Optimal Scheduling for Fair Resource Allocation in Ad Hoc Networks with Elastic and Inelastic Traffic

Juan José Jaramillo
Coordinated Science Laboratory and
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Email: jjjarami@illinois.edu

R. Srikant
Coordinated Science Laboratory and
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Email: rsrikant@illinois.edu

Abstract—This paper studies the problem of congestion control and scheduling in ad hoc wireless networks that have to support a mixture of best-effort and real-time traffic. Optimization and stochastic network theory have been successful in designing architectures for fair resource allocation to meet long-term throughput demands. However, to the best of our knowledge, strict packet delay deadlines were not considered in this framework previously. In this paper, we propose a model for incorporating the quality of service (QoS) requirements of packets with deadlines in the optimization framework. The solution to the problem results in a joint congestion control and scheduling algorithm which fairly allocates resources to meet the fairness objectives of both elastic and inelastic flows, and per-packet delay requirements of inelastic flows.

I. INTRODUCTION

As wireless networks become more prevalent, they will be expected to support a wide variety of services, including best-effort and real-time traffic. Such networks will have to serve flows that require quality of service requirements, such as minimum bandwidth and maximum delay constraints, while at the same time keeping the network queues stable for data traffic and guaranteeing throughput optimality. For the case of wireless networks with best-effort traffic only, optimization-based algorithms which naturally map into different layers of the protocol stack have been proposed in the last few years [1]–[6]; see [7] for a survey. However, these models do not take into account strict per-packet delay bounds.

Scheduling packets with strict deadlines has been studied in [8]–[11], but all of these papers provide approximate solutions. The model that we study in this paper builds upon the recent work in [12]–[14] on admission control and scheduling for inelastic flows in collocated wireless networks, i.e., networks where all links interfere with each other. Among the many contributions in these papers is a key modeling innovation whereby the network is studied in frames, where a frame is a contiguous set of time-slots of fixed duration. Packets with deadlines are assumed to arrive at the beginning of a frame and have to be served by the end of the frame. In this paper, we explore this modeling paradigm further to study the design of resource allocation algorithms for ad hoc networks. The

frame-based model allows us to incorporate delay deadlines in the optimization framework for very general network models, and somewhat surprisingly, allows us to design a common framework for handling both elastic and inelastic flows.

The main contributions of the paper are as follows:

1) We present an optimization framework for resource allocation in a wireless network consisting of both best-effort flows and flows that generate traffic with per-packet delay constraints. The framework allows for very general interference, channel and arrival models.
2) Using a dual decomposition approach, we derive an optimal scheduling and congestion control algorithm that fairly allocates resources and ensures that a required fraction of each inelastic flow's packets are delivered on time by appealing to connections between Lagrange multipliers, queues, and service deficits. The scheduling algorithm seamlessly integrates inelastic and elastic traffic into a unified max-weight scheduling framework, extending the well-known results in [15].
3) The convergence of the above algorithm in an appropriate stochastic sense is proved and it is also shown that the network is stable.

II. NETWORK MODEL

The network is represented by a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{L})$, where $\mathcal{N}$ is the set of nodes and $\mathcal{L}$ is the set of directional links such that for all $n_1, n_2 \in \mathcal{N}$ if $(n_1, n_2) \in \mathcal{L}$ then node n_1 can transmit to node n_2. The links are numbered 1 through $|\mathcal{L}|$, and by abusing notation, we sometimes use $l \in \mathcal{L}$ to mean $l \in \{1, 2, \ldots, |\mathcal{L}|\}$.

Traffic is assumed to be a mixture of elastic and inelastic flows, where an inelastic flow is one that has maximum per-packet delay requirements. In contrast, elastic flows do not have such requirements.

Time is divided in slots, where a set of T consecutive time slots makes a *frame*. We assume that packet arrivals only occur at the beginning of a frame, and every inelastic packet has a deadline of T time slots. If a packet misses its deadline it is discarded, and it is required that the loss probability at link $l \in \mathcal{L}$ due to deadline expiry must be no more than p_l. For elastic traffic we associate a utility function $U_l(x_l)$ which is

a function of the mean elastic arrival rate per frame x_l. We assume that $U_l(.)$ is a concave function.

For a given frame, we denote by the vector $a_i = (a_{il})_{l \in \mathcal{L}}$ the number of inelastic packet arrivals at every link, where a_{il} is a random variable with mean λ_l and variance σ_{il}^2. We further assume that arrivals are independent between different frames and that $Pr(a_{il} = 0) > 0$ and $Pr(a_{il} = 1) > 0$. The last two assumptions are used to guarantee that the Markov chain we define later is both irreducible and aperiodic, although these can be replaced by other similar assumptions. Similarly, we define $a_e = (a_{el})_{l \in \mathcal{L}}$ to be the number of elastic packet arrivals at every link in a given frame.

The channel state is assumed to be constant in a given frame, independent between different frames, and independent of arrivals. The vector $c = (c_l)_{l \in \mathcal{L}}$ denotes the number of packets link l can successfully transmit on a time slot in a given frame.

Depending on the wireless technology used, we can have some channel feedback before or after a transmission occurs. If channel estimation is performed before transmitting, we can determine the optimal rate at which we can successfully transmit. Alternatively, feedback from the receiver after the transmission can be used to detect if a transmission is successful or not. In this paper we try to capture both scenarios in the following cases:

1) Known channel state: It is assumed that c_l is a *non-negative* random variable with mean $\bar{c}_l$ and variance σ_{cl}^2, and we get to know the channel state at the *beginning* of the frame.

2) Unknown channel state: It is assumed that c_l is a *Bernoulli* random variable with mean $\bar{c}_l$ and we only get to know the channel state at the *end* of the frame.

In the known channel state case where we do channel estimation to determine the optimal transmission rate, we can potentially send more than one packet in a time slot at higher rates. This is captured by the fact that we make no assumptions on the values c_l can take since it will be determined by the particular wireless technology used. In the case of unknown channel state we assume that we only get the binary feedback of acknowledgments, which is reflected in the Bernoulli assumption on c_l. In this case, and without any loss of generality, we assume only one packet can be transmitted per time slot per link.

In the rest of the paper we will consider the known channel case first and then in Section VI we will highlight the differences in the analysis for the unknown channel case.

III. PROBLEM FORMULATION

We first formulate the problem as a static optimization problem. Using decomposition theory, we will then obtain a dynamic solution to this problem and prove its stability using stochastic Lyapunov techniques.

A feasible schedule $s = (s_{il,t}, s_{el,t})$ is such that $s_{il,t}$, $s_{el,t}$ respectively denote the number of inelastic and elastic packets that can be scheduled for transmission at link $l \in \mathcal{L}$ and time slot $t \in \{1, 2, \ldots, T\}$; thus, $s_{il,t} + s_{el,t} > 0$ means that link l is

scheduled to transmit in time slot t of the frame. Furthermore, for any t, if $s_{il_1,t} + s_{el_1,t} > 0$ and $s_{il_2,t} + s_{el_2,t} > 0$ then links l_1 and l_2 can be scheduled to simultaneously transmit without interfering with each other. Assuming the inelastic arrivals and the channel state are given by a_i and c respectively, we have the following constraints:

$$\sum_{t=1}^{T} s_{il,t} \leq a_{il} \text{ for all } l \in \mathcal{L} \text{ and} \tag{1}$$

$$s_{il,t} + s_{el,t} \leq c_l \text{ for all } l \in \mathcal{L} \text{ and } t \in \{1, 2, \ldots, T\}. \tag{2}$$

We denote by $\mathcal{S}(a_i, c)$ the set of all feasible schedules when the arrival state is a_i and the channel state is c; thus, $\mathcal{S}(a_i, c)$ captures any interference constraints we have on our network and satisfies (1) and (2).

At the beginning of any frame we must choose a feasible schedule to serve all links and decide how many elastic packets are allowed to be injected in the network. Therefore, our goal is to find a function $Pr(s|a_i, c)$ which is the probability of using schedule $s \in \mathcal{S}(a_i, c)$ when the inelastic arrivals are given by a_i and the channel state is c, subject to the constraint that the loss probability at link $l \in \mathcal{L}$ due to deadline expiry cannot exceed p_l. For elastic traffic, we want to select the vector a_e such that we maximize the network utility while keeping the queues stable.

To properly formulate the problem, let us first define $\mu_i(a_i, c)$ to be the expected number of inelastic packets served if the number of packet arrivals is given by a_i and the channel state is c. Similarly, $\mu_e(a_i, c)$ denotes the expected number of elastic packets that can be served. Therefore, we have the following constraints:

$$\mu_{il}(a_i, c) \leq \sum_{s \in \mathcal{S}(a_i, c)} \sum_{t=1}^{T} s_{il,t} Pr(s|a_i, c)$$

$$\mu_{el}(a_i, c) \leq \sum_{s \in \mathcal{S}(a_i, c)} \sum_{t=1}^{T} s_{el,t} Pr(s|a_i, c).$$

The expected service for mixed traffic at link l is then given by

$$\mu_{il} \overset{def}{=} \sum_{a_i} \sum_{c} \mu_{il}(a_i, c) Pr(c) Pr(a_i)$$

$$\mu_{el} \overset{def}{=} \sum_{a_i} \sum_{c} \mu_{el}(a_i, c) Pr(c) Pr(a_i)$$

and due to QoS requirements and capacity constraints, we require that

$$\mu_{il} \geq \lambda_l (1 - p_l) \text{ and } x_l \leq \mu_{el}.$$

We will focus on maximizing the following objective for some given vector $w \in \mathbb{R}_+^{|\mathcal{L}|}$:

$$\max_{\substack{\mu_i(a_i, c), \mu_e(a_i, c), \\ \mu_i, \mu_e, x, Pr(s|a_i, c)}} \sum_{l \in \mathcal{L}} U_l(x_l) + w_l \mu_{il} \tag{3}$$

subject to

$$\mu_{il}(a_i,c) \leq \sum_s \sum_{t=1}^T s_{il,t} Pr(s|a_i,c) \text{ for all } l \in \mathcal{L}, a_i, c$$

$$\mu_{el}(a_i,c) \leq \sum_s \sum_{t=1}^T s_{el,t} Pr(s|a_i,c) \text{ for all } l \in \mathcal{L}, a_i, c$$

$$\mu_{il} = \sum_{a_i} \sum_c \mu_{il}(a_i,c) Pr(c) Pr(a_i) \text{ for all } l \in \mathcal{L}$$

$$\mu_{el} = \sum_{a_i} \sum_c \mu_{el}(a_i,c) Pr(c) Pr(a_i) \text{ for all } l \in \mathcal{L}$$

$$\mu_{il} \geq \lambda_l(1-p_l) \text{ for all } l \in \mathcal{L}$$

$$0 \leq x_l \leq \mu_{el} \text{ for all } l \in \mathcal{L}$$

$$Pr(s|a_i,c) \geq 0 \text{ for all } s \in \mathcal{S}(a_i,c), a_i, c$$

$$\sum_s Pr(s|a_i,c) \leq 1 \text{ for all } a_i, c.$$

The vector w can be used to allocate additional bandwidth fairly to inelastic flows beyond what is required to meet their QoS needs. Other uses for w will be explored in the simulations section. We will assume that the arrivals and loss probability requirements are feasible and thus the optimization problem has a solution (x^*, μ_i^*).

IV. SOLUTION USING DUAL DECOMPOSITION

Using the definition of the dual function [16], we have that $D(\delta_i, \delta_e) =$

$$\max_{\substack{\mu_i(a_i,c), \mu_e(a_i,c), \\ \mu_i, \mu_e, x, Pr(s|a_i,c)}} \sum_{l \in \mathcal{L}} \left\{ \begin{array}{l} U_l(x_l) + w_l \mu_{il} - \delta_{el}[x_l - \mu_{el}] \\ -\delta_{il}[\lambda_l(1-p_l) - \mu_{il}] \end{array} \right\}$$

subject to

$$\mu_{il}(a_i,c) \leq \sum_s \sum_{t=1}^T s_{il,t} Pr(s|a_i,c) \text{ for all } l \in \mathcal{L}, a_i, c$$

$$\mu_{el}(a_i,c) \leq \sum_s \sum_{t=1}^T s_{el,t} Pr(s|a_i,c) \text{ for all } l \in \mathcal{L}, a_i, c$$

$$\mu_{il} = \sum_{a_i} \sum_c \mu_{il}(a_i,c) Pr(c) Pr(a_i) \text{ for all } l \in \mathcal{L}$$

$$\mu_{el} = \sum_{a_i} \sum_c \mu_{el}(a_i,c) Pr(c) Pr(a_i) \text{ for all } l \in \mathcal{L}$$

$$x_l \geq 0 \text{ for all } l \in \mathcal{L}$$

$$Pr(s|a_i,c) \geq 0 \text{ for all } s \in \mathcal{S}(a_i,c), a_i, c$$

$$\sum_s Pr(s|a_i,c) \leq 1 \text{ for all } a_i, c.$$

Slater's condition [17] states that, since the objective is concave and the constraints are affine functions, the duality

gap is zero and therefore $D(\delta_i^*, \delta_e^*) = \sum_{l \in \mathcal{L}} U_l(x_l^*) + w_l \mu_{il}^*$, where

$$(\delta_i^*, \delta_e^*) \in \operatorname*{arg\,min}_{\delta_{il} \geq 0, \delta_{el} \geq 0} D(\delta_i, \delta_e).$$

We are interested in finding (x^*, μ_i^*) but not the value $D(\delta_i^*, \delta_e^*)$, so if we rewrite the objective in the dual function as

$$\max_{\substack{\mu_i(a_i,c), \mu_e(a_i,c), \\ \mu_i, \mu_e, x, Pr(s|a_i,c)}} \left\{ \begin{array}{l} \sum_{l \in \mathcal{L}} U_l(x_l) - \delta_{el} x_l \\ + \sum_{l \in \mathcal{L}} (w_l + \delta_{il}) \mu_{il} + \delta_{el} \mu_{el} \\ - \sum_{l \in \mathcal{L}} \delta_{il} \lambda_l (1 - p_l) \end{array} \right\}$$

we notice that the problem can be decomposed into the following subproblems:

$$\max_{x_l \geq 0} U_l(x_l) - \delta_{el} x_l$$

and

$$\max_{\substack{\mu_i(a_i,c), \mu_e(a_i,c), \\ \mu_i, \mu_e, Pr(s|a_i,c)}} \sum_{l \in \mathcal{L}} (w_l + \delta_{il}) \mu_{il} + \delta_{el} \mu_{el} \qquad (4)$$

subject to

$$\mu_{il}(a_i,c) \leq \sum_s \sum_{t=1}^T s_{il,t} Pr(s|a_i,c) \text{ for all } l \in \mathcal{L}, a_i, c$$

$$\mu_{el}(a_i,c) \leq \sum_s \sum_{t=1}^T s_{el,t} Pr(s|a_i,c) \text{ for all } l \in \mathcal{L}, a_i, c$$

$$\mu_{il} = \sum_{a_i} \sum_c \mu_{il}(a_i,c) Pr(c) Pr(a_i) \text{ for all } l \in \mathcal{L}$$

$$\mu_{el} = \sum_{a_i} \sum_c \mu_{el}(a_i,c) Pr(c) Pr(a_i) \text{ for all } l \in \mathcal{L}$$

$$Pr(s|a_i,c) \geq 0 \text{ for all } s \in \mathcal{S}(a_i,c), a_i, c$$

$$\sum_s Pr(s|a_i,c) \leq 1 \text{ for all } a_i, c.$$

Furthermore, since we are interested in solving the problem for non-negative values of δ_{il} and δ_{el}, it must be the case that μ_i^* and μ_e^* are as large as the constraints allow, and since the upper bounds for $\mu_{il}^*(a_i,c)$ and $\mu_{el}^*(a_i,c)$ are expressed as a convex combination, and the objective function in (4) is linear, the problem can be decomposed into the following subproblems for fixed a_i and c:

$$\max_{s \in \mathcal{S}(a_i,c)} \sum_{l \in \mathcal{L}} \left\{ (w_l + \delta_{il}) \sum_{t=1}^T s_{il,t} + \delta_{el} \sum_{t=1}^T s_{el,t} \right\}.$$

This suggests the following iterative algorithm to find the solution to our optimization problem, where k is the step index and $X_{max} > \max_{l \in \mathcal{L}} x_l^*$ is a fixed parameter:

$$\tilde{x}_l^*(k) \in \operatorname*{arg\,max}_{0 \leq x_l \leq X_{max}} U_l(x_l) - \delta_{el}(k) x_l$$

$$\tilde{s}^*(a_i, c, k) \in$$

$$\operatorname*{arg\,max}_{s \in \mathcal{S}(a_i,c)} \sum_{l \in \mathcal{L}} \left\{ [w_l + \delta_{il}(k)] \sum_{t=1}^{T} s_{il,t} + \delta_{el}(k) \sum_{t=1}^{T} s_{el,t} \right\}$$

$$\tilde{\mu}_{il}^*(k) = \sum_{a_i} \sum_{c} \sum_{t=1}^{T} \tilde{s}_{il,t}^*(a_i, c, k) Pr(c) Pr(a_i)$$

$$\tilde{\mu}_{el}^*(k) = \sum_{a_i} \sum_{c} \sum_{t=1}^{T} \tilde{s}_{el,t}^*(a_i, c, k) Pr(c) Pr(a_i).$$

We update the Lagrange multipliers $\delta_i(k)$, $\delta_e(k)$ at every step according to the following equations:

$$\delta_{il}(k+1) = \{\delta_{il}(k) + \epsilon[\lambda_l(1 - p_l) - \tilde{\mu}_{il}^*(k)]\}^+$$

and

$$\delta_{el}(k+1) = \{\delta_{el}(k) + \epsilon[\tilde{x}_l^*(k) - \tilde{\mu}_{el}^*(k)]\}^+$$

where $\epsilon > 0$ is a fixed step-size parameter, and for any $\alpha \in \mathbb{R}$, $\alpha^+ \overset{def}{=} \max\{\alpha, 0\}$.

Making the change of variables $\epsilon\hat{d}(k) = \delta_i(k)$ and $\epsilon\hat{q}(k) = \delta_e(k)$, we have that our iterative algorithm can be rewritten as

$$\tilde{x}_l^*(k) \in \operatorname*{arg\,max}_{0 \le x_l \le X_{max}} \frac{1}{\epsilon} U_l(x_l) - \hat{q}_l(k) x_l$$

$$\tilde{s}^*(a_i, c, k) \in$$

$$\operatorname*{arg\,max}_{s \in \mathcal{S}(a_i,c)} \sum_{l \in \mathcal{L}} \left\{ [\frac{1}{\epsilon} w_l + \hat{d}_l(k)] \sum_{t=1}^{T} s_{il,t} + \hat{q}_l(k) \sum_{t=1}^{T} s_{el,t} \right\}$$

$$\tilde{\mu}_{il}^*(k) = \sum_{a_i} \sum_{c} \sum_{t=1}^{T} \tilde{s}_{il,t}^*(a_i, c, k) Pr(c) Pr(a_i)$$

$$\tilde{\mu}_{el}^*(k) = \sum_{a_i} \sum_{c} \sum_{t=1}^{T} \tilde{s}_{el,t}^*(a_i, c, k) Pr(c) Pr(a_i).$$

with update equations

$$\hat{d}_l(k+1) = [\hat{d}_l(k) + \lambda_l(1 - p_l) - \tilde{\mu}_{il}^*(k)]^+$$

$$\hat{q}_l(k+1) = [\hat{q}_l(k) + \tilde{x}_l^*(k) - \tilde{\mu}_{el}^*(k)]^+.$$

It should be noted that due to the change of variables $\hat{d}_l(k)$ can be interpreted as a queue that has $\lambda_l(1 - p_l)$ arrivals and $\tilde{\mu}_{il}^*(k)$ departures at step k; $\hat{q}_l(k)$ can have a similar queue interpretation. The dual decomposition approach only provides an intuition behind the solution, but the real network has stochastic and dynamic arrivals and channel state conditions. In the next section, we present the complete solution which takes into account these dynamics and we also establish its convergence properties.

V. DYNAMIC ALGORITHM AND ITS CONVERGENCE ANALYSIS

A. Scheduler and Congestion Controller

To implement the algorithm online, we propose the following congestion control algorithm in frame k, where the queue length at link l is given by $q_l(k)$:

$$\tilde{x}_l^*(k) \in \operatorname*{arg\,max}_{0 \le x_l \le X_{max}} \frac{1}{\epsilon} U_l(x_l) - q_l(k) x_l. \tag{5}$$

We need to convert this elastic arrival rate, which in general is a non-negative real number, into a non-negative integer indicating the number of elastic packets allowed to enter the network in a given frame. This conversion can be made in many different ways: we assume the elastic arrivals at link l, $\tilde{a}_{el}(k)$, are a random variable with mean $\tilde{x}_l^*(k)$ and variance upper-bounded by σ_e^2, and are such that $Pr(\tilde{a}_{el}(k) = 0) > 0$ and $Pr(\tilde{a}_{el}(k) = 1) > 0$ for all $l \in \mathcal{L}$ and all k. The last two assumptions are used to guarantee the Markov chain we define below is both irreducible and aperiodic, although these can be replaced by other similar assumptions.

Letting the number of inelastic arrivals be denoted by $a_i(k)$ and the channel state by $c(k)$, we propose the following scheduling algorithm:

$$\tilde{s}^*(a_i(k), c(k), d(k), q(k)) \in \tag{6}$$

$$\operatorname*{arg\,max}_{s \in \mathcal{S}(a_i(k),c(k))} \sum_{l \in \mathcal{L}} \left\{ [\frac{1}{\epsilon} w_l + d_l(k)] \sum_{t=1}^{T} s_{il,t} + q_l(k) \sum_{t=1}^{T} s_{el,t} \right\}.$$

The vectors $d(k)$ and $q(k)$ are updated from frame to frame as follows:

$$d_l(k+1) = [d_l(k) + \tilde{a}_{il}(k) - I_{il}^*(a_i(k), c(k), d(k), q(k))]^+$$

$$q_l(k+1) = [q_l(k) + \tilde{a}_{el}(k) - I_{el}^*(a_i(k), c(k), d(k), q(k))]^+,$$

where

$$I_{il}^*(a_i(k), c(k), d(k), q(k)) = \sum_{t=1}^{T} \tilde{s}_{il,t}^*(a_i(k), c(k), d(k), q(k))$$

$$I_{el}^*(a_i(k), c(k), d(k), q(k)) = \sum_{t=1}^{T} \tilde{s}_{el,t}^*(a_i(k), c(k), d(k), q(k))$$

and $\tilde{a}_{il}(k)$ is a binomial random variable with parameters $a_{il}(k)$ and $1 - p_l$. The quantity $\tilde{a}_{il}(k)$ can be generated by the network as follows: upon each inelastic packet arrival, toss a coin with probability of *heads* equal to $1 - p_l$, and if the outcome is *heads*, add a one to the deficit counter.

In our notation we make explicit the fact that for fixed ϵ and w, the optimal scheduler (6) is a function of $a_i(k)$, $c(k)$, $d(k)$, and $q(k)$. We interpret $d_l(k)$ as a virtual queue that counts the deficit in service for link l to achieve a loss probability due to deadline expiry less than or equal to p_l. This deficit queue was first used in the inelastic traffic context in [12] for the case of collocated networks; the connection to the dual decomposition approach now provides a Lagrange multiplier interpretation to it and allows the extension to general ad hoc networks. Note that $q_l(k)$ is just the queue size for elastic packets at link l.

B. Convergence Results

For readability, we present the main results in this section, but the proofs are deferred to the appendixes. We start by noting that $(d(k), q(k))$ defines an irreducible and aperiodic Markov chain. To prove that our dynamic algorithm achieves the optimal solution to the static problem (3) in some average sense and fulfills all links' requirements, we will first bound the expected drift of $(d(k), q(k))$ for a suitable Lyapunov function.

Lemma 1: Consider the Lyapunov function $V(d, q) = \frac{1}{2} \sum_{l \in \mathcal{L}} d_l^2 + q_l^2$. If $\mu_{il}^* > \lambda_l(1 - p_l)$ and $\mu_{el}^* > x_l^*$ for all $l \in \mathcal{L}$, then

$$E\left[V(d(k+1), q(k+1))|d(k) = d, q(k) = q\right] - V(d, q)$$
$$\leq B_1 - B_2 \sum_{l \in \mathcal{L}} d_l - B_3 \sum_{l \in \mathcal{L}} q_l - \frac{1}{\epsilon} \sum_{l \in \mathcal{L}} [U_l(x_l^*) - U_l(\tilde{x}_l^*(k))]$$
$$- \frac{1}{\epsilon} \sum_{l \in \mathcal{L}} w_l \mu_{il}^* - w_l E\left[I_{il}^*(a_i(k), c(k), d, q)\right]$$

for some positive constants B_1, B_2, B_3, any $\epsilon > 0$, where (x^*, μ_i^*) is the solution to (3), $\tilde{x}^*(k)$ is the solution to (5), and $I_i^*(a_i(k), c(k), d, q)$ is obtained from the solution to (6). $\diamond$

It is important to note that since the last two terms in the right-hand side of the inequality can be upper-bounded, Lemma 1 implies that $(d(k), q(k))$ is positive recurrent since the expected drift is negative but for a finite set of values of $(d(k), q(k))$. As a direct consequence of this fact, we note that the total service deficit and queue length have a $O(1/\epsilon)$ bound.

Corollary 1: If $\mu_{il}^* > \lambda_l(1 - p_l)$ and $\mu_{el}^* > x_l^*$ for all $l \in \mathcal{L}$, then the total expected service deficit and network queue length is upper-bounded by

$$\limsup_{k \to \infty} E\left[\sum_{l \in \mathcal{L}} d_l(k) + q_l(k)\right] \leq B_4 + \frac{1}{\epsilon} B_5$$

for all $l \in \mathcal{L}$ and

$$B_4 = \frac{B_1}{\min\{B_2, B_3\}}$$

and

$$B_5 \leq \frac{\sum_{l \in \mathcal{L}} \max_{0 \leq x_l \leq X_{max}} 2|U_l(x_l)| + w_l \lambda_l}{\min\{B_2, B_3\}}.$$

$\diamond$

This also implies that the scheduling and congestion control algorithm fulfills all links' inelastic requirements.

Corollary 2: If $\mu_{il}^* > \lambda_l(1 - p_l)$ and $\mu_{el}^* > x_l^*$ for all $l \in \mathcal{L}$, then the online algorithm fulfills all the inelastic constraints. That is:

$$\liminf_{K \to \infty} E\left[\frac{1}{K} \sum_{k=1}^{K} I_{il}^*(a_i(k), c(k), d(k), q(k))\right] \geq \lambda_l(1 - p_l)$$

for all $l \in \mathcal{L}$. $\diamond$

The above corollary simply states that the arrival rate into the deficit counter is less than or equal to the departure rate.

This result is an obvious consequence of the stability of the deficit counters and so a formal proof is not provided here.

Now we are ready to prove that our online algorithm is within $O(\epsilon)$ of the optimal value.

Theorem 1: For any $\epsilon > 0$, if $\mu_{il}^* > \lambda_l(1 - p_l)$ and $\mu_{el}^* > x_l^*$ for all $l \in \mathcal{L}$, then

$$\limsup_{K \to \infty} E\left[\sum_{l \in \mathcal{L}} U_l(x_l^*) + w_l \mu_{il}^* - \sum_{l \in \mathcal{L}} \frac{1}{K} \sum_{k=1}^{K} U_l(\tilde{x}_l^*(k))\right.$$
$$\left. - \sum_{l \in \mathcal{L}} \frac{1}{K} \sum_{k=1}^{K} w_l I_{il}^*(a_i(k), c(k), d(k), q(k))\right] \leq B\epsilon$$

for some $B > 0$, where (x^*, μ_i^*) is the solution to (3), $\tilde{x}^*(k)$ is the solution to (5), and $I_i^*(a_i(k), c(k), d(k), q(k))$ is obtained from the solution to (6). $\diamond$

In conclusion, there is a trade-off in choosing the parameter ϵ: smaller values will achieve a solution closer to the optimal, but at the same time the deficit in service at the links and the aggregate queue length increase. The statement and the proof of Theorem 1 follows the techniques in [3]. The result can also be derived, in a slightly different form, using the techniques in [4]. A closely related result can be obtained using the methods in [1].

VI. UNKNOWN CHANNEL STATE

The analysis for the unknown channel case is similar to the one we presented for the known channel case, so in this section we will only highlight the differences.

A feasible schedule $s = (s_{il,t}, s_{el,t})$ is such that $s_{il,t}$, $s_{el,t}$ respectively denote the number of inelastic and elastic packets that can be scheduled for transmission at link $l \in \mathcal{L}$ and time $t \in \{1, 2, \ldots, T\}$ without violating any interference constraints. Assuming the inelastic arrivals are given by a_i, and since we can only schedule at most one packet per link at every time slot, we have the following constraints:

$$\sum_{t=1}^{T} s_{il,t} \leq a_{il} \text{ for all } l \in \mathcal{L} \text{ and} \tag{7}$$

$$s_{il,t} + s_{el,t} \leq 1 \text{ for all } l \in \mathcal{L} \text{ and } t \in \{1, 2, \ldots, T\}. \tag{8}$$

We denote by $\mathcal{S}(a_i)$ the set of all feasible schedules for fixed arrivals, capturing any interference constraints we have on our network, and satisfying (7) and (8).

Our goal now is to find a function $Pr(s|a_i)$ which is the probability of using schedule $s \in \mathcal{S}(a_i)$ when the inelastic arrivals are given by a_i, subject to the constraint that the loss probability at link $l \in \mathcal{L}$ due to deadline expiry cannot exceed p_l. For elastic traffic, we still want to select the vector a_e such that we maximize the total utility while keeping the queues stable.

For a given distribution $Pr(s|a_i)$ we have that $\mu_{il}(a_i)$ is the expected number of attempted inelastic transmissions if arrivals are given by a_i. Similarly, $\mu_{el}(a_i)$ denotes the expected

number of times link l is scheduled to serve elastic packets in a given frame. As before, we have the following constraints:

$$\mu_{il}(a_i) \leq \sum_s \sum_{t=1}^{T} s_{il,t} Pr(s|a_i)$$

$$\mu_{el}(a_i) \leq \sum_s \sum_{t=1}^{T} s_{el,t} Pr(s|a_i)$$

When the (unknown) channel state is c, we have that $c_l \mu_{il}(a_i)$ is the expected number of successful inelastic transmissions per frame at link l for fixed arrivals, while $c_l \mu_{el}(a_i)$ is the expected service to link l for inelastic arrivals. Thus, the expected service for mixed traffic at link l is given by

$$\mu_{il} \stackrel{def}{=} \sum_{a_i} \sum_c c_l \mu_{il}(a_i) Pr(c) Pr(a_i)$$

$$\mu_{el} \stackrel{def}{=} \sum_{a_i} \sum_c c_l \mu_{el}(a_i) Pr(c) Pr(a_i).$$

Simplifying both expressions we get

$$\mu_{il} = \sum_{a_i} \bar{c}_l \mu_{il}(a_i) Pr(a_i)$$

$$\mu_{el} = \sum_{a_i} \bar{c}_l \mu_{el}(a_i) Pr(a_i).$$

Due to service requirements and capacity constraints we need that

$$\mu_{il} \geq \lambda_l (1 - p_l) \text{ and } x_l \leq \mu_{el}.$$

With the definitions and constraints stated above we can formulate the optimization problem in a similar way as in (3).

The only difference with the known channel state case is the scheduling algorithm. Assuming inelastic arrivals are given by $a_i(k)$ the scheduling algorithm is given by

$$\tilde{s}^*(a_i(k), d(k), q(k)) \in$$

$$\arg\max_{s \in \mathcal{S}(a_i(k))} \sum_{l \in \mathcal{L}} \left\{ [\frac{1}{\epsilon} w_l + d_l(k)] \bar{c}_l \sum_{t=1}^{T} s_{il,t} + q_l(k) \bar{c}_l \sum_{t=1}^{T} s_{el,t} \right\}.$$

The main difference in the scheduling algorithm compared to the known channel state case is that the network now uses the expected channel state in making scheduling decisions. Thus, the network needs to know or estimate $\bar{c}_l$ as in [12].

Similar results can be proved for this algorithm using the techniques developed in Section V-B, whereby one can show that the algorithm meets all the inelastic QoS constraints, the total expected service deficits and the queue lengths have a $O(1/\epsilon)$ bound, and the mean value of the objective is within $O(\epsilon)$ of the optimal value.
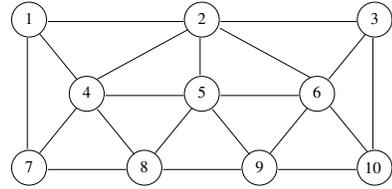

Fig. 1.   Interference graph used in the simulations

VII. SIMULATIONS

The purpose of this simulation study is to understand how the parameter ϵ and the link weights w_l impact the performance of the algorithm, and how a greedy heuristic can be used to implement the optimal scheduler. We simulate a 10-link network with an interference graph given by Fig. 1, where each node represents a link and each edge means that the two adjacent links cannot be scheduled simultaneously. For example, if link 1 is scheduled, then links 2, 4, and 7 cannot be activated. The required loss probability due to deadline expiry of inelastic packets is set to 0.1, the link arrivals are assumed to have a Bernoulli distribution with mean 0.6 packets/frame, and there are 3 time slots per frame. The channel for every link is assumed to have a Bernoulli distribution with mean 0.96, and we get to know the channel state at the beginning of the frame. We set $U_l(x_l) = \log(x_l)$ for all links. The simulation time was 10^6 frames.

As can be noted from (6), the max-weight scheduler requires that we do an exhaustive search to find the optimal schedule at every frame. For large networks this can become a burden due to the large search space; thus we explore a greedy heuristic and check how close it is to the optimal solution: at any given time slot, the greedy scheduler orders all links according to their weights. The greedy scheduler adds one of the links with the largest weight to the schedule, then removes all links that interfere with this link from the graph, then schedules a link with the largest weight among the remaining links, and so on. This procedure continues until no more links can be scheduled.

In Figs. 2, 3, and 4, we plot the expected values of the deficit counters and queues per link for various values of w_l, and compare their evolution for both the scheduler with optimal decisions and the greedy scheduler. The greedy scheduler seems to give lower deficit values than the optimal scheduler for larger values of w_l. We believe that the reason is that weights given to inelastic flows increase with increasing w_l and therefore, the greedy scheduler picks them first. However, our optimality goal is given by (3) which is determined by the rates received by the various flows. The rates achieved by the two schedulers are quite close in the simulations, which are not shown here due to space limitations; the interested reader can see these figures in [18].

We see that as w_l increases, the deficit counters become small. The upper bound in Corollary 1 only suggests that the sum of the deficit counters and queues is $O(1/\epsilon)$. Thus, it is interesting to note that by changing w_l, one can nearly eliminate the backlog in deficit for inelastic traffic while
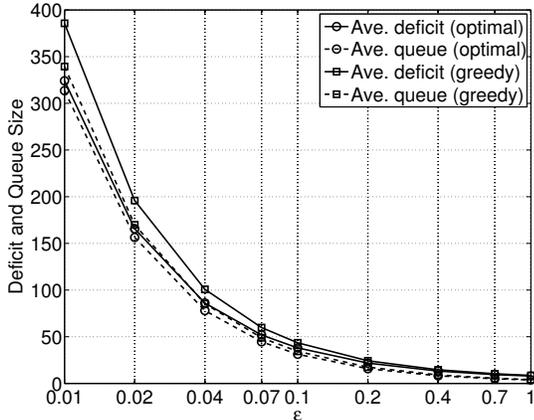
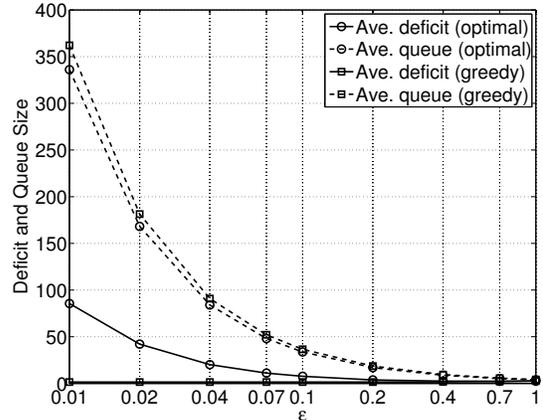Fig. 2.   Deficit size and queue length when $w_l = 0$

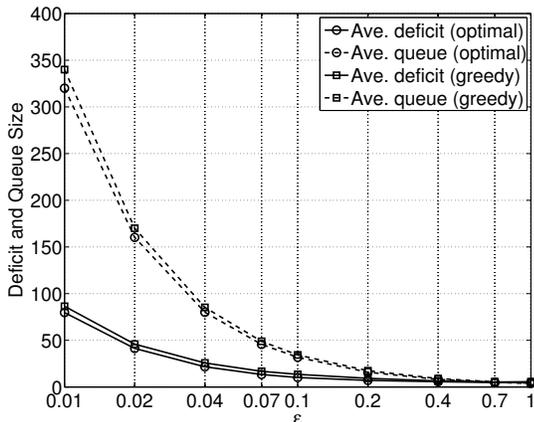Fig. 4.   Deficit size and queue length when $w_l = 6$

Fig. 3.   Deficit size and queue length when $w_l = 3$

maintaining the same order of queue sizes. The reason for this can be understood by examining the scheduling algorithm (6). Note that the algorithm gives priority to elastic traffic if queues are larger than counters. When w_l is small compared to ϵ, the effect of w_l is negligible in the scheduling algorithm. On the other hand, when w_l is $O(1)$, w_l/ϵ is $O(1/\epsilon)$ which is comparable to the queue lengths and hence, the deficit does not have to be large to provide service to inelastic traffic under algorithm (6).

It must be noted that small deficit counters mean that there is a small backlog in providing acceptable service to inelastic arrivals. For the case of real-time traffic this is a desirable property, since we do not want to have large variations in the service provided that could affect the perceived quality. Thus, even if fair allocation of bandwidth beyond the minimum is not required for inelastic flows, choosing w_l an order of magnitude larger than ϵ is desirable to maintain small deficits.

VIII. CONCLUSIONS

In this paper we have presented an optimization framework for the problem of congestion control and scheduling of elastic and inelastic traffic in ad hoc wireless networks. The model was developed for general interference graphs, general arrivals and time-varying channels. Using a dual function approach we presented a decomposition of the problem into an online algorithm that is able to make optimal decisions while keeping the network stable and fulfilling the inelastic flow's QoS constraints. A key result is that, through the use of deficit counters, one can treat the scheduling problem for elastic and inelastic flows in a common framework. It is also interesting to note that the deficit counters introduced in [12]–[14] have the interpretation of Lagrange multipliers. Simulations corroborate our results and show the dependency of the performance of the algorithm on the auxiliary parameter ϵ and its role into assigning resources to both elastic and inelastic traffic. We note that, in the unknown channel case, we do not consider channel feedback at the end of each time slot as in [12]–[14]. This will be addressed in future work.

APPENDIX A
PROOF OF LEMMA 1

To prove Lemma 1, we start by first stating two auxiliary lemmas and then stating a fact. The interested reader can see the proofs of the lemmas in [18].

Lemma 2: Given that at frame k we have the event $d(k) = d$ and $q(k) = q$, then

$$E\left[\frac{1}{2}\sum_{l\in\mathcal{L}}\{[d_l + \tilde{a}_{il}(k) - I_{il}^*(a_i(k), c(k), d, q)]^+\}^2\right] - \sum_{l\in\mathcal{L}}\frac{d_l^2}{2}$$

$$\leq B_6 + \sum_{l\in\mathcal{L}} d_l\lambda_l(1 - p_l)$$

$$- E\left[\sum_{l\in\mathcal{L}}\left(\frac{1}{\epsilon}w_l + d_l\right)I_{il}^*(a_i(k), c(k), d, q)\right]$$

$$-\sum_{l\in\mathcal{L}}\frac{1}{\epsilon}w_l I_{il}^*(a_i(k),c(k),d,q)\Bigg]$$

for some non-negative constant B_6, and where $I_{il}^*(a_i(k),c(k),d,q)$ is given by the solution to (6). ◇

Lemma 3: Given that at frame k we have the event $d(k)=d$ and $q(k)=q$, then

$$E\left[\frac{1}{2}\sum_{l\in\mathcal{L}}\{[q_l+\tilde{a}_{el}(k)-I_{el}^*(a_i(k),c(k),d,q)]^+\}^2\right]-\sum_{l\in\mathcal{L}}\frac{q_l^2}{2}$$

$$\leq B_7-\frac{1}{\epsilon}\sum_{l\in\mathcal{L}}[U_l(x_l^*)-U_l(\tilde{x}_l^*(k))]$$

$$-\sum_{l\in\mathcal{L}}q_l\left\{E\left[I_{el}^*(a_i(k),c(k),d,q)\right]-x_l^*\right\}$$

for some constant $B_7>0$, where x^* and $\tilde{x}^*(k)$ are the solutions to (3) and (5) respectively. ◇

Fact 1: The optimization in (6) can be performed over $\mathcal{S}(a_i(k),c(k))_{\mathcal{CH}}$, the convex hull of $\mathcal{S}(a_i(k),c(k))$; that is,

$$\max_{s\in\mathcal{S}(a_i(k),c(k))}\sum_{l\in\mathcal{L}}[\frac{1}{\epsilon}w_l+d_l(k)]\sum_{t=1}^{T}s_{il,t}+q_l(k)\sum_{t=1}^{T}s_{el,t}=$$

$$\max_{s\in\mathcal{S}(a_i(k),c(k))_{\mathcal{CH}}}\sum_{l\in\mathcal{L}}[\frac{1}{\epsilon}w_l+d_l(k)]\sum_{t=1}^{T}s_{il,t}+q_l(k)\sum_{t=1}^{T}s_{el,t}.$$

The reason for this comes from the fact that the objective function is linear and therefore there must be an optimal point $s^*(a_i(k),c(k),d(k),q(k))\in\mathcal{S}(a_i(k),c(k))$. ◇

Proof of Lemma 1: For the purpose of this proof, we define the capacity region for fixed arrival and channel states a_i and c as follows:

$$\mathcal{C}(a_i,c)\stackrel{def}{=}\left\{\begin{array}{l}(\bar{\mu}_{il},\bar{\mu}_{el})_{l\in\mathcal{L}}: \text{ there exists }\bar{s}\in\mathcal{S}(a_i,c)_{\mathcal{CH}},\\ \bar{\mu}_{il}\leq\sum_{t=1}^{T}\bar{s}_{il,t}\text{ and }\bar{\mu}_{el}\leq\sum_{t=1}^{T}\bar{s}_{el,t}\end{array}\right\}.$$

Then, the overall capacity of the network is defined as $\mathcal{C}\stackrel{def}{=}$

$$\left\{\begin{array}{l}(\mu_{il},\mu_{el})_{l\in\mathcal{L}}: \text{ there exists }(\bar{\mu}_{il}(a_i,c),\bar{\mu}_{el}(a_i,c))_{l\in\mathcal{L}}\in\\ \mathcal{C}(a_i,c)\text{ for all }a_i,c\text{ and }\mu_{il}=E[\bar{\mu}_{il}(a_i,c)],\\ \mu_{el}=E[\bar{\mu}_{el}(a_i,c)]\text{ for all }l\in\mathcal{L}\end{array}\right\}.$$

From Lemmas 2 and 3 we have:

$$E\left[V(d(k+1),q(k+1))|d(k)=d,q(k)=q\right]-V(d,q)$$

$$\leq B_1+\sum_{l\in\mathcal{L}}d_l\lambda_l(1-p_l)+\sum_{l\in\mathcal{L}}q_l x_l^*$$

$$-\frac{1}{\epsilon}\sum_{l\in\mathcal{L}}[U_l(x_l^*)-U_l(\tilde{x}_l^*(k))]$$

$$-E\left[\sum_{l\in\mathcal{L}}\left(\frac{1}{\epsilon}w_l+d_l\right)I_{il}^*(a_i(k),c(k),d,q)\right.$$

$$+\sum_{l\in\mathcal{L}}q_l I_{el}^*(a_i(k),c(k),d,q)\Bigg]$$

$$+\sum_{l\in\mathcal{L}}\frac{1}{\epsilon}w_l E\left[I_{il}^*(a_i(k),c(k),d,q)\right]$$

$$\leq B_1+\sum_{l\in\mathcal{L}}d_l\lambda_l(1-p_l)+\sum_{l\in\mathcal{L}}q_l x_l^* \qquad (9)$$

$$-\frac{1}{\epsilon}\sum_{l\in\mathcal{L}}[U_l(x_l^*)-U_l(\tilde{x}_l^*(k))]$$

$$-E\left[\sum_{l\in\mathcal{L}}\left(\frac{1}{\epsilon}w_l+d_l\right)\bar{\mu}_{il}(a_i(k),c(k))\right.$$

$$+\sum_{l\in\mathcal{L}}q_l\bar{\mu}_{el}(a_i(k),c(k))\Bigg]$$

$$+\sum_{l\in\mathcal{L}}\frac{1}{\epsilon}w_l E\left[I_{il}^*(a_i(k),c(k),d,q)\right]$$

$$=B_1-\sum_{l\in\mathcal{L}}d_l[\mu_{il}-\lambda_l(1-p_l)]-\sum_{l\in\mathcal{L}}q_l(\mu_{el}-x_l^*) \qquad (10)$$

$$-\frac{1}{\epsilon}\sum_{l\in\mathcal{L}}[U_l(x_l^*)-U_l(\tilde{x}_l^*(k))]$$

$$-\frac{1}{\epsilon}\sum_{l\in\mathcal{L}}w_l\mu_{il}-w_l E\left[I_{il}^*(a_i(k),c(k),d,q)\right]$$

where $B_1=B_6+B_7$, (9) follows for any $(\bar{\mu}_{il}(a_i(k),c(k)),\bar{\mu}_{el}(a_i(k),c(k)))_{l\in\mathcal{L}}\in\mathcal{C}(a_i(k),c(k))$ as was explained in Fact 1, and (10) holds for any $(\mu_{il},\mu_{el})_{l\in\mathcal{L}}\in\mathcal{C}$. It should be clear that $(\mu_{il}^*,\mu_{el}^*)_{l\in\mathcal{L}}\in\mathcal{C}$, where $(\mu_{il}^*,\mu_{el}^*)_{l\in\mathcal{L}}$ is the solution to (3). Thus we have the following:

$$E\left[V(d(k+1),q(k+1))|d(k)=d,q(k)=q\right]-V(d,q)$$

$$\leq B_1-B_2\sum_{l\in\mathcal{L}}d_l-B_3\sum_{l\in\mathcal{L}}q_l-\frac{1}{\epsilon}\sum_{l\in\mathcal{L}}[U_l(x_l^*)-U_l(\tilde{x}_l^*(k))]$$

$$-\frac{1}{\epsilon}\sum_{l\in\mathcal{L}}w_l\mu_{il}^*-w_l E\left[I_{il}^*(a_i(k),c(k),d,q)\right]$$

where

$$B_2=\min_{l\in\mathcal{L}}\left\{\mu_{il}^*-\lambda_l(1-p_l)\right\}$$

and

$$B_3=\min_{l\in\mathcal{L}}\left\{\mu_{el}^*-x_l^*\right\}.$$

∎

APPENDIX B
PROOF OF THEOREM 1

From Lemma 1 we know that

$$\frac{1}{\epsilon}\sum_{l\in\mathcal{L}}U_l(x_l^*)-U_l(\tilde{x}_l^*(k))$$

$$+\frac{1}{\epsilon}\sum_{l\in\mathcal{L}}w_l\mu_{il}^*-w_l E\left[I_{il}^*(a_i(k),c(k),d,q)\right]$$

$$\leq B_1-B_2\sum_{l\in\mathcal{L}}d_l-B_3\sum_{l\in\mathcal{L}}q_l+V(d,q)$$

$$-E\left[V(d(k+1),q(k+1))|d(k)=d,q(k)=q\right]$$

$$\leq B_1+V(d,q)$$

$$-E\left[V(d(k+1),q(k+1))|d(k)=d,q(k)=q\right]$$

since $B_2 \sum_{l \in \mathcal{L}} d_l + B_3 \sum_{l \in \mathcal{L}} q_l \geq 0$. Taking expectations:

$$\frac{1}{\epsilon} E \left[\sum_{l \in \mathcal{L}} U_l(x_l^*) - U_l(\tilde{x}_l^*(k)) \right.$$

$$\left. + \sum_{l \in \mathcal{L}} w_l \mu_{il}^* - w_l I_{il}^*(a_i(k), c(k), d(k), q(k)) \right]$$

$$\leq B_1 - E\left[V(d(k+1), q(k+1))\right] + E\left[V(d(k), q(k))\right].$$

Adding the terms for $k = \{1, \ldots, K\}$ and dividing by K we get:

$$\frac{1}{\epsilon} E \left[\sum_{l \in \mathcal{L}} U_l(x_l^*) + w_l \mu_{il}^* \right.$$

$$\left. - \sum_{l \in \mathcal{L}} \frac{1}{K} \sum_{k=1}^{K} U_l(\tilde{x}_l^*(k)) + w_l I_{il}^*(a_i(k), c(k), d(k), q(k)) \right]$$

$$\leq B_1 - \frac{E\left[V(d(K+1), q(K+1))\right]}{K} + \frac{E\left[V(d(1), q(1))\right]}{K}$$

$$\leq B_1 + \frac{E\left[V(d(1), q(1))\right]}{K} \tag{11}$$

where (11) follows from the fact that the Lyapunov function V is non-negative.

Assuming $E\left[V(d(1), q(1))\right] < \infty$ we get the following limit expression:

$$\limsup_{K \to \infty} E \left[\sum_{l \in \mathcal{L}} U_l(x_l^*) + w_l \mu_{il}^* - \sum_{l \in \mathcal{L}} \frac{1}{K} \sum_{k=1}^{K} U_l(\tilde{x}_l^*(k)) \right.$$

$$\left. - \sum_{l \in \mathcal{L}} \frac{1}{K} \sum_{k=1}^{K} w_l I_{il}^*(a_i(k), c(k), d(k), q(k)) \right] \leq B\epsilon$$

where $B = B_1$. ∎

REFERENCES

[1] A. Eryilmaz and R. Srikant, "Fair resource allocation in wireless networks using queue-length-based scheduling and congestion control," in *IEEE INFOCOM*, vol. 3, Miami, FL, USA, Mar. 13–17, 2005, pp. 1794–1803.

[2] X. Lin and N. B. Shroff, "Joint rate control and scheduling in multihop wireless networks," in *43rd IEEE Conference on Decision and Control (CDC)*, vol. 2, Atlantis, Paradise Island, Bahamas, Dec. 14–17, 2004, pp. 1484–1489.

[3] M. J. Neely, E. Modiano, and C.-P. Li, "Fairness and optimal stochastic control for heterogeneous networks," in *IEEE INFOCOM*, vol. 3, Miami, FL, USA, Mar. 13–17, 2005, pp. 1723–1734.

[4] A. Stolyar, "Maximizing queueing network utility subject to stability: Greedy primal-dual algorithm," *Queueing Systems*, vol. 50, no. 4, pp. 401–457, Aug. 2005.

[5] A. Eryilmaz and R. Srikant, "Joint congestion control, routing and mac for stability and fairness in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1514–1524, Aug. 2006.

[6] L. Chen, S. H. Low, M. Chiang, and J. C. Doyle, "Cross-layer congestion control, routing and scheduling design in ad hoc wireless networks," in *IEEE INFOCOM*, Barcelona, Catalunya, Spain, Apr. 23–29, 2006.

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

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

[9] V. Raghunathan, V. Borkar, M. Cao, and P. R. Kumar, "Index policies for real-time multicast scheduling for wireless broadcast systems," in *IEEE INFOCOM*, Phoenix, AZ, USA, Apr. 13–18, 2008, pp. 1570–1578.

[10] A. Dua and N. Bambos, "Downlink wireless packet scheduling with deadlines," *IEEE Trans. Mobile Comput.*, vol. 6, no. 12, pp. 1410–1425, Dec. 2007.

[11] Q. Liu, X. Wang, and G. B. Giannakis, "A cross-layer scheduling algorithm with QoS support in wireless networks," *IEEE Trans. Veh. Technol.*, vol. 55, no. 3, pp. 839–847, May 2006.

[12] I.-H. Hou, V. Borkar, and P. R. Kumar, "A theory of QoS for wireless," in *IEEE INFOCOM*, Rio de Janeiro, Brazil, Apr. 19–25, 2009, pp. 486–494.

[13] I.-H. Hou and P. R. Kumar, "Admission control and scheduling for QoS guarantees for variable-bit-rate applications on wireless channels," in *10th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc)*, New Orleans, LA, USA, May 18–21, 2009, pp. 175–184.

[14] ——, "Scheduling heterogeneous real-time traffic over fading wireless channels," arXiv, Tech. Rep. 0908.0587, 2009. [Online]. Available: http://arxiv.org/abs/0908.0587

[15] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Trans. Autom. Control*, vol. 37, no. 12, pp. 1936–1948, Dec. 1992.

[16] D. G. Luenberger, *Linear and Nonlinear Programming*, 2nd ed. Norwell, MA: Kluwer Academic Publishers, 2003.

[17] S. Boyd and L. Vandenberghe, *Convex Optimization*, 1st ed. New York, NY: Cambridge University Press, Mar. 2004.

[18] J. J. Jaramillo and R. Srikant, "Optimal scheduling for fair resource allocation in ad hoc networks with elastic and inelastic traffic," arXiv, Tech. Rep. 0907.5402, 2009. [Online]. Available: http://arxiv.org/abs/0907.5402