INVITED PAPER

# A Unified Approach to QoS-Guaranteed Scheduling for Channel-Adaptive Wireless Networks

*Terminals wirelessly-linked to an access point should be able to achieve near-optimal data rates, while maintaining required quality of service, by using simple scheduling algorithms.*

By Xin Wang, *Member IEEE*, Georgios B. Giannakis, *Fellow IEEE*, and Antonio G. Marques, *Member IEEE*

**ABSTRACT** | Scheduling amounts to allocating optimally channel, rate and power resources to multiple connections with diverse quality-of-service (QoS) requirements. It constitutes a throughput-critical task at the medium access control layer of today's wireless networks that has been tackled by seemingly unrelated information-theoretic and protocol design approaches. Capitalizing on convex optimization and stochastic approximation tools, the present paper develops a unified framework for channel-aware QoS-guaranteed scheduling protocols for use in adaptive wireless networks whereby multiple terminals are linked through orthogonal fading channels to an access point, and transmissions are (opportunistically) adjusted to the intended channel. The unification encompasses downlink and uplink with time-division or frequency-division duplex operation; full and quantized channel state information comprising a few bits communicated over a limited-rate feedback channel; different types of traffic (best effort, non-real-time, real-time); uniform and optimal power loading; off-line optimal scheduling schemes benchmarking fundamentally achievable rate limits; as well as on-line scheduling algorithms capable of dynamically learning the intended channel statistics and converging to the optimal benchmarks from any initial value. The take-home message offers an important cross-layer design guideline: judiciously developed, yet surprisingly simple, channel-adaptive, on-line schedulers can approach information-theoretic rate limits with QoS guarantees.

**KEYWORDS** | Adaptive modulation and coding; convex optimization; quality of service (QoS); scheduling and resource allocation; stochastic approximation

## I. CONTEXT

Over the last decade, we have witnessed a rapid growth in demand for fast and error-resilient telecommunication services such as e-mail, transfer of data files, voice over the Internet and video conferencing, to name a few. In accordance with this growth, broadband wireless networks have become an integral part of the global communication infrastructure. Beyond speed and robustness to errors induced by fading propagation, present and next-generation wireless networks are challenged to meet the diverse quality-of-service (QoS) requirements imposed by current and envisioned services. In addition to prescribed error rates, these requirements include e.g., minimum rates for file transfers and maximum delay bounds for voice and video conferencing. Critical to QoS

Manuscript received January 6, 2007; revised July 6, 2007. This work was supported in part by the U.S. Department of Defense Army Research Office under Grant W911NF-05-1-0283 and was prepared through collaborative participation in the Communications and Networks Consortium sponsored by the U.S. Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement DAAD19-01-2-0011. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon.

**X. Wang** is with the Department of Electrical Engineering, Florida Atlantic University, Boca Raton, FL 33431 USA (e-mail: xin.wang@fau.edu).
**G. B. Giannakis** is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: georgios@ece.umn.edu).
**A. G. Marques** is with the Department of Signal Theory and Communications, Rey Juan Carlos University, 28943 Madrid, Spain (e-mail: antonio.garcia.marques@urjc.es).

Digital Object Identifier: 10.1109/JPROC.2007.907120

provisioning is judicious utilization of the wireless channel as well as allocation of the available rate and power resources among multiple connections, i.e., communicating users. This is the task undertaken by the packet scheduler at the medium access control (MAC) layer.

A number of scheduling protocols are available for wireline networks [92], including the well-known schemes for fair queueing [21], virtual clock [93], self-clocked fair queueing [27], and earliest-due-date [37] ones. But these are not directly applicable to the wireless setting because wireless and wireline propagation media are inherently different: whereas wireline channels are deterministic, wireless channels are random and fade unpredictably, thus rendering link capacity location-dependent and time-varying. In an effort to maintain wireline schedulers operational in wireless links, incorporation of channel-dependent features has been attempted to account for fading [55], [60], [62]; see also [14] for a review. However, adhering to the conventional layered architecture of communication networks these scheduling algorithms simply model a wireless channel as being either "good" or "bad." And lacking a channel-aware physical layer that can capitalize on the fading nature of the wireless interface, they cannot effectively exploit the time-varying multiuser channel capacity which, relative to single-user fading links, is enhanced by the multiuser diversity gain [40].

Along with wireless scheduling algorithms, research interest and development efforts have increased over the last decade towards wireless systems that adapt to the underlying communication channel—a notion dating back 30 years ago [33]. Resources in these adaptive transmission systems are adjusted in accordance with the channel quality so that higher (lower) rate and power is allocated as the channel quality increases (respectively decreases). This adjustment is enabled at the physical layer through adaptive modulation and coding (AMC). Systems relying on AMC have been analyzed and successfully applied in various settings to cope with (and even exploit) the fading characteristics of the wireless channel [5], [28]–[30]; see also [17] for a unified view. Capitalizing on some form of channel state information at the transmitter (CSIT), AMC-based systems have well-documented merits over non-adaptive alternatives. They can, e.g., exhibit as high as 17 dB gain in spectral efficiency [28]. Testament to their success is further provided by the adoption of AMC in current and future wireless standards. Those include the CDMA2000 1× evolution for the downlink (1xEV-DO), the WCDMA standard with high-speed downlink packet access (HSDPA), the IEEE 802.16 broadband wireless access (WiMax) standard, as well as the IEEE 802.11 and HIPERLAN/2 wireless local area networks (WiFi) [1], [8], [23], [34], [35].

Benefits of adapting to the intended channel can also permeate to higher layers of the network stack through the recent so called cross-layer designs. AMC at the physical layer of these networks has been optimized jointly with scheduling parameters at the MAC layer, e.g., with the number of ARQ (re-) transmissions and the queue length [51]–[53], [82]. Such a cross-layer approach to scheduling is particularly attractive for wireless networks where, due to the broadcasting nature of the air-interface, interdependencies among different layers are heavier than those encountered with a wireline network. Cast in a cross-layer network utility maximization framework, these interdependencies have been also exploited to optimize the design of channel-adaptive wireless networks in [16], [44], [54], and [89]; see also [61] for a tutorial treatment. With the emergence of *network science* as a field encompassing multiple disciplines and having far-reaching implications [59], it is natural to foresee that channel adaptation and cross-layer optimization will play instrumental roles in the design of commercial and tactical wireless networks of tomorrow.

As far as wireless scheduling is concerned, the first steps in this direction were taken with the introduction of proportional fair scheduling (PFS)—the AMC-based algorithm used in Qualcomm's HDR (1xEV-DO) system [8], [81]. By exploiting opportunities for reliable connectivity provided by the multiuser fading channel, PFS effects multiuser diversity which enhances throughput while maintaining "proportional fairness" among users. The striking success of PFS prompted further studies on channel-aware "opportunistic" scheduling for best effort [3], [42], [73], non-real-time [6], [49], [50], and real-time traffic [2], [64], [65], [69], [74]. Built on cross-layer channel-adaptive approaches, the resultant algorithms broadened the scope of traditional schedulers by allowing a scheduler to perform not only user selection (channel assignment) but also rate allocation. In fact, going beyond channel and rate assignments, we contend that with sufficient CSIT available, the scheduler should also perform optimal power allocation. This augmentation brings scheduling design closer to the overall optimal resource allocation task, the fundamental limits of which have been explored by information theorists in the context of determining the capacity region (maximum achievable rates) of multiple access and broadcast fading channels [9], [32], [45], [46], [76].

Even though expectations regarding channel-aware wireless schedulers are great, optimality criteria, performance metrics, and algorithms are still in a rather primitive stage, especially when it comes to handling diverse QoS requirements and coping with heterogenous traffic. Current research is fragmented and is usually conducted in a disciplinary setup, necessitating systematic designs, rigorous analysis, and testable predictions. In this work, we aspire to provide such a unified view for the downlink and uplink scheduling of multiple connections with diverse QoS requirements [83]–[86], where each connection transmits using AMC over a wireless fading

channel. Aiming at optimal yet simple schedulers with affordable implementation, the unified framework encompasses both uniform and optimal power allocation, off-line optimal scheduling schemes benchmarking fundamentally achievable rate limits, as well as on-line scheduling algorithms capable of dynamically learning the intended channel statistics and converging to the optimal benchmarks from any initial value. Towards this objective, we will describe first the multiuser system and QoS model before outlining key advances in scheduling and resource allocation for wireless networks.

## II. SYSTEM MODELING

We will deal with scheduling in wireless packet access networks as the one depicted in Fig. 1. Similar configurations appear in wireless local area networks and broadband cellular standards, and have been also considered in [3], [6], [8], [42], [49], [50], [65], [73], and [74]. Multiple (here $K$) user terminals are connected to an access point (AP) over wireless links. For simplicity, we suppose only one connection (a.k.a. flow or session) per terminal. To accommodate multiple connections per terminal, scheduling can be implemented either on a per connection basis or on a per user basis. The former is possible by viewing multiple connections as multiple virtual users, while the latter requires each terminal to share the scheduling task with the AP. Specifically, each terminal requests aggregate rate and/or delay services from the AP and, upon receiving scheduling decisions, distributes the aggregate resource assignments among its own multiple connections. Because it reduces the burden of the AP, the user-based approach has been adopted by recent wireless standards including IEEE 802.16 [34].

We further assume time-division multiplexing (TDM) for the downlink and time-division multiple access (TDMA) for the uplink. Although the framework is detailed for TDM/TDMA systems, it carries over to any orthogonal channelization scheme, including those based on general orthogonal code-division multiplexing/multiple access (CDM/CDMA) and hybrids thereof, e.g., TDM/CDMA. Furthermore, the setup allows for either time- or frequency-division duplex (TDD or FDD) operation. Note that nonorthogonal user access may be motivated from a capacity perspective; however, orthogonal access is typically employed by practical wireless access because it is simple. (There is no need to deal with interference issues.)

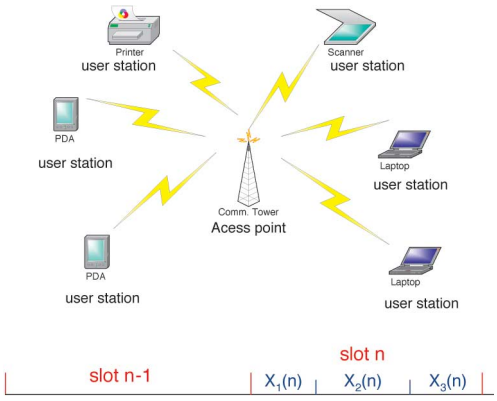



**Fig. 1.** *Cellular wireless packet access system.*

The wireless links are modeled as flat (nonselective) channels, each characterized by a random fading coefficient $\sqrt{h_k}$. However, the framework applies also to frequency-selective fading links provided that the aforementioned channelization is based on orthogonal frequency division multiplexing/multiple access (OFDM/OFDMA) [88]. This is possible because OFDM renders a frequency-selective channel equivalent to multiple flat fading subchannels [90]. Focusing henceforth on TDM/TDMA, the flat fading coefficients $\{\sqrt{h_k}\}_{k=1}^K$ remain invariant during a time slot $T_s$ but are allowed to vary from slot-to-slot (block fading model). With $^T$ denoting transposition, the resultant $K \times 1$ vector of channel gains $\boldsymbol{h} := [h_1, \ldots, h_K]^T$ is stationary and ergodic with continuous joint cumulative distribution function (cdf) $F(\boldsymbol{h})$; e.g., Rayleigh if $\{\sqrt{h_k}\}_{k=1}^K$ are jointly complex Gaussian. Note that practical scheduling algorithms should also be robust to channel nonstationarities.

In TDMA/TDM, the $K$ users can transmit/receive in uplink/downlink per slot over nonoverlapping *fractions* $\{\tau_k(\boldsymbol{h})\}_{k=1}^K$ whose duration depends on the channel realization $\boldsymbol{h}$. If we suppose without loss of generality (w.l.o.g.) that each slot has duration $T_s = 1$, then clearly $\sum_{k=1}^K \tau_k(\boldsymbol{h}) \in [0,1]$. Notice that the latter allows all users, or at the other extreme no user, transmitting over a given slot. Each user transmits using one AMC mode pair which comprises a modulation and an error control (i.e., channel) code. If scheduled, i.e., $\tau_k(\boldsymbol{h}) > 0$, user $k$ selects in each slot a modulation with rate $\rho_{k,m}^{(\mathrm{mod})}$ (e.g., 16-QAM) along with a channel code with rate $\rho_{k,m}^{(\mathrm{cod})}$ (e.g., a convolutional code with rate 1/2) to transmit with AMC rate $\rho_{k,m} := \rho_{k,m}^{(\mathrm{mod})} \rho_{k,m}^{(\mathrm{cod})}$. In addition to $\{\rho_{k,m}\}_{m=1}^{M_k}$ nonzero rates (AMC modes) that can differ per user $k = 1, \ldots, K$, we let $\rho_{k,0} := 0$ denote the case where user $k$ does not transmit.

The relationship between BER $(\epsilon_k)$, AMC rates $(\rho_{k,m})$, transmit-power $(p_k)$, and channel gain $(h_k)$ plays a major role in the design of adaptive schedulers. For constellation- and code-specific constants $\kappa_1$ and $\kappa_2$ and after assuming w.l.o.g. that the additive white Gaussian noise at the receiver has unit variance, this relationship can be accurately approximated (by exponential curve fitting) as [28, eq. (19)], [30, Ch. 9]

$$\epsilon_k(h_k p_k, \rho_{k,m}) = \kappa_1 \exp\left(-\frac{\kappa_2 h_k p_k}{2^{\rho_{k,m}} - 1}\right). \quad (1)$$

Given $h_k, p_k$ and a maximum allowable BER $\check{\epsilon}_k$, we can find using (1) the maximum AMC mode $m_k^*(h_k)$ so that $\epsilon_k(h_k p_k, \rho_{k,m_k^*}) \leq \check{\epsilon}_k$. If each user's set of AMC modes is infinite (a useful abstraction when exploring rate limits), then we can solve (1) for the rate $r_k^*(h_k) = \rho_{k,m_k^*(h_k)}$ which meets the prescribed BER with equality, i.e., $\epsilon_k(h_k p_k, r_k^*) = \check{\epsilon}_k$. This rate can be also expressed as $r_k^*(h_k) = \log_2(1 + h_k p_k/\Gamma)$, where $\Gamma := \kappa_2^{-1} \ln(\kappa_1/\check{\epsilon}_k) \geq 1$. With $\Gamma = 1$ the latter yields the maximum possible rate (Shannon's limit) and allows comparison of the practical AMC-based systems with fundamental capacity-achieving benchmarks.

In uplink, each user maintains a queue to store arriving packets from the network layer. If scheduled, a user transmits queued messages in the uplink slot in a first-in-first-out manner. In downlink, the AP maintains separate such queues for different connections. The channel-aware scheduler at the AP assigns time fractions to users and indicates the AMC mode indices (codewords) before a downlink/uplink slot. Users then transmit/receive with their rate/power adapted to CSIT.

### A. FDD and TDD

The two schemes commonly used for the uplink and downlink operation of wireless systems are TDD and FDD, each having different pros and cons. Uplink and downlink in FDD are carried over nonoverlapping frequency bands separated by a guard band to minimize cross-channel interference. Because it is simple to implement, FDD has been widely adopted by 2G and 3G systems, including those in the GSM, IS-95, WCDMA, and CDMA 2000 standards. TDD, on the other hand, separates uplink from downlink in the time domain. Its inherent flexibility to adjust data rates in the uplink and downlink by simply changing subframe durations makes TDD an attractive candidate for, e.g., the IEEE 802.20 wireless standard [15]. Furthermore, because uplink and downlink operate over the same frequency band in TDD, channel reciprocity allows for easy channel acquisition at the transmitter during the training phase of the reverse link, where transmitter and receiver exchange roles. (Reciprocity holds if the time interval between transmit- and receive-mode is selected not to exceed the channel coherence time.) Notice that reciprocity is not available in FDD where acquisition of CSIT necessitates feedback from the receiver to the transmitter. The limited number of bits carried by the feedback channel prompts one to distinguish between two forms of CSIT that we discuss next.

### B. F-CSIT and Q-CSIT

At the receiving end of the $k$th link, it is possible to obtain a practically perfect estimate of the channel $h_k$ via sufficiently many training symbols (pilots) known to both transmitter and receiver, i.e., we can always acquire essentially full (F) channel state information at the receiver. By reciprocity, this implies that F-CSIT can be always assumed available in TDD. In FDD though, where only the receiver can inform the transmitter about the channel state through a limited-rate feedback channel, the only pragmatic option is quantized (Q) CSIT comprising a codeword of a few bits; see, e.g., [56] and [57].

One can intuitively expect that the performance of channel-aware scheduling algorithms will depend critically on the form of the available CSIT. Whether at the AP (for downlink) or at the terminals (for uplink), optimal scheduling parameters will be selected using either F-CSIT or Q-CSIT. Either way, it is important to mention as a prelude that only Q-CSIT incurring a surprisingly small feedback overhead will be sufficient to implement the optimal schedule in both TDD and FDD operation.

### C. Traffic Types and QoS Requirements

Depending on the type of traffic, a unified approach to scheduling must account for three classes of services and the associated QoS requirements that are typical in wireless standards [1], [8], [34]:

- *Best effort* (BE) services entail applications such as e-mail and http web browsing. They come with a prescribed maximum allowable bit-error rate (BER) but pose no requirements on rate or delay guarantees.
- *Non-real-time* (nRT) services are for mission-critical but delay-tolerant applications such as file transfers (ftp). In addition to a maximum allowable BER, they require minimum rate (i.e., throughput) guarantees but do not impose any bound on delays.
- *Real-time* (RT) services such as video conferencing and streaming entail guarantees on BER, throughput, and latency. Since delayed packets are useless, the pertinent QoS metric is maximum delay for a given arrival rate.

These classes of services are bestowed different priorities, with RT traffic enjoying the highest and BE traffic receiving the lowest priority. A unified scheduling framework should certainly accommodate all three classes and the corresponding priorities.

## III. IT-PHY AND MAC PERSPECTIVES

In this section, we highlight key notions behind channel-aware resource allocation in wireless networks and outline the historical development of pertinent information-theoretic (IT) approaches at the physical (PHY) layer along with existing scheduling algorithms at the MAC layer. The goal is to provide a high-level description of the basic ideas paving the way towards a unified framework for QoS-guaranteed channel-aware approaches to scheduling.

### A. Multiuser Diversity and Proportional Fair Scheduling

The first IT-PHY concept is that of multiuser diversity introduced by [40] when studying the uplink sum-capacity

of fading channels. To appreciate its significance with multiuser fading systems it is instructive to consider the average capacity of a point-to-point fading channel with gain $h$, given by $\bar{C} := \mathbb{E}_h[\log(1+hP)]$, where $P$ denotes transmit-power; and compare it with the capacity $C$ of a nonfading channel having output SNR equal to $\mathbb{E}_h[hP]$. Concavity of the $\log(\cdot)$ function implies readily that $\bar{C} := \mathbb{E}_h[\log(1+hP)] \leq \log(1+\mathbb{E}_h[hP]) := C$, and verifies that fading leads to loss in (average) channel capacity of single-user channels. Consistent with this fact, the points in Fig. 2 where the curved line (capacity boundary for fading links) intersects the $\bar{R}_1$ and $\bar{R}_2$ axes correspond to the single-user case and thus lie below the corresponding point where the straight line (capacity boundary for nonfading links) intersects the $\bar{R}_1$ and $\bar{R}_2$ axes.

To be convinced that the contrary is possible with multiuser fading channels, consider a TDM downlink transmission with constant power $P$ serving two users with respective channel gains $h_1$ and $h_2$. If $h_1$, $h_2$ are nonfading, the maximum achievable weighted sum-rate is given by $w_1R_1 + w_2R_2 = w_1r_1^*\tau + w_2r_2^*(1-\tau) = w_1\tau\log(1+h_1P) + w_2(1-\tau)\log(1+h_2P)$, where the weights satisfy $w_1 + w_2 = 1$ and $\tau$ denotes the fraction of the slot that user 1 transmits. This maximum sum-rate is depicted by the hypotenuse of the orthogonal triangle (capacity region) in Fig. 2, every point of which corresponds to a value of the time-sharing parameter $\tau$. Notice that the AP can only schedule $\tau$ in this case. However, if $h_1[n]$, $h_2[n]$ are fading and thus change per slot $n$, we can attain a higher weighted sum-rate per realization, namely $\max(w_1r_1^*(h_1[n]), w_2r_1^*(h_2[n]))$, by selecting the most reliable of the two channels and assigning to the corresponding user the entire slot. And since this can happen for each $n$, the average capacity of the fading downlink channel will exceed that of the nonfading one as indicated by the shaded region of Fig. 2.

The corresponding gain emerges without increasing power or rate and becomes more pronounced as the number of users increases. Because such a means of enhancing average capacity of multiuser fading channels resembles that of multiantenna selection diversity systems with space-time coding [26], it is referred to as multiuser diversity gain. Note though that critical to effecting multiuser diversity gains is the availability of CSIT, not required by space-time coded multiantenna systems. This explains why opportunistic beamforming with "dumb antennas" outperforms Alamouti's space-time codes by 3 dB in required SNR [81].

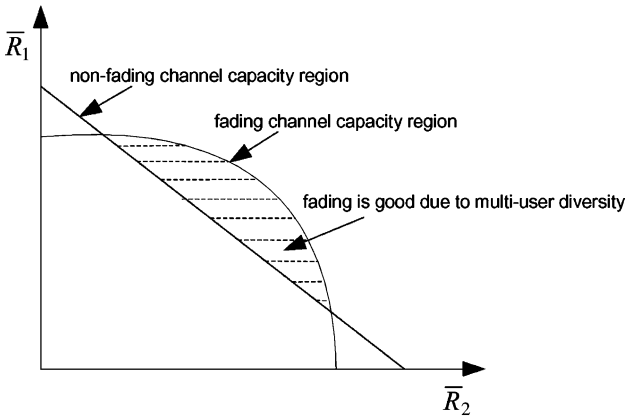



**Fig. 2.** ***Multiuser diversity enhances achievable average sum-rate.***

Although from an IT-PHY perspective multiuser diversity can turn the "curse" of fading to a "blessing" it was not until the introduction of proportional fair scheduling (PFS) [77] that its primary role was recognized for wireless scheduling of BE traffic at the MAC layer. In the PFS algorithm applied to the downlink, each user terminal $k$ determines and feeds back to the AP its rate request $r_k(h_k[n])$ adapted to its channel realization at slot $n$, as we discussed in the previous section [cf. (1)], i.e., either to meet a prescribed BER $\check{\epsilon}_k$ or to attain its maximum $r_k^*(h_k[n]) = \log_2(1+h_k[n]P)$. Having available the sample mean $\hat{\bar{r}}_k[n]$ of each user's rate averaged over previous slots, the AP finds the largest ratio $r_k(h_k[n])/\hat{\bar{r}}_k[n]$, and schedules only the corresponding user with index

$$k^*[n] = \arg\max_k r_k(h_k[n])/\hat{\bar{r}}_k[n] \tag{2}$$

to transmit over the entire slot $n$. At the same time, it updates the sample averages using standard stochastic approximation recursions $\forall k = 1, \ldots, K$ (see, e.g., [43])

$$\hat{\bar{r}}_k[n+1] = \hat{\bar{r}}_k[n] + \beta_n\big(\tau_k(h_k[n])r_k(h_k[n]) - \hat{\bar{r}}_k[n]\big) \tag{3}$$

where $\tau_{k^*[n]} = 1$ and $\tau_k(h_k[n]) = 0\ \forall k \neq k^*[n]$. The stepsize $\beta_n \in (0,1)$ implements a forgetting factor in the averaging and can be selected to be either asymptotically vanishing (e.g., $\beta_n = 1/n$) or constant ($\beta_n = \beta$). Similar to the least mean-square (LMS) algorithm, the "workhorse" of adaptive filtering, a constant stepsize gains robustness to channel nonstationarities; whereas $\beta_n \to 0$ ensures convergence of $\hat{\bar{r}}_k[n]$ to the ensemble average rate $\mathbb{E}_{h_k}[r_k(h_k)]$ when the channel process $h_k[n]$ is stationary [68].

The PFS algorithm capitalizes on multiuser diversity as it selects the user terminal with the channel having the highest peak relative to its own average; and since fading channel gains of different users fluctuate independently, most likely there will be a user near its relative peak at any slot. Picking that user in its own relative channel peak instead of the user with absolutely highest channel gain also results in "proportional fairness," hence the abbreviation PFS. Furthermore, since the probability of having a

relative channel peak is equal $\forall k = 1, \ldots, K$, users are served with equal probability regardless of their possibly different average channel quality [38].

Note also that PFS does not require knowledge of the underlying channel distribution. Only with CSIT available per slot, application of PFS in the CDMA 2000 1xEV-DO system has been found to even double the achievable rates of BE traffic in the downlink [91].

### B. Resource Allocation and Scheduling Algorithms

Effecting the multiuser diversity provided by multiuser fading channels boils down to judicious assignment of user rates—a task falling under the class of optimal resource allocation problems. This class has a rich history in IT which started with the definition of multiple access (uplink) and broadcast (downlink) communication channels in [4], [48], and [18], respectively. Early works on the capacity of fading multiple access and broadcast channels are [25] and [41]; while general results including optimal resource allocation can be found in [32], [47], [76], [45], and [46]; see also [9], where delay constraints were handled via dynamic programming; and [36], where a neat duality was established between multiple access and broadcast channels.

These IT-PHY approaches to channel-adaptive resource allocation rely on convex and nonlinear optimization tools and assume knowledge of the fading channel distribution to maximize achievable rates based on capacity-related criteria under (average or instantaneous and aggregate or individual) power constraints. To this end, IT-PHY schemes are implemented *off-line* to provide fundamental benchmarks and practical guidelines while putting aside MAC layer issues related to complexity and QoS guarantees.

On the other hand, *on-line* channel-aware resource allocation for wireless networks has received growing attention since the appearance of PFS, and a flux of wireless scheduling protocols have been reported from the MAC layer community. For BE traffic, PFS was shown to belong to a class of so called *gradient scheduling (GS)* recursive algorithms, which are asymptotically optimal in the sense that as the number of slots grows large they converge to the maxima of properly defined utility functions under fairly general conditions [3], [12], [42], [73]. For nRT traffic, linear utility-based opportunistic schedulers were developed in [49] and [50], while [6] used a token counter in GS to meet minimum and maximum rate requirements. Scheduling under absolute delay constrains for RT traffic was investigated in [2], [64], and [65]. Queueing aspects were also incorporated in [7] and [65], where a scheduling algorithm was termed *throughput-optimal* if it can keep all queues stable for any given arrival process with average rates within the interior of the channel capacity region; see also [74] for a greedy primal-dual algorithm maximizing queueing network utility subject to stability. Throughput-optimal schedulers maximizing the utility of average delays were derived in [69] and [70].

MAC layer approaches to channel-adaptive resource allocation rely on stochastic approximation tools to develop and analyze convergence of simple on-line scheduling algorithms that do not require *a priori* knowledge of the underlying fading channel distribution [3], [42], [73], [74]. With simplicity, robustness, and QoS provisioning as the major goals, optimality relative to IT-PHY benchmarks is often put aside.

### C. Towards a Unified Framework

Despite differences in the tools, implementation, criteria, and constraints, the common denominator of IT-PHY and MAC approaches to scheduling is channel-aware resource allocation over multiple wireless fading links. This common thread motivates pursuing a unified framework along the lines of Fig. 3, where both Q-CSIT and F-CSIT based scheduling algorithms are included. Schedulers under this umbrella should provide QoS for BE, nRT, RT, and heterogeneous traffic. The unified framework should also address the following questions.

- *Are there complexity-optimality tradeoffs?* If yes, these tradeoffs must be delineated quantitatively; if not, the unified framework should aim at QoS-guaranteed schedulers for all types of traffic that are as simple as heuristic ones while at the same time converge to a boundary point of the IT-PHY capacity region.
- *Is the fading distribution necessary?* If not, adaptive algorithms should be derived for QoS-guaranteed scheduling which "learn" the fading channel statistics on-line and are thus able to approach, as the number of slots grows large and from any initial value, the off-line optimal solutions of IT-PHY resource allocation schemes. (Adaptive algorithms converging from arbitrary initializations are also robust to channel nonstationarities.)
- *What if resources are more or less than enough to meet QoS?* If resources are more than sufficient to

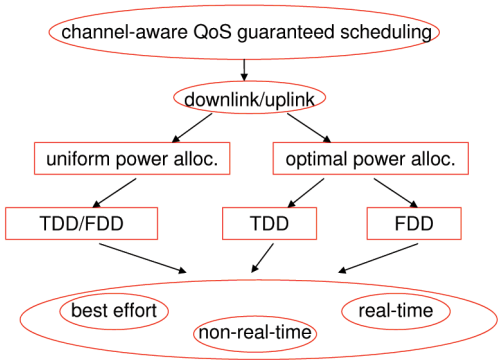



**Fig. 3.** ***Outline of the envisioned framework.***

satisfy QoS requirements, schedulers should automatically perform fair redistribution of residual resources; if not, then infeasibility of QoS provisioning should be detectable for the network to invoke a connection admission control scheme and drop users so that QoS can be guaranteed.

Recent advances outlined in the ensuing sections address many of these issues and thus push the envelope closer to the desirable unified framework for channel-aware QoS-guaranteed scheduling in wireless networks [83]–[86].

By cross-fertilizing ideas from the IT-PHY and MAC approaches, stochastic primal-dual (SPD) schemes are outlined in Sections IV and V for scheduling with uniform and optimal power allocation, respectively. It will be shown that these SPD-based schedulers boil down to solving a simple linear optimization problem involving instantaneous rates per slot. Likewise, even good (albeit suboptimal) on-line alternatives decompose the scheduling objective into a set of simple optimization subtasks involving instantaneous user rates per channel realization; see, e.g., [6], [53]. From this viewpoint, the SPD algorithms are as simple as any heuristic scheduling scheme. On the other hand, these SPD schemes provably learn the intended channel statistics online, and asymptotically converge from any initial value to the information-theoretic rate limits with average rate and/or average delay guarantees for heterogenous traffic [83], [84]. As a result, SPD schemes yield the simplest online optimal scheduling algorithms with QoS guarantees. Further considerations in Section VI will also delineate the advantages of SPD-based schedulers in detecting the (in-)feasibility of QoS provisioning and facilitating priorities associated with fairness and connection admission control.

## IV. SCHEDULING WITH UNIFORM POWER ALLOCATION

We will first consider channel-aware scheduling when the power is fixed and can thus be assumed w.l.o.g. uniform across users, i.e., $p_k = p \ \forall k = 1, \ldots, K$. Similar to schedulers adopted by wireless standards, bypassing the task of optimizing power allocation the schemes in this section are motivated by simplicity in implementation.

Given power $p$ and prescribed BER $\check{\epsilon}_k$, it is possible for the AMC-based connection $k$ to select per channel realization $\boldsymbol{h}$ the optimal AMC mode [cf. (1)]

$$m_k^*(\boldsymbol{h}) = \max\left\{m : \epsilon_k(h_k p, \rho_{k,m}) \leq \check{\epsilon}_k\right\} \tag{4}$$

with the highest rate $\rho_{k,m_k^*(\boldsymbol{h})}$ which satisfies automatically the maximum allowable BER. This AMC mode is chosen at the AP in the uplink and at the terminals in the downlink. But since in both cases the selection rule is common, optimality criteria for uplink and downlink scheduling are similar except for differences arising due to the type of service, as we discuss next using uplink for specificity and starting with BE traffic.

### A. BE Traffic

Having available $\boldsymbol{h}$ and $p$, the AP in TDMA finds via (4) the maximum allowable rate $r_k(h_k) = \rho_{k,m_k^*(h_k)}$ which meets each user's prescribed BER. With power and each user's maximum rate fixed, the AP can only select the channel (equivalently here the time) assignment across users; i.e., the AP looks for the time schedule (slot fractions) $\boldsymbol{\tau}(\boldsymbol{h}) := \{\tau_k(\boldsymbol{h})\}_{k=1}^K$, where $\tau_k(\boldsymbol{h})$ denotes the fraction allocated to user $k$. Clearly, the rate of user $k$ over its time fraction is $r_k(h_k)\tau_k(h_k)$. Hence, with weights $w_k \geq 0$, the optimal allocation maximizes the weighted average sum-rate subject to (s.to) the time allocation constraint, i.e.,

$$\max_{\boldsymbol{\tau}(\boldsymbol{h})} \mathbb{E}_{\boldsymbol{h}}\left[\sum_{k=1}^K w_k \tau_k(\boldsymbol{h}) r_k(h_k)\right]; \ \text{s.to} \ \sum_{k=1}^K \tau_k(\boldsymbol{h}) \leq 1, \ \forall \boldsymbol{h}. \tag{5}$$

For the index $k^*(\boldsymbol{h}) := \arg\max_k w_k r_k(h_k)$, it holds that $\sum_{k=1}^K w_k \tau_k(\boldsymbol{h}) r_k(h_k) \leq w_{k^*} r_{k^*}^*(h_{k^*}) \sum_{k=1}^K \tau_k(\boldsymbol{h}) \leq w_{k^*} r_{k^*}^*(h_{k^*})$, $\forall \boldsymbol{h}$, where the last upper-bound is achieved with equality if the entire slot is assigned to user $k^*$. This simple argument proves that for each realization $\boldsymbol{h}$ [and thus for the average rates in (5)], the optimal time allocation is $\tau_{k^*}^*(\boldsymbol{h}) = 1$ and $\tau_k^*(\boldsymbol{h}) = 0, \ \forall k \neq k^*$. This is a greedy allocation since *the winner user (i.e., the one with largest weighted rate) takes the entire slot*. The resultant schedule is also referred to as opportunistic because a terminal transmits when the opportunity of a reliable channel presents itself, allowing a high weighted rate over its wireless link.

Note that with the optimum user-time assignment available per realization $\boldsymbol{h}$, the ensemble average in (5) can be evaluated if the cdf $F(\boldsymbol{h})$ is known. Indeed, upon drawing sufficiently many (say $N_t$) training vectors $\boldsymbol{h}^{(t)}$ from $F(\boldsymbol{h})$, we can form the sample average $N_t^{-1} \sum_{t=1}^{N_t} \tau_k(\boldsymbol{h}) r_k(h_k)$, which thanks to the ergodicity of $\boldsymbol{h}$ approaches the ensemble average rate $\bar{r}_k(\boldsymbol{\tau}) := \mathbb{E}_{\boldsymbol{h}}[\tau_k(\boldsymbol{h}) r_k(h_k)]$.

*Weighted Average Rate Limits:* Proceeding to characterize the maximum achievable rates of TDMA under uniform power allocation, let $\mathcal{F}$ denote the set of all feasible time allocation policies, i.e., those satisfying $\sum_{k=1}^K \tau_k(\boldsymbol{h}) \leq 1$, $\forall \boldsymbol{h}$. Upon replacing $\rho_{k,m_k^*(h_k)}$ with Shannon's limit rate $r_k^*(h_k)$, the convex set of maximum achievable rates is $\mathcal{C} := \bigcup_{\boldsymbol{\tau}(\boldsymbol{h}) \in \mathcal{F}} \bar{\boldsymbol{r}}(\boldsymbol{\tau})$, where $\bar{\boldsymbol{r}}(\boldsymbol{\tau}) := [\bar{r}_1(\boldsymbol{\tau}), \ldots, \bar{r}_K(\boldsymbol{\tau})]^T$.

Associated with different weight vectors $\boldsymbol{w} := [w_1, \ldots, w_K]^T \geq 0$, each boundary point of $\mathcal{C}$ maximizes a corresponding weighted sum of average rates, i.e., it solves the convex optimization problem

$$\max_{\bar{\boldsymbol{r}}} \boldsymbol{w}^T \bar{\boldsymbol{r}}, \quad \text{s.to} \quad \bar{\boldsymbol{r}} \in \mathcal{C} \tag{6}$$

where the optimal $\bar{r}_k^* := \mathbb{E}_{\boldsymbol{h}}[\tau_k^*(\boldsymbol{h}) r_k(h_k)]$, and the optimal time allocation $\tau_k^*(\boldsymbol{h})$ is provided by the solution of (5). Clearly, if Shannon's limit is replaced by $\rho_{k,m_k^*(h_k)}$, the equivalence of (5) with (6) implies that the greedy solution of (5) yields the maximum achievable rates for TDMA under constant power and finite AMC constraints.

In addition to traversing the boundary of $\mathcal{C}$, the weights in this IT-PHY based approach can affect user fairness and priority. However, a desirable choice is only possible if sufficient channel statistics are known. For instance, if the means of the user channels $\bar{h}_k$ are known, then choosing offline the weights $w_k = \bar{h}_k$ leads to proportional fairness under which every user is served with equal probability [77]. Practical scheduling schemes on the other hand, certainly welcome on-line solutions that do not require *a priori* knowledge of $F(\boldsymbol{h})$ and are capable of adapting weights to effect desirable user fairness. This is possible with the class of gradient scheduling algorithms outlined next.

*Utility-Based GS Class:* Utility functions have been traditionally employed in economics to quantify the degree of satisfaction a user enjoys in using a certain resource. In networking, price or utility functions of rate, power and/or delay resources have been adopted recently to develop fair and efficient allocation as well as flow control schemes [12], [16], [44], [61]. Based on a reverse engineering approach, a family of "good" utility functions for different types of applications can be found in [44].

In our scheduling context, we select a concave and monotonically increasing utility function $U_{BE,k}(\bar{r}_k)$ for each user with BE traffic, and consider

$$\max_{\bar{\boldsymbol{r}}} \sum_{k=1}^{K} U_{BE,k}(\bar{r}_k), \quad \text{s.to } \bar{\boldsymbol{r}} \in \mathcal{C}. \tag{7}$$

Clearly, (7) includes (6) as a special case when $U_{BE,k}(\bar{r}_k) := w_k \bar{r}_k$. Aiming to replace the expectation in $\bar{r}_k$ over channel realizations drawn from $F(\boldsymbol{h})$ with averaging over time slots $n$, recall the on-line averaging performed in the PFS recursion (3). Substituting (3) into (7) and using Taylor's expansion with stepsize $\beta_n$ sufficiently small, we can write ($'$ denotes derivative)

$$\sum_{k=1}^{K} U_{BE,k}\big(\hat{\bar{r}}_k[n+1]\big) \approx \sum_{k=1}^{K} U_{BE,k}\big(\hat{\bar{r}}_k[n]\big) + \sum_{k=1}^{K} U'_{BE,k}\big(\hat{\bar{r}}_k[n]\big)\beta_n\big(\tau_k(\boldsymbol{h}[n]) r_k(h_k[n]) - \hat{\bar{r}}_k[n]\big). \tag{8}$$

Since $\hat{\bar{r}}_k[n]$ and thus $U_{BE,k}(\hat{\bar{r}}_k[n])$ as well as $U'_{BE,k}(\hat{\bar{r}}_k[n])$ are available at slot $n$, maximizing $\sum_{k=1}^{K} U_{BE,k}(\hat{\bar{r}}_k[n+1])$ amounts to solving [cf. (8)]

$$\begin{cases} \max\limits_{\boldsymbol{\tau}(\boldsymbol{h}[n])} & \sum_{k=1}^{K} U'_{BE,k}\big(\hat{\bar{r}}_k[n]\big)\tau_k(\boldsymbol{h}[n]) r_k(h_k[n]) \\ \text{s.to} & \sum_{k=1}^{K} \tau_k(\boldsymbol{h}[n]) \leq 1. \end{cases} \tag{9}$$

Defining the index $k^* := \arg\max_k U'_{BE,k}(\hat{\bar{r}}_k[n]) r_k(h_k[n])$ and repeating the argument we used to solve (5), yield the optimal policy per slot $n$ as: $\tau_{k^*}^*(\boldsymbol{h}[n]) = 1$ and $\tau_k^*(\boldsymbol{h}[n]) = 0$, $\forall k \neq k^*$. Using this time assignment in (3), average rates $\hat{\bar{r}}_k[n+1]$ can be found $\forall k$, and (9) can be subsequently solved for slot $n+1$. This scheme constitutes the GS class in [3] and [73]. The so obtained sequence of averages $\hat{\bar{r}}_k[n]$ converges in probability as $n \to \infty$, regardless of the initialization, to the ensemble $\bar{r}_k^*$ which solves (7); see [3], [42], and [73].

*Unification:* The optimal solution of both average rate maximization problems in (6) and (7) amounts to a greedy time allocation per slot. But the weights in (6) are fixed to $w_k$, whereas those in (7) are adapted per slot according to $U'_{BE,k}(\hat{\bar{r}}_k[n])$ [cf. (9)]. Upon convergence, the GS weights are $U'_{BE,k}(\bar{r}_k^*)$ and the solutions of (6) and (7) will coincide if $w_k = U'_{BE,k}(\bar{r}_k^*)$. (Of course, this selection is generally impossible beforehand since the limit $\bar{r}_k^*$ is not available.) Whereas the IT-PHY approach quantifies optimality in terms of fundamental limits, the GS class enjoys two attractive features: 1) its on-line schemes converge to the boundary of $\mathcal{C}$, thus achieving maximum average rates without requiring knowledge of the channel cdf and 2) a number of degrees of freedom becomes available through the selection of the utility functions which bring flexibility to design schedulers fulfilling additional desirable properties. For example, if $U_{BE,k}(\cdot) := \ln(\cdot)$, GS can learn the users' average channel quality $\bar{h}_k$ online and implement the PFS algorithm. More importantly, this flexibility allows accommodation of services other than BE as discussed in the next subsection.

### B. nRT Traffic

As mentioned earlier, nRT services entail BER as well as minimum rate requirements $\check{\boldsymbol{r}} := [\check{r}_1, \ldots, \check{r}_K]^T \geq \boldsymbol{0}$. In

fact, since $\check{\boldsymbol{r}} = \mathbf{0}$ corresponds to no rate guarantees, BE traffic can be seen as a special class of nRT traffic. The minimum rate vector $\check{\boldsymbol{r}}$ defines a hyper-paralleliped $\mathcal{S}_{\check{\boldsymbol{r}}} := \{\bar{\boldsymbol{r}} : \bar{r}_k \geq \check{r}_k, \forall k\}$, through which the maximum achievable rate region with nRT traffic can be defined as the intersection $\mathcal{C} \cap \mathcal{S}_{\boldsymbol{r}}$. In the IT-PHY approach with a given $\boldsymbol{w}$, optimal scheduling aims at [cf. (5)]

$$\begin{cases} \max\limits_{\boldsymbol{\tau}(\boldsymbol{h})} & \mathbb{E}_{\boldsymbol{h}}\left[\sum_{k=1}^{K} w_k \tau_k(\boldsymbol{h}) r_k(h_k)\right] \\ \text{s.to} & \sum_{k=1}^{K} \tau_k(\boldsymbol{h}) \leq 1, \forall \boldsymbol{h}; \mathbb{E}_{\boldsymbol{h}}[\tau_k(\boldsymbol{h}) r_k(h_k)] \geq r_k, \forall k. \end{cases} \tag{10}$$

Using the method of Lagrange multipliers, the rate constraints in (10) introduce dual variables (i.e., multipliers) $\boldsymbol{\lambda} := [\lambda_1, \ldots, \lambda_K]^T$. If we fix $\boldsymbol{\lambda}$ and $\boldsymbol{h}$, then similar to BE traffic, it can be shown that the user with index

$$k^*(\boldsymbol{\lambda}, \boldsymbol{h}) = \arg\max_k (w_k + \lambda_k) r_k(h_k) \tag{11}$$

yields the optimal time allocation: $\tau_{k^*}^*(\boldsymbol{\lambda}, \boldsymbol{h}) = 1$ and $\tau_k^*(\boldsymbol{\lambda}, \boldsymbol{h}) = 0$, $\forall k \neq k^*$, which is again a greedy one [83]–[86]. Notice though that with nRT traffic the weights $w_k$ of users with low average channel gains but high rate requirements are upgraded through the addition of the multipliers $\lambda_k$. (For the maximization in (10), the non-negative rate constraints imply that $\lambda_k \geq 0$ $\forall k$.)

To complete the optimal time allocation, the optimal $\boldsymbol{\lambda}^*$ must be specified in (11). Satisfying the average rate constraints, this is accomplished through iterations (indexed by $i$) for $k = 1, \ldots, K$

$$\lambda_k^{(i+1)} = \left[\lambda_k^{(i)} - \beta_i \left(\mathbb{E}_{\boldsymbol{h}}\left[\tau_k^*\left(\boldsymbol{\lambda}^{(i)}, \boldsymbol{h}\right) r_k(h_k)\right] - \check{r}_k\right)\right]^+ \tag{12}$$

where $[x]^+ := \max(x, 0)$ ensures that the Lagrange multipliers are always non-negative. The expected value in (12) is obtained offline by averaging as before over realizations drawn from the cdf $F(\boldsymbol{h})$. In practice, iterations are terminated when $|\lambda_k^{(i+1)} - \lambda_k^{(i)}| < \varepsilon_\lambda$ for a tolerance $\varepsilon_\lambda$, in which case the constraints $\mathbb{E}_{\boldsymbol{h}}[\tau_k^*(\boldsymbol{\lambda}^{(i)}, \boldsymbol{h}) r_k(h_k)] \geq \check{r}_k$ are met. If the rate requirements are feasible, i.e., $\check{\boldsymbol{r}} \in \mathcal{C}$, the set $\mathcal{C} \cap \mathcal{S}_{\check{\boldsymbol{r}}}$ is nonempty. In this case, (12) represents a standard sub-gradient projection update which converges fast to the unique optimum $\boldsymbol{\lambda}^*$ from any initial non-negative value, e.g., $\boldsymbol{\lambda}^{(0)} = \mathbf{0}$. This is guaranteed because the problem in (10) is convex [10], [67]. Upon convergence, the optimal offline scheduling scheme is obtained for nRT traffic with minimum average rates $\check{\boldsymbol{r}}$ guaranteed.

*Utility-Based Algorithm:* As with the GS algorithm, it is possible to formulate a utility maximization problem

$$\max_{\bar{\boldsymbol{r}}} \sum_{k=1}^{K} U_{nRT,k}(\bar{r}_k), \quad \text{s.to} \quad \bar{\boldsymbol{r}} \in \mathcal{C} \cap \mathcal{S}_{\check{\boldsymbol{r}}} \tag{13}$$

where the functions $U_{nRT,k}(\cdot)$ are chosen concave and monotonically increasing. Using again adaptive weights $U'_{nRT,k}(\hat{\bar{r}}_k[n])$, the winner user index is

$$k^* = \arg\max_k \left(U'_{nRT,k}\left(\hat{\bar{r}}_k[n]\right) + \hat{\lambda}_k[n]\right) r_k(h_k[n])$$

where $\hat{\bar{r}}_k[n]$ and $\hat{\lambda}_k[n]$ denote the estimated average rate and Lagrange multiplier associated with the rate constraint of terminal $k$ at the beginning of slot $n$. Arguing as before, the optimal time allocation per slot ends up being a greedy one with $\tau_{k^*}^*(\hat{\boldsymbol{\lambda}}[n], \boldsymbol{h}[n]) = 1$ and $\tau_k^*(\hat{\boldsymbol{\lambda}}[n], \boldsymbol{h}[n]) = 0$, $\forall k \neq k^*$.

Substituting this time assignment into (3), average rates $\hat{\bar{r}}_k[n+1]$ can be updated $\forall k$. Likewise for the estimates $\hat{\lambda}_k[n]$, LMS-like recursions can be used across slots after dropping the expectation in (12) to obtain for $k = 1, \ldots, K$

$$\hat{\lambda}_k[n+1] = \left[\hat{\lambda}_k[n] - \beta_n \left(\tau_k^*\left(\hat{\boldsymbol{\lambda}}[n], \boldsymbol{h}[n]\right) r_k(h_k[n]) - \check{r}_k\right)\right]^+. \tag{14}$$

This is a stochastic approximate of the subgradient projection update in (12). Without knowing $F(\boldsymbol{h})$, it learns the required expectation on-line. And since the utility based optimization for nRT traffic in (13) relies on such a stochastic primal-dual approach [83], it is abbreviated as nRT-SPD algorithm. Notice that if $\check{r}_k = 0$ in (14), the iterates $\hat{\lambda}_k[n]$ will converge to 0, since $\beta_n \tau_k^* \geq 0$; and the overall nRT-SPD algorithm will implement GS as it should. (Recall that without minimum rate guarantees BE traffic is equivalent to nRT traffic.)

For nRT traffic, if the token-based algorithm of [6] converges (an issue not resolved in [6]), it converges to the optimum solution. On the other hand, the provably convergent algorithm in [50] may converge to a suboptimum solution. Fortunately, this suboptimality can be detected and heuristic remedies have been reported which, as confirmed by simulations in [50], reach the optimum.

Without limiting applicability to linear utilities (adopted by [50]) and through the use of a stochastic subgradient scheme, the simple on-line nRT-SPD algorithm yields asymptotically the optimal scheduling with the required minimum average rate guarantees.

Specifically, with $\bar{r}_k^*$ denoting the solution of (13) and $int(\mathcal{C})$ the interior of $\mathcal{C}$, asymptotic optimality can be asserted as follows: *If* $\check{\boldsymbol{r}} \in int(\mathcal{C})$, *then the estimates* $\hat{\bar{r}}_k[n]$ *obtained recursively from* (3) *using any initial* $\hat{\bar{r}}_k[0] \geq 0$, *converge in probability to* $\bar{r}_k^*$ $\forall k$, *as* $n \to \infty$ *and* $\beta_n \downarrow 0$ [83]. Convergence of the nRT-SPD algorithm cannot be guaranteed when $\check{\boldsymbol{r}}$ is on the boundary of $\mathcal{C}$. This uncertainty is in the same spirit as the one encountered when proving stability at the boundary points of the capacity region in queueing analysis [39]. Furthermore, it can be shown that if $\check{\boldsymbol{r}} \notin int(\mathcal{C})$, the update in (14) diverges. Hence, infeasibility of scheduling nRT traffic can be detected. This feature is also useful in devising connection admission control policies as discussed in Section VI.

Being a generalization of GS, the nRT-SPD algorithm exhibits similar behavior in convergence. With a small but constant stepsize $\beta_n = \beta$, the nRT-SPD algorithm brings $\hat{\bar{\boldsymbol{r}}}[n]$ to a small neighborhood (with size $\mathcal{O}(\beta)$) of $\bar{\boldsymbol{r}}^*$ in $\mathcal{O}(1/\beta)$ iterations, uniformly for any initial state. Convergence to the exact $\bar{\boldsymbol{r}}^*$, is ensured with an asymptotically vanishing stepsize, e.g., $\beta_n = 1/n$ [42], [43], [50]. However, such a stepsize lowers robustness to channel nonstationarities [43], [73]. In fact, the robustness-convergence tradeoff is the same as the tracking-accuracy tradeoff encountered with the LMS algorithm [68].

*Example:* The plots in Fig. 4 depict the learning curves of nRT-SPD for $\beta_n = 0.001$ and $\beta_n = 0.01$, in a two-user TDM/TDMA packet access system where the user fading processes are uncorrelated and Rayleigh distributed. Clearly, using a smaller $\beta_n$ in SPD iterations results in faster convergence but larger variability. Using $\log_2(\bar{r}_k + 10^{-4})$, $k = 1, 2$, as utility functions and with prescribed rate requirements $\check{r}_1 = 20$ kb/s and $\check{r}_2 = 50$ kb/s, it is also observed that the nRT-SPD algorithm converges to a point satisfying both rate requirements.

### C. RT Traffic

Scheduling algorithms for RT services can be pursued with either absolute or average delay requirements. Optimal scheduling for RT traffic under absolute delay constraints has been developed based on dynamic programming, but incurs exponential complexity in the number of slots considered [2], [9]. A simpler design is derived in [83] based on an average delay criterion. For a given time allocation $\boldsymbol{\tau}(\cdot)$ and arrival rates $\bar{\boldsymbol{\alpha}} := [\bar{\alpha}_1, \ldots, \bar{\alpha}_K]^T$, let $\bar{\boldsymbol{d}}(\boldsymbol{\tau}) := [\bar{d}_1(\boldsymbol{\tau}), \ldots, \bar{d}_K(\boldsymbol{\tau})]^T$ denote the vector of average queueing delays. Using the set $\mathcal{F}$ of feasible schedules defined in Section IV-A, and paralleling the definition of the capacity region $\mathcal{C}$, the convex region of maximum achievable average delays can be defined as $\mathcal{D} := \bigcup_{\boldsymbol{\tau}(\cdot) \in \mathcal{F}} \bar{\boldsymbol{d}}(\boldsymbol{\tau})$. Given the arrival process and fading distribution, finding the average packet delay $\bar{\boldsymbol{d}}(\boldsymbol{\tau})$ is difficult. In many cases, even upper or lower delay bounds are not available [31], [39]. Likewise, $\mathcal{D}$ cannot be characterized analytically. Nonetheless, it is intuitively clear that average delays are affected by the size and stability of queues.

For RT traffic with arrival rates $\bar{\boldsymbol{\alpha}} \in int(\mathcal{C})$, scheduling ensures stability of queues (and thus bounded average delays) so long as (functions of) queue sizes are used to adapt the rate reward weights $w_k$ [7], [22], [66], [75]. Queueing-aware weight adaptation was linked to a utility function of average delays in [69] and [70]. While various algorithms adapted to queue-sizes deal with stability issues in scheduling RT traffic, their optimality has not been fully characterized under delay constraints. This is pursued in [83] through a utility-based on-line approach. Corresponding to a vector of maximum allowable average delay requirements $\check{\boldsymbol{d}} := [\check{d}_1, \ldots, \check{d}_K]^T$, consider the set $\mathcal{S}_{\check{\boldsymbol{d}}} := \{\bar{\boldsymbol{d}} : \bar{d}_k \leq \check{d}_k, \forall k\}$ based on which the maximum achievable delay region under average delay requirements is given by $\mathcal{D} \cap \mathcal{S}_{\check{\boldsymbol{d}}}$. As [69] dealt with unconstrained utility maximization to ensure stability,

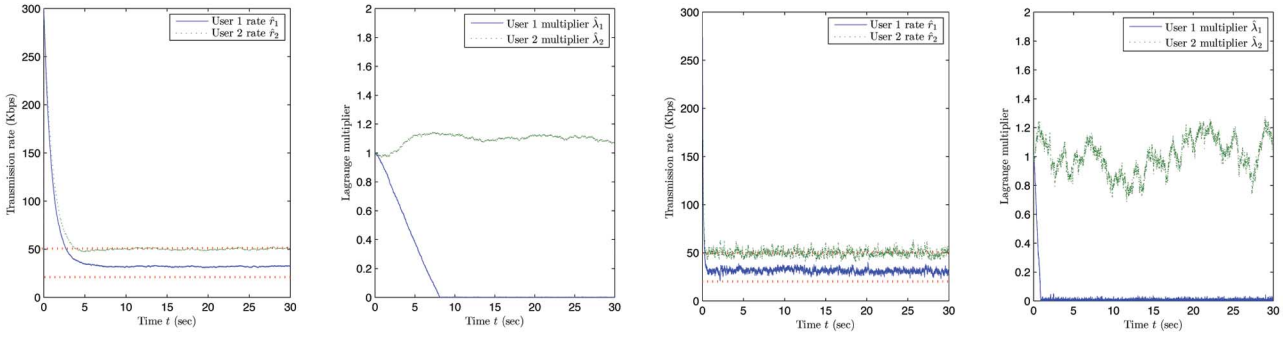



**Fig. 4.** ***Learning curves of nRT-SPD for $\beta = 0.001$ (left), and $\beta = 0.01$ (right); dashed lines indicate rate requirements $\check{r}_1 = 20$ kb/s and $\check{r}_2 = 50$ kb/s.***

[83] considers utility-based scheduling under the average delay constraints

$$\max_{\bar{\boldsymbol{d}}} \sum_{k=1}^{K} U_{RT,k}(\bar{d}_k), \ \text{s.to}\ \bar{\boldsymbol{d}} \in \mathcal{D} \cap \mathcal{S}_{\check{\boldsymbol{d}}} \tag{15}$$

where the functions $U_{RT,k}(\cdot)$ must be chosen here to be concave but monotonically decreasing, since the utility of user $k$ should decrease as $\bar{d}_k$ increases. Note that (15) is feasible for any arrival rate in the capacity region, if queues are stable. (If unstable even for one $k$, the constraint cannot be satisfied when the average delay requirements are finite.)

Assuming that each connection has a sufficiently large input queue, let $q_k[n]$ denote the queue size (in bits) of terminal $k$ at the beginning of slot $n$, and $\alpha_k[n]$ the number of arriving bits at slot $n$. With departure rate $\tau_k(\boldsymbol{h}[n])r_k(h_k[n])$, the queue size obeys the recursion

$$q_k[n+1] = q_k[n] - \min\{\tau_k(\boldsymbol{h}[n])r_k(h_k[n])T_s, q_k[n]\} + \alpha_k[n].$$

Similar to the average rate in (3), the average queue length can be updated on-line using: $\hat{\bar{q}}_k[n+1] = \hat{\bar{q}}_k[n] + \beta_n(q_k[n+1] - \hat{\bar{q}}_k[n])$. Little's result [39] on the other hand, asserts that with stable queues the average delay is given by the average queue length divided by the average arrival rate, i.e., $\bar{d}_k[n] = \bar{q}_k[n]/\bar{\alpha}_k$. This in turn leads to a recursive estimate of the average delay via

$$\hat{\bar{d}}_k[n+1] = \hat{\bar{d}}_k[n] + \beta_n\Big(\bar{\alpha}_k^{-1}(q_k[n] + \alpha_k[n] - \min\{\tau_k(\boldsymbol{h}[n])r_k(h_k[n])T_s, q_k[n]\}) - \hat{\bar{d}}_k[n]\Big). \tag{16}$$

Interestingly, (16) relates average delay estimates $\hat{\bar{d}}_k[n+1]$ with instantaneous rates $r_k(h_k[n])$. Such a relationship allows application of SPD principles to RT scheduling.

With $\hat{\boldsymbol{\lambda}}[n] := [\hat{\lambda}_1[n], \ldots, \hat{\lambda}_K[n]]^T$ denoting the estimated Lagrange multiplier vector corresponding to the average delay constraints at time slot $n$, maximization of $\sum_{k=1}^{K} U_{RT,k}(\hat{\bar{d}}_k[n+1])$ boils down to (after a first-order approximation of its Taylor's expansion) the linear programming problem [cf. (16)]

$$\begin{cases} \max\limits_{\boldsymbol{\tau}(\boldsymbol{h}[n])} & \sum_{k=1}^{K}\Big(-U'_{RT,k}\Big(\hat{\bar{d}}_k[n]\Big) + \hat{\lambda}_k[n]\Big)\tau_k(\boldsymbol{h}[n])r_k(h_k[n]) \\ \text{s.to} & \sum_{k=1}^{K}\tau_k(\boldsymbol{h}[n]) \le 1; \\ & \tau_k(\boldsymbol{h}[n])r_k(\boldsymbol{h}[n]) \le q_k[n]/T_s, \ \forall k. \end{cases} \tag{17}$$

To solve (17), users have to be sorted in descending order according to their weighted rates; i.e., if users are indexed via $\boldsymbol{u} := [u(1), \ldots, u(K)]^T$ it must hold that $-U'_{RT,u(i)}(\hat{\bar{d}}_{u(i)}[n]) + \hat{\lambda}_{u(i)}[n]r_{u(i)}(\boldsymbol{h}[n]) \ge (-U'_{RT,u(i+1)}(\hat{\bar{d}}_{u(i+1)}[n]) + \hat{\lambda}_{u(i+1)}[n])r_{u(i+1)}(\boldsymbol{h}[n])\forall i = 1, \ldots, K-1$. Based on this ordering, the scheduler should first consider assigning the entire slot to user $u(1)$ with maximum weighted rate. If only part of the slot is required to serve all the data in $u(1)$'s queue, the remaining time should be assigned to user $u(2)$. This allocation continues until the entire slot is assigned to users or data in all user queues are cleared. Using this time schedule $\tau_k^*(\boldsymbol{h}[n])$, the average delay $\hat{\bar{d}}_k[n+1]$ is updated via (16), and subsequently the Lagrange multiplier vector is updated using $\forall k$ the LMS-like recursion

$$\hat{\lambda}_k[n+1] = \Big[\hat{\lambda}_k[n] - \beta_n\big(\check{d}_k - q_k[n+1]/\bar{\alpha}_k\big)\Big]^+. \tag{18}$$

These updates together constitute an SPD algorithm for RT traffic, naturally abbreviated as RT-SPD.

The RT-SPD algorithm is also asymptotically optimal in the sense that if $\check{\boldsymbol{d}} \in int(\mathcal{D})$, then for $\beta_n \downarrow 0$, the estimated average delay $\hat{\bar{\boldsymbol{d}}}[n]$ converges to the optimal solution $\bar{\boldsymbol{d}}^*$ of (15), as $n \to \infty$. Different from the nRT case, the optimal time allocation for RT traffic depends on the queue sizes $q_k[n]$, and allows multiple users transmitting jointly over each slot (contrary to winner-takes-all).

When average delay requirements are arbitrarily large, i.e., $\check{d}_k \to \infty\ \forall k$, it follows from (18) that $\hat{\lambda}_k[n]$ will converge to 0. Furthermore, with $U_{RT,k}(\bar{d}_k) = -w_k\bar{d}_k^2/2$ the RT-SPD algorithm amounts to what one could term largest-weighted-average-delay-first (LWADF) scheduling, where users are served according to the order of $w_k\hat{\bar{d}}_k[n]r_k(\boldsymbol{h}[n])$. If in addition the average delay is estimated using the *instantaneous* delay, which is effected by setting $\beta_n = 1$, LWADF reduces to the largest-weighted-delay-first (LWDF) scheme reported in [7], [75], which minimizes in the limit the tail of the delay outage probability. Similarly, selecting different utility functions yields alternative queue-size based scheduling schemes including those reported in [22] and [66]. From this viewpoint, the RT-SPD method offers a general queue-based scheduler that can attain throughput optimality as $\check{d}_k \to \infty\ \forall k$, and is thus stable for any RT traffic with arrival rates $\bar{\boldsymbol{\alpha}} \in int(\mathcal{C})$. On the other hand, the RT-SPD approach is optimal in the sense of maximizing the utility of average delays (when queues are assumed stable) even for RT traffic with (even small) finite delay requirements.

Certainly, the ultimate goal in scheduling RT traffic should be to optimize absolute (rather than average) delay guarantees. But since absolute (a.k.a. deterministic) delay requirements may lead to an overly conservative scheduling policy, current research efforts focus on providing

statistical delay guarantees by constraining the maximum allowable outage probability [53], [75]. Interestingly, the RT-SPD scheme provides such guarantees by exploiting the relationship between delay outage probabilities and average delays. Specifically, given absolute delay requirements $\check{\boldsymbol{D}} := [\check{D}_1, \ldots, \check{D}_K]^T$ and maximum outage probability requirements $\check{\boldsymbol{\xi}} := [\check{\xi}_1, \ldots, \check{\xi}_K]^T$, Markov's inequality implies readily that outage probabilities for absolute delay $d_k$, $k = 1, \ldots, K$, must obey

$$\Pr(d_k \geq \check{D}_k) \leq \bar{d}_k / \check{D}_k.$$

Based on this inequality, it is possible to set the average delay requirement $\check{d}_k = \check{D}_k \check{\xi}_k$ and apply the RT-SPD scheme with utility functions $U_{RT,k}(\bar{d}_k / \check{D}_k)$ to obtain a scheduling algorithm delivering absolute delay guarantees.

### D. Heterogeneous Traffic

GS, nRT-SPD and RT-SPD algorithms suggest a unified approach to scheduling heterogeneous traffic, by maximizing the superposition of utilities for BE, nRT, and/or RT services subject to average rate and delay constraints, namely

$$\begin{cases} \max & \sum_i U_{BE,i}(\bar{r}_i) + \sum_j U_{nRT,j}(\bar{r}_j) + \sum_k U_{RT,k}(\bar{d}_k) \\ \text{s.to} & \bar{r}_j \geq \check{r}_j, \ \forall j; \\ & \bar{d}_k \leq \check{d}_k, \ \forall k. \end{cases}$$

An SPD algorithm for heterogeneous traffic (HET-SPD) can then be developed by simply combining the rules of GS, nRT-SPD, and RT-SPD algorithms, following these steps:

i) pick the BE, nRT, or RT user with largest weighted rate (weights for nRT and RT users are the sum of first derivatives of their utility functions plus the corresponding Lagrange multipliers), and let the winner take the entire slot;
ii) update primal variables via on-line time averages, and dual variables using stochastic subgradient projections.

Notice that the suboptimal winner-takes-all rule is used here for RT traffic. The asymptotic (near) optimality of SPD algorithms for nRT and RT traffic implies also the asymptotic (near) optimality of the HET-SPD algorithm with average rate and delay guarantees.

### E. Implementation and Overhead

The optimal resource allocation schemes described so far assume that F-CSIT is available at both AP and users, a case only possible in TDD systems. Notice however that F-CSIT is only needed to select the rate $r_k(h_k) = \rho_{k,m_k^*(h_k)}$, $\forall k$. For each slot, the codeword $\boldsymbol{m}^*(\boldsymbol{h}) := [m_1^*(h_1), \ldots, m_K^*(h_K)]^T$ contains the indices of the most bandwidth-efficient AMC modes users can support under their BER requirements. This observation implies that implementation of the chosen schedule does not require the analog-valued vector channel $\boldsymbol{h}$, i.e., the quantized AMC codeword $\boldsymbol{m}^*(\boldsymbol{h})$ is sufficient to implement channel-aware scheduling. If the number of AMC modes $M_k$ is finite $\forall k$, the typical case in practice, the binary codeword $\boldsymbol{m}^*(\boldsymbol{h})$ belongs to a set $\mathcal{M}$ with cardinality $\prod_{k=1}^K (M_k + 1)$, and can thus be described by a finite number of bits. We reiterate that based on $\boldsymbol{h}$, the AMC mode information contained in $\boldsymbol{m}^*(\boldsymbol{h})$ is decided at user terminals for the downlink and at the AP for the uplink.

*TDMA Uplink:* In this case, the channel vector $\boldsymbol{h}$ is known at the AP. After selecting the scheduling parameters (AMC modes) based on $\boldsymbol{h}$, the AP broadcasts the scheduled user-mode pair $(k^*, m_{k^*}^*(h_k))$. Since there are $\sum_{k=1}^K M_k$ different user-mode combinations plus one more when all the users are deferring, the feedback link from the AP to the users must carry up to $B = \lceil \log_2(\sum_{k=1}^K M_k + 1) \rceil$ bits for the schedule to be announced to the users ($\lceil \ \rceil$ denotes the ceiling operation). Note that this is a small number for practical systems. For example, in a network of 10–20 active users with each supporting $M_k = 5$ AMC modes, the required number of feedback bits per channel realization is $B = 6$.

*TDM Downlink:* Here each terminal $k$ knows $h_k$ and can select the AMC mode specific to its own channel gain. Since the schedule is decided by the AP, each terminal needs to feedback its chosen AMC codeword $m_{k^*}^*(h_k)$, which requires a feedback channel carrying up to $B_k = \lceil \log_2(M_k + 1) \rceil$ bits per user $k$. After collecting $m_{k^*}^*(h_k)$ $\forall k$, the AP decides the optimum index $k^*$ and transmits to the corresponding user using the AMC mode $m_{k^*}^*(h_{k^*})$. To notify the users about the scheduling decision, the AP only needs a feedback link with rate $B_k = \lceil \log_2(K) \rceil$ bits per slot, since based on $h_k$ each user terminal knows its optimal AMC mode $m_{k^*}^*(h_{k^*})$ in case it is selected. This operation is simple to implement, as testified by its adoption in standardized systems such as CDMA2000 1xEV-DO and WCDMA HSDPA.

It is important to stress that as far as implementation is concerned Q-CSIT suffices for channel-aware scheduling when the power is assumed constant as long as the feedback channel can carry per slot $B = \lceil \log_2(\sum_{k=1}^K M_k + 1) \rceil$ bits for the uplink or $B_k = \lceil \log_2(M_k + 1) \rceil$ bits per connection $k$ for the downlink. This low-rate feedback requirement makes scheduling with uniform power loading a perfect fit for FDD systems such as CDMA2000 1xEV-DO.

Summarizing, the Q-CSIT vector $\boldsymbol{m}$ must be fed back from terminals to the AP (downlink transmitter) *before* downlink scheduling; whereas in uplink, the AP only needs to feedback $m_{k^*}^*$ to the scheduled user $k^*$ (uplink transmitter) *after* scheduling. Feedback for the uplink only

incurs $\lceil \log_2 \max_k M_k + 1 \rceil$ bits overhead, less than that of $\lceil \log_2 \prod_{k=1}^{K}(M_k+1) \rceil$ bits for downlink. Moreover, no dedicated channel is required for feedback in uplink, simply because Q-CSIT is part of the scheduling decision $(k^*, m^*_{k^*})$.

## V. SCHEDULING WITH OPTIMAL POWER ALLOCATION

Assigning power uniformly across time (fading states), the scheduling algorithms in Section IV allocate optimally time and rate resources. If one optimizes also power allocation, then scheduling performance can only improve. This is the goal of this section which, unexpectedly, will be possible to accomplish with minimal extra complexity in optimizing and broadcasting the schedule.

To this end, it is useful to recognize first that except for the prespecified $\rho_{k,m}$ modes, it is also possible for each connection $k$ to support transmit-rates expressed as linear combinations of these AMC modes by time sharing their usage over the $k$th slot. Specifically, using the mode $m$ over $\zeta_{k,m}$ percentage of the $\tau_k$ slot, and letting $\tau_{k,m} := \zeta_{k,m}\tau_k$, connection $k$ can support rate

$$r_k(\boldsymbol{h})\tau_k(\boldsymbol{h}) = \sum_{m=0}^{M_k} \tau_{k,m}(\boldsymbol{h})\rho_{k,m} \tag{19}$$

where clearly $\sum_{k=1}^{K}\sum_{m=0}^{M_k}\tau_{k,m} \in [0,1]$. In accordance with each AMC mode $\rho_{k,m}$, a "power mode" $p_{k,m}$ can be obtained to meet the prescribed BER $\check{\epsilon}_k$ for a given realization $h_k$. Indeed, solving the BER function in (1) w.r.t. power yields $p_{k,m} = (1/h_k)\epsilon_k^{-1}(\check{\epsilon}_k, \rho_{k,m})$. By time-sharing, any linear combination of $\{\rho_{k,m}\}$ as in (19) gives rise to the same linear combination of corresponding powers $\{p_{k,m}(h_k)\}$, which meet the prespecified BER constraint $\check{\epsilon}_k$ for a given $h_k$; hence,

$$p_k(\boldsymbol{h})\tau_k(\boldsymbol{h}) = \sum_{m=0}^{M_k} \tau_{k,m}(\boldsymbol{h})p_{k,m}(h_k). \tag{20}$$

Since $\rho_{k,m}$ is related to $p_{k,m}$ one-to-one per realization $h_k$, it suffices to optimize scheduling only w.r.t. the power variables $\boldsymbol{p}(\boldsymbol{h}) := \{p_{k,m}(h_k), m = 1, \ldots, M_k\}_{k=1}^{K}$, and w.r.t. the time allocation variables $\boldsymbol{\tau}(\boldsymbol{h}) := \{\tau_{k,m}(\boldsymbol{h}), m = 1, \ldots, M_k\}_{k=1}^{K}$. (Recall that the $m = 0$ mode corresponds to no transmission; i.e., $\rho_{k,0} = p_{k,0} = 0\ \forall k$.)

Furthermore, since $\rho_{k,m}$ can be uniquely expressed in terms of $p_{k,m}$ for a given $h_k$, rates in (19) are related with powers in (20) via a piecewise linear function $r_k(\boldsymbol{h}) = R_k(p_k(\boldsymbol{h}))$, as depicted in Fig. 5. This function in the limit $M_k \to \infty$ approaches with capacity-achieving transmissions ($\Gamma = 1$) Shannon's formula, i.e., $R_k(h_k p_k) = \log_2(1 + h_k p_k)$. The latter provides also an one-to-one mapping between the fundamental limits of rate and power. (In this case, there is no need for time sharing and the optimization variables are $\boldsymbol{p}(\boldsymbol{h}) := \{p_k(h_k)\}_{k=1}^{K}$ and $\boldsymbol{\tau}(\boldsymbol{h}) := \{\tau_k(\boldsymbol{h})\}_{k=1}^{K}$.)

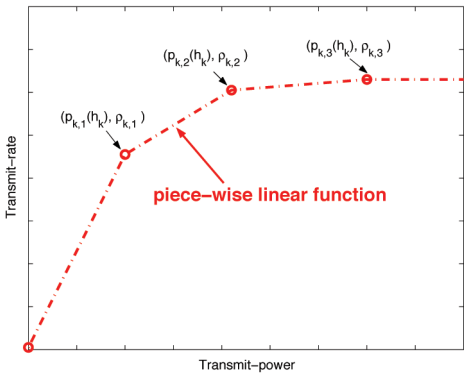



**Fig. 5.** *Piecewise linear function relating transmit-rate with transmit-power for a given channel gain when connections rely on finite AMC modes.*

Having specified the optimization variables, one can proceed to revisit the resource allocation problem starting with BE traffic and the TDM downlink for specificity.

### A. BE Traffic

In TDM downlink operation with finite AMC modes, the AP has an average sum-power requirement $\check{p}$ and seeks to solve

$$\begin{cases} \max\limits_{\boldsymbol{\tau}(\boldsymbol{h})} & \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{h}}\left[\sum_{m=0}^{M_k} w_k \tau_{k,m}(\boldsymbol{h})\rho_{k,m}\right] \\ \text{s.to} & \sum_{k=1}^{K}\sum_{m=0}^{M_k} \tau_{k,m}(\boldsymbol{h}) \le 1;\ \forall \boldsymbol{h}; \\ & \mathbb{E}_{\boldsymbol{h}}\left[\sum_{k=1}^{K}\sum_{m=0}^{M_k} \tau_{k,m}(\boldsymbol{h})p_{k,m}(h_k)\right] \le \check{p}. \end{cases} \tag{21}$$

This is a convex problem and the unique global optimum can be found using the method of Lagrange multipliers. The Lagrangian depends on the *instantaneous* reward function $\varphi_{k,m}(\lambda, \boldsymbol{h}) := w_k\rho_{k,m} - \lambda p_{k,m}(h_k)$; and the optimum user-mode pair for a given multiplier $\lambda$ (corresponding to the average power constraint) is given by [57], [84]

$$\left(k^*, m^*_{k^*}\right) := \arg\max_{(k,m)} \varphi_{k,m}(\lambda, \boldsymbol{h}). \tag{22}$$

For each $\lambda$, this winner user-mode pair once again is greedily assigned all the channel resources, i.e., $\tau^*_{k^*,m^*_{k^*}}(\lambda, \boldsymbol{h}) = \tau^*_{k^*}(\lambda, \boldsymbol{h}) = 1$, $p^*_{k^*}(\lambda, \boldsymbol{h}) = p_{k^*,m^*_{k^*}}(h_{k^*})$, $r^*_{k^*}(\lambda, \boldsymbol{h}) = \rho_{k^*,m^*_{k^*}}$, and $\tau^*_{k,m}(\lambda, \mathbf{h}) = \tau^*_k(\lambda, \boldsymbol{h}) = p^*_k(\lambda, \boldsymbol{h}) = r^*_k(\lambda, \boldsymbol{h}) = 0, \quad \forall (k,m) \neq (k^*, m^*_{k^*})$.

To determine the optimal multiplier $\lambda^*$ needed to complete the scheduler design, it suffices to recognize the similarity between (10) and (21). Both are convex problems, the first with an average rate and the second with an average power constraint. Therefore, mimicking (12), it is possible to find $\lambda^*$ as the limit of the subgradient iteration

$$\lambda^{(i+1)} = \left[\lambda^{(i)} - \beta_i\left(\check{p} - \mathbb{E}_{\boldsymbol{h}}\left[p_{k^*,m_{k^*}^*}(h_{k^*})\right]\right)\right]^+ \qquad (23)$$

where the form of the expected value has been simplified after taking into account the greedy nature of the optimal user-mode allocation. Similar to (12), this iteration requires the channel cdf to compute the expectation involved and is carried off-line.

With power uniformly fixed across time (fading states), the scheme for BE traffic in Section IV schedules for transmission the AMC mode $m_{k^*}^*(\boldsymbol{h})$ yielding the highest rate reward $w_{k^*}\rho_{k^*,m_{k^*}}$. Here, depending on the power price $\lambda$ the scheduler relies on the quality indicator $\varphi_{k,m}(\lambda,\boldsymbol{h})$ to select the user-mode pair yielding the highest *net* reward (rate reward minus power cost). If $m_{k^*}=0$ and the "winner" connection $k^*$ has to transmit with the zeroth mode (an indication of a deep fade), then all connections defer. Furthermore, note that the higher $h_k$ is the more likely it becomes to transmit at higher rate; or, the less power is required to meet the BER requirement for the same rate. And since the BER function is convex, the higher $h_k$ is the higher the quality indicator will be. These considerations indicate a discrete water-filling behind the present optimum resource allocation policy. In fact, with infinite capacity-achieving transmission modes, this policy converges to the classical water-filling procedure where power is allocated optimally per connection and the channel is assigned to the connection with the maximum net reward [84], [87]; see also [19], [45], [78].

*Weighted Average Rate Limits:* To characterize the maximum achievable rates, let $\bar{r}_k(\boldsymbol{\tau},\boldsymbol{p}) := \mathbb{E}_{\boldsymbol{h}}[\tau_k(\boldsymbol{h})R_k(p_k(h_k))]$, where $R_k(p_k(h_k))$ denotes either the piecewise linear function in Fig. 5 or Shannon's formula. With $\bar{\boldsymbol{r}}(\boldsymbol{\tau},\boldsymbol{p}) := [\bar{r}_1(\boldsymbol{\tau},\boldsymbol{p}),\ldots,\bar{r}_K(\boldsymbol{\tau},\boldsymbol{p})]^T$, the convex region of achievable rates in TDM downlink under optimum power allocation is $\mathcal{C} := \bigcup_{(\boldsymbol{\tau}(\boldsymbol{h}),\boldsymbol{p}(\boldsymbol{h}))\in\mathcal{F}} \bar{\boldsymbol{r}}(\boldsymbol{\tau},\boldsymbol{p})$, where the feasible set $\mathcal{F}$ here includes all policies satisfying constraints in (21). The boundary points and thus the entire region $\mathcal{C}$ can be determined by solving for all weight vectors $\boldsymbol{w}\geq\boldsymbol{0}$ the problem: $\max_{\bar{\boldsymbol{r}}}\boldsymbol{w}^T\bar{\boldsymbol{r}}$, s. to $\bar{\boldsymbol{r}}\in\mathcal{C}$. With a finite number of AMC modes, this is equivalent to (21). The same steps followed to solve (21) can also be used to solve the optimal resource allocation with capacity-achieving modes per connection.

*SPD Algorithm:* As with uniform power allocation, an on-line SPD scheduling algorithm can be devised through the utility based formulation in (7). With $U'_{BE,k}(\hat{\bar{r}}_k[n])$ replacing $w_k$ and estimates $\hat{\bar{\boldsymbol{r}}}[n]$, $\hat{\lambda}[n]$ available at slot $n$, the optimum user-mode pair is selected similar to (22) but with the net reward defined on-line as

$$\varphi_{k,m}\left(\hat{\lambda}[n],\boldsymbol{h}[n]\right) := U'_{BE,k}\left(\hat{\bar{r}}_k[n]\right)\rho_{k,m} - \hat{\lambda}[n]p_{k,m}(h_k[n]).$$

The entire slot is assigned to terminal $k^*$, and the AP transmits to $k^*$ with mode $m_{k^*}^*$ at power $p_{k^*}^*(\hat{\lambda}[n],\boldsymbol{h}[n]) = p_{k^*,m_{k^*}^*}(h_{k^*}[n])$. In accordance with this allocation, primal variables are updated as in (3) to find $\hat{\bar{\boldsymbol{r}}}[n+1]$, and the dual variables $\hat{\lambda}[n]$ are found using a stochastic subgradient projection [cf. (14)]

$$\hat{\lambda}[n+1] = \left[\hat{\lambda}[n] - \beta_n\left(\check{p} - p_{k^*,m_{k^*}^*}(h_{k^*}[n])\right)\right]^+.$$

These steps do not require knowledge of $F(\boldsymbol{h})$ and constitute an SPD scheduling algorithm for TDM broadcasting (BC). This BC-SPD scheme converges as $n\to\infty$ to the optimal solution of (21) with $w_k = U'_{BE,k}(\bar{r}_k^*)$.

*TDMA Uplink:* Both off-line and on-line SPD scheduling algorithms with optimum power allocation can be devised for TDMA uplink operation based on their TDM downlink counterparts. The sum-power constraint must be replaced by individual constraints $\{\mathbb{E}_{\boldsymbol{h}}[\tau_k(\boldsymbol{h})p_k(\boldsymbol{h})]\leq\check{p}_k\}_{k=1}^K$, where $\check{p}_k$ denotes the power requirement for the $k$th terminal. The scalar $\lambda$ is substituted by a vector of Lagrange multipliers $\boldsymbol{\lambda} := [\lambda_1,\ldots,\lambda_K]^T$, and the net reward $\varphi_{k,m}(\lambda_k,\boldsymbol{h})$ corresponding to terminal $k$ now depends on its own power price $\lambda_k$. Despite these differences, both off-line and on-line optimal scheduling amount to a greedy policy, where the winner user $k^*$ transmits with mode $m_{k^*}^*$ over the entire slot.

### B. QoS Guaranteed Scheduling

The SPD scheduling algorithm with optimal power allocation for BE traffic can be extended to provide QoS for nRT, RT, and heterogeneous traffic too. For specificity, consider the downlink scheduling task for nRT services.

As with uniform power allocation, the optimal on-line scheduler should be found as the solution of the utility problem in (13). The method of Lagrange multipliers entails here a dual variable vector $\boldsymbol{\lambda}^r := [\lambda_1^r,\ldots,\lambda_K^r]^T$ associated with the average rate constraints and a second scalar variable $\lambda^p$ corresponding to the average power constraint. Let $\hat{\bar{\boldsymbol{r}}}[n]$, $\hat{\boldsymbol{\lambda}}^r[n]$, and $\hat{\lambda}^p[n]$ denote estimates of $\bar{\boldsymbol{r}}$, $\boldsymbol{\lambda}^r$ and $\lambda^p$ at the beginning of slot $n$. Similar to the nRT-SPD and BC-SPD algorithms, the SPD algorithm for nRT

traffic in TDM downlink (abbreviated as BC-nRT-SPD) follows these steps per slot $n$:

i) with $\varphi_{k,m}(\hat{\lambda}^p[n], \hat{\lambda}^r_k[n], \boldsymbol{h}[n]) := (U'_{BE,k}(\hat{\bar{r}}_k[n]) + \hat{\lambda}^r_k[n])\rho_{k,m} - \hat{\lambda}^p[n]p_{k,m}(h_k[n])$, choose the user-mode pair as: $(k^*, m^*_{k^*}) = \arg\max_{(k,m)} \varphi_{k,m}(\hat{\lambda}^p[n], \hat{\lambda}^r_k[n], \boldsymbol{h}[n])$; assign the entire slot to user $k^*$; and let the AP transmit to this user using the $m^*_{k^*}$ AMC mode with power $p^*_{k^*}(\boldsymbol{h}[n]) = p_{k^*,m^*_{k^*}}(h_{k^*}[n])$;

ii) update the primal variables $\hat{\bar{\boldsymbol{r}}}[n]$ by averaging across slots, and the dual variables $\hat{\boldsymbol{\lambda}}^r[n]$ and $\hat{\lambda}^p[n]$ using a stochastic subgradient projection.

As in previous SPD algorithms, $\{U'_{nRT,k}(\hat{\bar{r}}_k[n])\}_{k=1}^K$ act as rate-reward weights to allocate time and power. However, to compensate for those users with poor average channel quality but high rate requirements, their relative priorities are promoted by adding (estimates of) the positive Lagrange multipliers $\hat{\lambda}^r_k[n]$ to their weights. One can then use as quality indicator the net reward $\varphi_{k,m}(\hat{\lambda}^p[n], \hat{\lambda}^r_k[n], \boldsymbol{h}[n]) := (U'_{BE,k}(\hat{\bar{r}}_k[n]) + \hat{\lambda}^r_k[n])\rho_{k,m} - \hat{\lambda}^p[n]p_{k,m}(h_k[n])$; and follow the optimal time and power allocation as well as updates of the dual variables similar to those in the BC-SPD algorithm. Again, it holds that if $\check{\boldsymbol{r}} \in int(\mathcal{C})$, then for $\beta_n \downarrow 0$, the BC-nRT-SPD algorithm is asymptotically optimal. Following related steps, it is also possible to derive QoS-guaranteed asymptotically optimal SPD scheduling schemes for RT and heterogeneous traffic in uplink and downlink operation.

### C. Implementation and Overhead—TDD Systems

For all types of services, the scheduling algorithms in this section include optimal power allocation and end up with a greedy allocation policy. This greedy format is common to the schemes in Section IV, where uniform power allocation was assumed. But since power allocation depends on the analog-valued $h_k\ \forall k$ differences arise too, both in selecting the scheduling parameters as well as in implementing the optimal schedule. Recall that selecting AMC rate parameters (modes $\{m^*_k\}_{k=1}^K$) under uniform power allocation required F-CSIT, which is available wherever needed in both TDD and FDD; while Q-CSIT was sufficient for implementing the optimal schedule throughout Section IV. In this section's optimal resource allocation, an additional scheduling parameter must be decided, namely power, which being dependent on the channel can be made available in full (analog-valued) form wherever needed, only for TDD systems.

In this subsection we deal with TDD systems, where F-CSIT $h_k$ is available at each terminal $k$ and $\boldsymbol{h}$ is known at the AP. Since $h_{k^*}$ is known also at terminal $k^*$, the transmit-power $p_{k^*,m^*_{k^*}}(h_{k^*})$ is available wherever needed. Whether in downlink or uplink, optimization and implementation of the schedule takes place at the AP and follows similar steps:

i) based on $\boldsymbol{h}$, the AP runs the SPD scheduling algorithm to optimize time and power allocation, and broadcasts the scheduled user-mode codeword $(k^*, m^*_{k^*})$;

ii) terminal $k^*$ transmits in uplink with rate corresponding to $m^*_{k^*}$ and power $p_{k^*,m^*_{k^*}}(h_{k^*})$; or, receives (i.e., decodes) knowing the AMC mode $m^*_{k^*}$ and power $p_{k^*,m^*_{k^*}}(h_{k^*})$ that the AP transmits in the downlink.

The induced overhead for broadcasting the schedule is clearly $B = \lceil \log_2(\sum_{k=1}^K M_k + 1) \rceil$ bits, the same as for uplink scheduling with uniform power allocation in Section IV.

The following toy-example illustrates differences of SPD scheduling under uniform versus optimal power allocation, and also compares their achievable rates against benchmarks.

*Example:* Consider downlink operation involving only two terminals. (Since the AP transmits over orthogonal TDM channels, the number of users is not critical; but having only two allows plotting the benchmark regions of maximum achievable rates.) Utility functions $\ln(\bar{r}_k + 10^{-4})$, $k = 1, 2$ are used to serve either BE or nRT traffic. For BE services, PFS is implemented with uniform power allocation (UPA) along with the BC-SPD algorithm relying on optimal power allocation (OPA). For nRT services, the nRT-SPD algorithm is implemented with UPA and requirements $\check{r}_1 = 20$ kb/s, $\check{r}_2 = 50$ kb/s; along with the BC-nRT-SPD algorithm based on OPA and $\check{r}_2 = 100$ kb/s, $\check{r}_2 = 100$ kb/s. Fig. 6 depicts using square, diamond, circle, and triangle markers the resulting

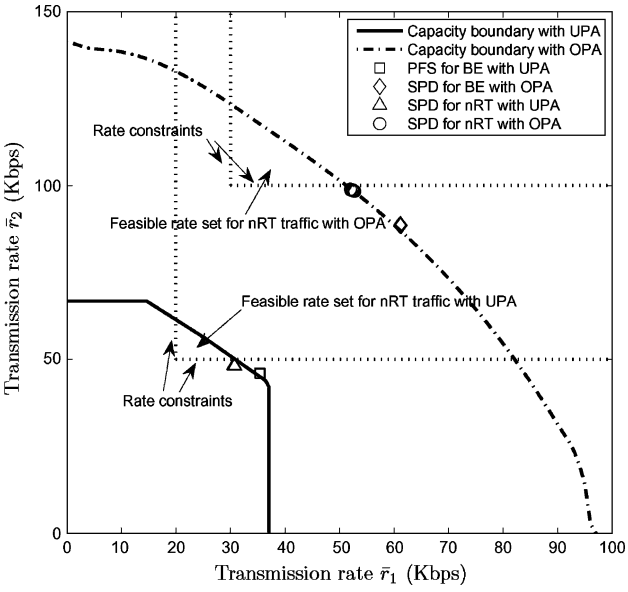



**Fig. 6.** ***Regions of maximum achievable rates (benchmarks) and achievable rates with SPD algorithms in a two-user downlink system under uniform power allocation (UPA) for TDD and FDD operation; and under optimal power allocation (OPA) for TDD operation only.***

average rates $\bar{r}_2^*$ versus $\bar{r}_1^*$ at the end of the simulation runs. Boundaries of the maximum achievable rate regions $\mathcal{C}$ are plotted in Fig. 6 with a solid line under UPA and a dash-dotted line under OPA. Rate requirements $\check{r}_1, \check{r}_2$ are also depicted in the figure with dotted horizontal and vertical lines.

Proximity of the points $(\bar{r}_1^*, \bar{r}_2^*)$ to boundary points of the $\mathcal{C}$ regions under UPA and OPA confirm the optimality of the corresponding SPD algorithms. For OPA, the area between the dash-dotted line and the dotted lines $\check{r}_1 = 30$ and $\check{r}_2 = 100$ which correspond to the rate requirements, represents the set $\mathcal{C} \cap \mathcal{S}_{\check{r}}$ for nRT traffic. The optimal $(\bar{r}_1^*, \bar{r}_2^*)$ point for nRT traffic (see, e.g., the circle marker) resides at the intersection of the $\mathcal{C}$ boundary (dash-dotted line) with the $\mathcal{S}_{\check{r}}$ boundary of the rate requirement (dotted line). Since the same utility functions are employed, the circle marker is the point of the set $\mathcal{C} \cap \mathcal{S}_{\check{r}}$ lying closest to the diamond marker which is the optimal $(\bar{r}_1^*, \bar{r}_2^*)$ point for BE traffic (and thus lies on the boundary of $\mathcal{C}$ but outside the set $\mathcal{S}_{\check{r}}$ since no rate constraints are imposed on BE services). Similar behavior is observed for UPA-based nRT-SPD algorithm.

It is also evident that scheduling with OPA outperforms considerably that with UPA. The sizeable gap between the two is for TDD where channel reciprocity holds. But as will discuss next, even for FDD where both UPA and OPA can only rely on Q-CSIT the gap is not reduced much.

### D. Implementation and Overhead—FDD Systems

As mentioned earlier, adaptive scheduling with optimal power allocation in FDD systems can only rely on Q-CSIT. Resource allocation however, becomes more challenging in this case. Indeed, F-CSIT is available in TDD wherever needed, and the optimal power scheduled to meet the BER $\check{\epsilon}_k \ \forall k$ can be drawn from a set of infinite cardinality, in accordance with the analog-valued realization $h_k$. But since the schedule is announced with a $B$-bit codeword ($B < \infty$), and in FDD $h_k$ is not available at individual terminals to find the power value adapted to $h_k$, users can only encode or decode using a power book (i.e., a set of power values) with finite cardinality $2^B < \infty$.

It is thus evident that if power assignment is optimized in scheduling based only on Q-CSIT (the only option in FDD), one must design a quantizer optimizing the $2^B$ codewords in the power-book jointly with the resource allocation task. Although heuristic quantizers can be found separately to form a suboptimum power-book, jointly optimized time allocation with power quantization ensures that the overall scheduling policy is optimum. Using the channel cdf $F(\boldsymbol{h})$, related jointly optimal designs have been explored in [56] and [57]. Interestingly, optimizing the quantizer (i.e., the power-book design) can be carried out off-line; and based on it, the optimal resource allocation can be performed on-line with surprisingly low complexity. This becomes possible if each transmit-mode is associated with a unique quantized rate-power pair [57]. Under this design condition, the possible transmit-configurations of user $k$ are the pairs $\{(\rho_{k,m}, p_{k,m})\}_{m=0}^{M_k}$, where both $\rho_{k,m}$ and $p_{k,m}$ are constant quantities known to both transmit- and receive-ends. The results in [57] show that the Q-CSIT based optimum scheduling using this quantizer results in SNR loss smaller than 2 dB relative to the F-CSIT based benchmark, provided that the feedback carries $B = \lceil \log_2(\sum_{k=1}^K M_k + 1) \rceil$ bits per slot. Using these operational conditions, uplink and downlink scheduling algorithms with optimal power allocation can be devised for FDD systems [58], as outlined next.

*TDMA Uplink:* Recall that in TDMA uplink, the F-CSIT vector $\boldsymbol{h}$ can be always made available at the AP but not at the users. The AP first finds for each user $k$ the set of AMC modes satisfying the prescribed BER, call it $\mathcal{M}_k(\boldsymbol{h}) := \{m : \epsilon_{k,m}(h_k p_{k,m}) \leq \check{\epsilon}_k\}$. Under the uniform power allocation schemes of Section IV, this set contains only a single optimal AMC mode per user. However, when power $p_{k,m}$ can be different across AMC modes, the optimum mode per user needs to be chosen by the scheduler from the set $\mathcal{M}_k(\boldsymbol{h})$ per $\boldsymbol{h}$. Based on $\mathcal{M}_k(\boldsymbol{h})$, the AP will again schedule following the winner-takes-all principle. In this case, given the optimal non-negative Lagrange multiplier vector $\boldsymbol{\lambda}^* := [\lambda_1^*, \ldots, \lambda_K^*]^T$, the AP first selects the optimum transmit-mode per user as

$$m_k^* = \arg\max_{m \in \mathcal{M}_k(\boldsymbol{h})} \left( w_k \rho_{k,m} - \lambda_k^* p_{k,m} \right) \tag{24}$$

and then schedules the user with index

$$k^* = \arg\max_k \left( w_k \rho_{k,m_k^*} - \lambda_k^* p_{k,m_k^*} \right) \tag{25}$$

to access the channel. As before, $\lambda_k^*$ can be interpreted as the price of power.

It is also possible to develop on-line SPD based scheduling algorithms in FDD uplink systems, for which the optimal schedule also follows (24) and (25). Note that here too the AP broadcasts the scheduled user-mode codeword $(k^*, m_{k^*}^*)$ per slot. With $p_{k,m} \ \forall k, m$ known to the users, the overhead is again minimal, just $B = \lceil \log_2(\sum_{k=1}^K M_k + 1) \rceil$ bits per slot.

*TDM Downlink:* In the TDM based downlink of FDD systems, each user $k$ knows $h_k$ and needs to feed $m_k^*$ back to the AP for scheduling. The optimization problem here is similar to the one in TDMA, but with a sum-power constraint $\check{p}$ replacing the individual power constraints. To find the optimal $\boldsymbol{\tau}^*(\boldsymbol{h})$, the AP needs per slot the mode indices in $\mathcal{M}_k(\boldsymbol{h}) \ \forall k$, which can be encoded and fed back using a finite number of bits. Using this information and

based on the power price (Lagrange multiplier) $\lambda^*$ corresponding to the sum-power constraint, the AP can similarly implement the scheduling as in (24) and (25), with $\lambda^*$ replacing $\lambda_k^*$ $\forall k$.

The overhead in the feedback channel is further reduced with on-line utility-based SPD algorithms in the FDD downlink. In this case, each terminal keeps track of the power price $\hat{\lambda}[n]$ and its own average rate $\hat{\bar{r}}_k[n]$ to adapt the corresponding weight (e.g., $w_k[n] = U'_{BE,k}(\hat{\bar{r}}_k[n])$ for BE traffic), whose updates depend only on the scheduling decision $(k^*, m_{k^*}^*)$. Hence, given identical initialization and stepsize $\beta_n$, these updates are consistent across all terminals and at the AP. Then based on $h_k$, each user $k$ can determine and feed back $m_k^*$ for the AP to select $k^*$. With this alternative, the Q-CSIT feedback reduces to $B_k = \lceil \log_2(M_k + 1) \rceil$ bits per user $k$, the same as in downlink scheduling based on uniform power allocation.

Following the steps sketched in this subsection, all the QoS-guaranteed SPD scheduling algorithms in Section IV can be rederived using Q-CSIT to incorporate also optimal power allocation in FDD systems.

## VI. FURTHER CONSIDERATIONS

In this section, we outline important manifestations of the unified SPD-based scheduler in facilitating adherence to priorities associated with various services, admission and interruption of services and pricing issues. These are also tested in a simulated IEEE 802.16 setup to gauge the potential of SPD scheduling for practical deployment.

*Respecting Service-Specific Priorities:* If the available power and rate resources are sufficient to satisfy the QoS requirements of heterogeneous traffic sessions, it is desirable and possible under the unified approach to schedule resources in accordance with service-specific priorities. As we mentioned in Section III, such a fairness in resource allocation is ensured by scheduling RT traffic first, nRT applications next, and BE services last. This flexibility is conveniently available by selecting utility functions with derivatives respecting the service-specific priorities. Indeed, if one selects a proper function $\bar{d} = f(\bar{r})$ to map average rates to average delays, and also chooses utilities with derivatives satisfying

$$\left|U'_{RT,k}(f(\bar{r}))\right| > \left|U'_{nRT,j}(\bar{r})\right| > \left|U'_{BE,i}(\bar{r})\right|$$

the resultant scheduling algorithm will weigh (and thus favor) RT more than nRT, and nRT more than BE. (Recall that derivatives of utility functions act as weights in weighted-sum-rate utility functions.) Respecting priorities contributes to fairness and better utilization of resources under identical QoS requirements.

*Connection Admission Control:* Besides its use for optimizing resources, the Lagrangian formulation of the unified scheduling framework provides guidelines for: (cac-1) admitting new services if resources are abundant; and (cac-2) dropping services in accordance to their priorities as resources are consumed. To implement (cac-1) and (cac-2) it suffices to monitor each session's non-negative Lagrange multiplier which must: i) be zero (or relatively small) if resources are plentiful (or relatively sufficient) to satisfy QoS constraints as strict inequalities and ii) grow to infinity (or are relatively large) as resources are completely consumed (or have been almost used up), thus rendering QoS requirements infeasible.

If the Lagrange multipliers of existing sessions are "small," then new services can be admitted. If they are "large," then services should be dropped. To quantitatively guide this control policy and also respect service-specific priorities we can heuristically select three positive thresholds (the highest corresponding to RT, the middle one to nRT and the lowest to BE). As resources are used up the Lagrange multipliers of RT and nRT applications will increase to meet their QoS requirements, and this will in turn reduce the average rates for BE services (eventually down to zero if the BE multipliers exceed the BE threshold). If the nRT threshold is exceeded, likewise nRT services will be dropped next, unless the average SNR of RT services is extremely poor. Clearly, RT services will be dropped last when the Lagrange multipliers exceed even the highest RT threshold.

*Lagrange Multiplier Based Pricing:* Before initiating a connection, the AP negotiates with the corresponding user to agree on a price for utilizing the network resources. Because utility-based scheduling entails a positive Lagrange multiplier in providing service, say to an RT or nRT user, the system actually designates resources for QoS provisioning to this user. It is thus reasonable when a user's Lagrange multiplier grows, for the AP to solicit a higher price, or, even withdraw service if the two parties cannot agree on the upgraded price.

*Potential for Adoption in IEEE 802.16 Standard:* The simplicity and optimality of the SPD scheduler in delivering QoS guarantees along with its flexibility to respect service priorities and provide admission control mechanisms, all for heterogeneous traffic, make it an attractive choice for adoption in future wireless standards. A preliminary test highlighting its potential is presented next for heterogeneous traffic in an IEEE 802.16a TDM downlink setup. Four QoS classes are included in the standard. In addition to BE, nRT, and RT, the highest-priority unsolicited grant service (UGS) supports constant

bit-rate connections for, e.g., voice over the Internet, and requires a fixed number of minislots per slot. The scheduling task is to allocate the remaining, say $N_r$, minislots to two RT, two nRT, and two BE admitted connections which transmit packets of fixed length using one of $M_k = 6$ AMC modes. The six fading channel coefficients are generated independently from a Nakagami-$m$ distribution with different average SNR values, and are modified according to the Doppler model in [52] to test variations due to mobility.

The running averages of delays $\hat{\bar{d}}_k(t)$ corresponding to RT connections $k = 1, 2$ and $\hat{\bar{r}}_k(t)$ corresponding to the average rates of nRT and BE connections $k = 3, 4, 5, 6$, are plotted in Figs. 7, 8, and 9 for $N_r = 3, 2$, and 1, respectively. Fig. 7 confirms that all $\hat{\bar{d}}_k(t)$ and $\hat{\bar{r}}_k(t)$ curves converge to an equilibrium with all QoS requirements satisfied. While user $k = 4$ receives "excess service" because its channel gain is high, both BE connections also receive a fair treatment. Fig. 8 illustrates the performance for $N_r = 2$ minislots, where one minislot is reduced as, e.g., the system has to reserve resources for new UGS connections. BE connection $k = 5$ with the worst channel quality is then dropped first, and user $k = 6$ is disconnected afterwards; whereas the QoS requirements for nRT and RT sessions are still fulfilled. With another minislot reserved for UGS, Fig. 9 depicts the performance for $N_r = 1$. Here, the service is stopped for nRT user $k = 3$, who has worse channel and tighter requirement than user $k = 4$. It is also seen that the delays $\hat{\bar{d}}_k$ for RT traffic exhibit large variations, which are due to the suboptimal rule (winner-takes-all) adopted in scheduling RT traffic. Simulations depicted in Figs. 7–9 demonstrate not only the asymptotic optimality and adherence to service-specific priorities, but also the robustness of the SPD algorithm to channel nonstationarities induced by mobility.

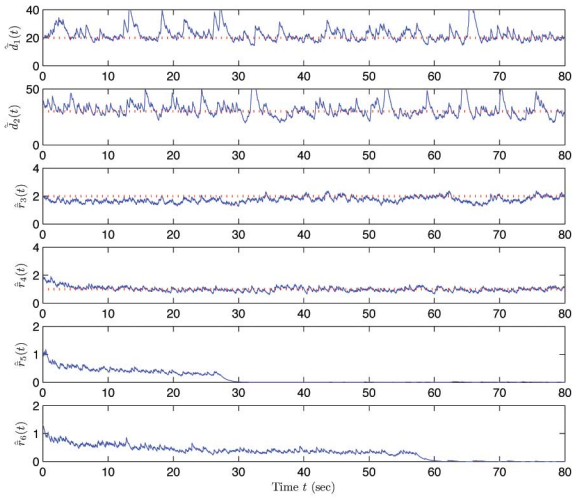



**Fig. 8.** *Running averages $\hat{\bar{d}}_1(t)$, $\hat{\bar{d}}_2(t)$, $\hat{\bar{r}}_3(t)$, $\hat{\bar{r}}_4(t)$, $\hat{\bar{r}}_5(t)$, $\hat{\bar{r}}_6(t)$ versus $t$ for $N_r = 2$ (delays in ms, and rates in Mb/s; dashed lines correspond to $\check{d}_1 = 20$ ms, $\check{d}_2 = 30$ ms, $\check{r}_3 = 2$ Mb/s and $\check{r}_4 = 1$ Mb/s).*

## VII. THE ROAD AHEAD

The unified framework for QoS-guaranteed scheduling has several attractive features with far reaching implications;

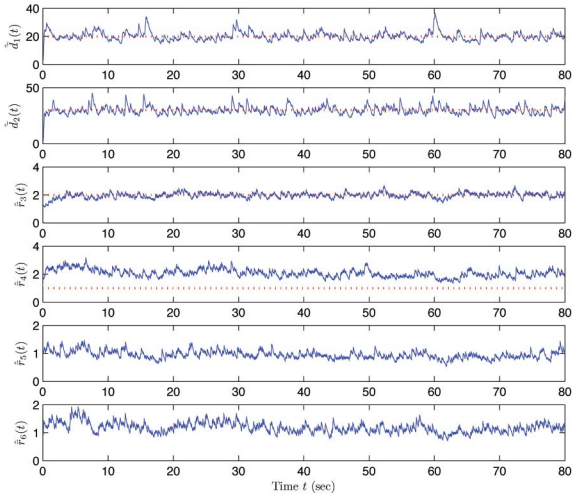



**Fig. 7.** *Running averages $\hat{\bar{d}}_1(t)$, $\hat{\bar{d}}_2(t)$, $\hat{\bar{r}}_3(t)$, $\hat{\bar{r}}_4(t)$, $\hat{\bar{r}}_5(t)$, $\hat{\bar{r}}_6(t)$ versus $t$ for $N_r = 3$ (delays in ms, and rates in Mb/s; dashed lines correspond to $\check{d}_1 = 20$ ms, $\check{d}_2 = 30$ ms, $\check{r}_3 = 2$ Mb/s and $\check{r}_4 = 1$ Mb/s).*

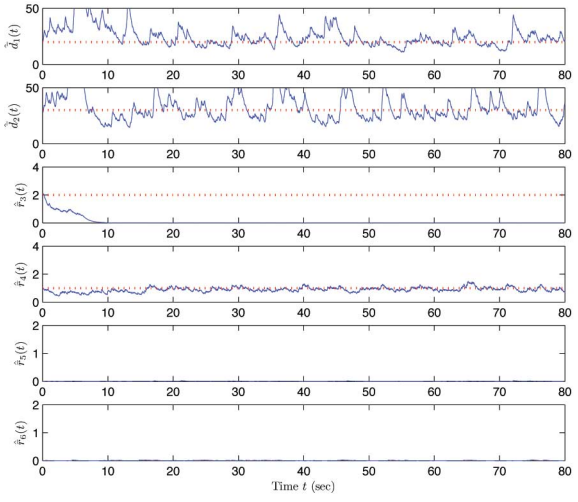



**Fig. 9.** *Running averages $\hat{\bar{d}}_1(t)$, $\hat{\bar{d}}_2(t)$, $\hat{\bar{r}}_3(t)$, $\hat{\bar{r}}_4(t)$, $\hat{\bar{r}}_5(t)$, $\hat{\bar{r}}_6(t)$ versus $t$ for $N_r = 1$ (delays in ms, and rates in Mb/s; dashed lines correspond to $\check{d}_1 = 20$ ms, $\check{d}_2 = 30$ ms, $\check{r}_3 = 2$ Mb/s and $\check{r}_4 = 1$ Mb/s).*

but at least as exciting, challenging and rewarding appears to be the research outlook it motivates in analytical as well as algorithmic investigations.

*Analytical Studies on Fundamental Limits:* The term capacity we used in the present TDM/TDMA context refers to maximum achievable rates within the class of orthogonal signalling schemes. The fundamental rate limit on the other hand, is provided by Shannon's capacity which for multiuser fading channels is achieved via superposition coding and successive decoding [19], [78]. It is thus interesting, at least for benchmarking purposes, to study capacity-optimal resource allocation under QoS constraints when the finite number of AMC modes is replaced by superposition coding at the transmitter(s) and successive decoding is employed at the receiver(s). For these nonorthogonal transmissions, it is further useful to develop the associated optimal on-line scheduling algorithms using SPD-based stochastic optimization techniques; see, e.g., [3], [50] for results in this direction. Particularly challenging to derive is also the counterpart of the capacity region in RT connections, namely the delay region, and the related task of off-line optimal resource allocation for RT traffic. In lieu of analytical expressions for the average delay, these problems are still open as are links between IT-PHY and queueing-theoretic approaches to scheduling altogether.

*Scheduling Frequency-Selective and Multiantenna Links:* As the demand for higher rates increases, coping with frequency-selective fading channels becomes increasingly important. OFDM and OFDMA are very popular for such wireless channels because they are flexible in rate allocation and offer low-complexity decoding. In addition, multiantenna links over multi-input multi-output (MIMO) channels can boost rates via spatial multiplexing and enhance error-resilience through spatial diversity. For these reasons, MIMO-OFDM/OFDMA systems have been adopted by (or proposed for) wireless packet access networks in the IEEE 802.11n WiFi, IEEE 802.16e WiMax, and UMTS WCDMA standards. Extension of the unified TDM/TDMA framework to general orthogonal space-time-frequency channels is worth pursuing because the increased degrees of freedom will provide extra flexibility in scheduling channel, power, and rate resources; see [71] and [72] for a preliminary IT-PHY approach to power allocation across subcarriers, and [88] for optimal joint subcarrier, power, and rate allocation in OFDMA scheduling. This flexibility will lead to SPD based on-line scheduling algorithms capable of attaining desirable tradeoffs among rate/error performance and software/hardware complexity.

*SPD Algorithms for Joint Network Utility Maximization:* Designing congestion control, routing, and scheduling policies jointly based on network utility maximization has drawn growing attention recently [16], [44], [54]. Different from the IT-PHY and MAC perspectives unified here for scheduling single-hop connections, these designs deal with multihop links and aim at fair and efficient allocation of network resources across layers. Interestingly, the utility maximization framework is common to both single-lop and multihop settings. This opens up a host of related convex optimization problems, and suggests investigation of QoS-guaranteed SPD-based adaptive algorithms under the broader framework of network utility maximization.

*Distributed SPD Algorithms for Ad Hoc Networks:* Distributed scheduling, flow control, and routing are also exciting research thrusts broadening the scope of the centralized algorithms considered here. These extensions are well fit for infrastructure-free and low-cost (e.g., sensor) networks which would welcome distributed on-line solutions as simple as those offered by SPD-based adaptive algorithms. Parallel and distributed computation tools [11] have been recently used to tackle distributed routing, congestion control, and scheduling problems [16], [63], [89]. Their adoption to develop on-line SPD-based algorithms for distributed resource allocation is well motivated for tactical mobile ad hoc networks [20] and commercial cognitive radios [24].

*Cross-Disciplinary Contributions to Network Science:* Due to the ubiquitous and rapidly growing dependence of our society on the interacting networks in a wide variety of domains, interest in network research has exploded in the past five years. Ideas from applied mathematics, engineering, and sociology have contributed to a new field—the science of networks, braced to explore fundamental concepts and rigorous analytical tools behind the currently fragmented research in networking [59]. Although the notion of general network science is evolving with limited understanding of its ultimate scope and content, it is commonly acknowledged that the evolution of the science for communication networks will likely rely on cross-layer and interdisciplinary approaches. In search of a unified scheduling approach for channel-adaptive wireless networks, the framework here is rooted on a rigorous analytical basis and bridges capacity metrics from information theory, utility models from economics, convex optimization of communication systems as well as stochastic approximation algorithms from adaptive signal processing and control. As a result, it leads to systematic designs, rigorous analysis, testable predictions, and a host of directions to build on.

## VIII. CLOSING REMARKS

We presented a unified framework for deriving and analyzing QoS-guaranteed channel-aware scheduling protocols for wireless packet access networks. The resultant

stochastic primal-dual (SPD) algorithms are not only as simple as any existing heuristic scheduling scheme, but also robust to channel nonstationarities and uniformly convergent from any initial value to the optimal off-line solutions provided by information-theoretic benchmarks. Since SPD algorithms are generalizations of gradient scheduling, which includes proportional fair scheduling as a special case, they can be seamlessly integrated for QoS provisioning in existing tactical and commercial systems including those in the 1xEV-DO standard. Furthermore, the underlying systematic approach to designing SPD schedulers can be potentially useful in optimizing (possibly across layers) the design of present and future-generation wireless standards tailored for sensor and cognitive radio networks, channel-adaptive mesh networks, and wireless mobile ad hoc networks.

Decades ago the focus of telecommunication system designers was placed on point-to-point links in the presence of noise, whereas today the main challenge lies in reliable end-to-end QoS provisioning for packet data and heterogeneous services over wireless networks. Tackling this formidable challenge calls for major advances in network science as a whole. It will certainly be gratifying to the authors if this attempt to unify channel-aware QoS-guaranteed scheduling will serve as a stepping stone in these advances.[1] ■

[1]The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government.

## REFERENCES


[1] 3GPP TR 25.848 V4.0.0, Physical Layer Aspects of UTRA High Speed Downlink Packet Access (Release 4), 2001.

[2] M. Agarwal and A. Puri, "Base station scheduling of requests with fixed deadlines," in *Proc. IEEE INFOCOM Conf.*, New York, Jun. 23–27, 2002, vol. 2, pp. 487–496.

[3] R. Agrawal and V. Subramanian, "Optimality of certain channel aware scheduling algorithms," in *Proc. 40th Allerton Conf. Communication, Control and Computing*, Monticello, IL, Oct. 2002.

[4] R. Ahlswede, "Multi-way communication channels," in *Proc. Int. Symp. Information Theory*, Tsahkadsor, U.S.S.R., 1971, pp. 103–135.

[5] M. S. Alouini and A. J. Goldsmith, "Adaptive modulation over Nakagami fading channels," *Kluwer J. Wireless Commun.*, vol. 13, no. 1–2, pp. 119–143, May 2000.

[6] M. Andrews, L. Qian, and A. L. Stolyar, "Optimal utility based multi-user throughput allocation subject to throughput constraints," in *Proc. IEEE INFOCOM Conf.*, Miami, FL, Mar. 13–17, 2005, vol. 4, pp. 2415–2424.

[7] M. Andrews, K. Kumaran, K. Ramanan, A. L. Stolyar, R. Vijayakumar, and P. Whiting, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, vol. 39, no. 2, pp. 150–154, Feb. 2001.

[8] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushyana, and A. Viterbi, "CDMA/HDR: A bandwidth efficient high speed wireless data service for nomadic users," *IEEE Commun. Mag.*, vol. 38, no. 7, pp. 70–77, Jul. 2000.

[9] R. Berry and R. G. Gallager, "Communication over fading channels with delay constraints," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1135–1149, May 2002.

[10] D. Bertsekas, *Nonlinear Programming*, 2nd ed. Nashua, NH: Athena Scientific, 1999.

[11] D. Bertsekas and J. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, 2nd ed. Nashua, NH: Athena Scientific, 1999.

[12] S. C. Borst and P. A. Whiting, "Dynamic rate control algorithms for HDR throughput optimization," in *Proc. IEEE INFOCOM Conf.*, Anchorage, AK, April 22–26, 2001, vol. 2, pp. 976–985.

[13] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[14] Y. Cao and V. Li, "Scheduling algorithms in broadband wireless networks," *Proc. IEEE*, vol. 89, no. 1, pp. 76–87, Jan. 2001.

[15] P. Chan, E. Lo, R. Wang, E. Au, V. Lau, R. Cheng, W. Mow, R. Murch, and K. Letaief, "The evolution path of 4G networks: FDD or TDD?" *IEEE Commun. Mag.*, vol. 44, no. 12, pp. 42–50, Dec. 2006.

[16] L. Chen, S. H. Low, M. Chiang, and J. C. Doyle, "Jointly optimal congestion control, routing, and scheduling for wireless ad hoc networks," in *Proc. IEEE INFOCOM Conf.*, Barcelona, Spain, Apr. 2006.

[17] S. T. Chung and A. J. Goldsmith, "Degrees of freedom in adaptive modulation: A unified view," *IEEE Trans. Commun.*, vol. 49, no. 9, pp. 1561–1571, Sep. 2001.

[18] T. M. Cover, "Broadcast channels," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 1, pp. 2–14, Jan. 1972.

[19] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley, 2006.

[20] DARPA, *The Next Generation (XG) Program*. [Online]. Available: http://www.darpa.mil/ato/programs/xg/index.htm

[21] A. Demers, S. Keshav, and S. Shenker, "Analysis and simulation of a fair queueing algorithm," in *Proc. ACM SIGCOMM*, Austin, TX, Sep. 1989, pp. 3–12.

[22] A. Eryilmaz, R. Srikant, and J. R. Perkins, "Stable scheduling policies for fading wireless channels," *IEEE/ACM Trans. Netw.*, vol. 13, no. 4, pp. 411–424, Apr. 2005.

[23] *Radio Equipment and Systems; High Performance Radio Local Area Network (HIPERLAN) Type 1*, ETSI, ETS, 300-652, Oct. 1996.

[24] FCC, "Facilitating opportunities for flexible, efficient, and reliable spectrum use employing cognitive radio technologies," FCC Report and Order, FCC-05-57A1, Mar. 2005.

[25] R. G. Gallager, "An inequality on the capacity region of multiple access multipath channels," in *Communications and Cryptography: Two Sides of One Tapestry*. Boston, MA: Kluwer, 1994, pp. 129–139.

[26] G. B. Giannakis, Z. Liu, X. Ma, and S. Zhou, *Space-Time Coding for Broadband Wireless Communications*. New York: Wiley, 2007.

[27] S. Golestani, "A self-clocked fair queueing scheme for broadband applications," in *Proc. IEEE INFOCOM Conf.*, Toronto, ON, Canada, Jun. 1994, pp. 636–646.

[28] A. J. Goldsmith and S. G. Chua, "Variable-rate variable power MQAM for fading channels," *IEEE Trans. Commun.*, vol. 45, no. 10, pp. 1218–1230, Oct. 1997.

[29] A. J. Goldsmith and S. G. Chua, "Adaptive coded modulation for fading channels," *IEEE Trans. Commun.*, vol. 46, no. 5, pp. 595–602, May 1998.

[30] A. J. Goldsmith, *Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[31] D. Gross and C. Harris, *Fundamentals of Queueing Theory*, 3rd ed. New York: Wiley, 2006.

[32] S. V. Hanly and D. Tse, "Multiaccess fading channels—Part II: Delay-limited capacities," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2816–2831, Nov. 1998.

[33] J. Hayes, "Adaptive feedback communications," *IEEE Trans. Commun.*, vol. 16, pp. 29–34, Feb. 1968.

[34] *IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems (Revision of IEEE Standard 802.16-2001)*, IEEE Standard 802.16 Working Group, 2004.

[35] *802.11: Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE, IEEE Standard 802.11 Working Group, 1997.

[36] N. Jindal, S. Vishwanath, and A. J. Goldsmith, "On the duality of Gaussian multiple-access and broadcast channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 5, pp. 768–783, May 2004.

[37] D. Kandlur, K. Shin, and D. Ferrari, "Real-time communication in multi-hop networks," in *Proc. 11th Int. Conf. Distributed Computer System*, Washington, DC, May 1991, pp. 300–307.

[38] F. Kelly, "Charging and rate control for elastic traffic," *Eur. Trans. Telecommun.*, vol. 8, pp. 33–37, 1997.

[39] L. Kleinrock, *Queueing Systems, Vol. I: Theory*. New York: Wiley, 1975.

[40] R. Knopp and P. Humblet, "Multiuser diversity," unpublished manuscript.

[41] R. Knopp and P. Humblet, "Information capacity and power control in single cell

multiuser communications," in *Proc. IEEE Int. Conf. Communications*, Seattle, WA, Jun. 1995.

[42] H. Kushner and P. A. Whiting, "Convergence of proportional-fair sharing algorithms under general conditions," *IEEE Trans. Wireless Commun.*, vol. 3, no. 4, pp. 1250–1259, Jul. 2004.

[43] H. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications*, 2nd ed. Berlin, Germany: Springer-Verlag, 2003.

[44] J. W. Lee, M. Chiang, and R. A. Calderbank, "Optimal MAC design based on utility maximization: Reverse and forward engineering," in *Proc. IEEE INFOCOM Conf.*, Barcelona, Spain, Apr. 2006.

[45] L. Li and A. J. Goldsmith, "Capacity and optimal resource allocation for fading broadcast channels—Part I: Ergodic capacity," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 1083–1102, Mar. 2001.

[46] L. Li and A. J. Goldsmith, "Capacity and optimal resource allocation for fading broadcast channels—Part II: Outage capacity," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 1103–1127, Mar. 2001.

[47] L. Li, N. Jindal, and A. J. Goldsmith, "Outage capacities and optimal power allocation for fading multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1326–1347, Apr. 2005.

[48] H. Liao, "A coding theorem for multiple access communications," in *Proc. Int. Symp. Information Theory*, Asilomar, CA, 1972.

[49] X. Liu, E. Chong, and N. Shroff, "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 10, pp. 2053–2064, Oct. 2001.

[50] X. Liu, E. Chong, and N. Shroff, "A framework for opportunistic scheduling in wireless networks," *Comput. Netw.*, vol. 41, pp. 451–474, 2002.

[51] Q. Liu, S. Zhou, and G. B. Giannakis, "Cross-layer combining of adaptive modulation and coding with truncated ARQ over wireless links," *IEEE Trans. Wireless Commun.*, vol. 3, no. 5, pp. 1746–1755, Sep. 2004.

[52] Q. Liu, S. Zhou, and G. B. Giannakis, "Queuing with adaptive modulation and coding over wireless links: Cross-layer analysis and design," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 1142–1153, May 2005.

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

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

[55] S. Lu, V. Bharghavan, and R. Srikant, "Fair scheduling in wireless packet networks," *IEEE/ACM Trans. Netw.*, vol. 7, no. 4, pp. 473–489, Aug. 1999.

[56] A. G. Marques, F. F. Digham, and G. B. Giannakis, "Power-efficient OFDM via quantized channel state information," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1581–1592, Aug. 2006.

[57] A. G. Marques, X. Wang, and G. B. Giannakis, "Optimizing energy-efficient of TDMA with finite rate feedback," in *IEEE Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Honolulu, HI, Apr. 16–20, 2007.

[58] A. Marques, X. Wang, and G. B. Giannakis, "Channel-adaptive resource allocation for cognitive OFDMA radios based on limited-rate feedback," in *Proc. XV Eur. Signal Process. Conf.*, Poznan, Poland, Sep. 3–7, 2007, (invited).

[59] National Research Council, *Network Science*. Washington, DC: The National Academies Press, 2005.

[60] T. Ng, I. Stoica, and H. Zhang, "Packet fair queueing algorithms for wireless networks with location-dependent errors," in *Proc. IEEE INFOCOM Conf.*, San Francisco, CA, Mar. 1998, pp. 1103–1111.

[61] D. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1439–1451, Aug. 2006.

[62] P. Ramanathan and P. Agrawal, "Adapting packet fair queueing algorithms to wireless networks," in *Proc. ACM/IEEE MOBICOM'98*, Dallas, TX, Oct. 1998, pp. 1–9.

[63] A. Ribeiro, N. D. Sidiropoulos, and G. B. Giannakis, "Distributed routing algorithms for wireless multihop networks," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, Honolulu, HI, Apr. 15–20, 2007.

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

[65] S. Shakkottai and A. Stolyar, "Scheduling algorithms for a mixture of real-time and non-real-time data in HDR," in *Proc. 17th Int. Teletraffic Congress*, Salvador da Bahia, Brazil, 2001, pp. 793–804.

[66] S. Shakkottai, R. Srikant, and A. L. Stolyar, "Pathwise optimality of the exponential scheduling rule for wireless channels," *Adv. Appl. Prob.*, vol. 36, no. 4, pp. 1021–1045, 2004.

[67] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*. New York: Springer-Verlag, 1985.

[68] V. Solo and X. Kong, *Adaptive Signal Processing Algorithms: Stability and Performance*. Englewood Cliffs, NJ: Prentice Hall, 1995.

[69] G. Song, "Cross-Layer resource allocation and scheduling in wireless multicarrier networks," Ph.D. dissertation, Georgia Inst. Technol., Apr., 2005.

[70] G. Song and Y. Li, "Utility-based resource allocation and scheduling in OFDM-based wireless broadband networks," *IEEE Commun. Mag.*, vol. 43, no. 12, pp. 127–134, Dec. 2005.

[71] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks—Part I: Theoretical framework," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 614–624, Feb. 2005.

[72] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks—Part II: Algorithm development," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 625–634, Feb. 2005.

[73] A. L. Stolyar, "On the asymptotic optimality of the gradient scheduling algorithm for multi-user throughput allocation," *Op. Res.*, vol. 53, no. 1, pp. 12–25, Jan.–Feb. 2005.

[74] A. L. Stolyar, "Maximizing queueing network utility subject to stability: Greedy primal-dual algorithm," *Queueing Syst.*, vol. 50, pp. 401–457, 2005.

[75] A. L. Stolyar and K. Ramanan, "Largest weighted delay first scheduling: Large deviation and optimality," *Annals Appl. Prob.*, vol. 11, no. 1, pp. 1–48, 2001.

[76] D. Tse and S. V. Hanly, "Multiaccess fading channels—Part I: Polymatroid structure, optimal resource allocation and throughput capacities," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2796–2815, Nov. 1998.

[77] D. Tse, "Multiuser Diversity and Proportional Fairness," U.S. Patent 6 449 490, Sep. 10, 2002, pp. 12–25.

[78] D. Tse and P. Viswanath, *Fundamentals of Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[79] E. Uysal-Biyikoglu, B. Prabhakar, and A. El Gamal, "Energy-efficient packet transmission over a wireless link," *IEEE/ACM Trans. Netw.*, vol. 10, no. 4, pp. 487–499, Aug. 2002.

[80] M. Vidysagar, *Nonlinear System Analysis*. Englewood Cliffs, NJ: Prentice Hall, 1978.

[81] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1277–1294, Jun. 2002.

[82] X. Wang, Q. Liu, and G. B. Giannakis, "Analyzing and optimizing adaptive modulation-coding jointly with ARQ for QoS-guaranteed traffic," *IEEE Trans. Veh. Technol.*, vol. 56, no. 2, pp. 710–720, Mar. 2007.

[83] X. Wang and G. B. Giannakis, "An adaptive signal processing approach to scheduling in wireless networks—Part I: Uniform power allocation and quantized CSI," *IEEE Trans. Signal Process.*, Jul. 2007.

[84] X. Wang and G. B. Giannakis, "An adaptive signal processing approach to scheduling in wireless networks—Part II: Optimal power allocation and full CSI," *IEEE Trans. Signal Process.*, Jul. 2007.

[85] X. Wang and G. B. Giannakis, "A stochastic framework for scheduling in wireless packet access networks," in *Proc. IEEE Int. Conf. Communications*, Glasgow, Scotland, Jun. 24–28, 2007.

[86] X. Wang and G. B. Giannakis, "Stochastic primal-dual scheduling subject to rate constraints," in *Proc. of Wireless Communications and Networking Conf.*, Hong Kong, China, Mar. 11–15, 2007.

[87] X. Wang and G. B. Giannakis, "Energy-efficient resource allocation in TDMA over fading channels," in *Proc. Int. Symp. Information Theory*, Seattle, WA, Jul. 9–14, 2006.

[88] X. Wang, G. B. Giannakis, and Y. Yu, "Channel-adaptive optimal OFDMA scheduling," in *Proc. 41st Conf. Info. Sciences and Systems*, NJ, Mar. 14–16, 2007.

[89] X. Wang and K. Kar, "Cross-layer rate control for end-to-end proportional fairness in wireless networks with random access," in *Proc. ACM MobiHoc*, Urbana-Champaign, IL, May 25–27, 2005.

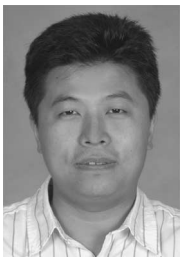
[90] Z. Wang and G. B. Giannakis, "Wireless multicarrier communications: Where Fourier meets Shannon," *IEEE Signal Process. Mag.*, vol. 17, no. 3, pp. 29–48, May 2000.

[91] Q. Wu and E. Esteves, "The CDMA2000 high rate packet data systems," in *Advances in 3G Enhanced Technologies for Wireless Communications*, J. Wang and T. Ng, Eds. Norwood, MA: Artech House, Mar. 2002, ch. 4.

[92] H. Zhang, "Service disciplines for guaranteed performance service in packet-switching networks," *Proc. IEEE*, vol. 83, no. 10, pp. 1374–1396, Oct. 1995.

[93] L. Zhang, "Virtual clock: A new traffic control algorithm for packet switching networks," in *Proc. of ACM SIGCOMM*, Philadelphia, PA, Sep. 1990, pp. 19–29.

## ABOUT THE AUTHORS



**Xin Wang** (Member, IEEE) received the B.Sc. degree and the M.Sc. degree from Fudan University, Shanghai, China, in 1997 and 2000, respectively, and the Ph.D. degree from Auburn University, Auburn, IL, in 2004, all in electrical engineering.

From September 2004 to August 2006, he was a Postdoctoral Research Associate with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis. Since September 2006, he has been an Assistant Professor in the Department of Electrical Engineering, Florida Atlantic University, Boca Raton. His research interests include medium access control, cross-layer design, resource allocation, and signal processing for communication networks.

**Georgios B. Giannakis** (Fellow, IEEE) received the Diploma in electrical engineering from the National Technical University of Athens, Greece, in 1981, and the MSc. degree in electrical engineering, the MSc. degree in mathematics, and the Ph.D. degree in electrical engineering from the University of Southern California (USC) in 1983, 1986, and 1986, respectively.

Since 1999, he has been a Professor with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, where he now holds an ADC Chair in Wireless Telecommunications. His general interests span the areas of communications, networking and statistical signal processing—subjects on which he has published more than 250 journal papers, 450 conference papers, two edited books, and two research monographs. Current research focuses on diversity techniques, complex-field and space-time coding, multicarrier, cooperative wireless communications, cognitive radios, cross-layer designs, mobile ad hoc networks, and wireless sensor networks.

Prof. Giannakis is the corecipient of six paper awards from the IEEE Signal Processing (SP) and Communications Societies including the G. Marconi Prize Paper Award in Wireless Communications. He also received Technical Achievement Awards from the SP Society (2000), from EURASIP (2005), a Young Faculty Teaching Award, and the G. W. Taylor Award for Distinguished Research from the University of Minnesota. He has served the IEEE in a number of posts.

**Antonio G. Marques** (Member, IEEE) received the degree in Telecommunication Engineering and the doctorate degree (B.Sc., M.Sc., and Ph.D. degrees in electrical engineering) with highest honors from the Universidad Carlos III de Madrid, Madrid, Spain, in 2002 and 2007, respectively.

In 2003, he joined the Department of Signal Theory and Communications, Universidad Rey Juan Carlos, Madrid, Spain, where he currently develops his research and teaching activities as an Assistant Professor. Since 2005, he has also been a Visiting Researcher in the Department of Electrical Engineering, University of Minnesota, Minneapolis. His research interests lie in the areas of communication theory, signal processing, and networking. His current research focuses on channel state information designs, energy-efficient resource allocation, and wireless ad hoc and sensor networks.

Dr. Marques' work brought him several awards in distinctive international conferences including MILCOM 2006 and ICASSP 2007.