The Scheduling Problem in Wireless Networks

Anna Pantelidou and Anthony Ephremides

(Invited Paper)

Abstract: **We describe the fundamental issue of scheduling the allocation of wireless network resources and provide several formulations of the associated problems. The emphasis is on scheduling transmission attempts. We place this problem in the context of existing approaches, like information theoretic and traditional network theoretic ones, as well as novel avenues that open up the possibility of addressing this issue for non-stationary and non-ergodic environments. We summarize concrete recent results for specific special cases that include unicast and multicast traffic, different objective functions, and reduced complexity versions of the problem. We conclude with some thoughts for future work. We identify and single out the cross-layer nature of the problem and include a simple physical-layer criterion in what is mostly a medium access control (MAC) problem.**

Index Terms: **Cross-layer design, minimum-length scheduling, rate control, scheduling, utility maximization.**

I. INTRODUCTION

In wireless networks, the problem of allocating transmission rights to subsets of network users at each time and under different channel qualities is known as the *scheduling problem*. It arises in wireless environments because of three main reasons related to the fundamental properties of the wireless medium. Specifically, scheduling is mandatory since in wireless environments (i) communication resources are *shared* amongst geographically separated users, (ii) transmissions *interfere* with each other, and (iii) transmissions undergo *impairments*, such as fading, attenuation, etc.

Simply put, the scheduling problem is about identifying the users that are allowed to transmit at any given time and their corresponding transmission power levels and rates. Traditionally, the scheduling problem has been studied in the context of medium access control (MAC) protocols that ignore the physical layer. Initial works in the scheduling problem employ simplistic channel models, such as, for instance, the collision channel. They further suppress the questions of power and rate control by assuming "packets" of traffic that can be transmitted during corresponding time slots, without worrying about the number of bits that can be accommodated in each packet. The transmission "range" is chosen arbitrarily and no interference is assumed to be possible outside that transmission range. It is worth noting that considering the case of multiple available channels, although a practical reality with several advantages and complicating consequences, does not enrich the basic problem in a fundamental way. Thus, for simplicity, it is convenient to assume a single channel that is shared amongst all users across the network.

In parallel, the ultimate capabilities of wireless networks, in terms of transmission rates, have been studied by the information theoretic community by means of extensions of Shannon's formulations without worrying about "scheduling" at all. In other words, in these formulations every user transmits at all times and never runs out of data to transmit. It is only a question, albeit a difficult one, of determining the achievable rates of simultaneous transmissions for reliable communication at given power levels for different channel models. It is important to note that these studies, in spite of the recent advances in the analysis of relay channels, multiple input multiple output (MIMO) technology, cooperative communication techniques, etc., have had limited success. Furthermore, they lead to formidably difficult problems, even without taking practical realities into account, such as bursty traffic, time-varying channel behavior, or the need to achieve finite end-to-end delays.

Recently, it has been observed that it is possible to incorporate an element of physical layer characteristics in network-oriented approaches. This is achieved by defining a transmission to be successful if the value of the signal to interference plus noise ratio (SINR) at the receiver exceeds a specific threshold value. This is a relatively accurate criterion in the case of a large number of network users and it emerged in the study of code division multiple access (CDMA) systems in cellular environments. However, it is more generally useful because, in all cases, although only approximately valid, it permits the coupling of the MAC with the physical layer in a simple, tractable, and meaningful way. In fact, it incorporates directly the effect of power control and attenuation/fading in the outcome of a transmission since the SINR depends directly on both of these factors. Furthermore, it couples the effect of the transmission rate and target bit error rate (BER) with the outcome of a transmission attempt since the value of the threshold is an increasing function of the rate and a decreasing function of the BER. Thus, it allows a rich and in-depth analysis and understanding of the effects of the wireless medium on the performance of MAC protocols. It will be this approach that we will focus mostly on in this paper, in which we describe the overall problem and present a description of solutions for selected special cases of interest.

Once the formulation of the scheduling problem uses the notion of slots and seeks to determine the best way of "grouping" subsets of transmitting users in each slot, the scheduling problem obtains a component of combinatorial complexity. Any ver-

Manuscript received July 12, 2009.

This work is supported by the Department of Defense under MURI grants W911NF-05-1-0246 and W911NF-08-1-0238 and by the National Science Foundation under the grant CCF0728966.

A. Pantelidou is with the Centre for Wireless Communications, University of Oulu, Finland, email: apantel@ee.oulu.fi.

A. Ephremides is with the Department of Electrical and Computer Engineering, University of Maryland, College Park, USA, email: etony@umd.edu.

sion of the problem that includes this combinatorial aspect has been shown to be NP-hard (see e.g., [1]) and thus it is customary to either invent heuristic solutions or limit the search space of transmission strategies in order to derive sub-optimal solutions that can be evaluated.

We start by a discussion of the cost functions and criteria that are appropriate for this optimization problem, continue with a description of the constraints that can be adopted, and then we review briefly an earlier, but powerful formulation of the problem that gave rise to the, so-called, "back-pressure" algorithm [2]. That formulation is based on the assumption that the traffic is bursty and the objective is to maximize the stable throughput region of the network, where a network is stable when all the queues do not grow without bound. We propose an alternative formulation of the scheduling problem that is based on minimizing the length of a schedule. This formulation has potential applicability to general wireless environments that do not have a stationary and ergodic time-variation. However, it has a fundamental combinatorial complexity limitation, which leads us then to review some special cases that are intentionally simplified for tractability. Albeit simplified, these formulations retain the essential trade-offs that are inherent in the scheduling problem and that yield insightful results and solutions. We conclude the paper with thoughts and suggestions about future research possibilities.

II. COST CRITERIA FOR THE SCHEDULING PROBLEM

It is not immediately clear what the objective should be when we decide which transmissions to enable in a wireless network. There are numerous possibilities. One reasonable objective is to maximize the *sum throughput*, namely the total, aggregate packet rate across the network. Another, slightly more general and versatile, objective is to achieve the largest *throughput region*, namely the multi-dimensional set of rate vectors that consist of the individual rate components for each source/destination pair. A further distinction can be made if we consider the subtle differentiation between the throughput region and the *stable throughput region*. The latter consists of the rate vectors that are sustainable under the requirement that the network remains stable, which further implies that the packets at the buffers of every node (source or relay) experience finite average delays. The comparison between these two regions depends on the specifics of the system under consideration. There are examples of systems in which the two regions are identical and there are examples where they are different. Furthermore, it is possible for one to contain the other or for them to be simply overlapping.

These performance objectives fall under the category of *rate* criteria that are similar to *capacity* criteria. However, we need to draw an important clarification here. Capacity is a quantity that has precise meaning in the information-theoretic sense and can be defined (or computed) for very few *network* cases. Notably, it is known for multiple access channels, broadcast channels, and, up to a degree, for the simple interference and relay channels. We will not be concerned with the capacity region for general networks because of the lack of exact knowledge and understanding of its nature in general networks. In fact, the relationship between the capacity region and the two throughput regions mentioned above is not known except in very few cases. In [3] there has been a study (far from complete and conclusive) of the relationship among these three regions with some generality. We should note that in the throughput case, as opposed to the stable throughput case, it is assumed that the source nodes have unlimited reservoirs of traffic and never face the possibility of underflow (i.e., lack of packets to transmit at any time). The case of infinite reservoir of traffic is referred to as *saturated*, or *backlogged*, user case.

A variation of this theme occurs when we realize that maximizing throughput may result in very uneven (and, hence, unfair) values of the individual source/destination rate components. This has given rise to the so-called *fairness criteria* that try to strike a balance between overall high throughput and individual throughput components. There are several definitions of fairness, including *max-min* fairness that tries to maximize the smallest rate component and *proportional fairness* that is equivalent to trying to maximize the product of the individual rate components. More generally, a *utility* function may be chosen, that is more general than any one of the rate-related quantities discussed so far, and that consists of individual components for each source/destination pair, just as in the case of throughput. In Section VI we will address such utility maximization formulations in some detail.

We should note at this point that the rate (i.e., throughput) related criteria presuppose that the network operates in a *stationary* and *ergodic* environment so that the respective rates are well-defined and make sense. However, most wireless networks are subject to arbitrary and capricious changes over time that include mobility, variable fading, addition and deletion of users as new users enter and leave the network, and, often, finite lifetime. In such cases, throughput values may not be well defined in general and, hence, a thoroughly different set of objectives needs to be considered. One possibility that attempts to identify an *ultimate capability* of a network, akin to *capacity* or *throughput*, is to consider an initial traffic volume loading on the source nodes and then find the scheduling policy that empties the network in minimum time. Such a formulation is attractive, not only because it makes sense in its own right, but also because it is consistent with what we call *minimum-length scheduling* in all cases (ergodic and stationary, or not) and that is almost equivalent to a form of throughput maximization.

In the case of minimum-length scheduling, for which there has been a significant amount of work (see e.g., [4]–[8]), the formulation of the problem allows for two equivalent choices. Either we assume that a certain rate demand is given (which we need to satisfy with minimum-length periodic scheduling actions) or a fixed volume of traffic, e.g., a file, is given that must be delivered to the destinations in minimum time (which is akin to the formulation described above).

All in all, we see that there is no unique way of formulating the most fundamental of problems in wireless network operation. There are similarities among the formulations, and among the solutions, but there are significant differences too. Our selected examples in Section V will provide insight into the dimensions of the minimum-length scheduling problem.

III. CONSTRAINTS IN THE SCHEDULING PROBLEM

Once the objective function is set, we need to identify the constraints that the solution, i.e., the scheduling policy, must satisfy. These constraints define the possibilities of simultaneous transmissions that are guaranteed to be successful. In the early days of the study of the scheduling problem the criterion of success was very simple and ignored completely the delicate nature of the wireless channel. It simply assumed that if a node is within a given and fixed range from a transmitting node, and if no other node at this or smaller range is simultaneously transmitting, then the message is successfully received at the node in question. In addition, if another node within range was also transmitting then all transmissions would be unsuccessful at the receiving node. This was known as the *disc* or *collision* channel model. It was simplistic, allowed for no recognition of the effects of power, fading, rate, etc., and hence focused exclusively on the MAC layer. Even so, the difficulties in solving the scheduling problem were formidable, primarily because of the inherent combinatorial complexity of the problem.

Subsequently, there have been refinements of the success criteria. In fact, it is possible to have a completely general definition of the criteria of success that simply specifies the sets of different links that can be simultaneously transmitting successfully, without explaining the underlying mechanism that makes any such set a "feasible" transmission set. Better yet, it is possible to incorporate explicitly known properties of the wireless communication channel to define a cross-layer criterion of success that, in turn, defines what these feasible sets are. One such example is the so called SINR criterion. Borrowing from the properties of the additive white Gaussian noise (AWGN) and its use in CDMA systems, we say that a transmission is successful if the ratio of the received signal power to the sum of the thermal noise power and the total received interference (SINR) exceeds a certain threshold γ. The value of this threshold depends on many parameters of the communication system but, most notably, it is a known decreasing function of the target BER, or an increasing function of the transmission rate, all the rest of the parameters being fixed [9]. This function can be quite complex. The simplest known form is for the binary phase shift keying (BPSK) modulation scheme without error control coding. It also becomes rather simple, but approximate, for the Shannon limit case. In the BPSK case γ is given by

$$\gamma = \frac{r}{2}\left[Q^{-1}(z)\right]^2$$

and in the Shannon theoretic case the threshold γ is given by

$$\gamma = 2^r - 1$$

where z is the target probability of bit error, $Q(z)$ is the Gaussian complementary cumulative distribution, r is the instantaneous transmission rate (bits/sec), and where the bandwidth is assumed to be equal to one. Note that, by successful communication in BPSK we mean that the probability that any bit is received in error is bounded by some maximum probability z. In the Shannon limit case this probability is zero.

It is worth mentioning here that in [2] an original formulation was considered that led to a fairly general solution to the scheduling problem under the objective of maximizing the stable throughput region for arbitrary link activation constraints. Note that, by stable throughput region we mean the set of arrival rates in the network such that the network queues do not grow without bound and that guarantee finite average delays for all packets residing at the network queues. The solution ended up being known as the "back-pressure algorithm" (BPA). The nature of the solution and its relative simplicity for a problem of considerable complexity has provided broad insights into the scheduling problem and led to several generalizations [10], including a recent one that employs network coding [11]. In simple terms, the scheduling policy that is guaranteed to *stabilize* a system, if the offered load rates are within the, unknown, stable throughput region, is governed by the following rule: Every node in the network maintains a set of queues, one for each possible destination, or "commodity," in the network. If it is enabled to transmit in a given time slot it should select the head of the line packet of the queue which has the largest backlog difference with respect to the queues of this particular commodity at its neighbors. Next, it must send this packet to the neighbor whose queue maximizes the backlog difference. What decides whether the node can be activated or not in the given time slot depends on the values of the weighted sums of these differential backlogs across all members of the feasible sets. The feasible set with the largest such cumulative differential backlog will be the one to be activated. For more details on this approach, which is rather esoteric and will not be part of our main development in this paper, see [2], [10], and [12].

IV. SYSTEM MODELING

In this section we describe the network model under which the various instances of the scheduling problem are being studied. Specifically, we consider wireless networks consisting of T sources of traffic and D destinations. We denote by $\mathcal{T} = \{1, 2, \cdots, T\}$ and $\mathcal{D} = \{1, 2, \cdots, D\}$ the sets of sources and destinations, respectively. The objective of this paper is to focus on the understanding of the fundamental concepts that are involved in the scheduling of transmissions in wireless networks. Thus, we will restrict our attention to a basic model of single-hop networks with time-invariant links. This model, albeit simplified, is non-trivial since it captures effectively the adverse effects of interference in multiple access.

We may consider multicast traffic, which captures the special cases of unicast and broadcast transmissions. In the case of unicast traffic, each source is associated with a distinct destination (receiver). In the case of multicast traffic, each source is associated with a multicast session and wishes to transmit to a set of destinations. However, in order to keep the level of complexity low we will limit ourselves to the unicast case in which $D = T$. Furthermore, to simplify notation we will assume T source/destination pairs, where the kth source wishes to communicate with the kth destination as shown in Fig. 1.

We denote the transmission power level of the kth source at time t by $P_k(t)$. We also denote by $\mathbf{P}(t) = (P_k(t), \forall k \in \mathcal{T})$ the T-dimensional power vector at all sources at time t. We assume that each network source k can either operate at its maximum transmission power $P_k^{\max}$ or remain silent. Thus, we have that

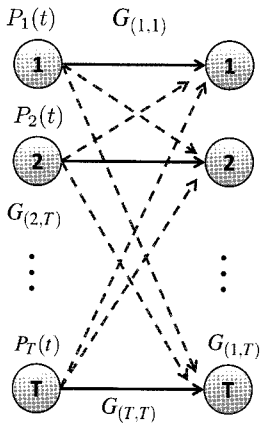

Fig. 1. A single-hop network of T source/destination pairs.

$P_k(t) \in \{0, P_k^{\max}\}$. We denote the thermal noise power level at every receiver $d \in \mathcal{D}$ by N_d.

We also denote by $\mathbf{G} = \{G_{(i,j)},\ i \in \mathcal{T},\ j \in \mathcal{D}\}$ the channel conditions between every source i and each destination j in the network, i.e., we assume that the channel effect is due to pure path loss. However, it is feasible to permit the variables $G_{(i,j)}$ to be random and time-varying, and in this way, incorporate the effects of fading or mobility.

As discussed previously, the scheduling constraints imposed by the physical layer are incorporated in our model through the SINR criterion. Specifically, under the SINR interference model a transmission from source k to destination k is successful if the ratio of the received signal power to the sum of the thermal noise and the total interference exceeds the required threshold. Although the exact value of this threshold depends on various communication related parameters, here we only focus on the dependence of this threshold on the *transmission rate*. Let $\gamma_{t,k}(r_k(t))$ denote the SINR threshold value at receiver k that must be met or exceeded in order to receive successfully at rate $r_k(t)$ at time t from source k. Then, at slot t when all sources operate at powers given by the power vector $\mathbf{P}(t)$, source k successfully transmits to its corresponding destination at rate $r_k(t)$ if

$$\mathrm{SINR}_{(k,k)}^{\mathbf{P}(t)}(t) := \frac{P_k(t)G_{(k,k)}}{N_k + \sum_{j=1,j\neq k}^{T} P_j(t)G_{(j,k)}} \geq \gamma_{t,k}(r_k(t)). \tag{1}$$

Our model is general and can capture the cases of receivers with multi-packet reception capabilities, so that at any given time a receiver can receive successfully from more than one transmitting source as long as all the corresponding SINRs exceed the required thresholds [13], [14]. Permitting multi-packet reception capability at the receivers allows us to assign several sources to the same destination, that is $T > D$. For simplicity, in this paper we restrict our attention to single-packet reception at the receivers, that is a destination can receive only from a *single* source at any given time.

Since the maximum transmission rate is an increasing function of the SINR threshold (see e.g., [9]) the following fundamental *trade-off* arises: If the transmission rates at the sources are increased, the corresponding minimum required values of the SINR thresholds consequently rise which restricts the number of successful transmissions that can occur concurrently (at higher rates). On the other hand, if the transmission rates at the sources are decreased, the required threshold values also decrease allowing a higher number of sources to operate concurrently (but at lower rates). Depending on the performance objective and the network parameters, such as the transmission powers, channel conditions, etc. it may be preferable to allow more concurrent transmissions (less time-sharing) at lower rates than allowing fewer concurrent transmissions (more time-sharing) at higher rates.

Under this restricted scheme of binary transmission, either at zero or maximum power, there exist $2^T - 1$ "maximal" feasible rate vectors, i.e., rate vectors whose components cannot be increased any further without violating the SINR condition. These correspond to taking all possible subsets of the set of sources that can be activated and to each such subset assigning the maximum transmission rates that allow them to jointly communicate successfully. We name these $2^T - 1$ rate vectors as *actions*. We denote the set of all actions by $\mathcal{R}$, where the cardinality of the set $\mathcal{R}$ satisfies $|\mathcal{R}| = 2^T - 1$. Under a given action, those sources that are assigned a zero rate are not activated. Thus, each action captures exactly which sources are activated and their corresponding rates and is therefore able to capture the scheduling decision at each time instant.

In the following sections, we first present an optimal scheduling solution for the minimum-length scheduling problem where the objective is to satisfy a given demand at the destinations (Section V) and then we proceed with scheduling under the objective of utility maximization (Section VI).

V. THE MINIMUM-LENGTH SCHEDULING PROBLEM

Let us assume that action i is employed for a duration τ_i. Then, a *schedule* $\mathcal{S}$ is defined to be the set $\mathcal{S} = \{(i, \tau_i), i \in \mathcal{R}\}$ specifying the duration for which each action in the set $\mathcal{R}$ is used. The *length* of a schedule is $L = \sum_{i\in\mathcal{R}} \tau_i$. In Fig. 2, a network of 2 sources and 2 destinations is depicted. Action 0 corresponds to *both* sources being active concurrently and action i, $i = 1, 2$, corresponds to *only* source i being active. Then, a schedule $\mathcal{S}$ is a sequence of actions j, $j \in \{0, 1, 2\}$ and corresponding durations τ_j. In this example, action 0 is used for 100 secs, action 1 for 150 secs, and action 2 for 40 secs and thus, the length of this schedule is $L = 290$ secs. Note that due to the assumption that the channel is time invariant only the overall duration that each action is used matters and not the exact location in time. Furthermore, for the case of time-invariant links the order in which the different actions are taken is immaterial.

In this section, we are interested in finding schedules of minimum length to satisfy a given demand at the destinations. As we mentioned previously, the demand can be in terms of a given volume of traffic (e.g., a file in bits) or in terms of a minimum rate requirement.

Furthermore, as mentioned earlier, we assume that the number of sources is equal to the number of destinations, i.e., $D =$

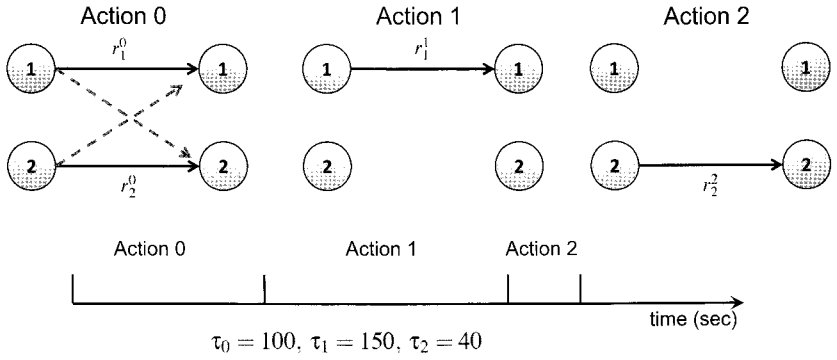

Fig. 2. A network of two source/destination pairs and a corresponding schedule. Action 0 is used for time $\tau_0 = 100$ secs, action 1 is used for $\tau_1 = 150$ secs, and action 2 is used for $\tau_2 = 40$ secs.

T. Moreover, we consider unicast traffic, where each source $k \in \mathcal{T}$ is paired up with a distinct destination in $\mathcal{D}$, forming T communication pairs. This is without loss of generality since, as it was recently exhibited in [15] and [16], our results can be generalized to multicasting as well.

Let us first assume that the set of feasible actions is restricted to contain actions activating a single source at any given time as in time division multiple access (TDMA). This corresponds to reducing the set of actions from $2^T - 1$ to simply T. Under this scheme, each transmitting source operates at its maximum possible rate due to the lack of interference. We denote the rate when only source k is activated by r_k. We first consider the case where the traffic demand is described in terms of a finite amount of data traffic (e.g., a file in bits). Let each network source k have d_k bits to deliver to its corresponding destination. Then, each source k must be activated for a duration of $\tau_k = d_k/r_k$ secs in order to deliver all of its traffic and the resulting schedule length is $L = \sum_{k\in\mathcal{T}} d_k/r_k$. This coincides with the minimum schedule-length. Thus, in this case there is nothing to solve for.

Alternatively, assume that the demand is with respect to a minimum rate requirement that needs to be met for every source. Let this rate for source k be R_k bits/sec. Note that R_k is the average rate and not the instantaneous rate in every time. Then, it must be true that $r_k\tau_k/L \geq R_k$ for every source k which can be rewritten as

$$\tau_k \geq \frac{R_k L}{r_k}, \quad \forall k \in \mathcal{T}.$$

By summing the above inequality over all the sources we obtain

$$L = \sum_{k\in\mathcal{T}} \tau_k \geq L \sum_{k\in\mathcal{T}} \frac{R_k}{r_k}$$

which finally yields

$$\sum_{k\in\mathcal{T}} \frac{R_k}{r_k} \leq 1.$$

This is a feasibility condition, that is, the minimum rate requirement is satisfied for every source k only if this condition holds. Hence, subject to the feasibility issue, easily settled by the above inequality, again there is nothing that needs to be solved.

Although, the solution to both formulations of the minimum-length scheduling problem seems to be trivial in the case of TDMA scheduling, this is not the case when we allow more general sets of scheduling actions. Let us now assume that the set $\mathcal{R}$ contains all of the $2^T - 1$ rate control and scheduling actions corresponding to activating all possible subsets of sources in the network. In this case, since different links have different rates, it is now non trivial which sequence of actions have to be used and for how long, so that the demand is met at the destinations. Let, under this generalized framework, the rate of source k under action j be r_k^j. Let us also restrict our attention to the minimum-length scheduling problem when the demand is in terms of a finite, fixed amount of data traffic. Then, the minimum-length scheduling problem can be stated as

$$\textbf{minimize}: \quad \sum_{i\in\mathcal{R}} \tau_i \tag{2}$$

$$\textbf{subject to}: \quad d_k \leq \sum_{i\in\mathcal{R}} \tau_i r_k^i, \qquad \forall k \in \mathcal{T} \tag{3}$$

$$\tau_i \geq 0, \qquad i \in \mathcal{R}, \tag{4}$$

i.e., the objective is to schedule the sources so that all traffic gets delivered to its intended destinations in minimum time.

This minimization problem is a linear program that has a relatively small number of constraints and a large number of variables as in the formulations in [5]–[7]. It is easy to see that the number of variables grows exponentially in the number of source/destination pairs in the network. Thus, although efficient algorithms exist that solve linear problems in polynomial time, still under this framework the solution to the linear program above remains prohibitively complex. Note also that the number of vertices at the boundary of the region over which the optimization takes place is extremely large, which renders their enumeration an extremely difficult task.

To simplify the minimum-length scheduling problem we are going to follow two approaches. First, we are going to discuss an approach based on column generation where the vertices at the boundary of the optimization region will be generated "as we go along," and thus provide a simplification to the problem. Then, we will consider an alternative that restricts the set of feasible

actions under consideration to only those that correspond to the two extreme modes of operation, namely "one-at-a-time" and "all-at-once."

The method of column generation [17] is a classical iterative algorithm that is known to solve efficiently large, linear, or integer, programming problems. It has been shown to reduce the computational complexity of solving "large" problems, i.e., problems that have a small number of constraints and a large number of variables.

The main idea behind the technique of column generation is to decompose the original *full master problem* into a problem of smaller dimension, the *restricted master problem* and a *subproblem*, also called as the *pricing* problem. The latter, can be a linear or an integer program itself, depending on the specifics of the original problem. The benefits of this decomposition is to iteratively solve two problems that are, typically, easier than the original full master problem. Since the restricted master problem has fewer constraints than the original, the optimal solution of the restricted problem is an upper bound on the optimal solution of the full master problem.[1] During each iteration, the algorithm tries to identify a column (action) from the ones that are not in the restricted problem that has the *most negative reduced cost*.[2] This column is added in the columns of the restricted problem and the optimal solution of the updated restricted problem is computed. This iterative procedure is repeated until the optimal solution of the restricted problem cannot be improved by adding new columns.

Since, it is expected that not all columns will be in the optimal solution the average running time of column generation is shorter than that of the original problem. However, the worst case performance can be significantly worse. Note that the above results can be easily generalized to account for routing as was shown in [7], [18].

An alternative approach to solving the minimum-length scheduling problem is to restrict the space of actions under consideration. For instance, instead of considering all the $2^T - 1$ rate control and scheduling decisions corresponding to activation of all possible subsets of sources, we focus on two simple schemes as in [16], [19]. Specifically, these schemes include two extreme modes of operation, namely (i) the simultaneous activation of *all* T sources operating successfully and at instantaneous rates that ensure that all SINR threshold inequalities are satisfied (we call this operation "all-at-once" or "action 0") and (ii) the individual activation of each source separately (we call this operation "one-at-a-time" or "action k" when source k is activated). The above two modes of operation yield a total of $T + 1$ actions.

Restricting attention to these two modes of operation is somewhat natural since it permits comparison between two extreme cases, namely the cases of "all-at-once" and "one-at-a-time" operation. Although under action 0 all sources operate concurrently, their individual rates will likely be low so that the SINR criterion is satisfied under the heavy interference that occurs when they all transmit simultaneously at their maximum power levels. On the other hand, although under action k the instantaneous rate of the kth source will likely be much higher (than the corresponding rate under concurrent operation), it may be lower on the average due to the time sharing. Although this represents a severe restriction of the action space, it is expected to provide an insight into the trade-off between concurrent and individual activation.

The minimum-length scheduling problem over the restricted action space is given by

$$\textbf{minimize}: \quad \sum_{i=0}^{T} \tau_i \tag{5}$$

$$\textbf{subject to}: \quad d_k \leq \tau_k r_k^k + \tau_0 r_k^0, \quad \forall k \in \mathcal{T} \tag{6}$$

$$\tau_i \geq 0, \quad i \in \{0, \cdots, T\}. \tag{7}$$

The constraints simply specify that the delivered traffic can be at least as large as the given volume demand. Clearly, the reduction in the state space by resorting to these two schemes is very substantial since from a space of $2^T - 1$ actions we retain only $T+1$ of them. So although such an approach does not hope for global optimality, it is a useful heuristic of polynomial time (instead of exponential) which yields a tractable and, hopefully, insightful problem. The following proposition characterizes an optimal scheduling and rate control policy that solves (5)–(7).

Proposition 1: Let $\tau_0^\star, \tau_1^\star, \cdots, \tau_T^\star$ be the optimal solution to (5)–(7). Then, we have the following:

1. If

$$\sum_{k=1}^{T} r_k^0 / r_k^k \leq 1,$$

Action k is chosen ($k = 1, \cdots, T$) for a duration of

$$\tau_k^\star = d_k / r_k^k$$

while action 0 is never employed, i.e.,

$$\tau_0^\star = 0.$$

2. If

$$\sum_{k=1}^{T} r_k^0 / r_k^k > 1,$$

then a subset $\mathcal{J}$ of sources is selected such that for every $k \in \mathcal{J}$, action k is chosen for a duration of

$$\tau_k^\star = d_k - \tau_0 r_k^0 / r_k^k$$

and action 0 is chosen for a period of

$$\tau_0^\star = \max_{i \in \mathcal{T} \setminus \mathcal{J}} d_i / r_i^0.$$

Thus, the quantity $\sum_{k \in \mathcal{T}} r_k^0 / r_k^k$ determines, in a sense, the relative degree of interference in the network. Clearly, for any transmitting source the instantaneous rate under concurrent operation is no greater than its corresponding rate under individual transmission. If it is also true that $\sum_{k \in \mathcal{T}} r_k^0 / r_k^k \leq 1$, then the transmitting sources interfere among themselves sufficiently, so that their rates under concurrent operation are much lower than the corresponding rates under individual operation. Hence, when

[1] Accordingly, when we solve a maximization problem it will be a lower bound.

[2] Similarly, if the problem is posed as a maximization then the column with the *most positive cost* is sought.

$\sum_{k\in\mathcal{T}} r_k^0/r_k^k \leq 1$, the optimal policy would never activate all sources concurrently ($\tau_0^\star = 0$); instead the optimal scheduling and rate control solution is to activate a single source at a time as in a TDMA scheme. On the other hand, if $\sum_{k\in\mathcal{T}} r_k^0/r_k^k > 1$, the interference among the sources when they concurrently transmit is not so severe, and hence, the individual rates under concurrent operation result in levels that are "comparable" to those achieved under individual operation. Thus, the optimal policy will employ action 0 ($\tau_0^\star > 0$). To completely characterize the policy we need to specify the set $\mathcal{J}$, which results from the following proposition.

Proposition 2: Consider an ordering of the sources in decreasing order of the values d_k/r_k^0 for every $k \in \mathcal{T}$. Let the corresponding indexing of the sources be $\{\ell_k\}_{k=1}^T$, that is $d_{\ell_1}/r_{\ell_1}^0 \geq \cdots \geq d_{\ell_T}/r_{\ell_T}^0$. Then, the set $\mathcal{J}$ contains those sources with the highest d_k/r_k^0 ratios and the cardinality $|\mathcal{J}|$ of the set $\mathcal{J}$ is given by

$$|\mathcal{J}| = \underset{k\in\{0,\cdots,T\}}{\arg\min} \left\{ \frac{d_{\ell_{k+1}}}{r_{\ell_{k+1}}^0} + \sum_{j=\ell_1}^{\ell_k} \frac{d_j r_{\ell_{k+1}}^0 - d_{\ell_{k+1}} r_j^0}{r_j^j r_{\ell_{k+1}}^0} \right\}.$$

From the above we conclude that the set $\mathcal{J}$ contains the sources with the $|\mathcal{J}|$ highest values of d_k/r_k^0, where $|\mathcal{J}|$ is given by Proposition 2. Hence, an optimal scheduling and rate control policy individually activates the sources that either have a very high initial demand or whose rates under concurrent operation are very low, e.g., due to excessive amounts of interference caused by other concurrent transmissions. Those sources must be further assisted towards emptying their queues by being granted individual access to the channel. This result is very intuitive and clarifies the quantitative trade-off between "all-at-once" and "one-at-a-time" activation.

A different formulation of the minimum-length scheduling problem is based on shortest paths. In [8] we considered networks of time-invariant channels where the channel effect is due to pure path loss. Moreover, instead of looking at *durations of time* where each action is used we considered a slotted-time model and our objective was to minimize the number of time slots required to empty a fixed amount of data (volume) from the buffers of the source nodes. Under these assumptions we formulated the minimum-length scheduling problem in terms of obtaining a shortest path on a directed acyclic graph (DAG). Furthermore, in [20] we extended our results to time-varying channels. We formulated the minimum-length scheduling problem through stochastic shortest paths and provided an optimal policy through dynamic programing. The notions of state space reduction are applicable in these models as well, as one can reduce the set of feasible actions associated with each state to reduce complexity. Clearly, by restricting further the set of feasible rate control actions we can obtain lower complexity algorithms at the cost of obtaining sub-optimal solutions. From our formulations in [8] and [20], it is easy to observe that the minimum-length scheduling problem is tightly related to problems of minimum draining time. Furthermore, although in [20] we assumed that the policy has exact knowledge of the current channel conditions before making a scheduling and rate control decision, this information in general may be unavailable. In [15] we extended our results by considering policies that *only* have access to a *probability distribution* of the channel process and formulated the minimum-length scheduling problem by employing the theory of partially observable Markov decision processes (POMDPs). In summary, we see that even one version of the scheduling problem (i.e., minimum-length scheduling for finite volume delivery over single-hop networks) is very polymorphic and admits a variety of solutions. Note that it is this formulation that holds promise towards evaluating the ultimate capabilities of networks that are non-ergodic and non-stationary. Of course, in that case, there will be no "rate" values except for very short time intervals and will be dynamically changing in time. We are far from having formulated the non-ergodic and non-stationary case precisely, but this approach holds promise.

VI. UTILITY MAXIMIZATION

In this section, we consider a different performance objective, that of maximizing the total received user utility. The network model is exactly as before with the exception that here we assume a slotted-time model. Again, although in the following section we consider unicast traffic, the multicast case can be just as easily accommodated (see e.g., [16]). Unlike the previous section, here we assume a saturated traffic model in which each multicast source always has enough data to send whenever it is activated. Note that, this model is to be distinguished from bursty traffic models where the aim is to maintain the stability of the queues in the network and guarantee finite expected delays.

One more note regarding the case of multicast traffic is the following. Although under unicast traffic, throughput is unambiguously defined as the rate at which data is delivered to a receiver successfully, this is not the case for multicast traffic. This is due to the fact that in a multicast transmission it is possible that *some*, but *not* all, of the receivers decode the transmission successfully. Further, depending on the characteristics of the underlying application (e.g., in the case of video traffic) a multicast packet may not be required to be received successfully by all of the receivers, which complicates the definition of successful transmission for multicast traffic. We avoid these complications here by focusing only on unicast traffic. Accordingly, when the transmission powers at time slot t at the sources are given by the power vector $\mathbf{P}(t)$ we say that source k transmits successfully at rate $r_k(t)$ if its intended destination d satisfies the SINR criterion of (1).

In a problem of utility maximization the objective is to solve the following problem:

$$\max_{\mathbf{r}\in\text{co}(\bar{\mathcal{R}})} \sum_{k\in\mathcal{T}} U(r_k) \tag{8}$$

where the region $\bar{\mathcal{R}}$ is the set of long-term average feasible rates, $\text{co}(\cdot)$ indicates the convex hull of the region $\bar{\mathcal{R}}$, $\mathbf{r} = (r_1, \cdots, r_T)$ is a long-term average feasible rate vector, and $U(\cdot)$ is a utility function. Note that the rate region $\bar{\mathcal{R}}$ can be arbitrary. It can be defined as the set of average rate vectors where the corresponding instantaneous rates are such that all activated sources jointly satisfy the SINR criterion. In addition, this model allows for a more accurate inclusion of the physical layer by appropriately redefining the region $\bar{\mathcal{R}}$. For instance, we can allow to selectively control the coding method, the target bit error rate goal, etc. and propagate these characteristics of the physical layer in

the MAC decisions. This extra degree of freedom can only improve the system performance by expanding the set of rates that the achievable rate region contains. Moreover, taking the convex hull of these rates guarantees that the optimization takes place over a convex set.

In such problems of maximizing a user utility, a commonly used utility function is that of $\alpha-$fairness. A utility function is called α-fair [21] if it satisfies

$$U^{\alpha}(r) = \begin{cases} \log(r), & \text{if } \alpha = 1 \\ (1-\alpha)^{-1} r^{1-\alpha}, & \text{otherwise.} \end{cases} \tag{9}$$

This is a useful utility function that has attracted a lot of attention recently since for different parameterizations of α it yields some commonly used performance objectives. In particular, the parameter α captures, in a sense, the amount of "fairness" that the utility function provides to the users. As α increases, the fairness in the optimal solution with respect to the different users increases as well. As an example, for $\alpha = 0$ it yields the objective of sum throughput maximization which tries to maximize the aggregate system efficiency (rate) and thus, treats the users facing poor channel conditions unfairly by prohibiting them from accessing the channel. When $\alpha = 1$ it leads to the objective of *proportional fairness*, and as α increases to infinity it leads to the ultimate fairness, that is the *max-min* fairness.

In the rest of this section we are going to focus on the objective of proportional fairness that corresponds to maximizing the sum of the logarithms of the users' rates (or equivalently the product of the individual rate components). Although this problem is convex, the number of possible actions, and hence constraints, increases exponentially in the number of sources, just as before. Therefore, although when the number of transmitting sources in the network is sufficiently small numerical solutions can be obtained (for example, through interior-point methods [22]), computing the optimal solution analytically is infeasible.

For this reason, here, instead of considering the $2^T - 1$ rate control and scheduling decisions we restrict the set of feasible actions to the ones obtained by the aforementioned simple schemes of communication "one-at-a-time" and "all-at-once." As before, we denote by r_k^k the instantaneous transmission rate of the kth source when it transmits individually (action k) and by r_k^0 its corresponding rate when all sources operate concurrently (action 0). We also define the vector $\boldsymbol{\omega} = (\omega_0, \cdots, \omega_T)$ to be a *probability distribution* over the restricted set of rate control and scheduling actions, i.e., $\omega_i \geq 0$, $\forall i \in \{0, \cdots, T\}$ and $\sum_{i=0}^{T} \omega_i = 1$. That is, we randomize the policy decision so that in every slot, action j is taken with probability ω_j for $j = 0, \cdots, T$. This formulation by-passes one aspect of combinatorial complexity that arises when we associate each action in a deterministic way with each slot. We assume that such probability distribution exists, e.g., by requiring ergodicity on the action selection. This is illustrated in Fig. 3 for a simple network of 2 source/destination pairs. As opposed to Fig. 2, here time is slotted and a probability distribution over the various actions is computed by dividing the number of time slots that each action is employed with the overall number of slots in the frame.

It is easy to see that under a time invariant channel the average or *effective rate* of each source k depends on the chosen probability distribution vector $\boldsymbol{\omega} = (\omega_0, \cdots, \omega_T)$ and can be written as

$$r_k(\boldsymbol{\omega}) = \sum_{j=0}^{T} r_k^j \omega_j.$$

We are interested in obtaining an optimal probability distribution over the restricted set of actions, so that the *effective* rate of each source k is assigned in a proportionally fair way. This can be obtained by solving the following problem:

$$\max_{\boldsymbol{\omega}} \sum_{k \in \mathcal{T}} \log(\omega_0 r_k^0 + \omega_k r_k^k) \tag{10}$$

s.t.

$$\omega_j \geq 0, \quad \forall j \in \{0, 1, \cdots, T\}, \tag{11}$$

$$\sum_{j=0}^{T} \omega_j = 1. \tag{12}$$

Before we characterize the optimal policy solving (10)–(12), we provide some useful definitions. Let $\mathcal{Z}$ be a subset of the set $\mathcal{T}$, such that for every $j \in \mathcal{Z}$ it is true that $\omega_j > 0$ and let $|\mathcal{Z}|$ denote the cardinality of the set $\mathcal{Z}$. Then, we have the following result:

Proposition 3: Let $\boldsymbol{\omega}^{\star} = (\omega_0^{\star}, \cdots, \omega_T^{\star})$ be the solution to (10)–(12) above. Then, we have:

1. If

$$\sum_{k \in \mathcal{T}} r_k^0 / r_k^k \leq 1,$$

each source $k \in \mathcal{T}$ is scheduled to transmit individually with probability

$$\omega_k^{\star} = \frac{1}{T}, \quad \forall k \in \mathcal{T},$$

and the probability of concurrent operation is zero , i.e., $\omega_0^{\star} = 0$.

2. If

$$\sum_{k \in \mathcal{T}} r_k^0 / r_k^k > 1,$$

the optimal policy is of a threshold type with threshold $\tilde{R}(\mathcal{Z})$ given by

$$\tilde{R}(\mathcal{Z}) = \frac{1 - \sum_{j \in \mathcal{Z}} r_j^0 / r_j^j}{T - |\mathcal{Z}|}. \tag{13}$$

Specifically,

(a) A source $j \in \mathcal{T}$ is scheduled to transmit individually with probability $\omega_j^{\star} > 0$ (i.e., j is individually activated and belongs in $\mathcal{Z}$) given by

$$\omega_j^{\star} = \frac{1}{T} \left(1 - \sum_{i \in \mathcal{T} \setminus \mathcal{Z}} \frac{r_j^0 / r_j^j}{1 - \sum_{j \in \mathcal{Z}} r_j^0 / r_j^j} \right), \tag{14}$$

if and only if

$$\frac{r_j^0}{r_j^j} < \tilde{R}(\mathcal{Z}). \tag{15}$$

(b) All sources operate concurrently with probability $\omega_0^{\star}$ given by

$$\omega_0^{\star} = \frac{T - |\mathcal{Z}|}{T \left(1 - \sum_{j \in \mathcal{Z}} r_j^0 / r_j^j \right)}. \tag{16}$$

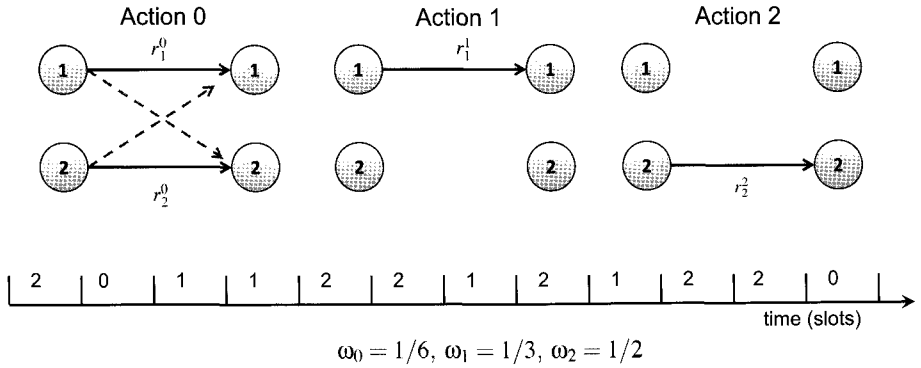

Fig. 3. A network of two source/destination pairs. Action 0 is used with probability $\omega_0 = 1/6$, action 1 is employed with probability $\omega_1 = 1/3$, and action 2 is used with probability $\omega_2 = 1/2$.

Again, if it is true that $\sum_{k\in\mathcal{T}} r_k^0/r_k^k \leq 1$, then we expect that concurrent transmissions cause excessive amounts of interference to each other and thus the optimal policy would not activate all sources concurrently ($\omega_0^\star = 0$). On the other hand, if $\sum_{k\in\mathcal{T}} r_k^0/r_k^k > 1$, concurrent transmissions do not cause excessive interference to each other, and thus, the optimal policy assigns a positive probability to action 0.

Proposition 3 characterizes the optimal solution based on the threshold function $\tilde{R}(\mathcal{Z})$ which itself is a function of the set $\mathcal{Z}$. Hence, in order to completely characterize the optimal policy we need to characterize the composition of $\mathcal{Z}$. Note that since the optimal policy is of threshold type, the cardinality $|\mathcal{Z}|$ of the "individually activated" set suffices to completely determine the set $\mathcal{Z}$ itself, provided we label the sources appropriately. To simplify the notation in the sequel we will write $\tilde{R}(j)$ to denote $\{\tilde{R}(\mathcal{Z}) : |\mathcal{Z}| = j\}$.

Let us reorder the sources with respect to their corresponding values of the ratios r_j^0/r_j^j, $j \in \mathcal{T}$ in increasing order, i.e.,

$$\frac{\tilde{r}_1^0}{\tilde{r}_1^1} \leq \frac{\tilde{r}_2^0}{\tilde{r}_2^2} \leq \cdots \leq \frac{\tilde{r}_T^0}{\tilde{r}_T^T} \tag{17}$$

where the rates $\tilde{r}_j^0$ and $\tilde{r}_j^j$ denote the quantities r_j^0 and r_j^j, respectively of the jth source under the new ordering. From now on, unless otherwise stated, the source j is the jth source under this new ordering. We will make use of the following property of the threshold function $\tilde{R}(j)$ to obtain the cardinality of the set $\mathcal{Z}$.

Proposition 4: The cardinality of the set $\mathcal{Z}$ under the optimal policy specified in Proposition 3 is given by the following:

$$|\mathcal{Z}| = \underset{\ell\in\{0,1,\cdots,T\}}{\arg\max} \frac{1 - \sum_{j=1}^{\ell} \tilde{r}_j^0/\tilde{r}_j^j}{T - \ell}. \tag{18}$$

Propositions 3 and 4 extend our prior work [19] where we had assumed that for every unicast source $j \in \mathcal{T}$ the rates under individual operation r_j^j were all equal to each other. Also, from Propositions 3 and 4 it follows that the set $\mathcal{Z}$ contains the $|\mathcal{Z}|$ sources with the lowest values of the ratios r_j^0/r_j^j for $j \in \mathcal{T}$. Hence, in the optimal solution the sources that are selected to be activated individually are the most "disadvantaged" sources, i.e., those that can only achieve very low rates under concurrent operation compared to individual transmission.

A generalization of the above results to time-varying channels was given in [13] where we presented an optimal on-line rate and power control algorithm, solving a general problem of utility maximization. The optimal policy in [13] is assumed to have access to perfect channel information at every scheduling decision instance. This assumption is relaxed in [14] where we introduced an optimal on-line, rate and power control policy operating under channel uncertainty. In the model in [14] the algorithm takes decisions based *only* on the stationary probability distribution with which the different channel states occur. Again, it can be seen that in this very different formulation of the scheduling problem a rich set of possibilities arises and leads to insightful observations and results even when we reduce the search space of the problem and give up optimality.

A. An Example

In this subsection, we present a few numerical results under the objective of proportional fairness. We consider a simplified network example with the following properties:

1. The maximum transmission powers $P_k^{\max}$ of all sources $k \in \mathcal{T}$ are equal to each other, i.e.,

$$P_k^{\max} := P, \qquad \forall k \in \mathcal{T}.$$

2. The path losses between every source/destination pair are all equal to each other, i.e.,

$$G_{(k,k)} := G^{\mathrm{S}}, \qquad \forall k \in \mathcal{T}, k \in \mathcal{D}.$$

3. The sum of the path losses from all sources at any given destination j is equal to G^{I}, i.e.,

$$\sum_{k\in\mathcal{T}} G_{(k,j)} := G^{\mathrm{I}}, \qquad \forall j \in \mathcal{D}.$$

4. The noise power levels at all receivers $d \in \mathcal{D}$ are equal to each other, i.e.,

$$N_d := N, \qquad \forall d \in \mathcal{D}.$$

Then, it follows that the instantaneous rates of all sources are equal to each other under each scheme of operation, i.e.,

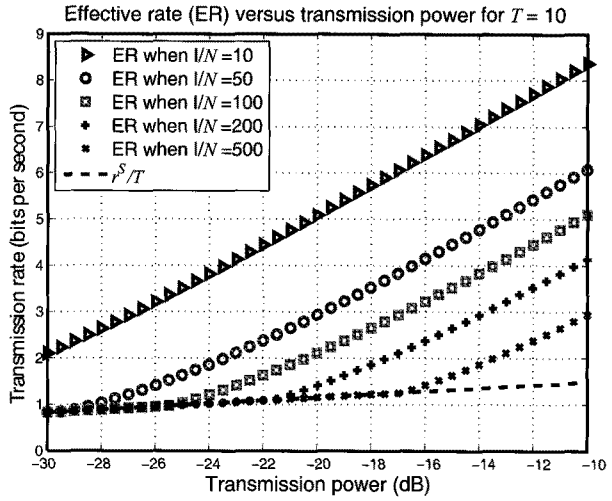

Fig. 4. Effective rate as a function of the transmission power when $T = 10$.

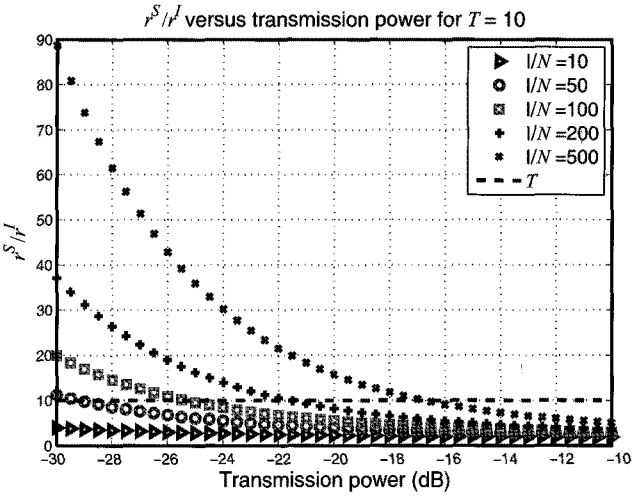

Fig. 6. Rate ratio as a function of the transmission power when $T = 10$.

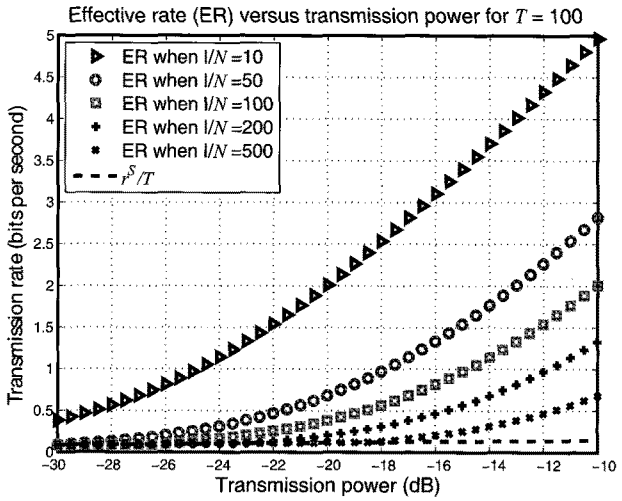

Fig. 5. Effective rate as a function of the transmission power when $T = 100$.

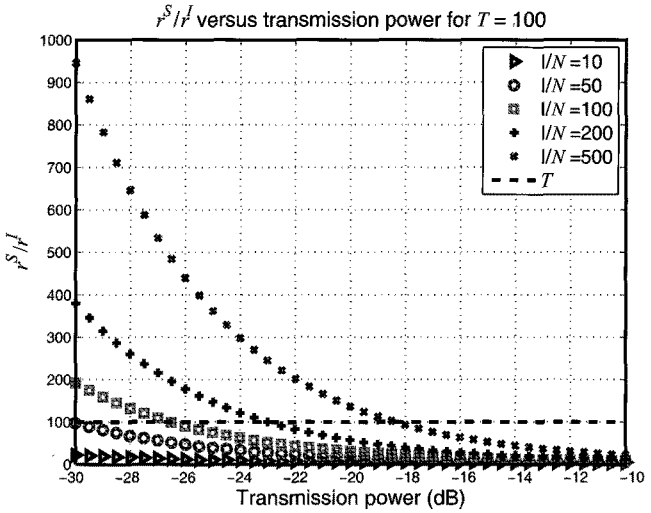

Fig. 7. Rate ratio as a function of the transmission power when $T = 100$.

1. The instantaneous rate of every scheduled source $r_k^k, k \in \mathcal{T}$, under operation "one-at-a-time" is equal to some value r^{S}, i.e.,

$$r_k^k = r^{\mathrm{S}}, \qquad \forall k \in \mathcal{T}.$$

2. The instantaneous rate of each source under operation "all-at-once" is equal to some value r^{I}, i.e.,

$$r_k^0 = r^{\mathrm{I}}, \qquad \forall k \in \mathcal{T}.$$

Note that the comparison of the term $\sum_{k\in\mathcal{T}} r_k^0/r_k^k$ with one in Proposition 3 becomes, in this simplified example, a comparison between the ratio $r^{\mathrm{S}}/r^{\mathrm{I}}$ and the horizontal line T. Figs. 4 and 5 depict the variation of the effective rate of an individual source (i.e., the quantity $\omega_0 r_k^0 + \omega_k r_k^k$ for every $k \in \mathcal{T}$) with respect to its transmission power (in dB) under different interference to noise (I/N) ratios, for 10 and 100 source/destination pairs respectively ($T = 10$ and $T = 100$). Note that here we use I to denote the interference from an individual source and not the overall interference. In our numerical experiments, the rates are computed based on the single user Shannon formula under the assumption of unit bandwidth. Furthermore, the path loss $G_{(k,k)}$ between each source/destination pair k is assumed to be equal to one and the noise power N to be equal to 3.34×10^{-6} Watts. In Figs. 4 and 5 we also plot r^{S}/T which is the effective rate of each source/destination pair when only operation "one at a time" is used (i.e., $\omega_0 = 0$). Figs. 6 and 7 simply plot the variation of the quantity $r^{\mathrm{S}}/r^{\mathrm{I}}$ with respect to the transmission power. From Figs. 4–7, we observe that the transmission power values at which the optimal action switches from operation "one-at-a-time" to operation "all-at-once" stays roughly the same for different values in the number of source/destination pairs T, although the effective rate of each pair decreases as the number of communication pairs increases. This is the point where the quantity $r^{\mathrm{S}}/r^{\mathrm{I}}$ crosses the line horizontal line T. In the case of operation "one-at-a-time," the decrease in the rate is due to the decrease in the time the source has access to the channel as more and more sources need to be scheduled. In the case of operation "all-at-once," the decrease in the rate is a result of the increased interference since more and more sources

now transmit simultaneously. Finally, as intuition suggests, under high levels of interference, a relatively higher transmission power is needed for the scheme "all-at-once" to be optimal.

VII. CONCLUSIONS

The scheduling problem has been studied extensively under various contexts in the literature. In this paper, we attempted to summarize our latest results and views on the scheduling problem through a physical-layer-aware perspective. We provided optimal policies under different performance objectives, such as the performance metrics of minimum-length scheduling and utility maximization. Realizing the inherent complexities of optimal scheduling, we provided two heuristic ideas to overcome them, namely the idea of reducing the action space of the scheduling decisions as well as the technique of column generation. In particular, we believe that the performance metric of obtaining schedules of minimum-length can be a useful alternative metric in non-ergodic and non-stationary environments where the commonly used criteria of stable throughput, delay, or utility maximization may not be well-defined. Although in this paper we considered only channels with stationary and ergodic behavior, we hope that our results can open new avenues to study the scheduling problem under non-ergodic and non-stationary environments. Furthermore, given the lack of centralized control in many network environments, obtaining decentralized solutions is naturally of great importance.

REFERENCES

[1] E. Arikan, "Some complexity results about packet radio networks," *IEEE Trans. Inf. Theory*, vol. 30, no. 4, July 1984.

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

[3] J. Luo and A. Ephremides, "On the throughput, capacity and stability regions of random multiple access," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, June 2006.

[4] B. Hajek and G. Sasaki, "Link scheduling in polynomial time," *IEEE Trans. Inf. Theory*, vol. 34, no. 5, Sept. 1988.

[5] S. A. Borbash and A. Ephremides, "Wireless link scheduling with power control and SINR constraints," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, Nov. 2006.

[6] S. Kompella, J. E. Wieselthier, and A. Ephremides, "A cross-layer approach to optimal wireless link scheduling with SINR constraints," in *Proc. IEEE MILCOM*, Oct. 2007.

[7] S. Kompella, J. E. Wieselthier, and A. Ephremides, "Revisiting the optimal scheduling problem," in *Proc. CISS*, Mar. 2008.

[8] A. Pantelidou and A. Ephremides, "Minimum schedule lengths with rate control in single-hop wireless networks," in *Proc. IEEE MILCOM*, Nov. 2008.

[9] A. Goldsmith. *Wireless Communications*. Cambridge University Press, 2005.

[10] L. Georgiadis, M. J. Neely, and L. Tassiulas, "Resource allocation and cross-layer control in wireless networks," *Foundations and Trends in Networking*, vol. 1, no. 11, 2006.

[11] T. Ho and H. Viswanathan, "Dynamic algorithms for multicast with intra-session network coding," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, Feb. 2009.

[12] M. J. Neely, E. Modiano, and C. E. Rohrs, "Dynamic power allocation and routing for time-varying wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 1, Jan. 2005.

[13] A. Pantelidou and A. Ephremides, "A cross-layer view of wireless multicast optimization," in *Proc. 46th Annual Allerton Conference on Communication, Control, and Computing*, Sept. 2008.

[14] A. Pantelidou and A. Ephremides, "A cross-layer view of wireless multicasting under uncertainty," in *Proc. ITW*, June 2009.

[15] A. Pantelidou and A. Ephremides, "Minimum-length scheduling for multicast traffic under channel uncertainty," in *Proc. IEEE GLOBECOM*, Nov. 2009.

[16] A. Pantelidou and A. Ephremides, "What is optimal scheduling in wireless networks?" in *Proc. WICON*, Nov. 2008.

[17] D. Bertsimas and J. N. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1997.

[18] M. Johansson and L. Xiao, "Cross-layer optimization of wireless networks using nonlinear column generation," *IEEE Trans. Wireless Commun.*, vol. 5, no. 2, Feb. 2006.

[19] A. Pantelidou and A. Ephremides, "Optimal rate control policies for proportional fairness in wireless networks," in *Proc. CISS*, Mar. 2008.

[20] A. Pantelidou and A. Ephremides, "Minimum-length scheduling and rate control for time-varying wireless networks," in *Proc. IEEE MILCOM*, Oct. 2009.

[21] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, 2000.

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

Anna Pantelidou received her B.S. degree in 2001 in Computer Science from the National and Kapodistrian University of Athens, Greece and her M.S. and Ph.D. degrees in Electrical and Computer Engineering in 2004 and 2009, respectively from University of Maryland at College Park. She is currently a Research Scientist at the Centre for Wireless Communications, University of Oulu, Finland. Her research interests are in the areas of cross-layer design and optimization of wireless networks, energy efficiency, network coding, and queuing systems.

Anthony Ephremides received his B.S. degree from the National Technical University of Athens (1967), and M.S. (1969) and Ph.D. (1971) degrees from Princeton University, all in Electrical Engineering. He has been at the University of Maryland since 1971, and currently holds a joint appointment as Professor in the Electrical Engineering Department and the Institute of Systems Research (ISR). He is co-founder of the NASA Center for Commercial Development of Space on Hybrid and Satellite Communications Networks established in 1991 at Maryland as an off-shoot of the ISR. He was a Visiting Professor in 1978 at the National Technical University in Athens, Greece, and in 1979 at the EECS Department of the University of California, Berkeley, and at INRIA, France. During 1985–1986 he was on leave at MIT and ETH in Zurich, Switzerland. He was the General Chairman of the 1986 IEEE Conference on Decision and Control in Athens, Greece. He has also been the Director of the Fairchild Scholars and Doctoral Fellows Program, an academic and research partnership program in Satellite Communications between Fairchild Industries and the University of Maryland. He won the IEEE Donald E. Fink Prize Paper Award (1992). He has been the President of the Inf. Theory Society of the IEEE (1987), and served on the Board of the IEEE (1989 and 1990). His interests are in the areas of communication theory, communication systems and networks, queuing systems, signal processing, and satellite communications.