Joint Rate Control and Scheduling for Providing Bounded Delay with High Efficiency in Multi-hop Wireless Networks

Soroush Jahromizadeh and Veselin Rakočević, *Member, IEEE*.

***Abstract*—We consider the problem of supporting traffic with elastic bandwidth requirements and average end-to-end delay constraints in multi-hop wireless networks, with focus on source rates and link data rates as the key resource allocation decisions. The network utility maximisation-based approaches to support delay-sensitive traffic have been predominantly based on either reducing link utilisation, or approximation of links as M/D/1 queues, which lead to inefficient link utilisation under optimal resource allocation, and mostly to unpredictable transient behaviour of packet delays. On the contrary, we present an alternative formulation where the delay constraint is omitted and sources' utility functions are multiplied by a weight factor. The alternative optimisation problem is solved by a scheduling algorithm incorporating a duality-based rate control algorithm at its inner layer, where link prices correlate with their average queueing delays. We then present an alternative strategy where the utility weight of each source is adjusted to ensure its desired optimal path prices, and hence the desired average path delays. Since the proposed strategy is based on solving a concave optimisation problem for the elastic traffic, it leads to maximal utilisation of the network capacity. The proposed approach is then realised by a scheduling algorithm that runs jointly with an integral controller whereby each source independently regulates the queueing delay on its paths at the desired level, using its utility weight factor as the control variable. The proposed algorithms are shown, using theoretical analysis and simulation, to achieve asymptotic regulation of end-to-end delay with good performance.**

***Index Terms*—Ad hoc networks, Quality of service, Cross layer design, Wireless networks.**

I. Introduction

MULTI-HOP wireless networks in essence use multi-hop routing, where autonomous nodes can act as relay for the traffic, in order to provide enhanced wireless network capacity, connectivity and power efficiency without the need for a fixed infrastructure and centralised control. However, supporting applications with high data rate requirements and strict delay constraints over such networks is a challenging design problem due to their dynamic and distributed nature. Specifically, since the wireless channel is shared by all nodes, their transmission interfere with each other and consequently the data rates of the wireless links are intertwined. As nodes join or leave the network, new links are set up or disappear, resulting in the interference level changes. Thus, the data rates of wireless links vary over time as a result of fluctuating interference as well as shadow fading. Furthermore, multi-hop routing and distributed control compromise the performance due to the additional communication overhead and time needed for coordination.

As pointed out in [1], approaches for quality of service (QoS) provision can be classified into bandwidth reservation, and best-effort schemes. Bandwidth reservation based schemes are suitable for supporting traffic with minimum bandwidth requirements. Inevitably, the rejection of some of the traffic results in inefficient utilisation of network capacity. More importantly, Bandwidth reservation based schemes lead to a significant communication overhead in networks with dynamic settings and as a result are unsuitable for multi-hop wireless networks. In contrast, best-effort schemes are suitable for supporting traffic whose demand for bandwidth is elastic, but their perceived QoS referred to as their utilities, are generally assumed to be an increasing and concave function of their transmission rates. As concluded in [2], for elastic traffic the overall network QoS is maximised by admitting all connections and allocating network capacity based on the connections' perceived QoS. This is the main design principle of best-effort schemes. The key advantage of the best-effort approach is that it enables QoS optimal allocation of network resources using simple and distributed algorithms [1], and as a result is a suitable approach for QoS provision in multi-hop wireless networks.

The problem of supporting traffic with high data rate requirements and average end-to-end delay constraints in multi-hop wireless networks is the main focus of this paper. We assume that all incoming traffic have elastic bandwidth requirements, but their perceived signal qualities, which exclude the effect of end-to-end delay, are increasing and concave functions of their transmission rates. Furthermore, all or some of the incoming traffic impose a strict limit on the average end-to-end queueing delay. With focus on source data transmission rates and link data rates as the key resource allocation decisions, our main objective is then to develop a source rate control and scheduling strategy that guarantees bounded average end-to-end queueing delay and maximises the overall signal quality of all incoming traffic, and furthermore can be implemented in the dynamic and distributed setting of multi-hop wireless networks. Scheduling here encompasses resource allocation decisions such as power assignment for each link which determines its data rate.

Recent research has shown the potential benefits of using network utility maximisation framework in systematic design

S. Jahromizadeh and V. Rakočević are with the School of Engineering and Mathematical Sciences, City University London, London EC1V 0HB, UK (e-mail: s.jahromizadeh@city.ac.uk; v.rakocevic@city.ac.uk)

Manuscript received October 19, 2012; revised July 18, 2013.

of hierarchical (or 'layered') distributed solutions of network resource allocation problems [3]. The benefits of this approach are even more profound in multi-hop wireless network design problems due to interdependencies between resource allocation decisions at different network modular layers, and consequently we also adopt such approach here to address the aforementioned objectives. As will be seen in Section III, the network utility maximisation based approaches to support delay-sensitive traffic have been predominantly based on either reducing link utilisation, or approximation of links as $M/D/1$ queues. The former approach includes methods such as using virtual data rates [4], [5] and minimising network congestion [6], which normally lead to nearly zero queue lengths in the long term due to reduced link utilisation, but provide no control over the transient behaviour of packet delays, as well as window-based flow control methods [7] which provide end-to-end delay guarantees and controlled transient behaviour of packet delay but utilise only less than half of the network capacity. The latter approach which is adopted in [8] and [9] is based on assumptions that contrast with realistic scenarios. Moreover, it also results in under-utilised links under optimal resource allocation. On the contrary, our prime objective here is to regulate average queueing delay with high accuracy, efficiency and performance.

Our proposed solution is based on an alternative formulation to the original optimisation problem, where the delay constraint is omitted and instead the utility function for each source is multiplied by a primarily unknown weight factor. We present a solution to the alternative optimisation problem which comprises a scheduling algorithm incorporating a duality-based rate control algorithm at its inner layer. The key feature of the proposed algorithm is that optimal link prices computed by the rate control algorithm correspond to the average link queueing delays. Based on the analysis of sensitivity of optimal path prices for each source to the variation of its weight factor, we present an alternative strategy where bounded average end-to-end queueing delay is provided through adjustment of sources' weights. In the proposed strategy, we still assume that the perceived signal quality of an incoming traffic is an increasing and strictly concave function of its transmission rate. However, the weight of its utility is adjusted to ensure the desired optimal path prices, and hence the desired average path delays, which can be interpreted as sources understating (overstating) their level of perceived signal quality from the allocated transmission rate in order to obtain their desired optimal path prices, and hence the desired average path delays at the equilibrium. This approach may result in slight reduction in the overall optimal perceived signal quality, due to possible different optimal rate allocation, but ensures the required average end-to-end delay and still leads to efficient utilisation of the network capacity. Given the dynamic and distributed nature of multi-hop wireless networks, the proposed strategy is realised by incorporating an integral controller in the scheduling algorithm whereby each source independently regulates the average queueing delay on its paths at the desired level, using its weight factor as the control variable. The proposed algorithm can be deployed in a distributed setting. Moreover, it is efficient and robust to the changes in the network configuration. The transient behaviour of the average end-to-end queueing delay can be further controlled by appropriate adjustment of the delay regulator parameter. We then study the conditions under which the proposed scheduling policy combined with the proposed integral controller achieve asymptotic regulation of end-to-end queueing delay. Simulation experiments further demonstrate the asymptotic regulation of end-to-end delay with good precision and performance.

The rest of this paper is organised as follows. In Section II the problem is introduced formally as a network utility maximisation problem, and the limitations of the $M/D/1$ approximation of links for delay estimation is discussed. In Section III the predominant network utility maximisation approaches to support delay sensitive traffic as well as their limitations are described. In Section IV-A the proposed alternative optimisation problem, its representation as a scheduling problem and the corresponding solution is presented. In Section V the result of sensitivity of delay to the variation of sources weights is presented followed by the proposed solution for providing bounded delay. Simulation results are given in Section VI, and Section VII concludes the paper.

II. Problem Definition

A. Assumptions and Notations

Throughout the text, vectors are denoted by boldface lowercase letters, and matrices and sets by capital letters. For simplicity, the same notations are used to denote the sets and their cardinality.

This paper considers the problem of rate control and scheduling for simultaneous transmissions of multiple delay-sensitive traffic over a multi-hop wireless network. Let S be the set of sources which generate the delay-sensitive traffic and L be the set of links which constitute the multi-hop wireless network. Each source $s \in S$ has multiple alternative paths to its destination denoted by I_s. For notational simplicity, let L, S and I_s also denote the total number of links, total number of sources, and total number of paths available to source s, respectively. The set of links used by each path $i \in I_s$ are defined by the $L \times 1$ vector R_i^s with elements

$$R_{l,i}^s = \begin{cases} 1 & \text{if path } i \in I_s \text{ uses link } l, \\ 0 & \text{otherwise.} \end{cases}$$

The $L \times I_s$ routing matrix for source s is subsequently defined by $R^s = [R_1^s \dots R_{I_s}^s]$, and the $L \times I$ routing matrix for the network, where $I = \sum_{s \in S} I_s$, by $R = [R^1 \dots R^S]$. We denote the lth row of R by R_l.

Let p_l be the power assignment, or any other resource control decisions such as activation/inactivation, and retransmission probability in random access MAC protocols, and c_l be the data rate at link l. Link data rates are assumed to be a function of global power assignments, i.e. $\boldsymbol{c} = u(\boldsymbol{p})$. Let Π be the set of feasible power assignments and $C = \{u(\boldsymbol{p}), \boldsymbol{p} \in \Pi\}$. The convex hull of C denoted by $\text{Co}(C)$ is assumed to be closed and bounded.

Let x_i^s be the data transmission rate on path $i \in I_s$, and $x_s = \sum_{i \in I_s} x_i^s$ be the aggregate data transmission rate of source s.

We assume that each source $s \in S$ gains a utility $f_s(x_s)$ at rate x_s, where f_s are twice continuously differentiable, strictly concave, and increasing for all $s \in S$. Furthermore $f_s'' < 0$.

We assume that the average delay experienced by a packet at link l is given by $\theta_l(c_l, y_l)$, where $y_l = R_l\boldsymbol{x}$ is the total traffic rate on link l. Furthermore, $\theta_l(c_l, y_l)$ are differentiable, decreasing in c_l and increasing in y_l, for all $l \in L$. Let $\boldsymbol{\theta}$ be the $L \times 1$ vector with elements $\theta_l(c_l, y_l)$, for all $l \in L$. The average end-to-end delay on path $i \in I_s$ of source $s \in S$ is then given by $R_i^{sT}\boldsymbol{\theta}$, and is assumed to be upper bounded by d_s. Let $\boldsymbol{d}$ be an $I \times 1$ vector of upper bounds on average end-to-end delay on each path with elements d_i^s. Since for every source $s \in S$ the average end-to-end delay upper bound is assumed to be an identical value of d_s for all paths $i \in I_s$, we have $d_i^s = d_s$, for all $i \in I_s$.

B. Problem Formulation

The optimisation objective is to find data transmission rates $\boldsymbol{x}$ and link data rates $\boldsymbol{c}$ such that

$$\max_{\boldsymbol{x},\boldsymbol{c}} \quad \sum_{s\in S} f_s(x_s) \tag{1}$$

$$\text{subject to} \quad R\boldsymbol{x} \le \boldsymbol{c} \tag{2}$$

$$\boldsymbol{c} \in \text{Co}(C) \tag{3}$$

$$R^T\boldsymbol{\theta} \le \boldsymbol{d} \tag{4}$$

$$\boldsymbol{x} \ge 0. \tag{5}$$

The optimisation objective (1) maximises the aggregate utility of all sources. Constraint (2) requires that the traffic rate entering each link not to exceed its allocated data rate. Constraint (3) restricts link data rates to the convex hull of feasible link data rates. Constraint (4) imposes an upper bound on the average end-to-end delay faced by a packet on individual paths.

Recall that $\theta(c_l, y_l)$ is increasing in y_l and decreasing in c_l. So the values of $\boldsymbol{d}$ for which the feasible set (2)-(5) is non-empty is given by $\sum_{l\in L} R_{l,i}^s \theta_l(c_l, 0) \le d_s$, $\forall i \in I_s$, and $\forall s \in S$, where $\boldsymbol{c}$ satisfies (3).

C. Limitations of approximation of Links as M/D/1 Queues

As will be discussed in Section III, estimation of the average delay experienced by a packet at a link, i.e. $\theta_l(c_l, y_l)$, $l \in L$, has been predominantly based on approximation of links as independent $M/D/1$ queues. This approach stems from the Kleinrock independence approximation [10], which is in principle based on assumptions that the traffic arrives at network entry points according to a Poisson process, and the network is densely connected. Using this approach the average packet delay $\theta_l(c_l, y_l)$ can be estimated as

$$\begin{aligned}\theta_l(c_l, y_l) &= \frac{1}{c_l} + \frac{y_l}{2c_l(c_l - y_l)} \qquad \forall l \in L \\ &\le \frac{1}{c_l} + \frac{1}{2(c_l - y_l)} \qquad \forall l \in L. \end{aligned} \tag{6}$$

The upper bound (6) is based on assumption that at optimality y_l is close to but not greater than c_l for all $l \in L$. Using the upper bound (6), the optimisation problem (1)-(5) becomes convex, and can be solved, for example, using primal decomposition [3] as follows

$$\max_{\boldsymbol{c}} \tilde{U}(\boldsymbol{c}) \quad \text{subject to} \quad (3) \tag{7}$$

where

$$\tilde{U}(\boldsymbol{c}) = \max_{\boldsymbol{x}} \sum_{s\in S} f_s(x_s) \quad \text{subject to} \quad (2), (4) \text{ and } (5). \tag{8}$$

Subproblem (8) can be solved using primal or dual algorithms proposed in [8]. Moreover, $\tilde{U}$ is concave (Proposition 3.4.3 in [11]), and therefore the set of optimal Lagrange multipliers associated with constraints (2) and (4) in subproblem (8) is the subdifferential of $\tilde{U}$ [11]. This property can then be used to develop algorithms for solving (7), when duality-based approaches are used to solve (8).

However, the $M/D/1$ queue approximation of links has several flaws. Firstly, the key assumptions behind the $M/D/1$ approximation (6) do not hold since the traffic at entry points are regulated by the rate controller and are deterministic, multi-hop networks are composed of mostly disjoint paths which comprise serial links, and the traffic entering the links can be further regulated to limit its burstiness [12], [13]. The delay caused by the burstiness of the arriving traffic at each link can therefore be assumed to be negligible and consequently the average delay a packet experiences at equilibrium is primarily a function of number of packets in the system at equilibrium, which is determined by the dynamics of the rate control and scheduling algorithms at their transient state. Secondly, in the approximation (6), as traffic rates at links approach their capacities, their delays grow exponentially. This implies that at optimality links are not efficiently utilised, in order to ensure bounded delay.

III. Related Work

A. Joint Rate Control and Scheduling for Elastic Traffic

The problem of joint rate control and scheduling for elastic traffic has been extensively studied [14], [15]. Dual optimisation-based approach has been the preferred solution strategy for this problem since it enables decomposition of the problem into the rate control and scheduling 'layers' coupled loosely through 'link prices'. Alternative formulations of this problem lead to different solutions, as described in following sections.

1) Link-Centric Formulation: The optimisation problem (1), (2), (3) and (5), which is the focus of this paper, is based on the link-centric formulation. The dual problem is given by

$$\min_{\boldsymbol{\lambda}\ge 0} D(\boldsymbol{\lambda}) \tag{9}$$

where

$$\begin{aligned} D(\boldsymbol{\lambda}) &= \max_{\boldsymbol{x},\boldsymbol{c}} L(\boldsymbol{x}, \boldsymbol{c}, \boldsymbol{\lambda}) \quad \text{subject to} \quad (3), (5) \\ &= \max_{\boldsymbol{x},\boldsymbol{c}} \sum_{s\in S} f_s(x_s) - \boldsymbol{\lambda}^T(R\boldsymbol{x} - \boldsymbol{c}) \text{ s.t. } (3), (5) \end{aligned} \tag{10}$$

and $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers associated with constraint (2). Using the shadow price interpretation of Lagrange variables [16], $\boldsymbol{\lambda}$ can also be interpreted as the link data

rate prices. The optimisation problem (10) can be decomposed into the following rate control and scheduling subproblems, respectively

$$D_1^s(\boldsymbol{\lambda}) = \max_{\boldsymbol{x}^s \geq 0} f_s(x_s) - \boldsymbol{\lambda}^T R^s \boldsymbol{x}^s \qquad \forall s \in S \qquad (11)$$

and

$$D_2(\boldsymbol{\lambda}) = \max_{\boldsymbol{c} \in u(\boldsymbol{p}), \boldsymbol{p} \in \Pi} \boldsymbol{\lambda}^T \boldsymbol{c} \qquad (12)$$

where x_s is the aggregate data transmission rate of source s, and $\boldsymbol{x}^s$ is $I_s \times 1$ vector of path transmission rates for source s. The dual problem (9) can be solved using the subgradient method as follows

$$\lambda_l(t+1) = \left[\lambda_l(t) + \beta \left(R_l \boldsymbol{x}(\boldsymbol{\lambda}(t)) - c_l(\boldsymbol{\lambda}(t))\right)\right]^+ \forall l \in L \quad (13)$$

where $\beta > 0$, R_l is the lth row of R, and $\boldsymbol{x}(\boldsymbol{\lambda}(t))$ and $\boldsymbol{c}(\boldsymbol{\lambda}(t))$ are the solutions of (11) and (12) given $\boldsymbol{\lambda}(t)$, respectively. The rate control subproblem (11) and the scheduling subproblems (12) are coupled via link prices $\boldsymbol{\lambda}(t)$. In (11), each source $s \in S$ adjusts its path rates $\boldsymbol{x}^s$ according to its path prices. In (12), link data rates $\boldsymbol{c}$ are updated based on the link prices. The link price algorithm (13) and rate control algorithm (11) can be performed in a distributed fashion by individual links and sources, respectively. The scheduling problem (12) is a computationally complex problem in general, since $u(\boldsymbol{p})$ is not concave in many cases and as a result convex programming methods cannot be used. Given the fact that link prices $\boldsymbol{\lambda}(t)$ are updated at every timeslot and therefore (12) has to be solved at every timeslot, finding an efficient, simple and distributed solution becomes crucial. Cases where the scheduling problem (12) is solvable include node-exclusive interference model, low-SINR model, and high-SINR model which are described in detail in [14]. Moreover, alternative suboptimal solutions that are simpler and enable distributed implementation have been developed and their efficiency have been studied for similar cases [14].

2) Node-Centric Formulation: In this formulation, links are denoted by node pairs (i,j), and the data rate on link (i,j) that is allocated for data towards destination d is denoted by c_{ij}^d. Furthermore, the source and destination nodes of sources s are denoted by e_s and d_s, respectively. The transmission rate of source s is denoted by x_s. The constraints (2) and (3) in the optimisation problem are then replaced by

$$\sum_{j:(i,j)\in L} c_{ij}^d - \sum_{j:(j,i)\in L} c_{ji}^d - \sum_{\substack{s:e_s=i,\\ d_s=d}} x_s \geq 0 \; \forall d, \; \forall i \neq d \quad (14)$$

$$\left[\sum_d c_{ij}^d\right] \in \mathrm{Co}(C) \qquad (15)$$

$$c_{ij}^d \geq 0 \; \forall (i,j) \in L, \; \forall d.$$

Consequently, dual decomposition results in the following rate control and scheduling subproblems, respectively

$$x_s = \arg\max \left(f_s(x_s) - x_s \lambda_{f_s}^{d_s}\right) \qquad \forall s \in S \qquad (16)$$

and

$$\boldsymbol{c} = \arg\max\nolimits_{\boldsymbol{c} \in u(\boldsymbol{p}), \boldsymbol{p} \in \Pi} \sum_{(i,j)\in L} c_{ij} \max_d \left(\lambda_i^d - \lambda_j^d\right) \qquad (17)$$

where λ_i^d is the Lagrange multiplier associated with constraint (d,i) in (14). After solving (17), for each link $(i,j) \in L$, $c_{ij}^d = c_{ij}$, if $d = \arg\max_{\hat{d}}(\lambda_i^{\hat{d}} - \lambda_j^{\hat{d}})$, and $c_{ij}^d = 0$, otherwise. The Lagrange multipliers are updated using the subgradient method by individual nodes as described in [14].

Compared with the link-centric formulation, the scheduling problem (17), which is referred to as maximum weight back-pressure algorithm, has the same form as (12) but it also incorporates routing decisions. Furthermore, the rate control problem (16) for each source $s \in S$ is only dependent on the price $\lambda_{f_s}^{d_s}$ at the source node, whereas the rate control problem (11) for each source $s \in S$ is dependent on the sum of the link prices along its paths q_i^s, $i \in I_s$.

B. Network Utility Maximisation Approaches to Support Delay Sensitive Traffic

1) Minimising Delay Using Virtual Data Rates: Since the algorithm (13) couples the link prices to their average queue lengths, it may lead to large queue lengths and hence large queueing delays at the equilibrium. As suggested in [4], [5], this can be avoided by using the slightly smaller 'virtual' link data rates in (13) instead of the actual link data rates. Specifically, $c_l(\boldsymbol{\lambda}(t))$, $l \in L$ in (13) is replaced by $\rho c_l(\boldsymbol{\lambda}(t))$, where ρ is a positive factor slightly smaller than 1. While the modified algorithm still leads to the link prices close to their optimal level, it results in zero equilibrium queue lengths, since links traffic loads are slightly less than their actual data rates at equilibrium. Main disadvantages of this approach are that it does not completely utilise network capacity and provides no control over the transient behaviour of packet delays.

2) Guaranteeing Bounded Delay Using Window-Based Flow Control: In [7], the problem of designing a joint rate control and scheduling algorithm that provide provable throughput and provable per-flow delay is considered. The node-centric formulation (1), (14), (15) and (5) is considered where the queue-length based back-pressure algorithm (17) or the developed low-complexity algorithms described in Section III-A have been shown to have poor delay performance under certain cases, and have difficult to quantify or control delay. A new distributed and low-complexity congestion control and scheduling algorithm is proposed where the packet transmissions are scheduled by a rate-based rather than a queue-length based scheduling algorithm, and congestion control is based on window flow control, which deterministically bounds the end-to-end delay backlog within the network. It is shown that by appropriately choosing the number of backoff mini-slots for the scheduling algorithm and the window size of each flow, the proposed algorithm can utilise close to half of the systems capacity under the one-hop interference constraint, and guarantee a per-flow expected delay upper bound that increases linearly with the number of hops. Furthermore, each flow's trade-off between throughput and delay can be individually controlled by the window size.

Similar to the approaches based on $M/D/1$ queue approximation and virtual data rates, the main limitation of this approach is that bounded end-to-end delay can only be guaranteed by under-utilisation of system capacity, i.e.

less than half under the one-hop interference constraint. In addition, there is a trade-off between the guaranteed delay bounds and throughput levels within this reduced capacity region. The guaranteed delay bound for each flow is also a factor of number of hops.

Our approach is similar to this approach in that it is too based on controlling the end-to-end backlog to control the delay both at equilibrium and the transient state. In our approach this is accomplished by correlating the link prices in a duality-based rate control algorithm similar to (13) with the average link queueing delays, and then using the utility weight coefficient of sources as control variables to asymptotically regulate path prices at the desired level. The transient behaviour of the average end-to-end queueing delay can be further controlled by appropriate adjustment of the delay regulator parameter. However, our approach differs from the approach presented in [7] in that it can guarantee any bound on the average end-to-end queueing delay, as long as R has full column rank, while ensuring maximal utilisation of the system available capacity. Hence, in our approach only packet end-to-end transmission time is lower bounded by a factor of number of hops, which is negligible when the equilibrium end-to-end backlog is modest. Rather than trade-off between delay and throughput, in our approach the average end-to-end queueing delay bounds are guaranteed by the appropriate choice of utility weight coefficient of sources that ensure the desired optimal path prices. This approach may result in the same or slightly different optimal rate allocation, but still leads to maximal utilisation of the network capacity.

3) Minimising Network Congestion: In [6] the primary objective is to find a joint link data rate and flow assignment strategy that supports maximum data rates and yields minimum end-to-end delay; however, for general queueing systems, this leads to an intractable problem formulation. Hence, an alternative problem formulation is proposed where the network congestion, defined as maximum link utilisation over all links

$$\Delta(\boldsymbol{c}, \boldsymbol{y}) = \max_{l \in L} \frac{y_l}{c_l} \tag{18}$$

is minimised while allowing communication between source and destinations at a given data rate. Since the objective function is quasi-convex, the resulting optimisation problem can be solved, for example, by a bisection algorithm that involves solving a sequence of convex feasibility problems. Evidently, this approach neither aims to efficiently utilise links nor can guarantee bounded delay.

4) Maximising Utility as a Function of Rate and Delay: In [8] the congestion control problem in networks supporting traffic with various levels of rate, delay and packet loss sensitivity, is studied. The proposed approach is based on incorporating the requirements for rate, delay and packet loss in the utility function of sources. It is assumed that each source transmits only one flow using a fixed path, and that link data rates $\boldsymbol{c}$ are fixed. The utility of each source $s \in S$ is subsequently defined by

$$U_s = a_s f_s(x_s) - b_s \sum_{l \in L} R_l^s \theta_l(y_l)$$

where coefficients a_s and b_s indicate the degree of sensitivity of the traffic to rate and delay, respectively. It is shown that similar to the basic congestion control problem for elastic traffic, the alternative optimisation problem can be solved using both primal and dual algorithms. The analysis is then applied to networks with mixed voice and data traffic, including the case where priority queueing is used. It is shown using simulation that priority queueing improves both the R-factor of voice traffic and the throughput of data traffic, at the expense of the packet delay of data traffic. However, the estimation of packet delay is based on approximation of links as independent $M/D/1$ queues which, as discussed in Section II-B, stems from unrealistic assumptions and results in under-utilised links under optimal resource allocation.

In [9] the congestion control problem for networks where traffic sources are heterogeneous with respect to their levels of sensitivity to both rate and delay is considered. It is assumed that source $s \in S$ incurs a delay cost $h_s d$ per unit of flow rate, where d is the average end-to-end delay experienced by a packet. The utility of each source $s \in S$ is subsequently defined by

$$U_s = f_s(x_s) - h_s x_s \sum_{l \in L} R_l^s \theta_l(y_l). \tag{19}$$

The resulting optimisation problem is shown to be non-concave in general and consequently may have several stationary points. Several variants of a primal rate control algorithm are shown to converge to a local maximum, but never to a saddle point. It is concluded that dynamic rate control algorithms such as TCP may not be able to attain efficient rate allocations and levels of delay that are acceptable to diverse classes of traffic, in the absence of differentiated services. In this paper we assume that the average packet end-to-end delay on each path is upper bounded in which case, as explained in Section II-B, the problem can be formulated as convex optimisation problem (1)-(5), given approximation (6).

This paper extends the ideas presented in our previous papers [17], [18]. In both papers we exploit the correlation between optimal link prices and equilibrium link average queueing delays in duality-based rate control algorithm, in order to provide bounded average end-to-end queueing delay. In [17] we first present an approach in which lower bounds on sources' transmission rates are derived in order to ensure the required bounded delay. This approach inevitably entails admission control. In the second approach, we introduce an alternative formulation where the delay constraint is omitted and instead the utility function for each source is multiplied by a weight factor. The proposed solution comprises a scheduling algorithm incorporating a duality-based rate control algorithm at its inner layer, and an algorithm that dynamically adjusts sources' weights to ensure the required bounded delay. In this paper we further develop the latter approach by designing a new scheduling algorithm and delay regulator that regulate average queueing delay with high accuracy and performance. Moreover, we provide a complete analysis of the stability of the proposed algorithms. In [18] we analyse the sensitivity of optimal path prices for each source to the variation of its weight factor, and present a delay regulator that is integrated

into the duality-based rate control and scheduling algorithms given in [14]. In this paper we use the sensitivity analysis results in [18] to develop a solution that regulates average queueing delay with higher accuracy and performance.

IV. Alternative Problem Formulation

The proposed solution for providing bounded delay is based on an alternative formulation for optimisation problem (1)-(5) which is presented in this section and its properties are examined.

A. The Alternative Optimisation Problem

The proposed alternative optimisation problem has the following form

$$\max_{\boldsymbol{x},\boldsymbol{c}} \sum_{s\in S} w_s f_s(x_s) \qquad \text{subject to} \quad (2),(3),(5) \tag{20}$$

where $w_s f_s(x_s)$ represents utility of source s, or preference over transmission rate x_s. Compared with original problem (1)-(5), in the alternative problem (20), delay constraint (4) has been omitted and instead the utility function for each source $s \in S$ is multiplied by the weight parameter w_s. A geometric interpretation of w_s is that higher (respectively, lower) values of w_s result in higher (respectively, lower) marginal increase in preference or utility of source s at a particular rate.

The dual problem for (20) is given by

$$\min_{\boldsymbol{\lambda}\geq 0,\boldsymbol{\mu}\geq 0} D(\boldsymbol{\lambda},\boldsymbol{\mu}) \tag{21}$$

where

$$\begin{aligned} D(\boldsymbol{\lambda},\boldsymbol{\mu}) &= \max_{\boldsymbol{x},\boldsymbol{c}} L(\boldsymbol{x},\boldsymbol{c},\boldsymbol{\lambda},\boldsymbol{\mu}) \qquad \text{subject to} \quad (3) \\ &= \max_{\boldsymbol{x},\boldsymbol{c}} \sum_{s\in S} w_s f_s(x_s) - \boldsymbol{\lambda}^T(R\boldsymbol{x}-\boldsymbol{c}) + \boldsymbol{\mu}^T\boldsymbol{x} \\ &\quad \text{subject to} \quad (3) \end{aligned} \tag{22}$$

and $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are the Lagrange multipliers associated with constraints (2) and (5), respectively. Using the shadow price interpretation of Lagrange variables [16], $\boldsymbol{\lambda}$ can also be interpreted as the link data rate prices.

The optimisation problem (20) is convex and constraints (2) and (5) are affine. Moreover, since C is a finite set, $\mathrm{Co}(C)$ is a polyhedral set which can be expressed by a set of affine inequalities and equalities. Thus, Slater's condition reduces to feasibility [16] and the optimal duality gap is zero. Let $(\boldsymbol{x}^*,\boldsymbol{c}^*)$ and $(\boldsymbol{\lambda}^*,\boldsymbol{\mu}^*)$ be the primal and dual optimal solutions, respectively. Let also $\boldsymbol{q}^* = R^T\boldsymbol{\lambda}^*$. It then follows from Karush-Kuhn-Tucker (KKT) optimality conditions [16] that

$$w_s f_s'(x_s^*) - q_i^{s*} + \mu_i^{s*} = 0 \qquad \forall i \in I_s,\ \forall s \in S \tag{23}$$

$$\lambda_l^*(R_l\boldsymbol{x}^* - c_l^*) = 0 \qquad \forall l \in L \tag{24}$$

$$\mu_i^{s*} x_i^{s*} = 0 \qquad \forall i \in I_s,\ \forall s \in S. \tag{25}$$

Equation (25) implies that $\mu_i^{s*} = 0$ for any $i \in I_s$ for which $x_i^{s*} > 0$. It then follows from (23) that

$$\begin{aligned} q_i^{s*} &= w_s f_s'(x_s^*) \qquad \forall i \in I_s,\ x_i^{s*} > 0,\ \forall s \in S \\ &\triangleq q_s^* \end{aligned} \tag{26}$$

which means that for each source $s \in S$ the values of q_i^{s*} associated with paths with positive flows are minimum and hence equal. Since the objective function in (20) is strictly concave with respect to $\{x_s\}$, $\{x_s^*\}$ is unique and it follows from (26) that $\boldsymbol{q}^*$ is also unique. However, (20) is not strictly concave in either $\boldsymbol{c}$, or $\boldsymbol{x}$ in general, since every $x_s = \sum_{i\in I_s} x_i^s$, $s \in S$ is a hyperplane of $\boldsymbol{x}$ for which $f_s(x_s)$ is identical. Hence neither $\boldsymbol{c}^*$ nor $\boldsymbol{x}^*$ may be unique. Furthermore, given that $\boldsymbol{q}^* = R^T\boldsymbol{\lambda}^*$, and R may have linearly dependent rows, $\boldsymbol{\lambda}^*$ may not be unique in general.

B. Representation as a Scheduling Problem

Optimisation problem (20) can be alternatively presented as the following equivalent form

$$\max_{\boldsymbol{c}} U_{\boldsymbol{w}}(\boldsymbol{c}) \qquad \text{subject to} \quad (3) \tag{27}$$

where

$$U_{\boldsymbol{w}}(\boldsymbol{c}) = \max_{\boldsymbol{x}} \sum_{s\in S} w_s f_s(x_s) \qquad \text{subject to} \quad (2),(5). \tag{28}$$

The key feature of the alternative form (27) is the decomposition of the problem into master scheduling problem (27), and the well-known rate control subproblem (28) with fixed link data rates. In addition, $U_{\boldsymbol{w}}$ is concave by Proposition 3.4.3 in [11], and therefore, as shown in Section 5.4.4 in [11], the set of optimal Lagrange multipliers associated with constraint (2) in subproblem (28) is the subdifferential of $U_{\boldsymbol{w}}$.

The dual problem for (28) is similar to (21) and (22) with fixed link data rates $\boldsymbol{c}$. Consequently, KKT conditions (23)-(25) as well as (26) also hold for problem (28). Let $\boldsymbol{x}(\boldsymbol{c})$ and $(\boldsymbol{\lambda}(\boldsymbol{c}),\boldsymbol{\mu}(\boldsymbol{c}))$ be the primal and dual optimal solutions of (28) given $\boldsymbol{c}$, respectively. Let also $\boldsymbol{q}(\boldsymbol{c}) = R^T\boldsymbol{\lambda}(\boldsymbol{c})$. Since the objective function in (20) is strictly concave with respect to $\{x_s\}$, $\{x_s(\boldsymbol{c})\}$ is unique and it follows from (26) that $\boldsymbol{q}(\boldsymbol{c})$ is also unique. However, as in the case of problem (20), $\boldsymbol{\lambda}(\boldsymbol{c})$ is not unique in general. Hence $\boldsymbol{\lambda}(\boldsymbol{c}) \in \Lambda(\boldsymbol{c})$, where $\Lambda(\boldsymbol{c})$ is the set of optimal Lagrange multipliers associated with constraint (2).

1) Solution of the Multipath Rate Control Subproblem: Rate control problem (28) has been extensively studied in the literature [19]. Here, we consider the duality-based solutions where Lagrange variables are updated according to

$$\dot{\lambda}_l = \frac{\beta}{c_l}[R_l\boldsymbol{x}(\boldsymbol{\lambda}) - c_l]^+_{\lambda_l} \qquad \forall l \in L \tag{29}$$

where $\beta > 0$, $\boldsymbol{x}(\boldsymbol{\lambda}) = \arg\max_{\boldsymbol{x}\geq 0} \sum_{s\in S} w_s f_s(x_s) - \boldsymbol{\lambda}^T R\boldsymbol{x}$, and $[g(x)]^+_x$ is defined by

$$[g(x)]^+_x = \begin{cases} g(x) & x > 0 \\ \max(g(x),0) & x = 0. \end{cases}$$

As discussed in [20], path rates of sources with multiple paths in (29) continuously oscillate and do not converge, since $f_s(\sum_{i\in I_s} x_i^s)$ is not strictly concave. To circumvent this problem Proximal Optimisation Algorithms [21] or the distributed algorithm proposed in [20] which is suitable for online implementation can be used. Since the objective function in (28) is not strictly concave in $\boldsymbol{x}$, by Proposition 6.1.1 in [11],

the dual function of (28) may not be differentiable at every point. Moreover, as shown in Section 6.1 in [11], the term within the brackets in (29) is a subgradient of the dual function and thus the term on right side of (29) is discontinuous.

The right-hand side of algorithm (29) corresponds to the β multiple of marginal increase in average queueing delay at link l, given the traffic rate entering link l is equal to $R_l\boldsymbol{x}(\boldsymbol{\lambda})$. While this condition holds at links at the network traffic entry points, the traffic rate at the other links are bounded by the data rates of the links connected to their source node. However, at the equilibrium, the right-hand side of (29) is the β multiple of marginal increase in average queueing delay for all links $l \in L$. This implies that, by the results from stability of systems with vanishing perturbation [22], if link prices are updated according to β multiple of the link average queueing delays (i.e. $\lambda_l(t) = \beta\frac{N_l(t)}{c_l}$, $l \in L$, where $N_l(t)$ is the number of packets in link l at time t), path rates $\boldsymbol{x}(\boldsymbol{\lambda})$ and link prices $\boldsymbol{\lambda}$ converge to the primal and dual optimal solutions of (28), respectively, given β is sufficiently small. In this case, link average queueing delays at equilibrium are equal to $\beta^{-1}\boldsymbol{\lambda}(\boldsymbol{c})$.

2) Solution of the Scheduling Problem: Using ideas from the gradient optimisation methods [11], we propose the following solution for scheduling problem (27)

$$\dot{\boldsymbol{c}} = \gamma(\tilde{\boldsymbol{c}} - \boldsymbol{c}) \tag{30}$$

where $\gamma > 0$, and

$$\tilde{\boldsymbol{c}} = \begin{cases} \boldsymbol{c} & \boldsymbol{c} = \arg\max_{\boldsymbol{\varsigma}\in \mathrm{Co}(C)} \boldsymbol{\lambda}(\boldsymbol{c})^T \boldsymbol{\varsigma} \\ \arg\max_{\boldsymbol{\varsigma}\in C} \boldsymbol{\lambda}(\boldsymbol{c})^T \boldsymbol{\varsigma} & \text{otherwise} \end{cases} \tag{31}$$

where $\boldsymbol{\lambda}(\boldsymbol{c})$ is the optimal Lagrange variable of (28) given $\boldsymbol{c}$. Since C is a finite set, $\mathrm{Co}(C)$ is a polyhedral set and hence by Proposition B.21 in [11], the optimisation problem in (31) attains a maximum at some extreme point of $\mathrm{Co}(C)$. Therefore, the solution space in (31) is reduced to C. At each step of (30), $\tilde{\boldsymbol{c}}$ is computed as follows. The optimisation problem in (31) is computed over the set C. If the solution results in the same objective value as the current link data rates $\boldsymbol{c}$, $\tilde{\boldsymbol{c}}$ equals $\boldsymbol{c}$ and (30) stops. Otherwise, $\tilde{\boldsymbol{c}}$ takes the value of the any of the solutions and (30) continues. The optimisation problem in (31) is of the same form as the well-known scheduling problem (12) and thus can be solved using distributed algorithms in some cases, as discussed in Section III-A. Efficient implementation of the equilibrium condition in (31) also depends on these solution algorithms.

The right-hand side of (30) may not be continuous in general, for example, when $\boldsymbol{\lambda}(\boldsymbol{c})$ is not unique, or when in (31) strict complimentary slackness condition does not hold at $\tilde{\boldsymbol{c}}$ (Theorem 3.2.2 in [23]), and as a result the existence of solutions is not guaranteed. In the analysis that follows we assume that the assumptions H1 (existence of solutions of (30)-(31)), and H2 (right-hand side of (30) is Lebesgue measurable and locally bounded) in [24] hold.

By Theorem 2.2.6 in [23], the mapping $\boldsymbol{q}(\boldsymbol{c})$ is continuous, and since $\boldsymbol{q}(\boldsymbol{c})$ is also unique, it is a continuous function. We assume that $\boldsymbol{q}(\boldsymbol{c})$ is also nonpathological [24].

Theorem 1: Algorithms (30)-(31) converge to an optimal solution of (27).

The proof is given in the Appendix.

V. Proposed Solution for Providing Bounded Delay

A. The Impact of Sources' Weights on Delay

As explained in Section IV-B1, if link prices in the duality-based algorithm (29) are instead updated proportionally to link average queueing delays, optimal link prices are then proportional to the equilibrium link average queueing delays. Furthermore, by (26) optimal path prices for each source $s \in S$ are equal to its marginal utility at its optimal aggregate data transmission rate, multiplied by its weight. This suggests an alternative approach to the original formulation (1)-(5), in which the delay bounds in (4) are instead guaranteed by adjusting the weight of sources.

The following lemma shows that for the alternative problem (20) the optimal path prices for each source $s \in S$, that is $q_s^* = R_i^{sT}\boldsymbol{\lambda}^*$, $i \in I_s$, grow as its weight w_s increases.

Lemma 1: Under the assumptions in Section II-A, upper and lower bounds on the sensitivity of $q_s^*(w)$ and $x_s^*(w)$ to the variation of parameters w_s are given by

$$0 < \frac{\partial q_s^*}{\partial w_s} \leq f_s'(x_s^*) \tag{32}$$

$$0 \leq \frac{\partial x_s^*}{\partial w_s} < -\frac{f_s'(x_s^*)}{w_s f_s''(x_s^*)} \tag{33}$$

for all $s \in S$.

The proof is given in the Appendix. As explained in Section IV-B1, if link prices in the duality-based rate control algorithm are updated according to β multiple of the link average queueing delays, link average queueing delays at equilibrium are equal to $\beta^{-1}\boldsymbol{\lambda}(\boldsymbol{c})$, and as a result path average queueing delays at equilibrium equal $\beta^{-1}\boldsymbol{q}(\boldsymbol{c})$. Consequently, Lemma 1 suggests an alternative strategy for providing bounded end-to-end delay based on the adjustment of sources' weights. Specifically, in the proposed strategy, we still assume that the perceived signal quality of an incoming traffic is an increasing and strictly concave function of its transmission rate. However, the weight of its utility is adjusted to ensure the desired optimal path prices, and hence the desired average path delays. This adjustment of the weights of the utilities can be interpreted as sources understating (overstating) their level of satisfaction, or perceived signal quality, from the allocated transmission rate in order to obtain their desired optimal path prices, and hence the desired average path delays at the equilibrium. This approach may result in the same or slightly different optimal rate allocation compared with the case with the initial source weights. Precisely, by (33), the optimal transmission rate of each source either remains unchanged or increases as its weight increases. In the latter case, the optimal transmission rate of some other sources decreases. Given the uniqueness of optimal source transmission rates, optimal source transmission rates for the perturbed objective function in this case is not optimal for the original problem. Consequently, this approach may lead to slight reduction in the overall perceived signal quality, but ensures the required optimal path prices and still leads to maximal utilisation of the network capacity.

B. Delay Regulation via Dynamic Adjustment of Sources' Weights

The main challenge in guaranteeing bounded delay through adjustment of sources' weights is that sources' weights that guarantee the required bounded delay generally vary for different network configurations. Clearly, a concave utility function always implies that a connection has elastic bandwidth requirements and as a result best-effort strategies may allocate different rates to the same connection (with the same weight) under various network configurations, in order to maximise the aggregate utility of all connections. This means that for every network configuration, sources' weights have to be recomputed to ensure the required bounded delay. Hence, the dynamic and decentralised nature of multihop wireless networks calls for a robust, responsive and distributed algorithm that can adjust sources' weights so as to ensure bounded end-to-end delay under modest parameter perturbations.

In order to design an algorithm that meets these requirements, recall that (32) implies that optimal path price q_s^* for each source $s \in S$ increase with its weight w_s. This suggests that for each source $s \in S$, utility weight w_s can be used as control variable to regulate its average path queueing delay. Thus, the following integral controller is proposed whereby each source adjust its weight w_s in proportion to the average end-to-end delay tracking error to regulate average end-to-end delay at the desired level

$$\dot{w}_s = \alpha \left[d_s - \frac{q_s(\boldsymbol{c}, \boldsymbol{w})}{\beta} \right]_{w_s}^{+} \qquad \forall s \in S. \tag{34}$$

Algorithm (34) is performed by each source independently and ensures bounded end-to-end delay under parameter perturbations that do not destabilise the system. Incorporating (34) in the scheduling algorithms (30)-(31) leads to continuous perturbation of utility weights $\boldsymbol{w}$ and thus the objective function in (27), which tends to zero as the algorithms approach the equilibrium. In this case, $\boldsymbol{\lambda}(\boldsymbol{c})$ can be interpreted as the perturbed supergradient of the new objective function $U_{\tilde{\boldsymbol{w}}}$ at $\boldsymbol{c}$, where $\tilde{w}$ is the perturbed utility weights. Intuitively, as long as the perturbation rate of the objective function in (27) is small compared with speed at which link data rates approach the optimal solution of the current (27), link data rates in algorithms (30)-(31) and (34) track the changing optimal solution of (27). The following theorem examines the conditions under which algorithms (30)-(31) combined with (34) achieve asymptotic regulation of end-to-end delay.

Theorem 2: Given R has full column rank and α in (34) is sufficiently small, algorithms (30)-(31) combined with (34) converge to an optimal solution of (27) with parameter $\boldsymbol{w}^*$, where $\boldsymbol{w}^*$ is the weight of sources that guarantees bounded delay specified in (4), if subproblem (28) is solved using duality-based algorithm (29), where link prices are instead updated as β multiple of link average queueing delays.

The proof is provided in the Appendix. Note that the full column rank condition for R ensures that (30)-(31) combined with (34) has an equilibrium for any $\boldsymbol{d}$. This condition is not very restrictive and holds for typical scenarios, such as the case when source-destination pairs are distinct.

VI. Simulation Results

The objective of the simulation experiments is twofold. Firstly, to illustrate that algorithms (30)-(31) converge to an optimal solution of (27), where link prices $\boldsymbol{\lambda}(\boldsymbol{c})$ are updated according to β multiple of link average queueing delays over a finite time. Secondly, to examine the accuracy of algorithms (30)-(31) combined with (34) in regulating packet end-to-end latency at the desired level, and to compare their performance against the commonly used virtual data rate approach described in Section III-A.

A. Network Model

For simulation experiments we consider the network topology in Fig.1, where there are two source-destination pairs $A \to C$ and $D \to E$. For source-destination pair $A \to C$, there are two alternative paths $A \to B \to D \to C$ and $A \to D \to C$. For source-destination pair $D \to E$, there is only a single path $D \to C \to E$.

Each active link is assumed to have a fixed data rate of c_0 packets per second. To model the scheduling constraint (3), we use the notions of contention graph and contention matrix [15]. In the contention graph, vertices represent links and edges represent the contention between the links. Maximal cliques of the contention graph embody the local contention among links; Links that belong to the same maximal clique cannot be active simultaneously. Let N be the number maximal cliques in the contention graph. The $N \times L$ contention matrix F is then defined by

$$F_{n,l} = \begin{cases} \frac{1}{c_0} & \text{if link } l \in L \text{ belongs to the maximal clique } n \\ 0 & \text{otherwise.} \end{cases}$$

Thus, a necessary condition for scheduling is given by

$$F\boldsymbol{c} \leq \mathbf{1}. \tag{35}$$

It can be shown that (35) is also a sufficient condition for scheduling if the contention graph is perfect [15].

We assume that each wireless node can only communicate with one other node at any time. This results in the contention graph shown in Fig.2. There are three maximal cliques: links (1,2,3), links (2,3,4), and links (4,5). Thus, on path $D \to C \to E$ only one link, either link 4 or 5 can be active. Similarly, on path $A \to D \to C$ only one link, either link 3 or 4 can be active. However, on path $A \to B \to D \to C$ links 1 and 4 can be active simultaneously. This provides incentives for using path $A \to B \to D \to C$, despite being longer, to increase the transmission rate of $A \to C$, since link 1 can be active while data is being transmitted on link 4 for any of the three flows. Since the contention graph in Fig.2 has no odd holes, it is perfect and therefore (35) is a sufficient scheduling constraint in this case.

The utility functions for both sources are assumed to be of the form $w_s f_s(x_s) = w_s \ln(x_s)$.

B. Experimental Results

We use SimEvents discrete-event simulation software for simulation experiments. As discussed in Section IV-B1, path

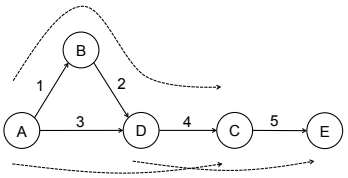

Fig. 1. Network topology and alternative paths for source-destination pairs $A \rightarrow C$ and $D \rightarrow E$

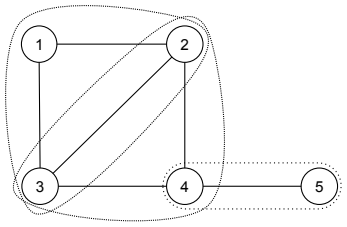

Fig. 2. Network contention graph and its maximal cliques

rates in (29) can be computed using Proximal Optimisation Algorithms [21] or the distributed algorithm proposed in [20]. Both of these algorithms are based on solving an equivalent optimisation problem where the objective function has an additional quadratic function of the difference between the auxiliary variables and the path rates. The equivalent problem is then solved using a two-level iterative algorithm, where at the outer level the auxiliary variables are updated, and at the inner level a strictly concave optimisation problem is solved given the current value of the auxiliary variables. At every step of the outer level, the auxiliary variable values become closer to the solution of the inner level optimisation problem and thus the added quadratic term becomes smaller and eventually converges to zero. Here, for simplicity we add a small quadratic term $-\delta \boldsymbol{x}^T \boldsymbol{x}$ in (22) in order to make it strictly concave. When δ is sufficiently small, the modified objective function remains increasing over the feasible rate region (2) and (5) for all $\boldsymbol{c}$ that satisfy (3). Moreover, the computed primal solutions are close to the those computed using Proximal Optimisation based algorithms when they run for finite time, since the added quadratic term in these algorithms does not completely vanish. Source delay regulation is based on algorithm (34), using current average end-to-end queueing delays as feedback. In all experiments it is assumed that $c_0 = 1$ packet/msec, $\gamma = 1 \times 10^{-2}$, $\beta = 1 \times 10^{-3}$, and $\delta = 3 \times 10^{-1}$.

In the first experiment, source weights are fixed at $w_1 = 2$ and $w_2 = 1$ and algorithms (30)-(31) are simulated where link prices $\boldsymbol{\lambda}(t)$ are updated for a fixed time period according to β multiple of link average queueing delays, i.e. $\lambda_l(t) = \beta \frac{N_l(t)}{c_l}$, $l \in L$, where $N_l(t)$ is the number of packets in link l at time t. The evolution of link data rates, path prices and path transmission rates are shown in Fig.3, Fig.4, and Fig.5, respectively. As seen in Fig.3, link data rates converge rapidly to the neighbourhood of their optimal values $(0.3333, 0.1496, 0.1838, 0.6665, 0.3333)$, although continue to oscillate slightly due to the fact that the rate control algorithm runs for finite time and thus path prices do not fully converge to their optimal values $\boldsymbol{\lambda}(\boldsymbol{c})$. Moreover, link data rates have a non-differentiable curve due to discontinuous nature of the right-hand side of (30).

The oscillation of link data rates also results in modest oscillation of path prices, and thus source aggregate rates, around their optimal levels $(5.9976, 5.9976, 3.0024)$ and $(0.3335, 0.33331)$, as seen in Fig.4, and Fig.5, respectively. Also, as seen in Fig.5, although path prices for source 1 must be equal (see Section IV-A) at every step of (30)-(31), they diverge at some points since the values of $\boldsymbol{\lambda}(\boldsymbol{c})$ are approximate. Specifically, when path rates are computed in a duality-based algorithm (see Section IV-B1), only paths with minimum prices have positive rates. Thus, the slight difference in path prices at some points caused by the approximate values of $\boldsymbol{\lambda}(\boldsymbol{c})$ results in significant oscillation of path rates of source 1, around the their optimal levels $(0.1496, 0.1838)$. The added quadratic term in this example, however, somewhat limits the extent of the divergence of path rates.

In the second experiment, algorithms (30)-(31), with the same setup as the first experiment, are simulated jointly with (34) with delay bounds $d_1 = d_2 = 1 \times 10^3$ milliseconds, and $\alpha = 2 \times 10^{-5}$. Moreover, algorithms (30)-(31) are simulated where link prices $\boldsymbol{\lambda}(t)$ are updated for a fixed time period according to (29), using virtual data rates with $\rho = 0.98$ to retain high link utilisation. The evolution of instantaneous packet end-to-end delays in both first and second experiments are compared in Fig.6. It can be seen that algorithms (30)-(31) combined with (34) with proper choice of parameter α regulate packet end-to-end delay at their upper bound levels with good precision in a relatively short time. The slight oscillation of delay is caused by the oscillations of link data rates and link prices in algorithms (30)-(31) described previously. The transient behaviour of packet end-to-end delay can be further controlled by adjustment of parameter α. Since their equilibrium is the solution to the optimisation problem (27) with equilibrium weights $\boldsymbol{w}^* \approx (0.37, 0.33)$ (Fig.7), they also lead to maximal utilisation of network capacity.

The approach based on virtual data rates reduces the delay to near zero in the long term, since it leads to under-utilised links at equilibrium, however, the rate at which it reduces the delay in inversely correlated to the level of link utilisation at its equilibrium. Thus, as seen in Fig.6, this approach can only achieve high link utilisation at the expense of slow reduction of end-to-end delays and unpredictable transient behaviour.

VII. Conclusion

We consider the problem of supporting traffic with elastic bandwidth requirements and average end-to-end delay constraints in multi-hop wireless networks, and present an approach where the utility weight of each elastic traffic is adjusted to ensure the desired average path delays, as well as maximal utilisation of network capacity and hence high level of overall perceived signal quality. The proposed solution comprises a scheduling algorithm incorporating a duality-based rate control algorithm at its inner layer, and an integral

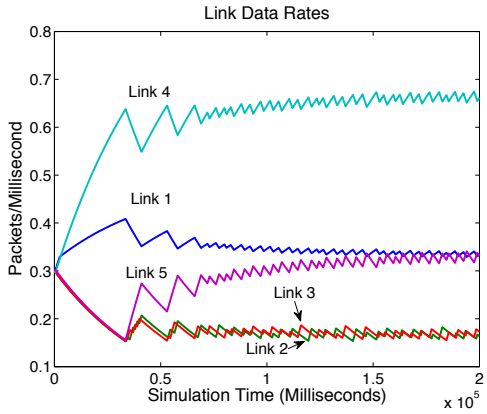

Fig. 3. Link data rates when source weights are fixed at $w_1 = 2$ and $w_2 = 1$ and algorithms (30)-(31) are simulated

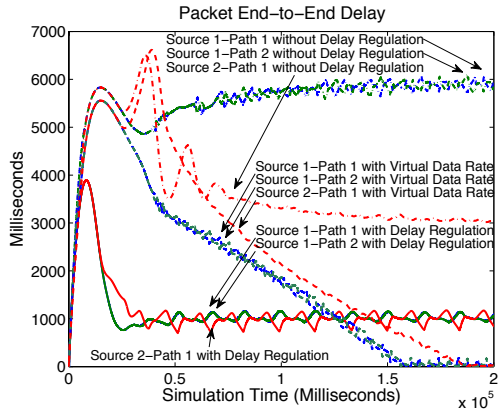

Fig. 6. Instantaneous packet end-to-end delay when algorithms (30)-(31) are simulated jointly with (34) with delay bounds $d_1 = d_2 = 1 \times 10^3$ msec

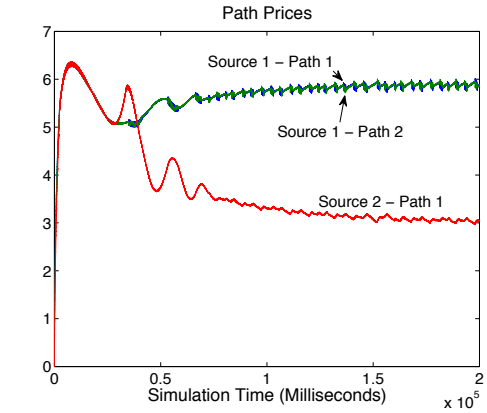

Fig. 4. Path prices when source weights are fixed at $w_1 = 2$ and $w_2 = 1$ and algorithms (30)-(31) are simulated

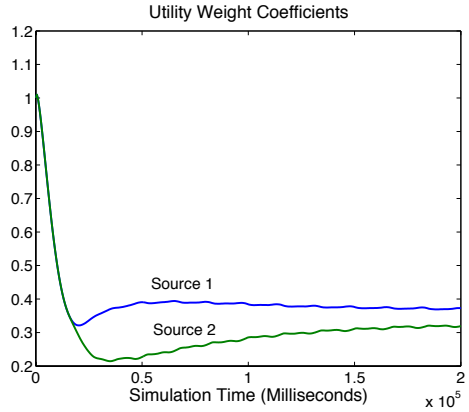

Fig. 7. Evolution of utility weight coefficients when algorithms (30)-(31) are simulated jointly with (34) with delay bounds $d_1 = d_2 = 1 \times 10^3$ msec

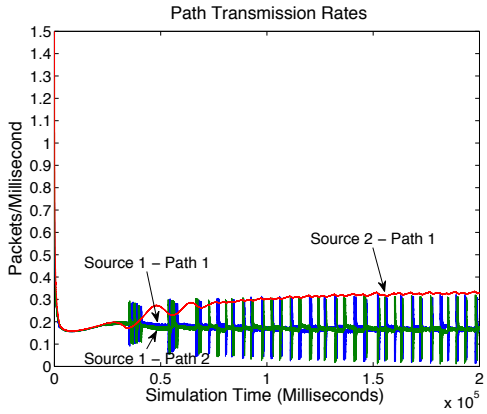

Fig. 5. Path transmission rates when source weights are fixed at $w_1 = 2$ and $w_2 = 1$ and algorithms (30)-(31) are simulated

controller whereby each source regulates average end-to-end queueing delay by using its utility weight as the control variable. Simulation experiments indicate that when the inner-layer rate control algorithm run over a finite time period, with proper choice of parameter α, the proposed algorithms regulate delay with good precision and transient performance compared with the commonly used virtual data rates approach, which under-utilises link capacities to reduce packet delay.

APPENDIX

Proof of Theorem 1: Let $\hat{\boldsymbol{c}}$ be an equilibrium point of (30)-(31). From (31) it follows that

$$\hat{\boldsymbol{c}} = \arg \max_{\boldsymbol{\varsigma} \in \mathrm{Co}(C)} \boldsymbol{\lambda}(\hat{\boldsymbol{c}})^T \boldsymbol{\varsigma}. \tag{36}$$

Since $\boldsymbol{\lambda}(\hat{\boldsymbol{c}})$ is a supergradient of U_w at $\hat{\boldsymbol{c}}$,

$$U_w(\boldsymbol{c}) \leq U_w(\hat{\boldsymbol{c}}) + \boldsymbol{\lambda}(\hat{\boldsymbol{c}})^T(\boldsymbol{c} - \hat{\boldsymbol{c}}). \qquad \forall \boldsymbol{c} \in \mathrm{Co}(C)$$

It follows from (36) that $\boldsymbol{\lambda}(\hat{\boldsymbol{c}})^T(\boldsymbol{c} - \hat{\boldsymbol{c}}) \leq 0$, so $U_w(\boldsymbol{c}) \leq U_w(\hat{\boldsymbol{c}})$, for all $\boldsymbol{c} \in \mathrm{Co}(C)$. This means that $\hat{\boldsymbol{c}}$ is an optimal solution of (27), i.e. $\hat{\boldsymbol{c}} \in C^*$, where C^* denotes the set of optimal solutions of (27). With a slight abuse of notation, here we define $\boldsymbol{q}$ as an $S \times 1$ vector with elements q_s.

Consider the Lyapunov function

$$V(\boldsymbol{c}) = \frac{1}{2} \left\| \boldsymbol{q}(\boldsymbol{c}^*) - \boldsymbol{q}(\boldsymbol{c}) \right\|_2^2$$

where $\boldsymbol{c}^* \in C^*$. Since $\boldsymbol{q}(\boldsymbol{c}^*) = \boldsymbol{q}^*$ is unique, $V(\boldsymbol{c}^*) = 0$ and $V(\boldsymbol{c}) > 0$, for all $\boldsymbol{c} \notin C^*$. Moreover, since $\boldsymbol{q}(\boldsymbol{c})$ is nonpathological, $V(\boldsymbol{c})$ is also nonpathological. Let $\dot{\overline{V}}$ be the nonpathological derivative of the map V with respect to (30)-(31) at $\boldsymbol{c} \in A_V$, where A_V and $\dot{\overline{V}}$ are defined in Definition 4 in [24]. Let $\boldsymbol{\psi}_s \in \partial_C q_s(\boldsymbol{c})$, $s \in S$, where $\partial_C q_s(\boldsymbol{c})$ is the

Clarke gradient of q_s at $\boldsymbol{c}$ [25]. Also let $\Psi = [\boldsymbol{\psi}_1 \cdots \boldsymbol{\psi}_S]^T$. Then

$$\begin{aligned}\dot{\overline{V}}(\boldsymbol{c}) &= -\left(\boldsymbol{q}(\boldsymbol{c}^*) - \boldsymbol{q}(\boldsymbol{c})\right)^T \dot{\overline{\boldsymbol{q}}}(\boldsymbol{c}) \\ &= -\left(\boldsymbol{q}(\boldsymbol{c}^*) - \boldsymbol{q}(\boldsymbol{c})\right)^T \Psi\, \dot{\boldsymbol{c}} \\ &= -\left(\boldsymbol{q}(\boldsymbol{c}^*) - \boldsymbol{q}(\boldsymbol{c})\right)^T \Psi\, \gamma(\tilde{\boldsymbol{c}} - \boldsymbol{c}) \\ &= -\gamma\left(\boldsymbol{q}(\boldsymbol{c}^*) - \boldsymbol{q}(\boldsymbol{c})\right)^T \Psi\, (\boldsymbol{c}^* - \boldsymbol{c} + \tilde{\boldsymbol{c}} - \boldsymbol{c}^*) \\ &= -\gamma\left(\boldsymbol{q}(\boldsymbol{c}^*) - \boldsymbol{q}(\boldsymbol{c})\right)^T\ \Psi\, (\boldsymbol{c}^* - \boldsymbol{c}) \\ &\quad - \gamma\left(\boldsymbol{q}(\boldsymbol{c}^*) - \boldsymbol{q}(\boldsymbol{c})\right)^T \Psi\, (\tilde{\boldsymbol{c}} - \boldsymbol{c}^*).\end{aligned}$$

Using the characterisation of Clarke gradient in equation A.11 in [25], it follows from Taylor's theorem that $\boldsymbol{q}(\boldsymbol{c}^*) - \boldsymbol{q}(\boldsymbol{c}) \approx \Psi\ (\boldsymbol{c}^* - \boldsymbol{c})$, as $\boldsymbol{c}$ approaches $\boldsymbol{c}^*$. Furthermore, since U_w is concave, by Proposition B.24 in [11], there exists $\hat{\boldsymbol{\lambda}} \in \Lambda(\boldsymbol{c}^*)$ such that

$$\hat{\boldsymbol{\lambda}}^T(\boldsymbol{c} - \boldsymbol{c}^*) \le 0 \qquad \forall \boldsymbol{c} \in \mathrm{Co}(C). \tag{37}$$

Let $\tilde{R}_{\boldsymbol{c}}$ be any $L \times S$ routing matrix that defines the links used by only one arbitrary path with positive optimal rate given $\boldsymbol{c}$ that is associated with each source. It then follows that

$$\dot{\overline{V}}(\boldsymbol{c}) \approx -\gamma \left\|\boldsymbol{q}(\boldsymbol{c}^*) - \boldsymbol{q}(\boldsymbol{c})\right\|_2^2 - \gamma(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}(\boldsymbol{c}))^T \tilde{R}_{\boldsymbol{c}}\, \Psi(\tilde{\boldsymbol{c}} - \boldsymbol{c}^*)$$

where $\hat{\boldsymbol{\lambda}}$ satisfies (37).

It can be shown that typically $\tilde{R}_{\boldsymbol{c}}\ \Psi \approx -k(\boldsymbol{c}) I_L$, where $k(\boldsymbol{c}) > 0$ and I_L is the identity matrix. To see this, consider the case where each source has only a single path, i.e. $I_s = 1$. In this case $\boldsymbol{x}(\boldsymbol{\lambda})$ is differentiable with respect to $\boldsymbol{\lambda}$ and $\frac{\partial \boldsymbol{x}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = \mathrm{diag}\left\{\frac{1}{w_s f_s''(x_s(\boldsymbol{\lambda}))}\right\} \tilde{R}_{\boldsymbol{c}}^T$ [26]. Evaluating the sensitivity equation (2.9) in [22] for dual algorithm (29) at its equilibrium point $\boldsymbol{\lambda}(\boldsymbol{c})$ yields

$$\begin{aligned}0 &= \mathrm{diag}\left\{\frac{\beta_l}{c_l}\right\} \tilde{R}_{\boldsymbol{c}} \frac{\partial \boldsymbol{x}(\boldsymbol{\lambda}(\boldsymbol{c}))}{\partial \boldsymbol{\lambda}} \frac{\partial \boldsymbol{\lambda}(\boldsymbol{c})}{\partial \boldsymbol{c}} - \mathrm{diag}\left\{\frac{\beta_l}{c_l^2} \tilde{R}_{\boldsymbol{cl}} \boldsymbol{x}(\boldsymbol{\lambda}(\boldsymbol{c}))\right\} \\ &= \mathrm{diag}\left\{\frac{\beta_l}{c_l}\right\} \tilde{R}_{\boldsymbol{c}}\, \mathrm{diag}\left\{\frac{1}{w_s f_s''(x_s(\boldsymbol{\lambda}(\boldsymbol{c})))}\right\} \tilde{R}_{\boldsymbol{c}}^T \frac{\partial \boldsymbol{\lambda}(\boldsymbol{c})}{\partial \boldsymbol{c}} \\ &\quad - \mathrm{diag}\left\{\frac{\beta_l}{c_l^2} \tilde{R}_{\boldsymbol{cl}} \boldsymbol{x}(\boldsymbol{\lambda}(\boldsymbol{c}))\right\} \\ &\approx \tilde{R}_{\boldsymbol{c}}\, \mathrm{diag}\left\{\frac{1}{w_s f_s''(x_s(\boldsymbol{\lambda}(\boldsymbol{c})))}\right\} \frac{\partial \boldsymbol{q}(\boldsymbol{c})}{\partial \boldsymbol{c}} - I_L.\end{aligned}$$

The last approximation is based on the fact that at optimality total flow on each link is near or equal its capacity. Thus, after factoring out $\mathrm{diag}\left\{\frac{\beta_l}{c_l}\right\}$, the second term on the right-hand side of the equality can be approximated as an identity matrix. Furthermore, it is assumed that the system operates at points where w_s and as a result x_s^* are close for all $s \in S$. Consequently, the values of $w_s f_s''(x_s(\boldsymbol{c}))$, $s \in S$ are close. Hence, $\tilde{R}_{\boldsymbol{c}} \frac{\partial \boldsymbol{q}(\boldsymbol{c})}{\partial \boldsymbol{c}} \approx -k(\boldsymbol{c}) I_L$, where $k(\boldsymbol{c}) \approx |w_s f_s''(x_s(\boldsymbol{c}))|$, $s \in S$.

From (31) it follows that

$$\boldsymbol{\lambda}(\boldsymbol{c})^T(\tilde{\boldsymbol{c}} - \boldsymbol{c}^*) \ge 0 \qquad \forall \boldsymbol{c} \in \mathrm{Co}(C).$$

Also, (37) implies

$$\hat{\boldsymbol{\lambda}}^T(\tilde{\boldsymbol{c}} - \boldsymbol{c}^*) \le 0.$$

Adding both inequalities yields

$$(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}(\boldsymbol{c}))^T(\tilde{\boldsymbol{c}} - \boldsymbol{c}^*) \le 0 \qquad \forall \boldsymbol{c} \in \mathrm{Co}(C). \tag{38}$$

Thus

$$\begin{aligned}\dot{\overline{V}}(\boldsymbol{c}) &= -\gamma \left\|\boldsymbol{q}(\boldsymbol{c}^*) - \boldsymbol{q}(\boldsymbol{c})\right\|_2^2 \\ &\quad + \gamma k(\boldsymbol{c}) I_L (\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}(\boldsymbol{c}))^T(\tilde{\boldsymbol{c}} - \boldsymbol{c}^*) \\ &\le -\gamma \left\|\boldsymbol{q}(\boldsymbol{c}^*) - \boldsymbol{q}(\boldsymbol{c})\right\|_2^2 .\end{aligned} \tag{39}$$

Furthermore, the largest weakly invariant set of points $\boldsymbol{c} \in A_v$ for which $\dot{\overline{V}}(\boldsymbol{c}) = 0$ is the set of equilibrium points C^*. Hence, by Proposition 3 in [24], every solution of algorithms (30)-(31) approaches the set of equilibrium points C^* as $t \to \infty$. ■

Proof of Lemma 1: It results from (26) that

$$\frac{\partial q_s^*}{\partial w_r} = \begin{cases} w_s f_s''(x_s^*) \frac{\partial x_s^*}{\partial w_s} + f_s'(x_s^*) & r = s \\ w_s f_s''(x_s^*) \frac{\partial x_s^*}{\partial w_r} & r \ne s \end{cases} \qquad \forall s \in S. \tag{40}$$

Let $\tilde{\boldsymbol{w}}$ be a perturbation of parameter $\boldsymbol{w}$ defined by

$$\tilde{w}_s = \begin{cases} w_s + dw_r & s = r \\ w_s & \text{otherwise} \end{cases} \qquad \forall s \in S$$

where $r \in S$ and $dw_r > 0$. If $x_r^*(\tilde{\boldsymbol{w}}) = x_r^*(\boldsymbol{w})$, taking the limit $dw_r \to 0$ yields $\frac{\partial x_r^*}{\partial w_r} = 0$, and from (40), $\frac{\partial q_r^*}{\partial w_r} = f_r'(x_r^*)$. If $x_r^*(\tilde{\boldsymbol{w}}) \ne x_r^*(\boldsymbol{w})$, given the strict concavity of f, $\{x_s^*(\boldsymbol{w})\}$ and $\{x_s^*(\tilde{\boldsymbol{w}})\}$ are the unique maximisers for problem (20) with parameters $\boldsymbol{w}$ and $\tilde{\boldsymbol{w}}$, respectively. So

$$\sum_{s \in S} w_s f_s(x_s^*(\boldsymbol{w})) > \sum_{s \in S} w_s f_s(x_s^*(\tilde{\boldsymbol{w}}))$$

and

$$\sum_{s \in S} \tilde{w}_s f_s(x_s^*(\tilde{\boldsymbol{w}})) > \sum_{s \in S} \tilde{w}_s f_s(x_s^*(\boldsymbol{w})).$$

Adding both inequalities results in

$$\sum_{s \in S} (\tilde{w}_s - w_s)(f_s(x_s^*(\tilde{\boldsymbol{w}})) - f_s(x_s^*(\boldsymbol{w}))) > 0.$$

Except for $s = r$, all the elements in the above summation are zero. Since $\tilde{w}_r - w_r = dw_r > 0$, $f_r(x_r^*(\tilde{\boldsymbol{w}})) > f_r(x_r^*(\boldsymbol{w}))$, which implies $x_r^*(\tilde{\boldsymbol{w}}) > x_r^*(\boldsymbol{w})$, since f is an increasing function. Thus

$$\frac{x_r^*(\tilde{\boldsymbol{w}}) - x_r^*(\boldsymbol{w})}{dw_r} > 0.$$

Taking the limit $dw_r \to 0$ yields the lower bound of (33).

It results from the optimality condition in Proposition 2.2.2 in [11] for optimisation problem (20) at $\boldsymbol{w}$ that

$$\sum_{s \in S} w_s f_s'(x_s^*(\boldsymbol{w}))(x_s^*(\tilde{\boldsymbol{w}}) - x_s^*(\boldsymbol{w})) \le 0.$$

Similarly, it results from the optimality condition for (20) at the perturbed $\tilde{\boldsymbol{w}}$ that

$$\sum_{s \in S} \tilde{w}_s f_s'(x_s^*(\tilde{\boldsymbol{w}}))(x_s^*(\boldsymbol{w}) - x_s^*(\tilde{\boldsymbol{w}})) \le 0.$$

Using definition (26), adding both inequalities and taking the limit $dw_r \to 0$ yields

$$\sum_{s \in S} \frac{\partial x_s^*}{\partial w_r} \frac{\partial q_s^*}{\partial w_r} \ge 0 \qquad \forall r \in S. \tag{41}$$

Let $S_d = \{s \in S, s \neq r | x_s^*(\tilde{\boldsymbol{w}}) < x_s^*(\boldsymbol{w})\}$. Since $x_r^*(\tilde{\boldsymbol{w}}) > x_r^*(\boldsymbol{w})$, S_d is non-empty, otherwise $\{x_s^*(\boldsymbol{w})\}$ would not be optimal. Since it is assumed that $f'' < 0$, $f_s'(x_s^*(\tilde{\boldsymbol{w}})) > f_s'(x_s^*(\boldsymbol{w}))$, and hence from (26), $q_s^*(\tilde{\boldsymbol{w}}) > q_s^*(\boldsymbol{w})$ for all $s \in S_d$. Taking the limit $dw_r \to 0$ results in $\frac{\partial x_s^*}{\partial w_r} < 0$ and $\frac{\partial q_s^*}{\partial w_r} > 0$, so

$$\sum_{s \in S_d} \frac{\partial x_s^*}{\partial w_r} \frac{\partial q_s^*}{\partial w_r} < 0 \qquad \forall r \in S.$$

Let $S_i = \{s \in S, s \neq r | x_s^*(\tilde{\boldsymbol{w}}) \geq x_s^*(\boldsymbol{w})\}$. Using a similar argument, $q_s^*(\tilde{\boldsymbol{w}}) \leq q_s^*(\boldsymbol{w})$, for all $s \in S_i$, so

$$\sum_{s \in S_i} \frac{\partial x_s^*}{\partial w_r} \frac{\partial q_s^*}{\partial w_r} \leq 0 \qquad \forall r \in S.$$

Hence, since it was assumed that $\frac{\partial x_s^*}{\partial w_s} > 0$, it follows from (41) that $\frac{\partial q_s^*}{\partial w_s} > 0$, for all $s \in S$.

In (40), since $f_s''(x_s^*) < 0$, the first and second terms on the right side of the equation are negative and positive, respectively. Since the term on the left side of the equation is positive

$$0 \leq -w_s f_s''(x_s^*) \frac{\partial x_s^*}{\partial w_s} < f_s'(x_s^*)$$

from which the upper bounds in (32) and (33) follow. ∎

Proof of Theorem 2: Similar to the proof of Theorem 1, we abuse the notation slightly and here define $\boldsymbol{q}$ as an $S \times 1$ vector with elements q_s. Given R has full column rank, $\beta^{-1}\boldsymbol{q}(\boldsymbol{c}) = \beta^{-1} R^T \boldsymbol{\lambda}(\boldsymbol{c}) = \boldsymbol{d}$ has a solution and hence the equilibrium of (30)-(31) combined with (34) exists for any $\boldsymbol{d}$. As explained in Section IV-B1, if link prices in the duality-based rate control algorithm are updated according to β multiple of the link average queueing delays, link average queueing delays at equilibrium are equal to $\beta^{-1}\boldsymbol{\lambda}(\boldsymbol{c})$, and as a result path average queueing delays at equilibrium equal $\beta^{-1}\boldsymbol{q}(\boldsymbol{c})$, which are bounded by $\boldsymbol{d}$ at the equilibrium $(\boldsymbol{c}^*, \boldsymbol{w}^*)$.

By Theorem 2.2.6 in [23], the mapping $\boldsymbol{q}^*(\boldsymbol{w})$ is continuous, and since $\boldsymbol{q}^*(\boldsymbol{w})$ is also unique, it is a continuous function. We assume that $\boldsymbol{q}^*(\boldsymbol{w})$ is also nonpathological. Consider the Lyapunov function

$$V(\boldsymbol{c}, \boldsymbol{w}) = \frac{1}{2} \left\| \boldsymbol{q}^*(\boldsymbol{w}^*) - \boldsymbol{q}^*(\boldsymbol{w}) \right\|_2^2 + \frac{1}{2} \left\| \boldsymbol{q}^*(\boldsymbol{w}) - \boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c}) \right\|_2^2$$

where $\boldsymbol{q}_{\boldsymbol{w}^*}(\boldsymbol{c}^*) = \boldsymbol{q}^*(\boldsymbol{w}^*) = \beta \boldsymbol{d}$. Therefore, $V(\boldsymbol{c}^*, \boldsymbol{w}^*) = 0$ and $V(\boldsymbol{c}, \boldsymbol{w}) > 0$, for all $(\boldsymbol{c}, \boldsymbol{w}) \neq (\boldsymbol{c}^*, \boldsymbol{w}^*)$. Moreover, since $\boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c})$ and $\boldsymbol{q}^*(\boldsymbol{w})$ are nonpathological, $V(\boldsymbol{c}, \boldsymbol{w})$ is also nonpathological. Let $\dot{\overline{V}}$ be the nonpathological derivative of the map V with respect to (30)-(31) and (34) at $(\boldsymbol{c}, \boldsymbol{w}) \in A_V$, where $\dot{\overline{V}}$ and A_V are defined in Definition 4 in [24]. Also let $\Phi = [\boldsymbol{\phi}_1 \cdots \boldsymbol{\phi}_S]^T$, where $\boldsymbol{\phi}_s \in \partial_C q_s^*(\boldsymbol{w})$, $s \in S$, and $\partial_C q_s^*(\boldsymbol{w})$ is the Clarke gradient of q_s^* at $\boldsymbol{w}$. Then

$$\begin{aligned}
\dot{\overline{V}}(\boldsymbol{c}, \boldsymbol{w}) &= -(\boldsymbol{q}^*(\boldsymbol{w}^*) - \boldsymbol{q}^*(\boldsymbol{w}))^T \dot{\overline{\boldsymbol{q}^*}}(\boldsymbol{w}) \\
&\quad + (\boldsymbol{q}^*(\boldsymbol{w}) - \boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c}))^T (\dot{\overline{\boldsymbol{q}^*}}(\boldsymbol{w}) - \dot{\overline{\boldsymbol{q}}}_{\boldsymbol{w}}(\boldsymbol{c})) \\
&= -(\boldsymbol{q}^*(\boldsymbol{w}^*) - \boldsymbol{q}^*(\boldsymbol{w}))^T \Phi \, \dot{w} \\
&\quad + (\boldsymbol{q}^*(\boldsymbol{w}) - \boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c}))^T \Phi \, \dot{w} \\
&\quad - (\boldsymbol{q}^*(\boldsymbol{w}) - \boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c}))^T \dot{\overline{\boldsymbol{q}}}_{\boldsymbol{w}}(\boldsymbol{c}) \\
&= -\frac{\alpha}{\beta} (\boldsymbol{q}^*(\boldsymbol{w}^*) - \boldsymbol{q}^*(\boldsymbol{w}))^T \Phi \, (\beta \boldsymbol{d} - \boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c})) \\
&\quad + \frac{\alpha}{\beta} (\boldsymbol{q}^*(\boldsymbol{w}) - \boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c}))^T \Phi \, (\beta \boldsymbol{d} - \boldsymbol{q}^*(\boldsymbol{w}) \\
&\quad + \boldsymbol{q}^*(\boldsymbol{w}) - \boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c})) - (\boldsymbol{q}^*(\boldsymbol{w}) - \boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c}))^T \dot{\overline{\boldsymbol{q}}}_{\boldsymbol{w}}(\boldsymbol{c}) \\
&= -\frac{\alpha}{\beta} (\boldsymbol{q}^*(\boldsymbol{w}^*) - \boldsymbol{q}^*(\boldsymbol{w}))^T \Phi \, (\beta \boldsymbol{d} - \boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c})) \\
&\quad + \frac{\alpha}{\beta} (\boldsymbol{q}^*(\boldsymbol{w}) - \boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c}))^T \Phi \, (\beta \boldsymbol{d} - \boldsymbol{q}^*(\boldsymbol{w})) \\
&\quad + \frac{\alpha}{\beta} (\boldsymbol{q}^*(\boldsymbol{w}) - \boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c}))^T \Phi \, (\boldsymbol{q}^*(\boldsymbol{w}) - \boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c})) \\
&\quad - (\boldsymbol{q}^*(\boldsymbol{w}) - \boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c}))^T \dot{\overline{\boldsymbol{q}}}_{\boldsymbol{w}}(\boldsymbol{c}).
\end{aligned}$$

Since $f'' < 0$, (40) implies that $\frac{\partial x_s^*}{\partial w_r} \frac{\partial q_s^*}{\partial w_r} \leq 0$, for all $r \neq s$. It then follows from (41) and (40) that

$$\frac{\partial x_s^*}{\partial w_s} \frac{\partial q_s^*}{\partial w_s} \geq \left| \sum_{\substack{r \in S \\ r \neq s}} \frac{\partial x_s^*}{\partial w_r} \frac{\partial q_s^*}{\partial w_r} \right| \quad \forall s \in S$$

$$\left(\frac{\partial q_s^*}{\partial w_s} - f_s'(x_s^*) \right) \frac{\partial q_s^*}{\partial w_s} \frac{1}{w_s f_s''(x_s^*)} \geq \left| \sum_{\substack{r \in S \\ r \neq s}} \left(\frac{\partial q_s^*}{\partial w_r} \right)^2 \frac{1}{w_r f_r''(x_r^*)} \right| \quad \forall s \in S.$$

We assume that the system operates at points where w_s and as a result x_s^* are close for all $s \in S$. Therefore, the values of $w_s f_s''(x_s(\boldsymbol{c}))$, $s \in S$ are close. Furthermore, in this case it can be assumed that $\frac{\partial q_s^*}{\partial w_s}$ has approximately the average value of the range given in (32), so $\left| \frac{\partial q_s^*}{\partial w_s} - f_s'(x_s^*) \right| \approx \frac{\partial q_s^*}{\partial w_s}$. Thus

$$\begin{aligned}
\frac{\partial q_s^*}{\partial w_s} &\gtrapprox \left(\sum_{\substack{r \in S \\ r \neq s}} \left(\frac{\partial q_s^*}{\partial w_r} \right)^2 \right)^{\frac{1}{2}} \\
&\geq \frac{1}{\sqrt{S-1}} \sum_{\substack{r \in S \\ r \neq s}} \left| \frac{\partial q_s^*}{\partial w_r} \right| \\
&\gtrapprox \sum_{\substack{r \in S \\ r \neq s}} \left| \frac{\partial q_s^*}{\partial w_r} \right|, \qquad \text{for small } S. \qquad (42)
\end{aligned}$$

Inequality (42) implies that Φ can be approximated as a strictly diagonally dominant matrix (Definition 6.1.9 in [27]). Moreover, off-diagonal elements of Φ are very small relative to the diagonal elements, and as a result Φ can be approximated as a symmetric matrix. Since by (32) the diagonal elements of Φ are positive, it then follows from Theorem 6.1.10 in [27] that all eigenvalues of Φ are real and positive and hence Φ is positive definite. Thus

$$\begin{aligned}
\dot{\overline{V}}(\boldsymbol{c}, \boldsymbol{w}) &= -\frac{\alpha}{\beta} (\boldsymbol{q}^*(\boldsymbol{w}^*) - \boldsymbol{q}^*(\boldsymbol{w}))^T \Phi \, (\beta \boldsymbol{d} - \boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c})) \\
&\quad + \frac{\alpha}{\beta} (\beta \boldsymbol{d} - \boldsymbol{q}^*(\boldsymbol{w}))^T \Phi \, (\boldsymbol{q}^*(\boldsymbol{w}) - \boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c})) \\
&\quad + \frac{\alpha}{\beta} (\boldsymbol{q}^*(\boldsymbol{w}) - \boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c}))^T \Phi \, (\boldsymbol{q}^*(\boldsymbol{w}) - \boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c})) \\
&\quad - (\boldsymbol{q}^*(\boldsymbol{w}) - \boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c}))^T \dot{\overline{\boldsymbol{q}}}_{\boldsymbol{w}}(\boldsymbol{c}) \\
&= -\frac{\alpha}{\beta} (\boldsymbol{q}^*(\boldsymbol{w}^*) - \boldsymbol{q}^*(\boldsymbol{w}))^T \Phi \, (\boldsymbol{q}^*(\boldsymbol{w}^*) - \boldsymbol{q}^*(\boldsymbol{w}))
\end{aligned}$$

$$+\frac{\alpha}{\beta}(\boldsymbol{q}^*(\boldsymbol{w})-\boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c}))^T\Phi\left(\boldsymbol{q}^*(\boldsymbol{w})-\boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c})\right)$$
$$-(\boldsymbol{q}^*(\boldsymbol{w})-\boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c}))^T\dot{\overline{\boldsymbol{q}}}_{\boldsymbol{w}}(\boldsymbol{c})$$
$$<\frac{\alpha}{\beta}(\boldsymbol{q}^*(\boldsymbol{w})-\boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c}))^T\Phi\left(\boldsymbol{q}^*(\boldsymbol{w})-\boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c})\right)$$
$$-(\boldsymbol{q}^*(\boldsymbol{w})-\boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c}))^T\dot{\overline{\boldsymbol{q}}}_{\boldsymbol{w}}(\boldsymbol{c}).$$

In the first equality, the second term on the right-hand side results from the assumption that Φ is symmetric. The above inequality results from the positive definiteness of Φ. It follows from Geršgorin Theorem (Theorem 6.1.1 in [27]) and inequalities (42) and (32) that eigenvalues of Φ are upperbounded by $2f_s'(x_s^*)$, $s\in S$. Applying Rayleigh-Ritz Theorem (Theorem 4.2.2 in [27]) to the first term, and using the upperbound (39) for the second term on the right-hand side of the above inequality then yields

$$\dot{\overline{V}}(\boldsymbol{c},\boldsymbol{w})<\left(\frac{2\alpha}{\beta}\max_{s\in S}f_s'(x_s^*)-\gamma\right)\|\boldsymbol{q}^*(\boldsymbol{w})-\boldsymbol{q}_{\boldsymbol{w}}(\boldsymbol{c})\|_2^2 \quad (43)$$

Consequently, if

$$\alpha<\frac{\beta\gamma}{2\max_{s\in S}f_s'(x_s^*)} \quad (44)$$

then $\dot{\overline{V}}(\boldsymbol{c},\boldsymbol{w})\le 0$ for all $(\boldsymbol{c},\boldsymbol{w})$. Furthermore, the largest weakly invariant set of points $(\boldsymbol{c},\boldsymbol{w})\in A_V$ for which $\dot{\overline{V}}(\boldsymbol{c},\boldsymbol{w})=0$ is the set of equilibrium points $(\boldsymbol{c}^*,\boldsymbol{w}^*)$. Hence, by Proposition 3 in [24], every solution of algorithms (30)-(31) combined with (34) approaches the set of equilibrium points $(\boldsymbol{c}^*,\boldsymbol{w}^*)$ as $t\to\infty$. ■

REFERENCES

[1] B. Wydrowski and M. Zukerman, "QoS in best-effort networks," *IEEE Commun. Mag.*, pp. 44–49, Dec. 2002.

[2] S. Shenker, "Fundmental design issues for the future internet," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 7, pp. 1176–1188, Sep. 1997.

[3] M. Chiang, S. H. Low, A. R. Calderbank, and J. C. Doyle, "Layering as optimization decomposition: A mathematical theory of network architectures," *Proc. IEEE*, vol. 95, no. 1, pp. 255–312, Jan. 2007.

[4] X. Lin and N. B. Shroff, "Joint rate control and scheduling in multihop wireless networks," in *Proc. IEEE 43rd Conf. Decision and Control*, Atlantis, Paradise Island, Bahamas, Dec. 2004, pp. 1484–1489.

[5] F. Paganini, Z. Wang, J. C. Doyle, and S. H. Low, "Congestion control for high performance, stability, and fairness in general networks," *IEEE/ACM Trans. Netw.*, vol. 13, no. 1, pp. 43–56, Feb. 2005.

[6] T. Yoo, E. Setton, X. Zhu, A. Goldsmith, and B. Girod, "Cross-layer design for video streaming over wireless ad hoc networks," in *Proc. IEEE 6th Workshop Multimedia Signal Processing*, Sienna, Italy, Oct. 2004, pp. 99–102.

[7] P. Huang, X. Lin, and C. Wang, "A low-complexity congestion control and scheduling algorithm for multihop wireless networks with order-optimal per-flow delay," in *Proc. IEEE INFOCOM*, Shanghai, China, Apr. 2011, pp. 2588–2596.

[8] Y. Li, M. Chiang, and A. R. Calderbank, "Congestion control in networks with delay sensitive traffic," in *Proc. IEEE INFOCOM*, 2007, pp. 2746–2751.

[9] S. Stidham, "Pricing and congestion management in a network with heterogeneous users," *IEEE Trans. Autom. Control*, vol. 49, no. 6, pp. 976–981, Jun. 2004.

[10] D. P. Bertsekas and R. Gallager, *Data Networks*. Englewood Cliffs, NJ: Prentice Hall, 1992.

[11] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA: Athena Scientific, 1999.

[12] R. L. Cruz, "A calculus for network delay, part I: Network elements in isolation," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 114–131, Jan. 1991.

[13] ——, "A calculus for network delay, part II: Network analysis," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 132–141, Jan. 1991.

[14] 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, Aug. 2006.

[15] L. Chen, S. H. Low, and J. C. Doyle, "Joint congestion control and media access control design for ad hoc wireless networks," in *Proc. IEEE INFOCOM*, Mar. 2005, pp. 2212–2222.

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

[17] S. Jahromizadeh and V. Rakocevic, "Rate control for delay-sensitive traffic in multihop wireless networks," in *Proc. 4th ACM Workshop Performance Monitoring and Measurement of Heterogeneous Wireless and Wired Networks*, Tenerife, Canary Islands, Spain, Oct. 2009, pp. 99–106.

[18] ——, "Joint rate control and scheduling for delay-sensitive traffic in multihop wireless networks," in *Proc. IEEE 73rd Vehicular Technology Conference*, Budapest, Hungary, May 2011, pp. 1–5.

[19] R. Srikant, *The Mathematics of Internet Congestion Control*. Boston, MA: Birkhäuser, 2004.

[20] X. Lin and N. B. Shroff, "Utility maximization for communication networks with multipath routing," *IEEE Trans. Autom. Control*, vol. 51, no. 5, pp. 766–781, May 2006.

[21] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Belmont, MA: Athena Scientific, 1997.

[22] H. K. Khalil, *Nonlinear Systems*. Upper Saddle River, NJ: Prentice Hall, 2002.

[23] A. V. Fiacco, *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*. NY: Academic Press, 1983.

[24] A. Bacciotti and F. Ceragioli, "Nonpathological Lyapunov functions and discontinuous carathéodory systems," *Automatica*, vol. 42, pp. 453–458, 2006.

[25] ——, "Nonsmooth optimal regulation and discontinuous stabilization," *Abstr. and Appl. Anal.*, vol. 20, pp. 1159–1195, 2003.

[26] S. H. Low and D. E. Lapsley, "Optimization flow control–I: Basic algorithm and convergence," *IEEE/ACM Trans. Netw.*, vol. 7, no. 6, pp. 861–874, Dec. 1999.

[27] R. A. Horn and C. R. Johnson, *Matrix Analysis*. NY: Cambridge University Press, 1990.

Soroush Jahromizadeh received the B.Sc. degree in industrial engineering from Amirkabir University of Technology, Tehran, Iran, in 1995, the M.Sc. degree in systems engineering from University College London, London, UK, in 2002, the M.Sc. degree in general engineering from City University London, London, UK, in 2005, and the Ph.D. degree in electrical engineering from City University London, London, UK, in 2013.

Since 1995 he has worked for engineering consultancy firms, where he has performed operations research and financial analysis for industrial projects. Between 2012 and 2013 he was researcher at City University London, UK, where he worked on optimisation of energy networks. His research interests include design, optimisation and control of networks.

Veselin Rakočević (M'01) received the Dipl.Ing. degree in electronic engineering from the University of Belgrade, Serbia, in 1998, and the Ph.D. degree in electronic engineering from Queen Mary, University of London, UK, in 2002.

He works as Senior Lecturer in Electronic Engineering at City University London, UK, where he has been since 2002. His main research interest is in the operation of multihop wireless networks, especially in the problem of optimal scheduling, rate control and quality of service. During his career he worked on a number of industry-led research projects in the area of wireless network control and energy network control.