Rate Control and Dynamic Dimensioning of Multihop Wireless Networks

Aditya Karnik, Ravi Mazumdar and Catherine Rosenberg
Dept. of Electrical and Computer Engineering,
University of Waterloo,
Waterloo, Canada N2L 3G1

Abstract—This paper is concerned with the problem of allocation of data transfer rates to the elastic applications in a multihop wireless network. While this is a well understood problem in wired networks, wireless networks present substantially different dimension to it. For one the very concept of link capacity, though obvious for wired links, is not so for wireless links in a network. Moreover, unlike a wired network, the traffic capacity of a wireless network can be varied simply by modifying certain protocol parameters; thus it is possible to dimension the network dynamically. The wireless networks we consider are so-called infrastructure-based networks implying that they are deployed by service providers by creating hot-spots through access points whereas the end applications running on user devices act as the end users. This permits the network and the users to be seen as separate economic entities competing and collaborating in response to each other. The aim of this paper is to precisely understand these issues and address the appropriate formulations based on them.

I. INTRODUCTION

This paper is concerned with the problem of allocation of data transfer rates to the elastic users, i.e., the end applications which can tolerate variable rate of data transfer and delay but are sensitive to loss, of a communication network. This problem has been extensively studied and fairly well understood in the context of wired Internet. In the case of wireless networks, however, it remains to be completely settled and hence they will be the focus of this paper.

To a large extent, resolution of this rate allocation (or control) problem in the case of wired networks was facilitated by viewing it as a special case of fair and efficient allocation of commodities or resources to consumers, a problem which forms the center piece of the economic theory. There is now substantial literature addressing it by employing a variety of tools from the economic and the game theory, such as utility maximization, bargaining, Nash games, etc. (e.g. [1], [2], [3]). The utility framework has been, in particular, quite successful (and popular) since it was shown that the versions of TCP are essentially primal-dual algorithms try to maximize certain utility functions ([4]). Economic formulations are compelling even in the case of wireless networks. However, certain peculiarities of the wireless operation of communication links imply that these networks require a different economic viewpoint than their wired counterparts.

Firstly, the very concept of link capacity, though obvious for wired links, is difficult to define for wireless links in a network. In fact, whether a link can be associated with a number denoting its link capacity depends on how the interference due to other transmitting devices in the network to the transmissions on the link is managed. In networks which employ randomized MAC protocols, assigning capacities to individual links is impossible. This means that in such wireless networks, the rate allocation problem can not be simply cast as that of dividing individual link capacities, as in the wired networks.

Secondly, with the advent of software controlled radios, modifying the radio transmission parameters (e.g., transmission power, modulation/FEC schemes, back-off parameters, etc.) “on the fly” has become extremely easy. Since the traffic carrying capacity of a wireless network is determined by these parameters, this essentially means that, rather than taking the network capacity as fixed and given, it can be dimensioned dynamically, i.e., at the time-scale of rate allocations by varying the “network parameters”. Thus in contrast to the wired networks, in which perhaps the only way to change network capacity is to revamp the infrastructure, a wireless network can potentially be “fitted” to its environment and users. The difficulty with this, however, is that a precise characterization the traffic capacity in terms of all the transmission/protocol parameters is difficult to obtain and would be unavailable in general.

Finally, the wireless technologies have evolved in such a way that it is possible to deploy and operate wireless networks in two significantly different ways. The first, which we call the infrastructure-based networks, are those with proper infrastructure (subscriber stations, access points, etc.) deployed with the intent of providing backbone service to the users; for example, IEEE 802.16
broadband wireless networks. In these networks, the networking functionality will be built into the network devices whereas the applications running on user devices will act as the end users. In view of the previous point, this permits the network and the users to be seen as separate economic entities influencing the rate allocations and network dimensioning through mutual interaction. In the second type, popularly known as the wireless ad hoc networks (e.g., IEEE 802.11 ad hoc mode), there is no distinction between the network and the users; the user devices will not only act as application endpoints but also as routers, i.e., they will not only transmit their own packets but will forward packets for other devices as well. Thus the user devices themselves constitute a network, thereby, implying that the rate allocations and network dimensioning will result from the competition and collaboration among the users.

As a consequence of these, interesting difficulties as well as possibilities are opened up in formulating and solving the problem in the context of wireless networks. This paper addresses these issues in the context of infrastructure-based networks. This has two ramifications. First, owing to the separation of the network and the users, network dimensioning in response to user rate requirements and vice versa is not a “cross-layer optimization” issue; this allows us to include any protocol parameters (amenable to on the fly modification) affecting the network capacity in our formulations. This is in contrast to the earlier papers (e.g., [5], [6] [7]) which have invariably focused on the ad hoc networks and cross-layer optimization (TCP-MAC, TCP-PHY). Secondly, in an economic context it allows the users to be seen as consumers, their data rates as produced commodities, the network as a producer and the protocol parameters as factors of production or input commodities.

We proceed as follows. In Section II we discuss characterization of the achievable rates in the network (not in an information-theoretic sense) and formulation of rate constraints. This actually pertains to wireless networks in general and not infrastructure-based networks specifically. For lack of space, we discuss in Section III the problem formulations only in the context of a subclass of infrastructure-based networks. We conclude in Section IV.

II. UNDERSTANDING THE RATE CONSTRAINTS

The capacity of a communication link, whether wired or wireless, is the maximum rate at which data units (packets, bits, etc.) can be transported across it successfully\(^1\)

\(^1\)In this non-information-theoretic setup, communication errors can not be completely eliminated; hence “success” is in the sense of achieving a specified bit or packet error rate.

if a continuous supply of them is made available to the link; we refer to this rate by “link capacity” (measured in say packets transmitted successfully per sec). The link capacity together with the rate at which data units are fed into it by a source determines the actual data transfer rate obtained by the source. For a wired link its capacity is essentially fixed. The characteristics of wireless medium, on the other hand, are prone to frequent changes; hence the capacity of a wireless link may vary with time. This can be possibly countered by adapting its radio (PHY) parameters to the channel conditions (channel gain, fading, co-channel interference). To keep the discussion short, in this paper we will assume that the channel condition are time-invariant.

What differentiates a network of wired links and that of wireless links is the fact that the link capacity of a wired link is invariant to what transpires on the other links in the network whereas such is not the case for a wireless link owing to the inherent broadcast nature of the wireless medium. This means that transmissions on a link, which otherwise would have been successful, may fail due to interference from other simultaneous transmission(s) on the links in its vicinity. This necessitates some sort of coordination among the wireless links.

If this coordination is achieved through a pre-computed (i.e., independent of the traffic being carried by the links), conflict-free (i.e., whenever a link transmits, it can transmit successfully in the sense mentioned earlier) link transmission schedule, then it is possible to assign each link a capacity. This link capacity, obviously not exceeding that “in isolation”, is simply the link capacity in isolation multiplied by fraction of time allocated to the link for transmission in the schedule\(^2\). With well defined link capacities, the user rate constraints can be split into individual link constraints as in the wired networks. In this sense, the wireless networks operated thus, which we refer to as scheduled networks, are the wireless counterparts of high speed wired networks. However, the possibility of dynamic configuration of parameters discussed in Section I, allows the link constraints to be seen as parameterized by the transmission schedule and PHY parameters. Examples of scheduled networks are IEEE 802.16 PMP or Mesh with centralized scheduling.

If, on the other hand, transmissions are coordinated dynamically in response to the traffic, while success may be guaranteed on transmission, the transmission attempts on a link themselves depend on the traffic on the other links implying that its capacity can be defined only by keeping the conditions on the other links fixed. Practically, in fact, even guaranteeing conflict

\(^2\)We have implicitly assumed that a schedule is a fixed repetitive pattern of link activations.
free transmissions is difficult owing to the distributed mechanisms (randomized MAC protocols) which result in contention. The MAC protocol at each node tries to avoid incidences of transmission failure by regulating the channel access by attempting or deferring transmissions based on its arriving data load and its perception of the load on the shared channel due to arriving traffic at the other transmitting nodes. Thus the transmission attempts and successes on a link depend on the traffic on the other links, how they respond to their traffic and their perception of the channel, and which PHY parameters, in particular transmission power, they employ. This, first of all, means that that link capacities cannot be defined for all the links in the network simultaneously and secondly, being so intricately coupled, the data rates can only be constrained “jointly” through the stability region of the network, not per link; it is the stability region that is parameterized by the network (PHY and MAC) parameters. We refer to these wireless networks as the unscheduled wireless networks; e.g., IEEE 802.16 Mesh with uncoordinated scheduling.

In contrast to the above, the earlier papers (e.g., [5], [6] [7]) characterize the rate constraints by arguing as follows. Any MAC scheme can be seen as a particular link activation schedule and as is the case with a link schedule, it would result in a certain assignment of capacities to the links (and data rates to the users). However, the schedule associated with a MAC is unknown a priori. Hence instead of insisting on a particular MAC, the idea is to derive a condition on the link capacities that would guarantee existence of a link schedule realizing them; such a condition is rendered by appropriately tightening the constraints on the link rates resulting from clique dependencies in the link contention graph. Inducting this condition in the formulation then obviates the need to specify a particular MAC (schedule) but simultaneously guarantees the existence of a “feasible MAC”. We think the problems with this approach are: first, a MAC scheme does not generate a link schedule a priori independent of traffic; by assuming so the above argument is essentially treating the case of scheduled networks. In fact, its link activation pattern emerges in response to the traffic carried by the links depending on the protocol parameters and hence link rates cannot be associated with the links in the network a priori. Secondly, it is not clear how this reasoning can be used to dynamically configure parameters of a specific MAC.

III. Formulations for Scheduled Networks

The discussion in Section II shows that rate allocation in an unscheduled network is harder to handle since the network can influence rate constraints only through the stability region which is difficult to characterize. In a scheduled network, on the other hand, link capacity constraints are valid, the only twist being that they can be parameterized. This is the case we discuss in this section. The following development is particularly motivated by the IEEE 802.16 standard ([8]) which allows allocation of fractions of a single frame to different network nodes for transmission. Modulation schemes and transmission power can be chosen adaptively. Moreover, data can be fragmented and packed to a much finer scale; thus justifying treatment of data as fluid.

Consider a wireless network of $M$ nodes and $L$ links; the set of links (resp. nodes) is denoted by $\mathcal{L}$ (resp. $\mathcal{N}$). $\mathcal{L}_o$ denotes the set of links controlled by (outgoing from) node $i$. For link $l \in \mathcal{L}$, $l_o$ and $l_d$ denote the originating and the destination nodes resp. The network serves a fixed number, $N$, of users (sources), $\mathcal{N}$ denotes the set of users and $x_r$ denotes the flow rate of user $r$, $r = 1, \ldots, N$ which perceives its rate performance through a utility function $U_r(.)$. $U_r(.)$ is assumed to be a continuously differentiable, strictly concave, increasing function of $x_r$. For each $r$, $x_i$ is constrained to $[m_r, M_r]$ where $0 \leq m_r$ and $M_r < \infty$. $\mathcal{N}_l$ denotes the set of users sharing link $l$ and $\mathcal{L}_r$ the set of links used by user $r$. The network has $Q$ configurable parameters denoted by $\alpha_i$, $i = 1, \ldots, Q$; the exact nature of these parameters is specified below. $\mathcal{A}$ denotes the parameter space (a compact subset of $R^Q$). Under a configuration $\alpha = (\alpha_1, \ldots, \alpha_Q)$, $h(\alpha)$ denotes the operating (production) cost for the network; $h(\alpha)$ is assumed to be a continuously differentiable strictly convex function of $\alpha$. Further $h(0) = 0$ and $h(\alpha_i, \alpha_{-i})$ is strictly increasing in $\alpha_i$, $i = 1, \ldots, L$.

Let $P_l$, $l \in \mathcal{L}$ denote the transmission power on link $l$. $\mathbf{P} := (P_1, \ldots, P_L)$ denotes the power vector and $\mathcal{P}$ the set of feasible power vectors (e.g., $0 \leq P_l \leq P_l^\star$, $l \in \mathcal{L}$, $i = 1, \ldots, M$ for some $P^\star_i$’s). $z_l$ denotes the modulation-coding scheme used on link $l$; $\mathcal{Z}$ denotes the available modulation-coding schemes. $z := (z_1, \ldots, z_L)$. A bit error rate specification (for success) essentially translates into a SINR threshold, $\beta_l(z_l)$ for each $z_l$ on link $l$. The SINR on link $l$, $\gamma_l$, at a given time instant is given by $\frac{G_{i_l}P_l}{N_0 + \sum_{l'}G_{i_{l'}}P_{l'}}$, where the summation in the denominator is over the simultaneously transmitting links. $G_{il}$ (resp. $G_{i_{l'}}$) denotes the channel gain on link $l$ (resp. from link $l'$ to $l$ where it is understood that it refers to the gain from $l'_o$ to $l_d$). Recall that we have assumed time-invariant channel conditions (or slowly varying in comparison to the rate allocation time-scale). We further assume that $G_{i_{l'}}$’s are known (so if fading is Rayleigh its sample has been realized and known). Given $\mathbf{P}$, let $\mathcal{Z}_l(P_l) \subseteq \mathcal{Z}$ denote the set of modulation-coding schemes which can
achieve the SINR threshold in the absence of co-channel interference on link \( l \); it is clear that under \( P \) only schemes in \( Z_l(P_l) \) may be feasible. \( Z_l(P) := \Pi Z_l(P_l) \).

Now given \( P \in \mathcal{P} \) and \( z \in Z(P) \), the condition for successful transmission, \( \gamma_l > \beta_l(z_l) \) on each link \( l \) essentially translates into its “conflicting set” \( D_l(z, P) \). Each \( D \in D_l(z, P) \) is a subset of \( L \) with the interpretation that if all the links in \( D \) are transmitting simultaneously with link \( l \) (using transmission powers given by \( P \)) then the transmission on link \( l \) fails. Therefore, to guarantee successful transmission on link \( l \), at least one link from each \( D \in D_l(z, P) \) must be silent. The link contention graph (or matrix) formulation considered in the literature is a special case of the above formulation; contention graph implies that each \( D \in D_l(z, P) \) is a singleton for each link \( l \). Without loss of generality we assume that the time interval over which a non-conflicting schedule is to be constructed is \([0, 1]\). Let \([t_f^l, t_f^l] \) denote the time segment during which link \( l \) is scheduled to transmit; \( 0 \leq t_f^l \leq t_f^l \leq 1 \). We will call \((t_f^l, t_f^l, \ldots, t_f^l, t_f^l)\) a transmission schedule and denote it by \( \theta \). Then the set of non-conflicting schedules under \( D_l(z, P) \) is

\[
\Theta(z, P) = \{(t_f^l, t_f^l, \ldots, t_f^l, t_f^l) | 0 \leq t_f^l \leq t_f^l \leq 1, (t_f^l, t_f^l) \cap \cap_{r \in D} (t_f^l, t_f^l) = \emptyset, D \in D_l(z, P), l \in L \}
\]

A choice of \( \theta \in \Theta(z, P) \) now determines the \( c_l(z_l) \) the capacity of each link \( l \); it simply equals \( c_l(z_l)(t_f^l - t_f^l) \), where \( c_l(z_l) \) is the bit rate on the link resulting from actual bit clocking and modulation-coding scheme \( z_l \).

We denote the link capacity of \( z_l \) by \( c_l(z_l, P, \theta) \) to make its dependence on \( z, P \) and \( \theta \in \Theta(z, P) \) explicit. These are the configurable network parameters denoted by \( \alpha \). The formulations we consider are as follows.

**Social Welfare Maximization (SWM):** This problem of a social planner’s interest jointly optimizes the data rates and the network parameters so as to maximize welfare.

\[
\max \sum_{r=1}^{N} U_r(x_r) - h(z, P, \theta) \quad \sum_{r \in N_l} x_r \leq c_l(z_l, P, \theta) \quad l = 1, \ldots, L \\
\quad x_r \in [m_r, M_r] \quad r = 1, \ldots, N \\
\quad \theta \in \Theta(z, P) \\
\quad z \in Z(P), P \in \mathcal{P}
\]

**Bilevel Optimization (BO):** While the users choose rates by maximizing the sum-utilities given the network parameters, i.e., by solving,

\[
\max \sum_{r=1}^{N} U_r(x_r) \quad \sum_{r \in N_l} x_r \leq c_l(z_l, P, \theta) \quad l = 1, \ldots, L \\
\quad x_r \in [m_r, M_r] \quad r = 1, \ldots, N
\]

the network configures the parameters so as to maximize the revenue generated from the (proportionally fair) prices the users pay; hence it solves the following.

\[
\max \sum_{l=1}^{L} \lambda_l^*(z, P, \theta) c_l(z_l, P, \theta) - h(z, P, \theta) \quad \theta \in \Theta(z, P) \\
\quad z \in Z(P), P \in \mathcal{P}
\]

where, \((\lambda_1^*(z, P, \theta), \ldots, \lambda_L^*(z, P, \theta))\) are the prices (or dual variables) of the user optimization problem which is parameterized by \((z, P, \theta)\).

**Monopoly Problem (MP):** Assuming that the network is deployed by a single service provider, the situation is best modelled as monopoly. Given prices \( p := (p_1, \ldots, p_N) \), user \( r \) solves the following.

\[
\max \sum_{r \in [m_r, M_r]} U_r(x_r) - p_r x_r
\]

Denoting by \( x^*(p) \) the optimal solution of the user problems for a price vector \( p \), the network sets parameters and prices so as to maximize its net revenue keeping the network stable, i.e., by solving,

\[
\max \sum_{l=1}^{L} \lambda_l^*(z, P, \theta) c_l(z_l, P, \theta) - h(z, P, \theta) \quad \theta \in \Theta(z, P) \\
\quad z \in Z(P), P \in \mathcal{P} \\
\quad p \in R^N
\]

An optimal solution to all three problems exists. We believe that in this set up, BO and MP are more compelling than SWM. Nevertheless, SWM provides a solution against which the other two can be compared. Interestingly, in BO and MP rate control on the user-side does not demand any special implementation- it is based on price or congestion signal feedback and may be viewed as a version of TCP. Below we discuss a special case of SWM which leads to a similar implementation. Assume that \( P \) and \( z \in Z(P) \) are fixed. This results in a major simplification since the conflicting sets \( D_l(z, P) \) are now fixed. Moreover, \( c_l(z_l, P, \theta) = c_l(t_f^l - t_f^l) \) for some fixed \( c_l \). SWM then reduces to the following.

\[
\max \sum_{r=1}^{N} U_r(x_r) - h(\theta) \quad \sum_{r \in N_l} x_r \leq c_l(t_f^l - t_f^l) \quad l = 1, \ldots, L \\
\quad x_r \in [m_r, M_r] \quad r = 1, \ldots, N \\
\quad \theta \in \Theta(z, P)
\]

There exist prices \( p^*_r \geq 0, l = 1, \ldots, L \) such that the optimal solution \((x^*, \theta^*)\) of (1) can be realized as the following economic equilibrium point.

(i) Each user \( r = 1, \ldots, N \) maximizes its surplus, i.e.,

\[
x^*_r = \arg \max_{x_r \in [m_r, M_r]} U_r(x_r) - \left( \sum_{l \in L_r} p^*_l \right) x_r
\]
\( \theta^* = \arg \max_{\theta \in \Theta(z, P)} \sum_{l=1}^{L} p^*_l c_l(t^*_l - t^*_l) - h(\theta) \) \hspace{1cm} (2)

(iii) Supply equals demand at positive price, i.e.,
\[ p^*_l \left( c_l(t^*_l - t^*_l) - \sum_{r \in N_l} x^*_r \right) = 0 \]

Thus, in this special case of SWM, maximizing welfare is amenable to a distributed algorithm; this is a fairly standard fact. The novel part is solving (2).

For this discussion, assume that \( h(.) = 0 \) for all \( \theta \in \Theta(z, P) \). In such a case, the optimal solution is that if a link transmits it transmits for the full duration, i.e., if \( t^*_l - t^*_l > 0 \), then \( t^*_l = 1 \). This essentially converts (2) into a combinatorial problem—of deciding which links transmit and which do not. Let \( u_l = 1 \) if link \( l \) transmits and 0 otherwise. Then the maximization problem in (2) can be stated as
\[ \max_{u} \sum_{l=1}^{L} p^*_l u_l \] (3)
\[ u_l \Pi_l u_l = 0 \quad D \in D_l(z, P), \quad l = 1, ..., L \] (4)
\[ u_l \in \{0, 1\} \quad l = 1, ..., L \]

Note that (3) is a generalization of the weighted maximal independent set problem. This means that (3) is an NP-complete problem. We discuss briefly the idea behind an exact algorithm for it; details can be found in [9]. Let \( u^* \) denote the optimal solution of (3). Note that the integer constraints in (3) can be relaxed \((0 \leq u_l \leq 1)\) without affecting the optimality of \( u^* \). However, even if seen as a continuous optimization problem (due to relaxed integer constraints), (2) is difficult to solve because constraint qualification (linear independence (LICQ) or Mangasarian-Fromowitz (MFCQ)) may not hold at \( u^* \). The idea is then to relax (4) and consider for some \( 0 < \epsilon < 1 \)
\[ \max_{u} \sum_{l=1}^{L} p^*_l u_l \] (5)
\[ u_l \Pi_l u_l \leq \epsilon \quad D \in D_l(z, P), \quad l = 1, ..., L \]
\[ 0 \leq u_l \leq 1 \quad l = 1, ..., L \]

Let \( \bar{u} \) denote the optimal solution of (5). Then it can be shown that (i) \( \bar{u}_l \in \{\epsilon, 1\} \) for each \( l \) (ii) MFCQ holds at \( \bar{u} \), and (iii) \( u^* = 0 \leftrightarrow \bar{u} = \epsilon, u^* = 1 \leftrightarrow \bar{u} = 1 \). It follows that solving (5) leads to the solution of (3). Interestingly, (5) is a geometric program ([10]), which implies that it can be solved efficiently using its associated convex problem. It turns out that in this convex program the constraints are linear. In [9] we also discuss its possible distributed implementation.

IV. CONCLUSION

In this paper we addressed the problem of allocation of data transfer rates to the elastic users of a wireless network. Noting the peculiarities of wireless networks, we argued that the economic formulations need a fresh perspective in this context. We then discussed the formulation of data rate constraints and argued that in wireless networks, where transmissions are orchestrated to be conflict-free through pre-computed schedules (scheduled networks) the rate constraints are similar to the wired networks. If, on the other hand, this is not the case (unscheduled networks) then the only way to characterize the rate constraints is through the stability region of the network. We provided three economic formulations for scheduled networks, namely, welfare maximization, bilevel optimization and monopoly. We believe that bilevel optimization and monopoly are the appropriate formulations from the perspective of rate control and dynamic dimensioning of these networks. Details of this work will be reported in [11].

REFERENCES