User-centric RRM and Optimizable Protocol Design for beyond-4G RANs

Petar Djukic, Halim Yanikomeroglu, Jietao Zhang

***Abstract*—We argue that achieving very high data rates can be accomplished by moving away from system-centric radio resource management (RRM), used in the cellular radio access network (RAN) architecture, and toward advanced user-centric RRM, obtained with top-down RAN protocol design. System-centric RRM uses a divide-and-conquer approach, which assigns resources to cells first and to users second. On the other hand, user-centric RRM assigns resources to users first and then finds radio ports to provide these resources, making it potentially more efficient. Top-down optimizable RAN protocol design provides a way to design RAN protocols, which support a wide range of RRM algorithms, ensuring that RRM is not restricted by RAN protocols. Combining the user-centric RRM with the top-down optimizable RAN protocol design approach seems to provide promising approach for advanced RRM in beyond-4G RANs.**

***Index Terms*—Beyond-4G RANs, User-centric RRM, Cross-layer Optimization.**

I. Introduction

The main requirement for the advanced RAN is that it should cost-effectively provide ubiquitous high data rate coverage, when and where required [1]. High data rate coverage is not required uniformly in time and space over the coverage area of the RAN, but only where there are users and when those users need it. As the limiting factor in achieving high data rates is the distance between the transmitter and the receiver, the only way to achieve a ubiquitously *available* high data coverage is with an advanced RAN architecture providing a dense mesh of radio ports. Clearly, using the cellular RAN architecture is not feasible from the cost perspective, so an advanced RAN architecture is required. Many of the RAN elements required for this architecture, such as distributed antenna ports, femto base-stations, and various types of relays, are either already available, or are coming online in the current standardization activities. The question, then, is: How should the new RAN be utilized to achieve ubiquitously high data rate, when and where required?

We argue that achieving this type of ubiquitous high data rates can be accomplished by moving away from the system-centric radio resource management (RRM), used in cellular RAN architecture, and toward advanced *user-centric* RRM. System-centric RRM uses a divide-and-conquer approach, which assigns resources to cells first and to users second. On the other hand, user-centric RRM assigns resources to users first and then finds radio ports to provide these resources. By first assigning resources to users, user-centric RRM solves a fundamental inefficiency with the system-centric RRM – in system-centric RRM resources are *assigned* to cells, while they are *used* by users.

User-centric RRM is enabled by the flexibility of Orthogonal Frequency Division Multiple Access (OFDMA) and requires a sophisticated RAN architecture, which provides a high density of radio ports. Instead of managing the network as a set of independent base-stations, user-centric RRM manages a network of inter-dependent radio ports. User-centric RRM dynamically shifts assignment of radio resources through the RAN coverage area as the users move to obtain virtually ubiquitous coverage. Since the user may be multiple radio hops away from the core network, user-centric RRM needs to consider cross-layer RRM techniques between the network and lower layers in the architecture.

The cross-layer RRM nature of the advanced RANs brings up two important questions for user-centric RRM: (1) how to devise RAN protocols, which support a wide array of RRM algorithms, and (2) what is the relationship between RAN protocols and RRM algorithms. As was observed in the traffic engineering research [2], existing network protocols make it very hard, or impossible, to design effective network management algorithms. It is therefore advisable to consider potential RRM algorithms, while designing RAN protocols. The relationship between the protocols and the algorithms is tightly coupled, since the various layers must communicate for the algorithms to be effective. In an advanced RAN, the tight coupling means that information may be transferred between multiple layers in a single network element and across multiple layers of many network elements.

We argue that recent research in top-down optimizable protocol design should be used to devise user-centric RRM algorithms and RAN protocols, which implement them. Top-down protocol design starts with a global utility maximization problem, which optimizes the rates of individual users. The objective function of the optimization is chosen so that, at the optimum (equilibrium) point, the user rates satisfy some criterion specified by the network operator. The optimization is subject to the availability of radio resources, which can support the optimum user rates. Then, one uses mathematical decomposition to devise distributed algorithm and a corresponding protocol, which solves the problem. We believe that this approach can be successfully used for design of advanced RRM algorithms and suitable RAN protocols for the advanced RANs.

The rest of this contribution is organized as follows: Sec-

P. Djukic and H. Yanikomeroglu are with the Department of Systems and Computer Engineering, Carleton University, Canada, e-mail: {djukic, halim}@sce.carleton.ca

J. Zhang is with the Huawei Wireless Research, China, e-mail: jtzhang@huawei.com.

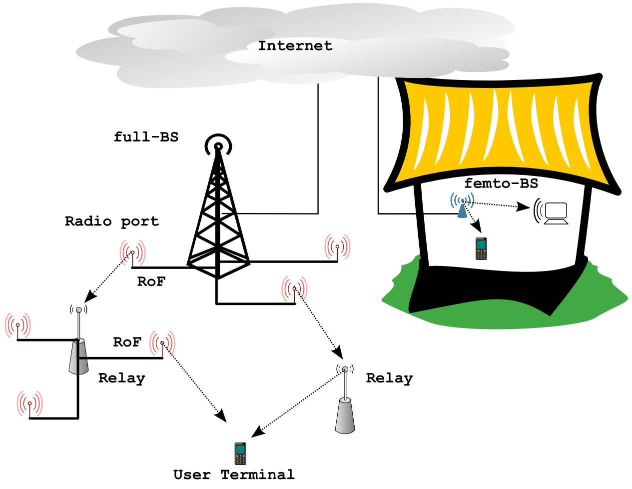


Fig. 1. Integration of the advanced RAN.

tion II discusses various elements of the advanced RANs and how they interact with each other; Section III introduces user-centric RRM; Section IV discusses top-down RAN protocol design and its relationship to RRM algorithms.

II. Advanced RAN Architecture

We now review the advanced RAN architecture, required for beyond-4G networks [1]. The architecture consists of many network elements: distributed antenna ports, femto base-stations, and various types of relays (Fig. 1).

The essential part of the advanced RAN are radio ports. A radio port may be located independently or collocated with a more complex RAN element, such as a base-station or a relay. If the radio port is located as a part of a more complex element it may connect to the rest of the element, which owns it, with a direct hardware interface. Otherwise, the radio port may be physically separated from the element, which owns it. In this case the radio port is connected to the element with Radio-over-Fiber (RoF). Since radio ports are deployed densely throughout the RAN coverage area, it may be possible for the user terminals to simultaneously send (and receive) radio signals to (and from) multiple radio ports. This technique has many names in the literature, such as "distributed antenna ports" [3], "coordinated multi-point transmission and reception (CoMP)", and in the 3GPP standardization process "Multi-cell MIMO", "Network MIMO", "Network Cooperative MIMO" [4], to mention a few.

The base-station (RAN anchor) is an important element of the advanced RAN. It manages multiple radio ports and has a wired connection to the Internet. RAN anchors do not require radio resources to provide backhaul services. We distinguish two types of RAN anchors: full base-station (full-BS) and femto base-station (femto-BS). A full-BS is a gateway to the Internet for multiple RAN elements, while a femto-BS is a gateway to the Internet for indoor elements.

In addition to the various types of base-stations, the RAN also uses many types of relays. Unlike base-stations, which are directly connected to the Internet, relays connect to the Internet through direct wireless connections to RAN anchors, or through multi-hop wireless connections over other relays, which connect directly to RAN anchors. A relay may have multiple radio ports attached to it, as a base-station would, and may have to participate in hand-off and other RRM procedures, as a base-station would. However, a relay is not connected to the RAN with a wired connection – it must at least connect to a base-station to get to the Internet and it may connect to the base-station in the network layer by using multi-hop transmissions through other relays.

The advanced RAN contains various types of relays, which vary in complexity. For example a relay may be a fairly simple amplify-and-forward relay, which does not examine the data flow, or a much more complex decode-and-forward relay, which examines and forwards packets. Since a relay may route and forward traffic, it can be more complex than a base-station. However, even the most complex type of relay does not need an expensive wired connection to the backhaul, nor does it need to be housed in a building; it only needs a power connection and weather-proof enclosure.

Typically a relay is also expected to have a shorter range than a base-station, so it requires a lower power amplification and thus cheaper power-amplifier than a base-station. Cheaper power amplifier circuitry also makes relays cheaper than base-station, from an engineering point of view.

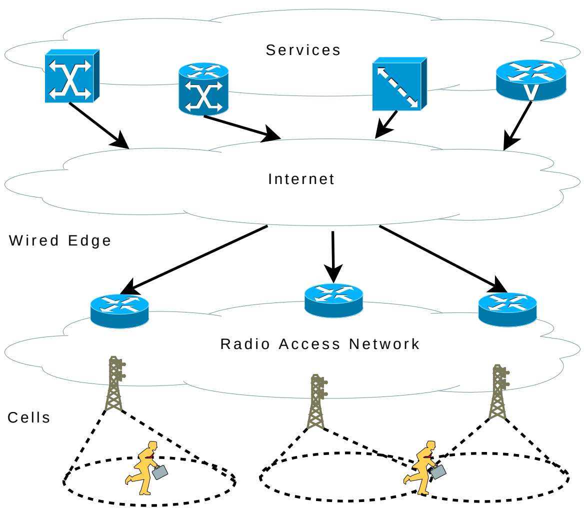
The advanced RAN is a *mesh* of RAN elements, where any one element can connect to any other element. Due to the flat-hierarchy in the RAN, RRM does not belong to any given RAN element. In the advanced RAN, RRM is a network-wide set of RAN protocols, which implement or facilitate RRM algorithms.

III. User centric RRM for advanced RANs

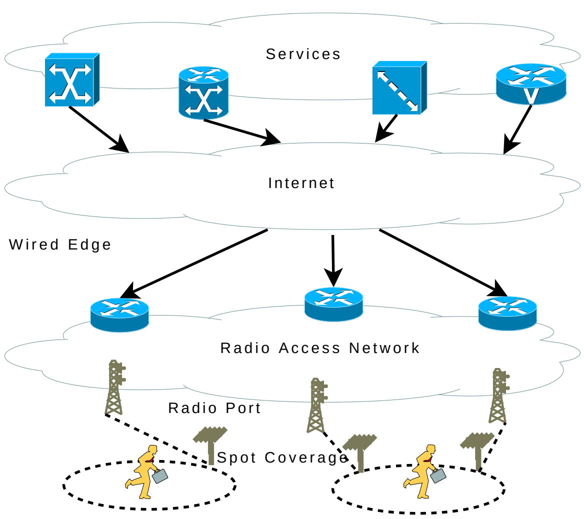
So far, we have discussed the hardware implementation of the advanced RANs, which is a facilitator for user-centric RRM proposed in this contribution. We now give an overview of user-centric RRM. We overview the top-down approach for RAN protocol design next.

In 3G and 4G networks, the global radio resources are divided among *cells*. A cell manages its radio resources and assigns them to users in the cell (Fig. 2a). As a user moves through the RAN coverage area it is handed-off from one cell to another, a process during which equivalent resources are assigned to the user in the new cell. The two main problems with this RAN architecture is that the assignment of resources to cells is semi-static and not easily modified to track the needs of the users and that cell splitting required for high data rates has a large cost.

Instead of basing RRM on the cellular concept, as in 4G and previous RAN architectures, we propose an RRM based on the concept of a *spot coverage* (Fig. 2b). While in the cellular RRM radio resources are assigned to cells, in the advanced RRM, radio resources are dynamically assigned in time, frequency, and space to users through temporary coverage spots. A coverage spot is associated with a single user and follows the user as it moves through the RAN coverage area. The advanced RAN provides radio ports, which



(a) "Classical" RAN (system-centric)



(b) Advanced RAN (user-centric)

Fig. 2. Conceptual difference between RANs.

are used to create spot coverage and connect the user to the Internet.

The concept of the coverage spot is similar to the concepts of "active zones" [5], "virtual cells" [3] and the "slide group cells" [6]. In each of those concepts, system resources track the user as it moves through the network. We use the term "coverage spot" to distinguish our usage of user-centric RRM, which is not used in the previous approaches.

Formally, a user-centric RRM optimization is

$$\max_{x_1,\ldots,x_m \in \mathcal{S}} \quad \sum_{l=1}^{m} U_l(x_l), \tag{1}$$

where $x_1, \ldots, x_m$ are the rates of the m users in the network, $U_l(\cdot)$ is the utility of user l and the optimization maximizes the total system utility subject to the existence of user rates, $x_1, \ldots, x_m \in \mathcal{S}$, where $\mathcal{S}$ is the set of all m-tuples of feasible rate assignments.

The utility function is chosen to represent the "satisfaction" of the user with the service (rate) it is getting. There are many utility functions, which correspond to different types of user satisfaction with the network. With a proper choice of utility functions [7], one may have an optimization that maximizes the total "weighted proportional fairness", a game theoretic optimum, "max-min fairness", which eliminates starvation, or simply "maximum total throughput". It is also possible to have utility functions that take the combination of traffic and profit into account [8].

Without getting into the mathematical details of utility functions, one can still mention that the Transmission Control Protocol (TCP) was recently reverse engineered as a utility optimization, which makes the network proportionally fair [9]. So, utility optimization gets its motivation from current network protocols, although the current network protocols do not necessarily get their motivation from utility optimizations [2]. By changing the utility function, one can easily design RAN protocols, which achieve different objectives in the RAN.

A system-centric optimization is

$$\max_{x_1,\ldots,x_m \in \mathcal{S}} \quad U_s(x_1, \ldots, x_m), \tag{2}$$

where $x_1, \ldots, x_m$ are the rates of the m users in the network, $U_s(\cdot)$ is the utility for the system, and the optimization maximizes the total system utility subject to the existence of user rates, $x_1, \ldots, x_m \in \mathcal{S}$. In this case, the utility function is chosen to represent the "satisfaction" of the network operator with the rates. For example, "maximum area spectral efficiency", or "maximum profit" may be one utility for the system. We note that in 3G and 4G networks the space of available user rates is limited, since all radio resources are not available in all cells (they are assigned to cells in advance). However, in general one can use a re-use factor of 1 and allow all radio resources to be available everywhere for a more efficient optimization.

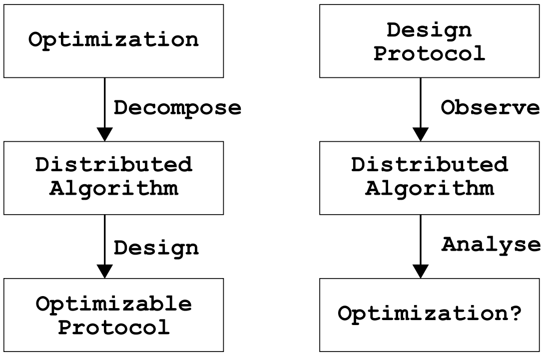
User-centric RRM seems more appropriate for advanced RANs for two reasons. First, it refocuses the network to user needs, and away from the network operator needs.[1] Second, since the optimization function is composed of many individual and independent components, the user-centric optimization is more amenable to decomposition and decentralized algorithm implementations. Indeed, this is why TCP is able to achieve a global network optimum, despite the fact that TCP flows do not exchange information with each other.

IV. Top-down RAN Protocol Design

Top-down protocol design was recently proposed as a methodology to ensure that new Internet protocols support network management techniques [2]. The motivation for this approach comes from the realization that many network management problems are complex due to the nature of network protocols, rather than some inherent difficulty in the problem itself.

Taking the top-down approach (Fig. 3a) first involves formulating a global optimization problem such as (1) or (2). Since the advanced RAN includes transmissions that span multiple

[1] It is also possible to use utility functions, which combine network operator needs with user needs.



(a) Top-Down (b) Bottom-up

Fig. 3. Protocol design approaches.

hops, this optimization considers multiple layers in the network stack. The second step is to solve the global optimization with a distributed optimization algorithm. Finally, a protocol is designed to implement the distributed optimization algorithm. In practice, a single protocol may be easily parameterized to perform multiple optimizations, based on its parameters [2].

The alternative is to devise algorithms and protocols bottom-up (Fig. 3b), by first devising RAN protocols and then devising RRM algorithms, based on the protocols. The problem with that approach is that it may result in very hard RRM problems, which may not be easily solvable or distributed. Since the advanced RAN has many non-heterogeneous elements, the protocol design may also get bogged down in working out the various inter-operability issues in the RAN, rather than focus on how to achieve the ultimate goal of providing the ubiquitous high data rate coverage with advanced RRM.

One example of bottom-up protocol design is TCP, which was widely successful in wired environments, but did not work well in wireless environments [10]. As it turns out, TCP could also be developed in the top-down way [9], which spurred the recent effort in research of top-down protocol design [11], [12] with promising results.

We now outline three factors, which we believe to be beneficial in top-down protocol design: restrictions on OFDMA schedules, implicit load-balancing in the backhaul, and the fact that scheduling access is different from scheduling the backhaul.

A. Restrictions on OFDMA Schedules

In the utility RRM optimizations (1) and (2), the utility function can be chosen arbitrarily, so the difficult part of either optimization is the constraint set over which the optimum should be found, $x_1, \ldots, x_m \in \mathcal{S}$. The problem is that (a) this set is usually not convex, making the optimization hard, and (b) due to the inter-dependence of wireless links, much of the information that forms the constraint set is distributed throughout the RAN, horizontally across network elements and vertically within the network elements.

With respect to (a), we note that the set of available rates depends on "optimum" OFDMA scheduling, which usually requires an exhaustive search over the available schedules. Since the search looks for OFDMA schedules that separate interfering transmissions in time, frequency and space, it is related to difficult graph colouring problems.

With respect to (b), we note that unlike wireline networks where link rates are constant and independent of other links, in wireless networks, link rates are dependent on other (interfering) links. Changes in the upper layers require scheduling changes in the wireless medium, so it is necessary to perform rate control optimizations (across many network elements) jointly with wireless scheduling (vertically in each element). It is therefore important to design RAN access mechanisms and RRM protocols, which make the constraint set convex and distribute the required information efficiently through the RAN.

Consider OFDMA schedules for three concurrent transmissions, which must be separated in time and frequency (Fig. 4). The first schedule (Fig. 4a) allows any OFDMA pattern and can be viewed as a graph colouring problem. In addition to being computationally expensive, that schedule is also not very practical due to the large amount of information required to distribute it. The second schedule (Fig. 4b) is restricted to data regions, which span in time and frequency, corresponding to 802.16j type schedules. Finding these types of schedules is even harder than graph colouring [13]. The third schedule (Fig. 4c), restricts transmissions to data strips, spanning in time, but not in frequency.

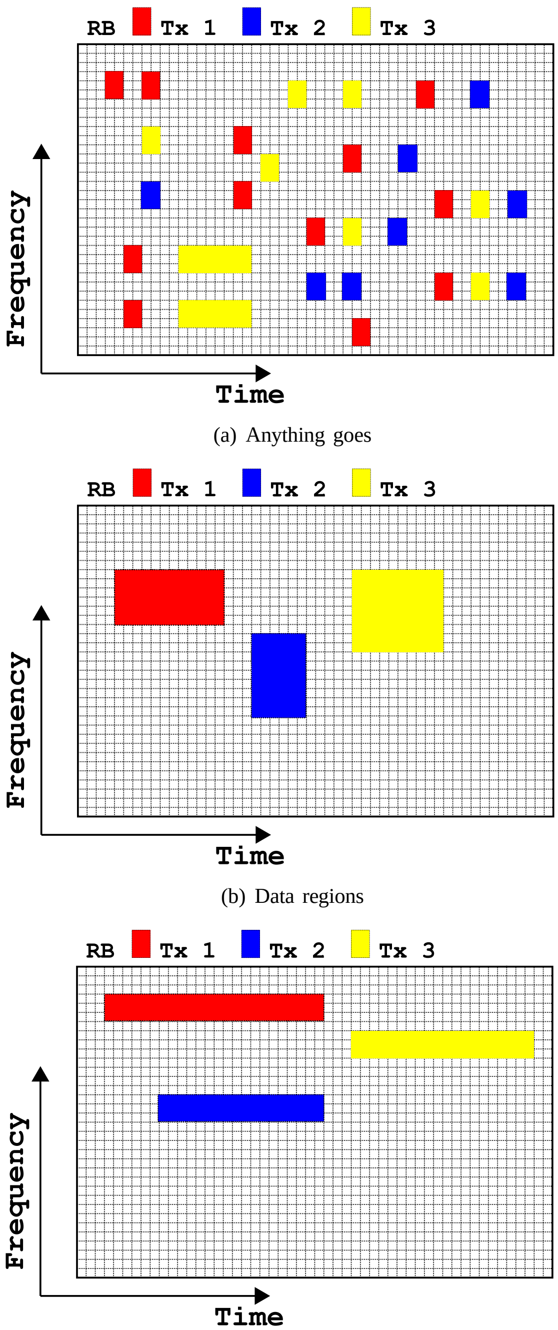
As it turns out, under some circumstances finding the third type of schedules takes polynomial time [14] and can be easily distributed [15]. Of course, the advantage in easier implementation is somewhat offset with lower performance. Nevertheless, to take advantage of distributed scheduling protocols, one should design RAN Medium Access Control (MAC) protocols which allow good OFDMA frame patterns.

B. Load-Balancing in the Backhaul

Future OFDMA based MACs are expected to be scheduled (as opposed to 802.11's best-effort), even in multi-hop situations. This means that the end-to-end delay can be controlled at the ingress part of the network and does not vary with the competing end-to-end traffic. Without getting into the details of this type of network, which is sometimes referred to as a "stop-and-go" queuing network [16], we can say that since the wireless access is scheduled throughout the network, hop-by-hop load-balancing is achieved implicitly by simply forwarding traffic.

The lack of load-balancing is a major reason why many wired network traffic management problems are difficult [2]. Without load-balancing a network traffic management optimization must, in addition to optimizing end-to-end traffic ensure, that all traffic only traverses one path between the source and the destination. The requirement on the solution to only use one path makes the optimization a more difficult "unsplittable flow" problem [17].

Due to it use of scheduled MACs in the advanced RAN, network wide RRM can be simplified with implicit load-balancing. One can consider optimization problems, which result in multi-path routed solutions. In a wireless network, a



(a) Anything goes

(b) Data regions

(c) Data strips

Fig. 4. Types of restrictions on OFDMA schedules.

network layer solution using multiple paths also benefits from using multiple radio ports, thus increasing diversity. Under this scenario, one can truly talk about "network MIMO". One way to achieve all of the above goals is to use a multi-hop MAC protocol, which includes a multi-path routing component may be a good approach to combine network layer and MAC layer functionality and take advantage of the load-balancing.

C. Scheduling Access is Different from Scheduling Backhaul

Our final observation is that the access part of the network is different from the backhaul part of the network, both in terms of the wireless channel and the offered traffic. The wireless channel in the backhaul varies more slowly than the wireless channel in the access network. Traffic patterns also change more slowly in the backhaul than in the access network, due to the static nature of relays. So, backhaul RRM algorithms and the resulting RAN protocols can be more accurate, although they may be slower to converge, than in the access part of the network.

V. Conclusion

We propose a promising new approach for combined RAN protocol design and RRM algorithm design for beyond-4G RANs. The new approach uses top-down methodology of formulating RRM optimizations, solving them with distributed algorithms and then designing RAN protocols, which implement the algorithms. It is possible to design parameterized RAN protocols, which depending on the parameters solve different RRM optimizations. We outline several factors, which should be considered in the future standardization efforts to make the top-down approach more viable.

References

[1] H. Yanikomeroglu and J. Zhang, "Beyond-4G cellular networks: Advanced radio access network (RAN) architectures, advanced radio resource management (RRM) techniques, and other enabling technologies," in *WWRF Meeting 21*, Stockholm, Sweden, 13–15 October 2008.

[2] J. He, J. Rexford, and M. Chiang, "Don't optimize existing protocols, design optimizable protocols," *ACM SIGCOMM Computer Communication Review*, vol. 37, no. 3, pp. 53–58, 2007.

[3] L. Dai, *Distributed Antenna Systems: Open Architecture for Future Wireless Communications*. CRC Press, 2007, ch. Optimal Resource Allocation of DAS, pp. 169–200.

[4] Ericsson, "3GPP TSG-RAN WG1 ♯53 R1-082024 a discussion on some technology components of LTE-Advanced," Kansas, MO, 2008.

[5] W. Lee, "Smaller cells for greater performance," *IEEE Commun. Mag.*, vol. 29, no. 11, pp. 19–23, November 1991.

[6] X. Xu, C. Wu, X. Tao, Y. Wang, and P. Zhang, "Maximum utility principle access control for beyond 3G mobile system," *Wireless Communications and Mobile Computing*, vol. 7, pp. 951–959, 2007.

[7] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, pp. 556–567, 2000.

[8] A. Elwalid, D. Mitra, and Q. Wang, "Distributed nonlinear integer optimization for data-optical internetworking," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1502–1513, August 2006.

[9] F. P. Kelly, "Fairness and stability of end-to-end congestion control," *European Journal of Control*, vol. 9, pp. 159–176, 2003.

[10] S. Xu and T. Saadawi, "Does the IEEE 802.11 MAC protocol work well in multihop wireless ad hoc networks," *IEEE Commun. Mag.*, vol. 39, no. 6, pp. 130–137, June 2001.

[11] 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, August 2006.

[12] M. Chiang, S. H. Low, A. R. Calderbank, and J. C. Doyle, "Layering as optimization decomposition: A mathematical theory of network architectures," *Proc. IEEE*, 2007.

[13] R. Cohen and L. Katzir, "Computational analysis and efficient algorithms for micro and macro OFDMA scheduling," in *INFOCOM*, 2008.

[14] P. Djukic and S. Valaee, "Link scheduling for minimum delay in spatial re-use TDMA," in *Proceedings of INFOCOM*, 2007.

[15] ——, "Distributed link scheduling for TDMA mesh networks," in *Proceedings of ICC*, 2007.

[16] S. J. Golestani, "A framing strategy for congestion management," *IEEE J. Sel. Areas Commun.*, vol. 9, no. 7, pp. 1064–1077, September 1991.

[17] Y. Dinitz, N. Garg, and M. X. Goemans, "On the single-source unsplittable flow problem," in *IEEE Symposium on Foundations of Computer Science*, 1991.