes result in similar delays for live video users due to preferential assignment. However this results in large delays for the streaming video users for both non-iterative and queue based schemes: close to 11x and 8x respectively compared to HoL delay based scheduling in terms of 95th percentile delays. Thus, leveraging the approximate packet delays obtained via our method leads to significant performance improvement over queue based

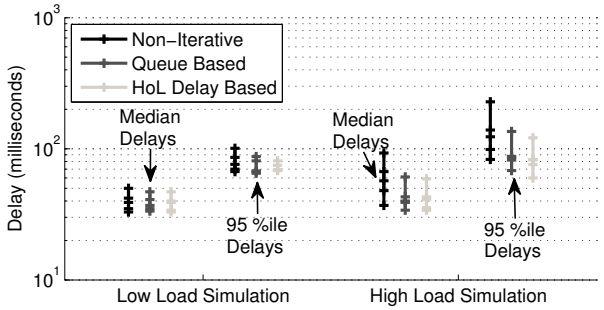

Fig. 4. Live video users: delay performance

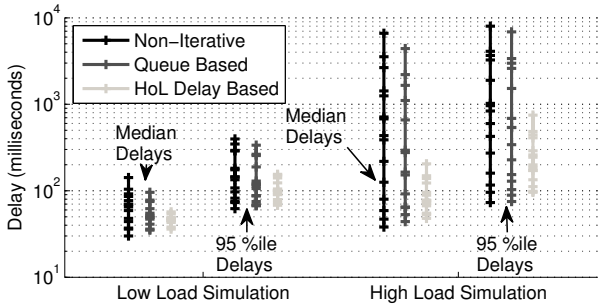

Fig. 5. Streaming video users: delay performance

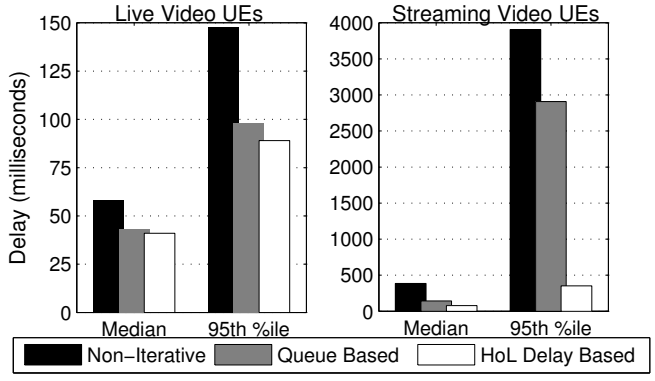

Fig. 6. Cell-wide delay performance of all packets in macro cell simulation

scheduling. Moreover, even for the queue based scheduler, the computational methods in this paper are very useful.

2) **Micro cell Topology:** In order to compare these scheduling schemes in a smaller cell topology, we ran a second simulation with 20 UEs located within a region with path loss 107-115 dB from the base station. Each UE, in this simulation, carries streaming video traffic with the mean data rate randomly selected between 300-2000 Kbits/sec. Decoupling the mean traffic rate with the path loss highlights the relative performance of the scheduling algorithms in real deployments where prior knowledge of user demand is rarely known. Individual and cell wide delay numbers are shown in Figure 7, which shows that 95th percentile delays for non-iterative and queue based schemes are 1.8x and 1.4x more than those for the HoL delay based scheduling.

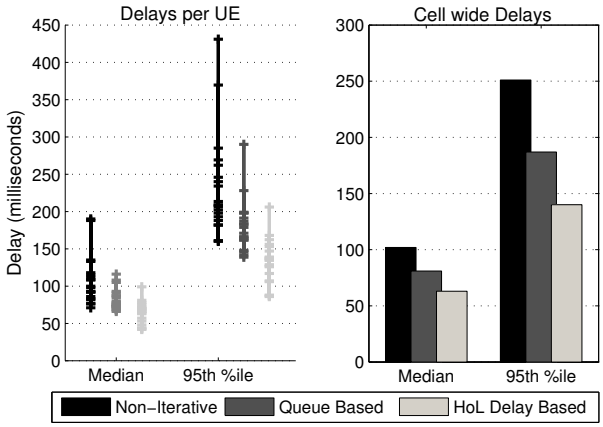

Fig. 7. Individual and Cell-wide Delay performance for micro cell simulation

VII. Frequency Selective Resource Allocation

We extend the analysis in [10] for frequency selective fading to concave functions f_i (such as the delay based reward function) which are thrice continuously differentiable everywhere except at L points where they are only continuous. We can re-write such a function as

$$f_i(r_i) = \sum_{l=1}^{L} f_{il}\left(\min\left(\rho_l - \rho_{l-1}, [r_i - \rho_{l-1}]_+\right)\right)$$

where $0 \leq \rho_1 < \ldots < \rho_L$ are the points of non-differentiability and $f_{il} : \mathbb{R}_+ \mapsto \mathbb{R}$ are thrice continuously differentiable concave functions defined as

$$f_{il}(x) = f_i(\rho_{l-1} + x) - f_i(\rho_{l-1}),\ x \in [0, \rho_l - \rho_{l-1}]$$

with $l \geq 1, \rho_0 = 0$, and satisfy

$$f'_{il}(x) < f'_{i,l-1}(y), x \in [0, \rho_l - \rho_{l-1}],\ y \in [0, \rho_{l-1} - \rho_{l-2}].$$

We also assume $x\psi^{-1}(y/x)$ is concave for all $(x, y) > 0$; this is true for example, when ψ is the Shannon capacity formula, and for practical M-QAM schemes.

Consider the following convex optimization problem over $\tilde{r}_{il}$'s, r_{ij}'s (rate for user i on sub-band j), and b_{ij}'s (bandwidth for user i on sub-band j):

$$\begin{aligned}
\max.\ & \sum_{i=1}^{N}\sum_{l=1}^{L} f_{il}(\tilde{r}_{il}), \\
\text{s.t.}\ & \sum_{l=1}^{L} \tilde{r}_{il} \leq \sum_{j=1}^{M} r_{ij},\ \forall i,\ \tilde{r}_{il} \leq \rho_l - \rho_{l-1},\ \forall i, l \\
& \sum_{i=1}^{N} b_{ij} = B,\ \forall j, \\
& \sum_{j=1}^{M} \frac{b_{ij}(N_0 + I_j)}{G_{ij}} \psi^{-1}\left(r_{ij}/b_{ij} - 1\right) \leq P,\ \forall i, \\
& r_{ij} \leq b_{ij}\psi\left(\frac{G_{ij}\gamma_{ij}}{N_0 + I_j}\right),\ r_{ij}, b_{ij} \geq 0,\ \forall i, j.
\end{aligned} \tag{7}$$

The first constraint implies that the total rate for a user is the sum of rates over sub-bands, the second constraint is on total bandwidth allocation in a sub-band, third constraint is on peak power at the UE in a subframe, and the fourth constraint

models fractional power control. The following lemma follows easily from the construction of the f_{il}s:

Lemma 7.1: If $(r^\star_{ij}, b^\star_{ij})$ is a solution to the optimization problem (7), then $\sum_i f_i \left(\sum_j r^\star_{ij}\right)$ is the maximum sum reward for any feasible resource allocation.

General purpose interior points methods to solve the above optimization problem have a complexity of $O(NM + NL)^3$ per iteration – we exploit the structure to reduce it to $O(N(L^2 + M^2))$. Note that in practice L and M are much smaller than N. In order to construct a solution for which the bandwidth allocation is contiguous in frequency to satisfy the SC-FDMA requirements, we can use the heuristic in [10]. The main computation to solve (7) is to determine the Newton step at each iteration which entails solving a set of linear equations of the form (we omit the details due to lack of space, the exact expressions can be obtained following the steps in [25]):

$$\left[\begin{array}{ccc|c} H_1 & & & A^T \\ & \ddots & & \\ & & H_N & \\ \hline & A & & 0 \end{array}\right] \left[\begin{array}{c} x_1 \\ \vdots \\ x_N \\ y \end{array}\right] = \left[\begin{array}{c} a \\ b \end{array}\right]$$

where $H_i \in \mathbb{R}^{(L+M)\times(L+M)}$, $A \in \mathbb{R}^{M\times N(L+M)}$, $x_i \in \mathbb{R}^{L+M}$, $a \in \mathbb{R}^{N(L+M)}$, $y, b \in \mathbb{R}^M$. We first eliminate the x_i's as

$$x_i = H_i^{-1}\left(a - A^T_{(L+M)(i-1)+1:(L+M)(i)} y\right)$$

where $A^T_{k:m}$ is the submatrix of A^T given by rows k to m. We invert H_i^{-1} in $O(L^2 + M^2)$ time, solve for y in $O(M^3)$ time (M linear equations in M variables), and back-substitute y to obtain x. To invert H_i, we note that it decomposes as

$$H_i = \left[\begin{array}{ccccccc} K_1 & & & & & & \\ & \ddots & & & & & \\ & & K_M & & & & \\ & & & g_1 & & & \\ & & & & \ddots & & \\ & & & & & g_L & \end{array}\right] + g_i g_i^T + \left[\begin{array}{cc} h_i h_i^T & 0 \\ 0 & 0 \end{array}\right] + \left[\begin{array}{cc} 0 & 0 \\ 0 & c_i c_i^T \end{array}\right]$$

where $g_i \in \mathbb{R}^{L+M}$, $h_i \in \mathbb{R}^L$, $c_i \in \mathbb{R}^M$. Using the matrix inversion lemma we can invert H_i in $O(L^2 + M^2)$ time.

VIII. Conclusions

We designed a general computational framework in this paper to enable a wide array of online scheduling policies in a computationally efficient manner. We modeled the constraints due to fractional power control, and formulated an optimization problem with non-differentiable objective function. We showed how to estimate the packet delays on the uplink via the BSR reports, and proposed a novel scheduling policy based on packet delays. Numerical results demonstrated that using packet delay estimates for the uplink can lead to significant reduction in packet delays as compared with a queue length based scheduler. There are many interesting directions for future work. For example, we can further study the connections with the work in [11], [12]. In terms of implementation, an interesting question is whether we can design approximation algorithms for the uplink bandwidth packing problem which are optimal according to some metric.

References

[1] E. Dahlman, S. Parkvall, J. Skold, and P. Beming, *3G Evolution, Second Edition: HSPA and LTE for Mobile Broadband*. Academic Press, 2008.

[2] H. J. Kushner and P. Whiting, "Convergence of proportional-fair sharing algorithms under general conditions," *IEEE Trans. Wireless Commun.*, 2004.

[3] A. Stolyar, "On the asymptotic optimality of the gradient scheduling algorithm for multi-user throughput allocation," *Oper. Res.*, 2005.

[4] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar, and P. Whiting, "Scheduling in a queueing system with asynchronously varying service rates," *Prob. Engg. & Inform. Sci.*, vol. 18, pp. 191–217, 2004.

[5] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *CDC*, 1990.

[6] S. Shakkottai and A. Stolyar, "Scheduling for multiple flows sharing a time-varying channel: The exponential rule," *American Mathematical Society Translations, Series 2,A*, vol. 207, 2002.

[7] B. Sadiq and G. de Veciana, "Large deviations sum-queue optimality of a radial sum-rate monotone opportunistic scheduler," *IEEE Trans. Info. Th.*, 2010.

[8] C. Castellanos, D. Villa, C. Rosa, K. Pedersen, F. Calabrese, P.-H. Michaelsen, and J. Michel, "Performance of uplink fractional power control in UTRAN LTE," *VTC Spring*, 2008.

[9] J. Huang, V. Subramanian, R. Agrawal, and R. Berry, "Joint scheduling and resource allocation in uplink OFDM systems for broadband wireless access networks," *IEEE JSAC*, 2009.

[10] R. Madan and S. Ray, "Uplink resource allocation for frequency selective channels and fractional power control," *To appear in ICC*, 2011.

[11] M. Sharma and X. Lin, "OFDM downlink scheduling for delay-optimality: Many-channel many-source asymptotics with general arrival processes," *ITA*, 2011.

[12] S. Bodas, S. Shakkottai, L. Ying, and R. Srikant, "Scheduling for small delay in multi-rate multi-channel wireless networks," in *INFOCOM*, 2011.

[13] Q. Zhang and S. Song, "Exact expression for the coherence bandwidth of rayleigh fading channels," *IEEE Trans. Commun.*, vol. 55, no. 7, pp. 1296 –1299, 2007.

[14] IEEE, "IEEE 802.11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications."

[15] G. T. 36.814, "Further advancements for E-UTRA Physical layer aspects," http://www.3gpp.org/.

[16] A. Baid, R. Madan, and A. Sampath, "Delay Estimation and Fast Iterative Scheduling Policies for LTE Uplink," 2012. [Online]. Available: http://arxiv.org/abs/1201.3215

[17] G. Fodor, C. Koutsimanis, A. Rcz, N. Reider, A. Simonsson, and W. Mller, "Intercell interference coordination in OFDMA networks and in the 3GPP Long Term Evolution system," *Journal of Commun.*, 2009.

[18] R. Madan, S. P. Boyd, and S. Lall, "Fast algorithms for resource allocation in wireless cellular networks," *IEEE/ACM Trans. Netw.*, 2010.

[19] B. Sadiq, R. Madan, and A. Sampath, "Downlink scheduling for multiclass traffic in lte," *EURASIP Journal on Wireless Communications and Networking Special Issue on 3GPP-LTE*, 2009.

[20] S. Kwon and N.-H. Lee, "Uplink QoS Scheduling for LTE System," in *Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd*, may 2011, pp. 1–5.

[21] N. Shor, K. Kiwiel, and A. Ruszcaynski, *Minimization methods for non-differentiable functions*. Springer-Verlag, 1985.

[22] V. Farias and R. Madan, "The irrevocable multiarmed bandit problem," *Operations Research*, vol. 59, no. 2, 2011.

[23] K. Brueninghaus, D. Astely, T. Salzer, S. Visuri, A. Alexiou, S. Karger, and G.-A. Seraji, "Link performance models for system level simulations of broadband radio access systems," *PIMRC*, 2005.

[24] G. C.-. V. 1.0, "CDMA2000 evaluation methodology," http://www.3gpp2.org/.

[25] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.