Great Expectations: The Value of Spatial Diversity in Wireless Networks

SUHAS N. DIGGAVI, MEMBER, IEEE, NAOFAL AL-DHAHIR, SENIOR MEMBER, IEEE,
A. STAMOULIS, MEMBER, IEEE, AND A. R. CALDERBANK, FELLOW, IEEE

Invited Paper

In this paper, the effect of spatial diversity on the throughput and reliability of wireless networks is examined. Spatial diversity is realized through multiple independently fading transmit/receive antenna paths in single-user communication and through independently fading links in multiuser communication. Adopting spatial diversity as a central theme, we start by studying its information-theoretic foundations, then we illustrate its benefits across the physical (signal transmission/coding and receiver signal processing) and networking (resource allocation, routing, and applications) layers. Throughout the paper, we discuss engineering intuition and tradeoffs, emphasizing the strong interactions between the various network functionalities.

Keywords*—Ad hoc networks, channel estimation, diversity, fading channels, hybrid networks, information theory for wireless communications, multiple-antenna communications, multiuser diversity, real-time applications, resource allocation, scheduling, signal processing for communications, space–time coding (STC), transport protocols, wireless networks.*

I. INTRODUCTION

Great expectations start with the mobile terminal and the promise of 3G[1] data rates, real-time internet protocol (IP) services[2] (including voice), and hours of useful battery life. The challenge here is that Moore's Law does not seem to apply to rechargeable battery capacity, and though the density of transistors on a chip has consistently doubled every 18 mo, the energy density of batteries only seems to double every 10 years. This need to conserve energy (see [2] and references therein) leads us to focus on what is possible when signal processing at the terminal is limited. Throughout this paper, we use the cost and complexity of the receiver to bound the resources available for signal processing. Wireless spectrum itself is a valuable resource that also needs to be conserved given the economic imperative of return on multibillion-dollar investments by wireless carriers [1].

The fundamental distinction of wireless communication is the ability to communicate on the move [71]. This represents both a freedom to the end user and also a challenge to the system designer. Wireless communication hinges on transmitting information riding on radio (electromagnetic) waves and, hence, the information undergoes attenuation effects (fading) of radio waves (see Section II for more details). These attenuation effects could also vary with time due to user mobility, making wireless a challenging communication medium.

The distinction from wired (twisted-pair, coaxial cable, optic fiber) communication is that the transmission channel is unpredictable (random) and could vary over very short time scales (order of microseconds). Given the limited power resources, this leads to new challenges in signal transmission (see, for example, [207]). Furthermore, since the medium is inherently shared by several users, the equitable sharing of limited resources is an important challenge. In addition, the freedom of mobility also implies that end users need to be located in order to deliver information to them. This leads to a challenge in dealing with the changing topology of the wireless network. Finally, the wireless device would typically connect to the wired Internet infrastructure and a major challenge is the interoperability of these disparate media. These constitute formidable challenges to the designer and

Manuscript received October 16, 2003; revised November 5, 2003.

S. N. Diggavi was with the AT&T Shannon Labs, Florham Park, NJ 07932 USA. He is now with Swiss Federal Institute of Technology, Lausanne, CH 1015 Switzerland (e-mail: suhas.diggavi@.epfl.ch).

N. Al-Dhahir was with the AT&T Shannon Labs, Florham Park, NJ 07932 USA. He is now with University of Texas at Dallas, Richardson, TX 75083 USA (e-mail: aldhahir@utdallas.edu).

A. Stamoulis is now with Qualcomm, San Diego, CA 92121 USA (e-mail: astamoul@qualcomm.com).

A. R. Calderbank was with the AT&T Shannon Labs, Florham Park, NJ 07932 USA. He is now with the Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544 USA (e-mail: calderbank@math.princeton.edu).

Digital Object Identifier 10.1109/JPROC.2003.821914

[1]G: Generation. Current wireless systems are 2G or 2.5G, where rudimentary data services are being deployed.

[2]See [65] for a discussion of architecture and philosophy.

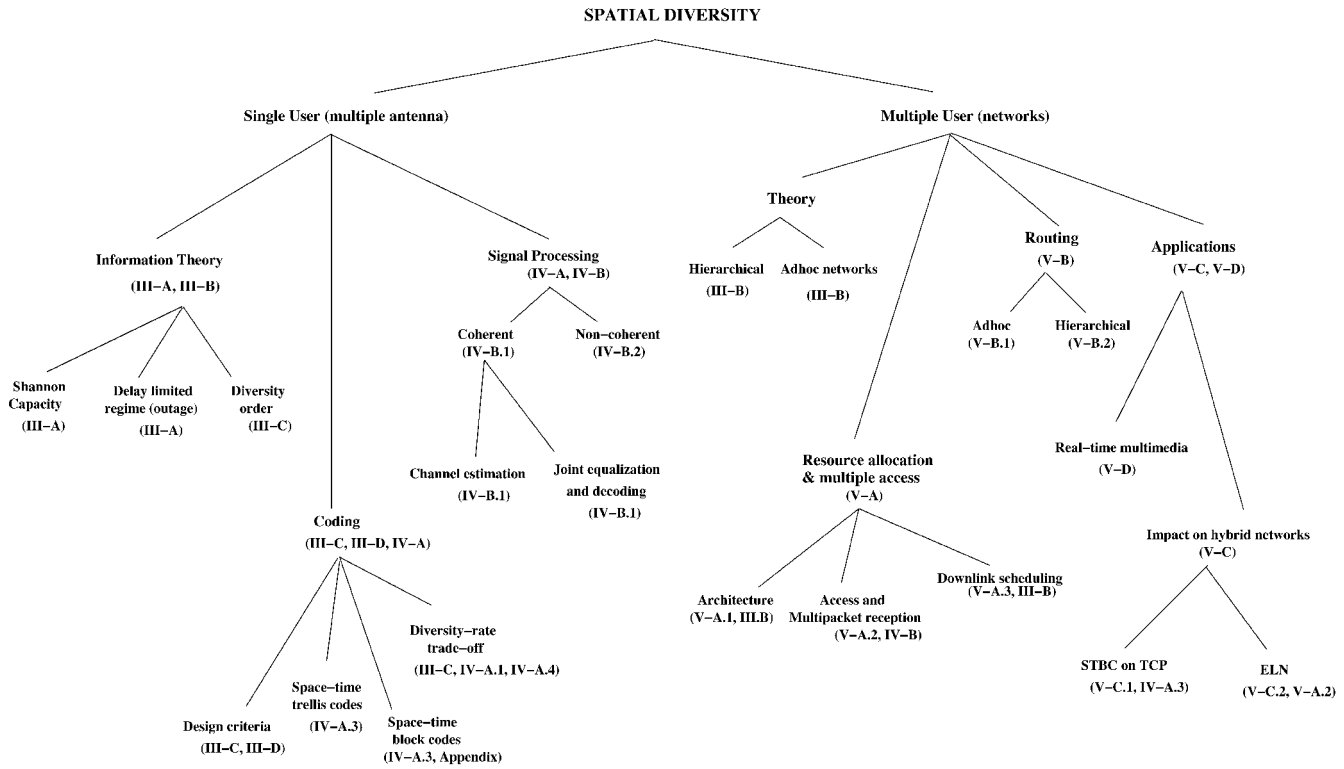

Fig. 1. Roadmap to the paper.

this paper explores some modern analysis and design tools that can help surmount these challenges to meet the great expectations.

The main characteristic of wireless communication is the uncertainty (randomness): randomness in users' transmission channels and randomness in users' geographical locations. These, in turn, lead to random signal attenuation independently across users (see Section II). This randomness can also be used as a tool to enhance performance through *diversity*. We broadly define *diversity* as the method of conveying information through multiple independent instantiations of these random attenuations. The diversity tool we utilize is that of *spatial diversity*, through *multiple antennas* and *multiple users*. Spatial diversity has emerged as the single most important tool for reliable wireless communication. The focus in this paper is on spatial diversity, through spatially separate antennas, which in their simplest instantiation are used to provide replicas of the transmitted signal to the receiver. The independent links (due to independent locations) of different users can also be used as a form of *multiuser diversity*. In addition, diversity also can be utilized through temporal and frequency domains. The topic of this paper is the various methods by which spatial diversity can contribute to reliable wireless communication and is summarized in Fig. 1.

Information theory [70], [112] provides tools to establish fundamental limits on information transmission. In Section III we review such fundamental limits on spatial diversity, both in single-user and multiple-user environments. We review the results on single-user multiple-antenna capacity established by Telatar [245] and by Foschini [103]. Most of our emphasis is on delay-bounded services, hence, on outage capacity, where the usual assumption is that fading is quasi-static, that is constant over a frame of data, and then changing in an independent manner from frame to frame. Recall that a 3% outage capacity (a typical assumption for voice services) is the transmission rate that can be achieved 97% of the time. Moreover, multiple-user environments bring to fore fundamental questions of resource sharing. We examine how diversity impacts per-user performance also in Section III.

Two antennas at the wireless access point (base station) provide two independent paths from the base station to the mobile. By distributing information across the two paths, and by appropriate signal processing at the receiver, we in effect construct a single channel that is better than either path. Superposition of fading on these two paths at the receiver reduces the variation in received signal strength at the mobile. Reduced channel fluctuations allows smoother and more efficient power control, since the base station is continually adjusting transmit power on the basis of reported signal strength at the mobile. At a systems level, this means that the base station requires less power to support existing users, or that more users can be supported for a given constraint on radiated signal power at the base station. The gains are significant; for example, system studies by Parkvall *et al.* [199] show up to 100% increase in the number of users that can be supported on the wideband code-division multiple access (WCDMA) downlink from the base station to the mobile terminal.

The earliest form of spatial transmit diversity is the delay diversity scheme proposed by Wittneben [282] where a

signal is transmitted from the second antenna, then delayed one time slot, and transmitted from the first antenna. Signal processing is used at the receiver to decode the superposition of the original and time-delayed codewords. Viewing multiple-antenna diversity as independent conduits of information, more sophisticated transmission (coding) schemes can be designed to get closer to the fundamental limits discussed in Section III. Using this approach, we focus on space–time codes defined by Tarokh *et al.* [242] and Alamouti [14] that introduce correlation of signals across the different transmit antennas. Space–time codes can provide additional coding gain beyond the diversity gain associated with delay diversity. In only a few years, space–time codes have progressed from invention to adoption in the major wireless standards. For WCDMA, where short spreading sequences are used, transmit diversity provided by space–time codes represents the difference between data rates of 100 and 384 kb/s. Section IV describes modern multiple-antenna coding and signal processing methods that take steps toward the theoretically predicted gains of Section III. Our emphasis is on solutions that include channel estimation, joint decoding and equalization, and where the complexity of signal processing is bounded. The new world of multiple transmit and receive antennas requires significant modification of techniques developed for single-transmit single-receive communication. Since receiver cost and complexity is an important consideration, our treatment of innovation in signal processing is grounded in systems with one, two, or four transmit antennas and one or two receive antennas. For example, the interference cancellation techniques presented in Section V-A.2 enable transmission of 1 Mb/s over a 200-kHz GSM/EDGE channel using four transmit and two receive antennas. This means that our limitation on numbers of antennas does not significantly dampen user expectations.

The functional modularities and abstractions of the IP protocol stack layering [65], [159] contributed significantly to the success of the wired Internet infrastructure. The layering achieves a form of information hiding, providing only interface information to higher layer, and not the details of the implementation. The physical layer is dedicated to signal transmission, while the datalink layer implements functionalities of data framing, arbitrating access to medium and some error control. The network layer abstracts the physical and datalink layers from the upper layers by providing an interface for end-to-end links. Hence, the task of routing and framing details of the link layer are hidden from the higher layers (transport and application layers).

In wireless networks, limited resources (bandwidth, power) imply the importance of efficient and equitable resource allocation techniques among the multiple users of the shared wireless medium. Because of channel fading and user mobility, channel conditions and error rates vary both from user to user and in time for a given user. The user-location-dependent errors and transmission rates (combined with users' mobility) lead to variations in connectivity information that raise fundamental questions on how to share the available resources in a fair manner. This requirement of wireless networks leads to a reexamination of the functional separation and partition of the traditional network protocol stack [65], [159]. We discuss this in Section V in the context of spatial diversity in wireless networks. Section V also addresses ad hoc wireless networks where there might be multiple wireless hops, and where mobility has a significant impact on capacity [81], [126]. Here, the definition of cross-layer protocols, specifically the balance between transmission to update routing information and transmission of information, is central to design.

The main issue to note in investigating these interlayer interactions is that of the *time scales* involved. The wireless channel variations (see Section II for details) occur at several time scales: *fast* variations occur due to Rayleigh fading (time scale of microseconds and upwards), such variations are perhaps best handled at the physical transmission layer and have interactions with the resource allocation schemes. There are slower variations that occur due to shadowing caused from changes in the multipath environment (time scale of hundreds of milliseconds and upwards). Such variations would be in the time scale of the resource allocation schemes and perhaps have impact on network topology (affecting routing). The slowest variations occur due to the distance effect of path loss, which occurs due to mobility of the nodes causing them to move further away from the receivers (time scale of seconds and upwards). Such variations will impact the network topology. Therefore, the time scale becomes important in understanding which "interlayer" interactions arise. Some cross-layer protocols, where performance information at one layer is made visible to other layers, are covered in Section V. We believe that the definition of functionality and interaction between layer abstractions is central to the design of next-generation wireless networks.

The interoperability of wireless networks with wired Internet infrastructure is a vast topic [203]. In Section V, we focus our discussion on the impact of spatial diversity on the end-to-end throughput of such hybrid networks. This is done through a short study of the impact of space–time coding (STC) on end-to-end throughput when users connect to the wired Internet through an IEEE 802.11 wireless network. We also study a complementary example of improving end-to-end throughput by disambiguating the source of losses in the wired and wireless hops.

We are just beginning on the path of meeting the great expectations of wireless communication. The goal of this paper was not to be encyclopedic, but to bring together topics from several different research areas on which spatial diversity has an impact. Hence, the exposition (almost by definition) had to be incomplete and biased from this point of view. A discussion of the steps described in this paper is given in Section VI.

II. Wireless Channel Characteristics

Wireless communication utilizes modulation of electromagnetic (radio) waves with a carrier frequency varying from a few hundred megahertz to several gigahertz depending on the system. Therefore, the behavior of the wireless channel

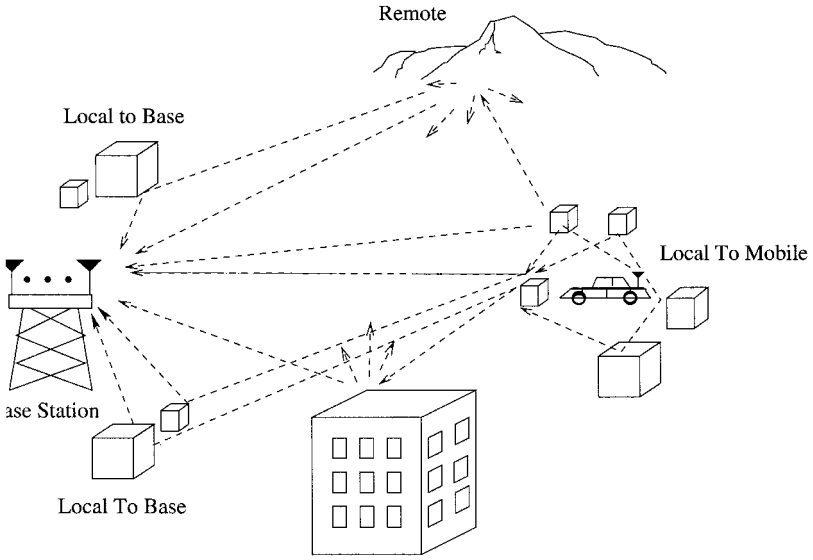

Fig. 2. Radio propagation environment.

is a function of the radio propagation effects of the environment [148], [213].

A typical outdoor wireless propagation environment is illustrated in Fig. 2 where the mobile wireless node is communicating with a wireless access point (base station). The signal transmitted from the mobile may reach the access point directly (line of sight) or through multiple reflections on local scatterers (buildings, mountains, etc.). As a result, the received signal is affected by multiple random attenuations and delays. Moreover, the mobility of either the nodes or the scattering environment may cause these random fluctuations to vary with time. Time variation results in the random waxing and waning of the transmitted signal strength over time. Finally, a shared wireless environment may incur undesired interference (due to concurrent transmissions) to the transmitted signal. The combined effect of these factors makes wireless a challenging communication environment.

The attenuation incurred by wireless propagation can be decomposed in three main factors: a signal attenuation due to the distance between communicating nodes (*path loss*), attenuation effects due to absorption in local structures such as buildings (*shadowing loss*) and rapid signal fluctuations due to constructive and destructive interference of multiple reflected radio wave paths (*fading loss*).

The signal attenuation or path loss $\mathcal{D}$ is due to the fact that, when radio waves transmitted from an omnidirectional antenna propagate through an environment, the fraction of the transmitted power intercepted by a receiver at a distance r away from the receiver decreases with r. Empirically [148], [213], this attenuation $\mathcal{D}$ behaves as $\mathcal{D} \sim r^{-\alpha_{\mathrm{DL}}}$. The exponent α_{DL} takes values typically in the range [2], [4] and captures environment effects such as ground reflections of the propagation waves. The time scale over which $\mathcal{D}$ varies depends on how the distance r varies with time as determined by the mobility of the communication nodes. Typically path loss is the slowest varying among the attenuation effects.

The shadowing loss $\mathcal{S}$ is much more difficult to model, since it is due to absorption of the radio waves by scattering structures. It can be thought of as a random variable that takes a different value for each environment. Empirically, $\mathcal{S}$ is found to be well modeled by a log-normal distribution, i.e., the attenuation behaves as $\mathcal{S} \sim A_{\mathrm{SL}}$, where $\log(A_{\mathrm{SL}})$ follows a Gaussian distribution [213]. In other words, the attenuation measured in decibels is Gaussian distributed. Furthermore, time variation occurs when the scattering environment changes, for example, when a mobile turns a corner on a street. The time variation of the shadowing loss is quite difficult to model, and there exist few empirical studies. Qualitatively, we may state that shadowing loss varies with time faster than the path loss, but slower than the fading loss described below.

The fading loss $\mathcal{F}$ occurs due to the constructive and destructive interference of multiple reflected radio wave paths (see Fig. 2). Therefore, $\mathcal{F}$ varies when we sample the radio wave in different spatial locations. In a dense multipath environment, when there are enough scatterers, the complex amplitude is empirically well modeled by a Gaussian distribution [148], [213]. If there is a line of sight available, the mean of the Gaussian distribution is nonzero leading to an attenuation (absolute value of the complex amplitude) which is Ricean distributed and, hence, termed *Ricean fading*. When there is no line of sight, as would be the case in dense urban environments, we would have a complex amplitude which is zero-mean Gaussian distribution, leading to an attenuation which is Rayleigh distributed, and, hence, we have *Rayleigh fading* channels. We discuss how the fading loss affects a discrete-time system after we introduce the respective model.

In the case where we have multiple antennas, the relative attenuations between the antennas becomes an important issue. When the receive antennas are not far apart (of the order wavelengths of carrier frequency), the main variation arises due to the fading effects [209]. The relative response depends on the behavior of arrival directions of the reflected radio waves with respect to the receive antennas,

which is characterized by the *angular spread* of the multiple paths [148], [213]. Another attenuation effect that we did not mention depends on the polarization of radio waves and this can be used to obtain additional diversity (see, for example, [163] and references therein).

The overall radio propagation attenuation is the combination[3] of all three effects. For a transmitted signal $s(t)$, the continuous time received signal $y_c(t)$ can be expressed as

$$y_c(t) = \int h_c(t;\tau)s(t-\tau)d\tau + z(t) \tag{1}$$

where $h_c(t;\tau)$ is the response at time t of the time-varying channel if an impulse is sent at time $t-\tau$, where[4] $s(t) = g(t) * x(t)$ with $g(t)$ the transmission shaping filter, $x(t)$ the information signal, and $z(t)$ the additive Gaussian noise. The channel impulse response (CIR) depends on the combination of all three propagation effects $\mathcal{D}$, $\mathcal{S}$, $\mathcal{F}$ and in addition contains the delay induced by the reflections. Moreover, it may include the effect of a transmitter/receiver filter of a communication system, as described in [208].

To collect discrete-time sufficient statistics[5] of the information signal $x(t)$ we need to sample (1) faster than the Nyquist rate [208]. That is, we sample (1) at a rate larger than $2(W_I + W_s)$, where W_I is the input bandwidth and W_s is the bandwidth of the channel time variation [154]. A careful argument about the sampling rate required for time-varying channels can be found in [183]. In this paper, we assume that this criterion is met and, therefore, we focus on the following discrete-time model:

$$y(k) = y_c(kT_s) = \sum_{l=0}^{\nu} h(k;l)x(k-l) + z(k) \tag{2}$$

where $y(k)$, $x(k)$, and $z(k)$ are the output, input, and noise samples at sampling instant k, respectively, and $h(k;l)$ represents the sampled time-varying channel impulse response of finite length ν. Any loss in modeling the channel as having a finite duration impulse response can be made small by appropriately selecting ν [183].

As mentioned earlier, the channel response $\{h(k;l)\}$ depends on all three radio propagation attenuation factors $\mathcal{D}, \mathcal{S}, \mathcal{F}$. However, for many of the discussions in Sections III-A, III-C, and IV, in the time scales of interest, $\mathcal{D}$ remains constant. Therefore, for these discussions, the main varying effect arises from the fading loss, which is well modeled as a complex Gaussian variable. In the rest of this section we discuss the effects at this time scale further.

[3]For example, if there are L distinct (specular) delayed paths, then $h_c(t;\tau) = \sum_{l=1}^{L} \gamma_l(t)\delta(\tau - \tau_l)$, where $\gamma_l(t)$ depends on $\mathcal{D}(t)$, $\mathcal{S}_l(t)$, and $\mathcal{F}_l(t)$.

[4]Here $*$ represents convolution [208].

[5]The term sufficient statistics refers to a many-to-one function which does not cause loss of information about the random quantity of interest. When the channel is known at the receiver, one can design a whitened matched filter receiver [102], [208] and sample at the symbol rate to collect sufficient statistics. Here symbol rate refers to the underlying discrete-time information symbols modulating the carrier signal.

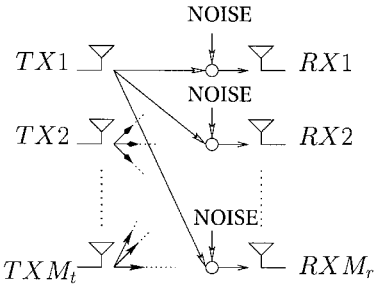

Fig. 3. MIMO channel model.

For the discrete-time model in (2) and the general case of a multiple-input multiple-output (MIMO) system (cf. Fig. 3), fading loss can be further refined as leading to time selectivity, frequency selectivity, and space selectivity.

Time selectivity arises from mobility, frequency selectivity arises from broadband transmission, and space selectivity arises from the spatial interference patterns of the radio waves. The three respective key parameters in the characterization of mobile broadband wireless channels are *coherence time*, *coherence bandwidth*, and *coherence distance*. The coherence time is the time duration over which CIR can be assumed constant. It is approximately equal to the inverse of the Doppler frequency.[6] The channel is said to be *time selective* if the symbol period is larger than the channel coherence time. The coherence bandwidth is the frequency duration over which the channel frequency response can be assumed flat. It is approximately equal to the inverse of the channel delay spread.[7] The channel is said to be *frequency selective* if the symbol period is smaller than the delay spread of the channel. Likewise, the coherence distance is the maximum spatial separation over which the channel response can be assumed constant and depends on the angular spread. The channel is said to be *space selective* between two antennas if their separation is larger than the coherence distance.

As seen from (2), the quantity $h(k;l)$ represents the sampled time-varying CIR (which combines the transmit filter with the physical channel).[8] The channel memory causes interference among successive transmitted symbols that results in significant performance degradation unless corrective measures (known as equalization) are implemented. In this paper, we shall use the terms *frequency-selective channel*, *broadband channel*, and *intersymbol interference (ISI) channel* interchangeably.

The introduction of M_t transmit and M_r receive antennas (cf. Fig. 3) leads to the following generalization of the basic channel model:

$$\mathbf{y}(k) = \sum_{l=0}^{\nu} \mathbf{H}(k;l)\mathbf{x}(k-l) + \mathbf{z}(k) \tag{3}$$

[6]The Doppler frequency is a measure of the frequency spread experienced by a pure sinusoid transmitted over the channel. It is proportional to the ratio of the mobile speed to the carrier wavelength.

[7]The channel delay spread is a measure of the time spread experienced by a pure impulse transmitted over the channel.

[8]The transmit filter may introduce correlation between the overall channel taps.

where the $M_r \times M_t$ complex[9] matrix $\mathbf{H}(k;l)$ represents the lth tap of the channel matrix response with $\mathbf{x} \in \mathbb{C}^{M_t}$ as the input and $\mathbf{y} \in \mathbb{C}^{M_r}$ as the output (cf. Fig. 3). The input vector may have independent entries to achieve high throughput (e.g., through spatial multiplexing) or correlated entries through coding or filtering to achieve high reliability (better distance properties, higher diversity, spectral shaping, or desirable spatial profile, see Sections III and IV). Throughout this paper, the input is assumed to be zero mean and to satisfy an average power constraint, i.e., $\mathbb{E}[||\mathbf{x}(k)||^2] \leq P$. The vector $\mathbf{z} \in \mathbb{C}^{M_r}$ models the effects of noise and interference.[10] It is assumed to be independent of the input and is modeled as a complex additive circularly symmetric Gaussian vector with $\mathbf{z} \sim \mathbb{C}\mathcal{N}(0, \mathbf{R}_{zz})$, i.e., a complex Gaussian vector with mean $\mathbf{0}$ and covariance $\mathbf{R}_{zz}$.

Finally, we modify the basic channel model to accommodate a block or frame of N consecutive symbols. Now, (3) can be expressed in matrix notation as follows:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{z} \tag{4}$$

where $\mathbf{y}, \mathbf{z} \in \mathbb{C}^{N.M_r}$, $\mathbf{x} \in \mathbb{C}^{M_t(N+\nu)}$, and $\mathbf{H} \in \mathbb{C}^{N.M_r \times M_t(N+\nu)}$. In each input block, we insert a guard sequence of length equal to the channel memory ν to eliminate interblock interference (IBI). In practice, the most common choices for the guard sequence are the all-zeros sequence (also known as *zero stuffing*) and the *cyclic prefix* (CP). When the channel is known at the transmitter, it is possible to increase throughput by optimizing the choice of the guard sequence, as shown in [13].

In practice, block interleaving (where successive data blocks are time interleaved by a depth larger than the channel coherence time) causes successive blocks to experience (nearly) independent fading. An alternative way to achieve independent fading between successive blocks is through carrier frequency hopping between block transmissions, where the separation between carrier hops is larger than the channel coherence bandwidth. Some of the results in Section IV use the assumption of independent fading from block to block; however, this is not crucial for the results in Section III.[11]

The channel model in (4) includes several popular channel models as special cases. First, the quasi-static channel model follows by assuming the channel time-invariant within the transmission block. In this case, using the cyclic prefix makes the channel matrix $\mathbf{H}$ *block-circulant*, hence, diagonalizable using the fast Fourier transform (FFT). Second, the flat-fading channel model follows by setting $\nu = 0$ which renders the channel matrix $\mathbf{H}$ a *block diagonal* matrix. Third, the channel model for single-antenna transmission, reception, or both follows directly by setting M_t, M_r, or both equal to one, respectively.

The single-user model of (4) can be easily generalized to the situation where there are multiple users. In Section III-B, the multiuser model is developed in the context of a *communication graph*. In such a graph, the wireless links between users can be modeled as edges with the users being the vertices. Therefore, the channel model developed in this section is general enough to handle a variety of possible propagation scenarios.

For illustration purposes, we consider the following two transmission scenarios.

- **Scenario A** characterized by a wide transmission bandwidth of 20 MHz, a channel delay spread less than 800 ns, and no mobility. Therefore, the channel memory is $\nu \leq 15$. This scenario represents channels with severe frequency selectivity and no time selectivity. This is applicable to indoor systems such as wireless local area networks (WLANs) (802.11 suite of systems) and outdoor systems such as wireless local loop (or fixed wireless systems).
- **Scenario B** characterized by a narrower transmission bandwidth of 1.25 MHz, a typical urban channel delay spread less than 5 μs, and high mobility. Therefore, the channel memory is $\nu \leq 5$. This scenario represents channels with severe time selectivity and mild frequency selectivity. This is applicable to high-mobility systems such as envisaged in 3G/4G cellular wireless systems (including wideband CDMA systems).

These two scenarios will be threaded throughput the paper to elucidate the main concepts and tradeoffs involved in the various system design choices.

III. Theoretical Considerations

In this section, we review the information-theoretic results that guide the way to recent developments in reliable high data-rate wireless networks. The major recent development in the past few years is the emergence of space–time (multiple-antenna) techniques and we present the results from this perspective. Another recent focus of interest is on the capacity of multiuser wireless networks. We also present some recent results on this topic.

The capacity of fading channels was studied in the 1960s (see, for example, [112, Sec. 8.6]). At the same time, the foundations of network information theory were being laid out by characterizing the information-theoretic capacity of channels with many users ([70, Ch. 14]). Since the wireless medium is an inherently shared medium, these results are of relevance for wireless networks. Most problems in network information theory remain open, but a notable exception is the multiple-access channel [70], which characterizes communication from many terminals to a single terminal or base station.

The work on fading channels has been pursued in parallel by communication engineers and information theorists, with important insights flowing both ways (see, for example,

[9]In passband communication, a complex signal arises due to in-phase and quadrature phase modulation of the carrier signal [208].

[10]Including cochannel interference, adjacent channel interference, and multiuser interference.

[11]The presence of channel correlation across blocks would require a longer observation interval (session time) to realize the same diversity performance gains as in the independent-channel-from-block-to-block case. This, in turn, places constraints for real-time delay-sensitive applications. On the other hand, this interblock correlation, if present, could be used to improve the convergence of channel tracking algorithms.

[197], [208], [226] and references therein). The importance of diversity to communication in the presence of fading was recognized quite early (see, for example, [45]). Frequency, time, and space (multiple antennas) were identified as different means of providing independent realizations of the fading channel. The early focus was on military applications with antenna arrays, where the different fading instantiations were collected at the receiver and the transmitted signal was detected using multiple "looks" through the diversity realizations. Transmit diversity was much less common, and typically the *same* signal was transmitted over the diversity branches (for example, see [283] and references therein). This is a form of repetition coding. The idea of coding across fading diversity branches (realizations) at the transmitter was proposed in the mid-1990s by several researchers ([103], [242], [245]) from different points of view. Foschini [103] and Telatar [245] focused on channel capacity and demonstrated significant theoretical improvements over single transmit, single receive channels. Tarokh *et al.* [242] focused on reliability, and introduced simple but nondegenerate correlation in the information streams across the multiple transmit antennas. These approaches led to the area of multiple-antenna coding which is still a rapidly evolving research topic.

We organize this survey of recent developments in spatial diversity as follows. In Section III-A we review results in single-user multiple-antenna communication for fading wireless channels. The unique problems associated with multiple-user channels are treated in Section III-B where the concept of *multiuser diversity* is introduced. Section III-C contrasts minimization of error probability with maximization of rate. Section III-D distills insights from these theoretical considerations that go into the design of practical techniques for realizing the promised performance gains.

A. Information-Theoretic Results for Multiple-Antenna Channels

The concept of capacity was introduced by Shannon in 1948 [227], where he showed that even in noisy channels, one can transmit information at positive rates with the error probability going to zero asymptotically in the coding block size. Perhaps the most famous illustration of this idea was the the formula derived in [227] for the capacity C of the additive white Gaussian noise (with variance N_0) channel with an input power constraint P, namely

$$C = \frac{1}{2}\log\left(1+\frac{P}{N_0}\right). \tag{5}$$

For scalar[12] flat-fading channels (where $\nu = 0$ in (2) of Section II), and a coherent receiver (where the receiver uses perfect channel state information), the capacity was shown to be [98]

$$C = \frac{1}{2}\mathbb{E}\left[\log\left(1+\frac{|h|^2P}{N_0}\right)\right] \tag{6}$$

where[13] the expectation is taken over the fading channel $\{h(k)\}$ and the channel is assumed to be stationary and ergodic. This is called the *ergodic channel capacity* [197]. This is the rate at which information can be transmitted if there is *no feedback* of the channel state ($\{h(k)\}$) from the receiver to the transmitter. If there is feedback available about the channel state, one can do slightly better through optimizing the allocation of transmitted power by "waterfilling" over the fading channel states. The problem of studying the capacity of channels with causal transmitter side information was introduced by Shannon in [228], where a coding theorem for this problem was proved. When the transmitter has noncausal side information (more suitable for data-hiding problems [55]) the capacity was characterized by [116]. Using these techniques, transmitter side information in the context of fading channels was explored in [123]. The capacity with instantaneous channel state feedback was shown to be

$$C = \frac{1}{2}\max_{S(\gamma)\in\mathcal{S}} \mathbb{E}_\gamma \log\{1+\gamma S(\gamma)\} \tag{7}$$

where $\gamma = |h|^2P/N_0$ is the varying signal-to-noise ratio (SNR), and the power control strategy is $S(\gamma)$, for a long-term power constraint $\mathcal{S} = \{S(\gamma) : \int S(\gamma)p(\gamma)d\gamma = P\}$. However, for fast time-varying channels the instantaneous feedback could be difficult, resulting in an outdated estimate of the channel being sent back [274]. The case where partial information about the channel state at the transmitter (and at the receiver) is studied in [48]. There has also been work in understanding the impact of imperfect transmitter channel state information on performance [38], [270]. The understanding of achievable performance over scalar fading channels is far deeper than what is covered here and we refer the interested reader to [41]. Also note that when we are dealing with complex channels (as is usual in communication with in-phase and quadrature-phase transmissions), the factor of 1/2 in the equations above disappears [193]. From this point on, we assume complex channels as set up in Section II. Also, unless otherwise stated, we will deal with the case where the transmitter does not have access to the channel state information.

In the presence of spatial diversity, the results above can be easily modified. The simplest case is when we have receive diversity, i.e., when $M_t = 1$, for a given M_r and flat-fading channels ($\nu = 0$). Here it is shown [197] that the capacity is given by

$$C = \mathbb{E}\left[\log\left(1+\frac{\|\mathbf{h}\|^2P}{N_0}\right)\right] \tag{8}$$

where $\{\mathbf{h}(k)\}$ is now the channel vector response [see (3) in Section II]. A similar expression when $M_r = 1$ can be derived; however, we will illustrate it as a special case of the general M_r, M_t channel capacity.

Recent results in [103], [141], [210], [242], and [245] suggest significant advantages in using both transmitter and re-

[12]In this paper we refer to single-antenna and multiple-antenna channels as scalars and vectors (or matrices), respectively.

[13]In this paper, $|h|^2 = \bar{h}h$ where $\bar{h}$ denotes complex conjugation, and for a vector $\mathbf{h}$ we denote its 2-norm by $\|\mathbf{h}\|^2 = \mathbf{h}^*\mathbf{h}$, where $\mathbf{h}^*$ denotes the Hermitian transpose.

ceiver spatial diversities, i.e., M_r, $M_t > 1$. In [245], the capacity of multiple-antenna fading channels was first established.[14] In the flat-fading case where $\nu = 0$, the mutual information[15] for a block of N time samples can be written as

$$\begin{aligned}
&\frac{1}{N} I(\mathbf{X}^{(N)}; \mathbf{Y}^{(N)}, \mathbf{H}^{(N)}) \\
&= \frac{1}{N}\left[I(\mathbf{X}^{(N)}; \mathbf{H}^{(N)}) + I(\mathbf{X}^{(N)}; \mathbf{Y}^{(N)} | \mathbf{H}^{(N)}) \right] \\
&\underset{(a)}{=} \frac{1}{N} \mathbb{E}_{\mathcal{H}}[I(\mathbf{X}^{(N)}; \mathbf{Y}^{(N)} | \mathcal{H}^{(N)} = \{\mathbf{H}^{(N)}\})] \\
&\underset{(b)}{=} \mathbb{E}_{\mathcal{H}}\left[\log\left(\frac{|\mathbf{R}_z + \mathbf{H}\mathbf{R}_x\mathbf{H}^*|}{|\mathbf{R}_z|} \right) \right] \qquad (9)
\end{aligned}$$

where (a) follows from the fact that the input $\{\mathbf{x}(k)\}$ is independent of the fading process [as the transmitter does not have channel state information (CSI)] and (b) follows from the memoryless property of the vector Gaussian channel obtained by conditioning on $\mathbf{H}(k)$[16] and also due to the assumption that $\{\mathbf{H}(k)\}$ is[17] i.i.d. over k. We also use i.i.d. Gaussian input $\{\mathbf{x}(k)\}$ with $\mathbf{R}_x = \mathbb{E}[\mathbf{x}(k)\mathbf{x}^*(k)]$, as this maximizes the mutual information conditioned on $\mathbf{H}$. In general, it is difficult to evaluate (9) except for some special cases. If we assume $\mathbf{R}_z = \sigma^2\mathbf{I}$, and if $\mathbf{H}(k)$ consists of i.i.d. Gaussian elements, [245] showed that

$$C = \mathbb{E}_{\mathcal{H}}\left[\log\left(\left| \mathbf{I} + \frac{P}{M_t\sigma^2}\mathbf{H}\mathbf{H}^* \right| \right) \right] \qquad (10)$$

is the capacity of the fading matrix channel, where $\mathbf{R}_x = P/M_t\mathbf{I}$. Further, it was shown in [245] that the capacity could be numerically computed using Laguerre polynomials [94], [186], [245].

Theorem 3.1: The capacity C of the channel with M_t transmitters and M_r receivers and average power constraint P is given by (11) at the bottom of the page, where $T_{\max} = \max(M_t, M_r), T_{\min} = \min(M_t, M_r)$, and $L_k^m(\cdot)$ is the generalized Laguerre polynomial of order k with parameter m [125].

■

This result allows us to numerically compute the capacity of the spatial diversity channel. This also shows an important connection between capacity of multiple-antenna channels and the mathematics related to eigenvalues of random matrices [94]. In [103] it was observed that when $M_t = M_r = M$, the capacity C grows linearly in M as $M \to \infty$.

Theorem 3.2: (Foschini [103]) For $M_t = M_r = M$, the capacity C given by (10) grows asymptotically (at least) linearly in M, i.e.,

$$\lim_{M\to\infty} \frac{C}{M} \geq \text{constant} . \qquad (12)$$

■

It is shown in [61] that this result is also robust under some restricted forms of correlation in the channel response matrix, i.e., the independence assumption on the elements of $\mathbf{H}(k)$ is not crucial.

To achieve the capacity given in (10), we require joint optimal [maximum-likelihood (ML)] decoding of all the receiver elements which could have large computational complexity. The channel model in (3) resembles a multiuser channel [268] with user cooperation. A natural question to ask is whether the simpler decoding schemes proposed in multiuser detection would yield good performance on this channel. A motivation for this is seen by observing that for i.i.d. elements of the channel response matrix (flat-fading) the normalized cross-correlation matrix decouples (i.e., $\lim_{M_r\to\infty} 1/M_r\mathbf{H}^*(k)\mathbf{H}(k) \to \mathbf{I}_{M_t}$). Therefore, since nature provides some decoupling, a simple "matched filter" receiver [268] might perform quite well. However, a tension arises between the decoupling of the channels and the added "interference" from the other antennas, as the number of antennas grow. We can show that the two effects exactly cancel each other and this detector still retains the linear growth rate of the optimal decoding scheme [84]. However, in the rate R_I achievable for this simple decoding scheme, we do pay a price in terms of rate growth with SNR.

Theorem 3.3: If $\mathbf{H}_{\mathrm{i,j}}(k) \sim \mathbb{CN}(0,1)$, then

$$\lim_{\substack{M_t\to\infty \\ M_t - |\alpha M_r|}} \frac{1}{M_t} I(\mathbf{Y}, \mathbf{H}; \mathbf{X}) \geq \lim_{\substack{M_t\to\infty \\ M_t - |\alpha M_r|}} \frac{R_I}{M_t} = \log\left(1 + \frac{\frac{P}{\sigma^2\alpha}}{1 + \frac{P}{\sigma^2}} \right).$$

■

Multiuser detection [143], [268] is a good analogy to understand receiver structures in MIMO systems. The main difference is that unlike multiple-access channels, the space–time encoder allows for cooperation between "users." Therefore, the encoder could introduce correlations that can simplify the job of the decoder. Such encoding structures using space–time block codes are discussed further in Section IV-A.3. An example of using the multiuser

[14] In [103], a similar expression was derived without illustrating the converse to establish that the expression was indeed the capacity.

[15] Mutual information (see [70]) is a lower bound on the achievable reliable information rate over a noisy channel.

[16] For a matrix $\mathbf{A}$, we denote its determinant as $det(\mathbf{A})$ and $|\mathbf{A}|$, interchangeably.

[17] The assumption that $\{\mathbf{H}(k)\}$ is independent identically distributed (i.i.d.) is not crucial. This result is (asymptotically) correct even when the sequence $\{\mathbf{H}(k)\}$ is a mean ergodic sequence [197].

$$C = \int_0^\infty \log\left(1 + \frac{P\lambda}{M_t}\right) \sum_{k=0}^{T_{\min}-1} \lambda^{T_{\max}-T_{\min}} \left[L_k^{T_{\max}-T_{\min}}(\lambda) \right]^2 \frac{k!}{k + T_{\max} - T_{\min}} e^{-\lambda} d\lambda \qquad (11)$$

detection approach is the result in Theorem 3.3 where a simple matched filter receiver is applied. Using more sophisticated linear detectors, such as the decorrelating receiver and the MMSE receiver [268], one can improve performance while still maintaining the linear growth rate. The decision feedback structures (successive interference cancellation, or onion peeling [68], [200], [284]) can be shown to be optimal, i.e., achieves the capacity, when an MMSE multiuser interference suppression is employed and the layers are peeled off [64], [265]. However, decision feedback structures inherently suffer from error propagation (which is not taken into account in the theoretical results) and could, therefore, have poor performance in practice, especially at low SNR. Thus, examining structures without decision feedback is important in practice. Recently several similar results have been developed in the context of random signature sequences for CDMA systems and asymptotic behavior of multiuser linear detector structures [254], [255].

All of the above results illustrate that there are significant gains in information rate (capacity) possible using multiple transmit and receive antennas. The gain is (asymptotically) linear in the number of antennas ($\min(M_r, M_t)$). Note that as the number of receive antennas grows, so does the channel gain, i.e., any norm of $\mathbf{H}$ also grows, reflecting the extra radio energy gathered by the antennas. One question examined in [84] is whether significant gains occur with multiple antennas even without these energy gathering effects (i.e., assuming a "passive" channel). It is shown that for channels which do not get these energy gathering gains (on the average), the rate grows linearly with SNR, asymptotically as the number of receive (and transmit) antennas grow. In this case, even though the capacity does not become unbounded for large number of receive (and transmit) antennas, the capacity grows linearly instead of logarithmically with SNR, i.e., $\lim_{M_r=M_t\to\infty} C \sim$ SNR. This is reminiscent of the infinite-bandwidth fading channel ([112, Ch. 8]). The intuition for the gains with multiple transmit and receive antennas is that there are a larger number of communication modes over which the information can be transmitted. This is formalized by the observation [84], [103], [290] that the capacity as a function of SNR $C(\text{SNR})$ grows linearly in $\min(M_r, M_t)$, even for finite number of antennas, asymptotically in the SNR.

Theorem 3.4:

$$\lim_{\text{SNR}\to\infty} \frac{C(\text{SNR})}{\log(\text{SNR})} = \min(M_r, M_t). \tag{13}$$

■

In the results above, the fundamental assumption was that the receiver had access to *perfect* channel state information, obtained through training or other methods. When the channel is slowly varying, the estimation error could be small, since we can track the channel variations and the effect of such estimation errors was investigated in [184]. As a rule of thumb, it is shown in [165] that if the estimation error is small compared to $1/\text{SNR}$, these results would hold. Another line of work assumes that the receiver does not have *any* channel state information. The question of the information rate that can be reliably transmitted over the multiple-antenna channel without channel state information was introduced in [141] and has also been examined in [290]. The main result from this line of work shows that the capacity growth is again (almost) linear in the number of transmit and receive antennas, as stated formally next.

Theorem 3.5: If the channel is block fading with block length T and we denote $K = \min(M_t, M_r)$, then for $T > K + M_t$, as SNR $\to \infty$, *the capacity*[18] *is*

$$C(\text{SNR}) = K\left(1 - \frac{K}{T}\right)\log(\text{SNR}) + c + o(1)$$

where c is a constant depending only on M_r, M_t, T.

■

In fact, [290] goes on to show that the rate achievable by using a training-based technique is only a constant factor away from the optimal, i.e., it attains the same capacity–SNR slope as in Theorem 3.5. Further results on this topic can be found in [136]. Therefore, even in the noncoherent block-fading case, there are significant advantages in using multiple antennas.

Fading ISI Channels: There has been a large body of work devoted to data transmission over frequency-selective fading channels [208]. Transmitter and receiver diversity in time-invariant ISI channels have been examined in [24], [44], [210], [286], and references therein. In [44], [210], and [286], the *time-invariant* MIMO ISI channel is studied when *both* the transmitter and the receiver have perfect side-information of the channel state. Since we are dealing with *time-invariant* channels, the Fourier basis is the eigenbasis of the channel and one can easily develop a capacity argument using standard techniques [112]. Using such an argument, it is shown in [210] that the capacity scales linearly in the number of antennas ($\min\{M_t, M_r\}$) at high SNR. Therefore, since the Fourier basis is optimal in this case, orthogonal frequency division multiplexing (OFDM) is a good choice as a modulation technique for the MIMO channel. In the time-varying channel as well, if the transmitter has perfect channel state information, then the channel can be decomposed into parallel channels [40], [83], [160] and optimal waterfilling spectrum can be found. Alternatively, in time-varying channels only the receiver might know the fading channel state [197]. However, if the channel is assumed to be quasi-static, i.e., time-invariant over the transmission block, the Fourier basis is again the eigenbasis for the channel transmission block and, therefore, is the optimal transmission basis [197]. This assumption is suitable for slowly time-varying channels where transmission bandwidth is much greater than the Doppler spread. The rate of reliable information for the scalar ISI channel has been derived in [197] in terms of the expected mutual information. This argument can easily be generalized [84] to the multiple-antenna block-fading ISI channel as

$$\lim_{n\to\infty} R_n = \mathbb{E}\left[\int_0^{2\pi} \log\left(\left|\mathbf{I} + \frac{\mathbf{H}(f)\mathbf{S}(f)\mathbf{H}^*(f)}{\sigma^2}\right|\right) df\right] \tag{14}$$

[18]Here the notation $o(1)$ indicates a term that goes to zero when SNR $\to \infty$.

where $\mathbf{H}(f)$ is the Fourier transform of the channel matrix and $\mathbf{S}(f)$ is the input power spectral density. In general, maximizing (14) with respect to $\mathbf{S}(f)$ is a hard problem and further simplifying assumptions need to be made for solving this problem. However, schemes using a flat input spectrum in both time and "space" may be practical [14], [103], [242].

On time-varying channels, there is an inherent conflict between increasing the transmission block length (for coding arguments) and the block time-invariance assumption. Therefore, as the block length increases, there could be time variation within a transmission block. In this case, the Fourier basis is no longer the optimal basis for transmission, and would suffer loss of orthogonality, since it would not be the eigenbasis for the time-varying channel. The general question of a characterization of capacity for such a channel is open. Some lower bounds for capacity have been developed [10], [84].

Outage Capacity: In all of the above results, the error probability goes to zero asymptotically in the length of the coding interval. Therefore, coding is assumed to take place *across* fading blocks,[19] and, hence, it inherently uses the *ergodicity* of the channel variations. This approach would clearly entail large delays and, therefore, [197] (see also [41]) introduced a notion of outage, where the coding is done (in the extreme case) just across one fading block. Here the transmitter sees only one block of channel coefficients and, therefore, the channel is *nonergodic*, and the strict Shannon-sense capacity is zero. However, one can define an outage probability which is the probability with which a certain rate R is possible. Therefore, for a block time-invariant channel $\mathbf{H}(k) = \mathbf{H}, \forall k$, the outage probability can be defined as follows.

Definition 3.1: The outage probability for a transmission rate of R and a given transmission strategy $p(\mathbf{X})$ is defined as

$$P_{\text{outage}}(R, p(\mathbf{X})) = \mathbb{P}\left\{\mathbf{H} : I(\mathbf{X}; \mathbf{Y}|\mathbf{H}(k) = \mathbf{H}) < R\right\}. \tag{15}$$

■

Therefore, if one uses a white Gaussian codebook ($\mathbf{R}_x = P/M_t\mathbf{I}$), then (abusing notation by dropping the dependence on $p(\mathbf{X})$) we can write the outage probability at rate R as

$$P_{\text{outage}}(R) = \mathbb{P}\left\{\log\left(\left|\mathbf{I} + \frac{P}{M_t\sigma^2}\mathbf{H}\mathbf{H}^*\right|\right) < R\right\}. \tag{16}$$

This is a concept that was also used by Foschini in [103] in order to numerically demonstrate the potential advantages of multiple-antenna coding. It has been shown in [291] that at high SNR the outage probability is the same as the frame-error probability in terms of the SNR exponent. Therefore, to evaluate the optimality of practical coding techniques, one can compare, for a given rate, how far the performance of the technique is from that predicted through an outage analysis. Moreover, the frame-error rates and outage capacity comparisons in [242] can be also formally justified through this argument.

[19]A block is a contiguous set of transmitted symbols and we use it interchangeably with frame.

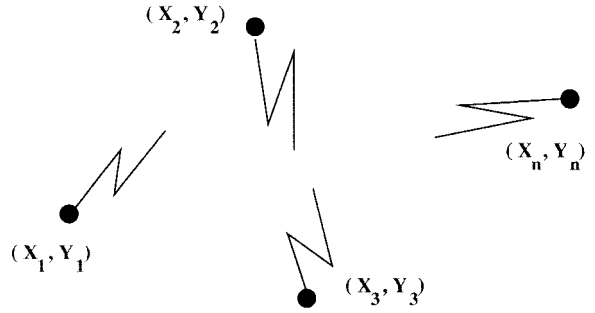

Fig. 4. General multiuser wireless communication network.

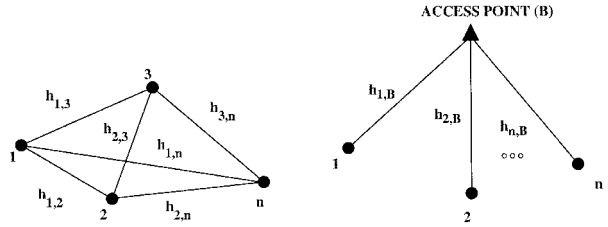

Fig. 5. Graph representation of communication topologies.

B. Multiuser Channels

The wireless medium is inherently shared and this directly motivates a study of multiuser communication techniques. The general communication network (illustrated in Fig. 4) consists of n nodes trying to communicate with each other. At each instant of time t, the ith node sends a symbol $x_i(t)$ that could in general depend on the message it wants to send and the entire history of received symbols $Y_i(\cdot)$. As mentioned earlier, multiuser information theory (or network information theory) has been an area of inquiry with a large number of open questions. One way of abstracting the multiuser communication problem is through embedding it in an underlying *communication graph* $\mathcal{G}_C$ where the n nodes are vertices of the graph and edges of the graph represent a channel connecting the two nodes. The graph could be directed with constraints and channel transition probability depending on the directed graph. A general multiuser network is, therefore, a fully connected graph with the received symbol at each node described as a conditional distribution dependent on the messages transmitted by all other nodes. Such a graph is illustrated in Fig. 5.

In the scalar wireless channel, the received symbol $Y_i(t)$ at the ith node is given by

$$Y_i(t) = \sum_{\substack{j=1 \\ j \neq i}}^{n} h_{i,j} X_j(t) + Z_i(t) \tag{17}$$

where $h_{i,j}$ is determined by the fading attenuation between nodes i and j, and the transmitted symbols could depend on the entire history of the received symbols on that node. Even with this restriction, the general problem has very few completely answered questions. We term this general structure an *ad hoc communication topology* where the graph structure is not *a priori* fixed. We further restrict graph (or node *topology*) by only allowing certains links to be active. One

formulation is obtained by giving special status to one of the nodes as the base station or access point. The other nodes can *only* communicate to the base station. We call such a topology the *hierarchical communication topology* and this is common in current commercial cellular wireless systems [237]. From a systems perspective, we can view the ad hoc communication topology as a situation where there is no separation between "infrastructure" (e.g., access point) and the end users.

In the hierarchical topology, we distinguish between two types of problems; the first is the *uplink* channel where the nodes communicate to the access point (many-to-one communication or the *multiple-access channel* [70]) and the second is the *downlink* channel where the access point communicates to the nodes (one-to-many communication or the *broadcast channel* [70]). Given that there are multiple users, multiuser communication problems investigate the rate-tuple that can be simultaneously achieved by the users (achievable *rate region*). The characterization of this rate region is central to all network information theory problems. The rate region of the multiple-access fading channel is quite well understood both in the scalar case as well as the case with multiple antennas at the transmitter [41], [271]. However, the rate region of the fading broadcast channel is less well understood. For the single-antenna case, the rate region has been characterized in [168]. For the multiple-antenna broadcast channel, the rate region has not been completely characterized. Some recent progress in terms of characterizing the sum capacity of this channel has been made [49], [269], [272], [289]. The rate-sum point is achieved using a coding technique termed "dirty-paper coding" [67]. We discuss a little more about this in Section III-D. Our coverage of the information-theoretic aspects of multiuser fading channels has been quite superficial, and we point the interested reader to [41] for an excellent survey of this topic.

Another reason for studying the wireless channel from a multiuser perspective is that we can view the different users as a form of diversity, called *multiuser diversity*. This is because each user potentially has independent channel conditions and local interference environment. This implies that in Fig. 5, the fading links between users are random and independent of each other. Therefore, this diversity in channel and interference conditions can be exploited by treating the independent links from *different* users as conduits for information transfer.

The idea of multiuser diversity can be further motivated by looking at the scalar fading multiple-access channel. The rate region for the uplink channel for a single cell was characterized in [161] where it was shown that in order to maximize the total information capacity (the sum rate), it is optimal to transmit *only* to the user with the best channel. For the scalar channel, the channel gain determines the best channel. The result in [161] when translated to rapidly fading channels results in a form of time-division multiple access (TDMA), where the users are not preassigned time slots, but are scheduled according to their respective channel conditions. If the users' channels are varying independently, then this strategy is a form of *multiuser diversity* where the diversity is viewed across users. Here the multiuser diversity (which arises through independent channel realizations across users) can be harnessed using an appropriate scheduling strategy. A similar result also holds for the scalar fading broadcast channel [168], [256]. This idea is inherently used in the downlink scheduling algorithm used in IS-856 [34], [149] [also known as the High Data Rate (HDR) 1xEV-DO system], while ensuring fairness between users. Note that this requires feedback from the users to the base station about the channel conditions. The feedback could be just the received SNR.

In an ad hoc communication topology (network), one need not transmit information directly from source to destination, but one can use other users which act as relays to help communication of information to its ultimate destination. Such multihop wireless networks have rich history (see, for example, [144] and references therein). Therefore, for such an information transmission strategy, understanding the rate region of the relay channel is an important component [70]. The relay channel was introduced in [263] and the rate region for special cases was presented in [69]. While the general rate region has not been characterized, there is some recent understanding in the context of wireless networks, asymptotically in the network size [115], [130], [285].

In an important step toward systematically understanding the capacity of wireless networks, [129] took an alternate approach, and explored the behavior of wireless networks asymptotically in the number of users while using simple coding strategies. In this case, they placed n nodes independently and randomly at locations $\{X_i\}$ in a finite geographical area (a scaled unit disk). Also $\Theta(n)$[20] source and destination pairs were randomly chosen. Instead of assuming sophisticated coding (so that mutual information rate is achievable), they explored a simpler scenario where transmissions were of a constant rate R bits, and where a successful transmission occurred when the signal-to-interference-plus-noise ratio (SINR) was above a certain threshold β, hence treating interference as noise. Note that this need not be an information-theoretically optimal strategy. In order to represent wireless signal transmission, the signal strength variation was modeled only through path loss (see Section II) with exponent α. Therefore, if $\{P_i\}$ were the powers at which the various nodes transmitted, then the SINR from node i to node j is defined as

$$\text{SINR} = \frac{\frac{P_i}{|X_i - X_j|^\alpha}}{N_0 + \sum_{\substack{k \in \mathcal{T} \\ k \neq i}} \frac{P_k}{|X_k - X_j|^\alpha}} \tag{18}$$

where $\mathcal{T}$ is the subset of users simultaneously transmitting at some time instant.

Definition 3.2: If there is a scheduling and relay policy π, and if $M_i^\pi(t)$ is the number of packets from source node i to its destination node $D(i)$ successfully delivered at time t,

[20]We use the notation $f(n) = \Theta(g(n))$ to denote $f(n) = O(g(n))$ as well as $g(n) = O(f(n))$. Here $f(n) = O(g(n))$ means $\limsup_{n\to\infty} |f(n)/g(n)| < \infty$.

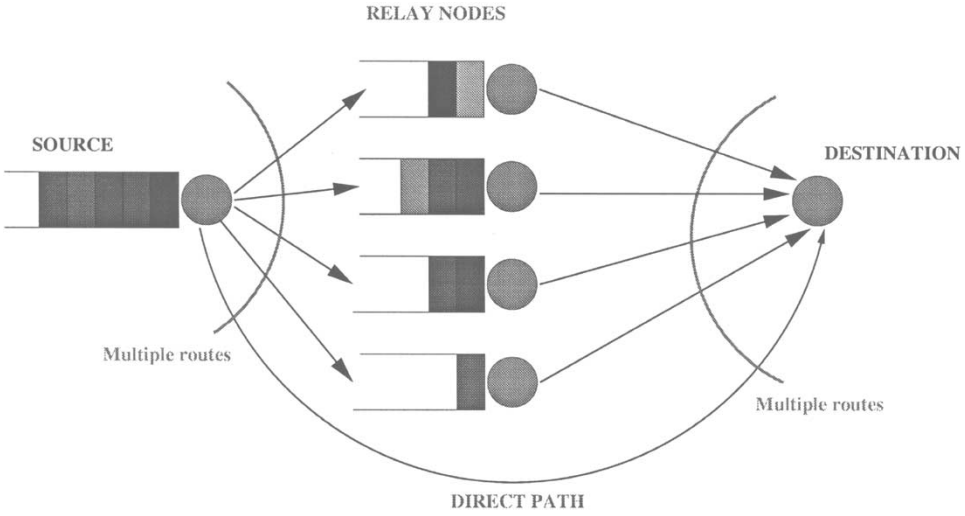

Fig. 6. Multiuser diversity through relays.

then a long-term throughput $\tilde{\lambda}(n)$ is feasible if there exists a policy π such that for every *source-destination pair*

$$\liminf_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}M_i^{\pi}(t)\geq\tilde{\lambda}(n). \tag{19}$$

We define the throughput $\lambda(n)$ as the highest achievable $\tilde{\lambda}(n)$.

■

Note that $\lambda(n)$ is a random quantity which depends on the node locations of the users. One of the main results in [129] was the following.

Theorem 3.6: There exists constants c_1 and c_2 such that

$$\lim_{n\to\infty}\mathbb{P}\left\{\lambda(n)=\frac{c_1R}{\sqrt{n\log n}}\text{ is feasible}\right\}=1$$
$$\lim_{n\to\infty}\mathbb{P}\left\{\lambda(n)=\frac{c_2R}{\sqrt{n}}\text{ is feasible}\right\}=0. \tag{20}$$

■

Therefore, the long-term per-user throughput decays as $O(1/\sqrt{n})$, showing that high per-user throughput may be difficult to attain in large scale (fixed) wireless networks. One way to interpret this result is the following. If n nodes are randomly placed in a unit disk, nearest neighbors (with high probability) are at a distance $O(1/\sqrt{n})$ distance apart. It is shown in [129] that it is important to schedule a large number of simultaneous short transmissions, i.e., between nearest neighbors. Therefore, if randomly chosen source-destination pairs are $O(1)$ distance apart and we can only schedule nearest neighbor transmissions, information has to travel $O(\sqrt{n})$ hops to reach its destination. Since there can be at most $O(n)$ simultaneous transmissions at a given time instant, this imposes a $O(1/\sqrt{n})$ upper bound on such a strategy. This is an intuitive argument, and a rigorous proof of Theorem 3.6 is given in [129] among other interesting results. Note that the coding strategy was simple and the interference was treated as part of the noise. An open question concerns the throughput when we use sophisticated codes and interference is *not* treated as just noise, i.e., multiuser detection [268] is used. Moreover, in Theorem 3.6 the traffic pattern was assumed to be uniform, i.e., source-destination pairs are chosen uniformly throughout the communication region. Perhaps by introducing more structured traffic patterns, the result might be less pessimistic.

In [126], node mobility was allowed and the locations $\{X_i(t)\}$ vary in a stationary and ergodic manner over the entire disk. In the presence of such symmetric (among users) and "space-filling" mobility patterns, a surprising result was established.

Theorem 3.7: There exists a scheduling and relaying policy π and a constant $c>0$ such that

$$\lim_{n\to\infty}\mathbb{P}\left\{\lambda(n)=cR\text{ is feasible}\right\}=1. \tag{21}$$

■

Therefore, node mobility allows us to achieve a per-user throughput of $O(1)$. The main reason this was attainable was that packets are relayed only through a finite number of hops by utilizing node mobility. Thus, a node carries packets over $O(1)$ distance before relaying it and, therefore, [126] shows that with high probability if the mobility patterns are space filling, the number of hops needed from source to destination is bounded instead of growing as $O(\sqrt{n})$ in the case of fixed (nonmobile) wireless networks [129]. One natural question is whether the mobility model assumed is generous and is the reason for the impressive result. In [81], it is shown that even with much more restricted user mobility, $O(1)$ throughput is possible. In particular, the mobility patterns were restricted to random line segments and once chosen, the configuration of line segments are fixed for all time. Therefore, given the configuration, the only randomness arose through user mobility along these line segments. In this case, it is possible to construct configurations that do not allow $O(1)$ per-user throughput. However, if the n line segments are chosen independently and randomly, then the following result can be established [81].

Theorem 3.8: Given a configuration $\mathcal{C}$, there exists a scheduling and relaying policy π and a constant $c > 0$ such that

$$\lim_{n\to\infty} \mathbb{P}\left\{\lambda(n) = cR \text{ is feasible}|\mathcal{C}\right\} = 1 \tag{22}$$

for almost all configurations $\mathcal{C}$ as $n \to \infty$, i.e., the probability of the set of configurations for which the policy achieves a throughput of λ goes to 1 as $n \to \infty$.

■

Therefore, $O(1)$ throughput is possible under restrictive mobility models as well. Mobility allows source-destination pairs to be able to relay information through several independent relay paths, since nodes have changing nearest neighbors due to mobility. This method of relaying information through independent attenuation links which vary over time is also a form of multiuser diversity. One can see this by observing that the transmission occurs over several realizations of the communication graph $\mathcal{G}_C$. The relaying strategy which utilizes mobility schedules transmissions over appropriate realizations of the graph. Conceptually, this use of independent relays to transmit information from source to destination is illustrated in Fig. 6, where the strategy of Theorems 3.7 and 3.8 is used. In stationary nodes as in the case of Theorem 3.6, it was shown [129] that nearest neighbor relaying was important to achieve this rate. Though the link gains between users do not change over time, they are independent of one another. This relaying strategy can be thought of as a weaker form of multiuser diversity.

Thus, the concept of diversity can be used in connection with frequency, time, and multiple antennas, but also arises through multiple users. If the users are distributed across geographical areas, their channel responses would be different depending on their local environments. Therefore, even if a particular user at the current time might be in a deep fade, there could be another user who has good channel conditions. We have seen two forms of this multiuser diversity, one using a scheduler in single-hop networks and the other using relays in multihop networks. This form of spatial diversity is important in the context of multiuser communication in wireless networks and can be exploited, since it is readily available. The issues of fairness and how to distribute resources in such a case are important and will be a topic we visit in Section V-A.3.

In ad hoc networks, multiuser diversity can be exploited by relaying information from source to destination via other users in the network. This was shown to be crucial in Theorems 3.6 and 3.7 to obtain the higher per-user throughput. In order to relay information, routing becomes an important problem and this is a topic we discuss in Section V-B.

C. Diversity Order

In Section III-A the focus was on achievable transmission rate. A more practical performance criterion is probability of error. This is particularly important when we are coding over a small number of blocks (low-delay) where the Shannon capacity is zero [197] and, therefore, we need to design for low error probability. By characterizing the error probability, we can also formulate design criteria for space–time codes. The frame error probability is also intimately connected to the outage probability introduced in Definition 3.1 and as discussed, the two are approximately equivalent at high SNR.

Since we are allowed to transmit a coded sequence, we are interested in the probability that an erroneous codeword[21] $\mathbf{e}$ is mistaken for the transmitted codeword $\mathbf{x}$. This is called the *pairwise error probability* (PEP) [90] and is then used to bound the error probability. This analysis relies on the condition that the receiver has perfect channel state information. However, a similar analysis can be done when the receiver does not know the channel state information, but has statistical knowledge of the channel [142].

For simplicity, we shall first present the result for a flat-fading channel[22] (where $\nu = 0$) and when the channel matrix contains i.i.d. Gaussian elements. Many of these results can be easily generalized for correlated fading, and other fading distributions. Consider a codeword sequence $\mathbf{x} = [\mathbf{x}^T(0), \ldots, \mathbf{x}^T(N-1)]$, where $\mathbf{x}(k) = [\mathbf{x}_1(k), \ldots, \mathbf{x}_{M_t}(k)]^T$ as defined in (3). In the case when the receiver has perfect channel state information, we can bound the PEP between two codeword sequences $\mathbf{x}$ and $\mathbf{e}$ (denoted by $P(\mathbf{x} \to \mathbf{e})$) as follows [128], [242]:

$$P(\mathbf{x}\to\mathbf{e}) \le \prod_{m=1}^{M_r}\left[\prod_{n=1}^{M_t}\frac{1}{1+\frac{E_s}{4N_0}\lambda_n}\exp\left(-\frac{K_{m,n}\frac{E_s}{4N_0}\lambda_n}{1+\frac{E_s}{4N_0}\lambda_n}\right)\right] \tag{23}$$

where $K_{m,n}$ is the Ricean coefficient [213], [242], $E_s = P/M_t$ is the power per transmitted symbol, λ_n are the eigenvalues of the matrix $\mathbf{A}(\mathbf{x},\mathbf{e}) = \mathbf{B}^*(\mathbf{x},\mathbf{e})\mathbf{B}(\mathbf{x},\mathbf{e})$, and

$$\mathbf{B}(\mathbf{x},\mathbf{e}) = \begin{pmatrix} \mathbf{x}_1(0)-\mathbf{e}_1(0) & \cdots & \mathbf{x}_{M_t}(0)-\mathbf{e}_{M_t}(0) \\ \vdots & \vdots & \vdots \\ \mathbf{x}_1(N-1)-\mathbf{e}_1(N-1) & \cdots & \mathbf{x}_{M_t}(N-1)-\mathbf{e}_{M_t}(N-1) \end{pmatrix}. \tag{24}$$

Therefore, for the Rayleigh fading channel ($K_{m,n} = 0$), we have

$$P(\mathbf{x}\to\mathbf{e}) \le \left[\frac{1}{\prod_{n=1}^{M_t}\left(1+\frac{E_s}{4N_0}\lambda_n\right)}\right]^{M_r}. \tag{25}$$

If q denotes the rank of $\mathbf{A}(\mathbf{x},\mathbf{e})$, (i.e., the number of nonzero eigenvalues) then we can bound (25) as

$$P(\mathbf{x}\to\mathbf{e}) \le \left[\prod_{n=1}^{q}\lambda_n\right]^{-M_r}\left(\frac{E_s}{4N_0}\right)^{-qM_r}. \tag{26}$$

Thus, we define the notion of diversity order as follows.

[21]For an information rate of R bits per transmission and a block length of N, we define the codebook as the set of 2^{NR} codeword sequences of length N.

[22]For scalar (single-antenna) fading channels the pairwise error probability was derived in [90]. For a treatment of error probability analysis for more general fading distributions, see [230].

Definition 3.3: A coding scheme which has an average error probability $\bar{P}_e(\mathrm{SNR})$ as a function of SNR that behaves as

$$\lim_{\mathrm{SNR}\to\infty} \frac{\log(\bar{P}_e(\mathrm{SNR}))}{\log(\mathrm{SNR})} = -d \tag{27}$$

is said to have a diversity order of d.

■

In words, a scheme with diversity order d has an error probability at high SNR behaving as $\bar{P}_e(\mathrm{SNR}) \approx \mathrm{SNR}^{-d}$. Given this definition, we see that the diversity order in (26) is at most qM_r. Moreover, in (26) we notice that we also obtain a coding gain of $(\prod_{n=1}^{q} \lambda_n)^{1/q}$.

Note that in order to obtain the average error probability, one can calculate a naive union bound using the pairwise error probability given in (26). However, this bound may not be tight and a more careful upper bound for the error probability can be derived [231], [291]. However, if we ensure that *every* pair of codewords satisfies the diversity order in (26), then clearly the average error probability satisfies it as well. This is true when the transmission rate is held constant with respect to SNR. Therefore, code design for diversity order through pairwise error probability is a sufficient condition, although more detailed criteria can be derived based on a more accurate expression for average error probability.

The error probability analysis can easily be extended to the case where we have quasi-static ISI channels with channel taps modeled as i.i.d. zero-mean complex Gaussian random variables (see, for example, [292] and references therein). In this case, the PEP can be written as

$$P(\mathbf{x}\to\mathbf{e}) \le \left[\frac{1}{\prod_{n=1}^{M_t\nu}\left(1+\frac{E_s}{4N_0}\tilde{\lambda}_n\right)}\right]^{M_r} \tag{28}$$

where the eigenvalues $\tilde{\lambda}_n$ are those of $\tilde{\mathbf{A}}(\mathbf{x},\mathbf{e}) = \tilde{\mathbf{B}}^*(\mathbf{x},\mathbf{e})\tilde{\mathbf{B}}(\mathbf{x},\mathbf{e})$

$$\tilde{\mathbf{B}}(\mathbf{x},\mathbf{e}) = \begin{pmatrix} \tilde{\mathbf{x}}^T(0) - \tilde{\mathbf{e}}^T(0) \\ \vdots \\ \tilde{\mathbf{x}}^T(N-1) - \tilde{\mathbf{e}}^T(N-1) \end{pmatrix} \tag{29}$$

and

$$\tilde{\mathbf{x}}(k) = [\mathbf{x}^T(k),\ldots,\mathbf{x}^T(k-\nu)]^T. \tag{30}$$

Since $\tilde{\mathbf{A}}(\mathbf{x},\mathbf{e})$ is a square matrix of size $M_t\nu$, clearly the maximal diversity order achievable for quasi-static ISI channels is $M_rM_t\nu$.

Finally, if we have a time-varying ISI channel, we can generalize (28) to

$$P(\mathbf{x}\to\mathbf{e}) \le \left[\frac{1}{\left|\mathbf{I}_{M_rNM_t\nu} + \frac{E_s}{4N_0}\mathbf{F}(\mathbf{R}_h \otimes \mathbf{I}_{M_rM_t\nu})\right|}\right] \tag{31}$$

where $\otimes$ denotes a Kronecker product, $\mathbf{R}_h$ is the $N \times N$ correlation matrix of the channel tap process, and $\mathbf{F} = \mathrm{diag}\{\mathbf{C}(0),\ldots,\mathbf{C}(N-1)\}$ with

$$\mathbf{C}(k) = \left[\tilde{\mathbf{x}}^T(k) - \tilde{\mathbf{e}}^T(k)\right] \otimes \mathbf{I}_{M_r}. \tag{32}$$

Again, it is clear that the maximal diversity attainable is $M_rM_t\nu N$, but for a given channel tap process, N is replaced by the number of dominant eigenvalues N_{dom} of the fading correlation matrix. This parameter is related to the Doppler spread of the channel and the block duration.

A natural question that arises is, how many codewords can we have which allow us to attain a certain diversity order? For a flat Rayleigh fading channel, this has been examined in [242] and the following result was obtained.[23]

Theorem 3.9: If we use a constellation of size 2^b and the diversity order of the system is qM_r, then the rate R that can be achieved is bounded as

$$R \le \frac{1}{N}\log\left[A_{2^{bN}}(M_t,q)\right] \tag{33}$$

where $A_{2^{bN}}(M_t,q)$ is the maximum size of a code of length M_t with minimum Hamming distance q defined over an alphabet size 2^{bN}.

■

One consequence of this result is that for maximum (M_tM_r) diversity order we can transmit at most b b/s/Hz. This result can be easily extended to the quasi-static ISI channel, where the maximal diversity order is $M_tM_r\nu$. Here too the maximal rate that can be transmitted with diversity order $M_tM_r\nu$ is b b/s/Hz. Therefore, the ISI channel provides a higher diversity order, but not a higher transmission rate at this level of diversity.

An alternate viewpoint in terms of the rate-diversity tradeoff has been explored in [291] from a Shannon-theoretic point of view. Here the authors are interested in the multiplexing gain of a transmission scheme.

Definition 3.4: A coding scheme which has a transmission rate of $R(\mathrm{SNR})$ as a function of SNR is said to have a multiplexing gain r if

$$\lim_{\mathrm{SNR}\to\infty} \frac{R(\mathrm{SNR})}{\log(\mathrm{SNR})} = r. \tag{34}$$

Therefore, the system has a rate of $r\log(\mathrm{SNR})$ at high SNR. One way to contrast it with the statement in Theorem 3.9 is that the constellation size is also allowed to become larger with SNR. However, note that the naive union bound of the pairwise error probability (25) has to be used with care if the constellation size is also increasing with SNR. There is a tradeoff between the achievable diversity and the multiplexing gain, and $d^*(r)$ is defined as the supremum of the diversity gain achievable by *any* scheme with multiplexing gain r. The main result in [291] states the following.

[23]A constellation size refers to the alphabet size of each transmitted symbol. For example, a QPSK modulated transmission has constellation size of 4.

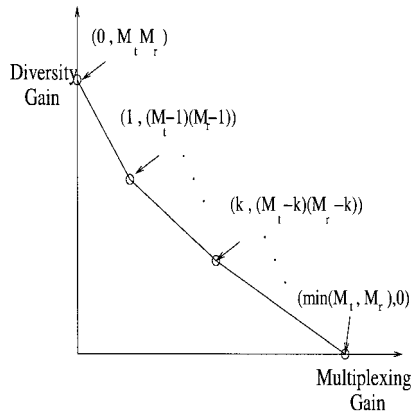

Fig. 7. Typical rate-diversity tradeoff curve.

Theorem 3.10: For $N > M_t + M_r - 1$, *and* $K = \min(M_t, M_r)$, *the optimal tradeoff curve* $d^*(r)$ *is given by the piecewise linear function connecting points in* $(k, d^*(k))$, $k = 0, \ldots, K$ *where*

$$d^*(k) = (M_r - k)(M_t - k). \tag{35}$$

■

If $r = k$ is an integer, the result can be interpretted as using $M_r - k$ receive antennas and $M_t - k$ tranmit antennas to provide diversity while using k antennas to provide the multiplexing gain. Clearly this result means that one can get large rates which grow with SNR if we reduce the diversity order from the maximum achievable. This diversity-multiplexing tradeoff implies that a high multiplexing gain comes at the price of decreased diversity gain and is a manifestation of a corresponding tradeoff between error probability and rate. This tradeoff is depicted in Fig. 7. Therefore, as illustrated in Theorems 3.9–3.10, the tradeoff between diversity and rate is an important consideration both in terms of coding techniques (Theorem 3.9) and in terms of Shannon theory (Theorem 3.10). We revisit this question of practical techniques in terms of this tradeoff in Section IV-A.4.

D. Design Considerations

The most important lesson we can learn from the theoretical results of Sections III-A–III-C is the significant impact of spatial diversity on wireless communication. The impact is both through increased transmission rates and/or increased reliability (lower error probability) in transmissions. Moreover, the theoretical results give us insight into efficient techniques that can achieve near-optimal performance. The purpose of this section is to bring together some of the insights gained from the theoretical considerations and also to discuss some of the many open questions in this rapidly evolving research area.

As discussed in Section III-A, in order to achieve the rate given in (10), we need to code *across* fading blocks to achieve the ergodic capacity. In a communication system, delay is an important parameter and, hence, we would be interested in maximizing performance for smaller coding blocks. In this case, the Shannon-theoretic capacity is not meaningful and the concept of outage is more relevant. Moreover, the utility of diversity order becomes apparent in the nonergodic scenario.

In order to design practical codes that achieve a performance target, we need to glean insights from the analysis to state design criteria. For example, in the flat-fading case of (26) we can state the following rank and determinant design criteria [128], [242].

Design Criteria for Space–Time Codes Over Flat-Fading Channels:

- *Rank criterion:* In order to achieve maximum diversity $M_t M_r$, the matrix $\mathbf{B}(\mathbf{x}, \mathbf{e})$ from (24) has to be full rank for any codewords $\mathbf{x}, \mathbf{e}$. If the minimum rank of $\mathbf{B}(\mathbf{x}, \mathbf{e})$ over all pairs of distinct codewords is q, then a diversity order of qM_r is achieved.
- *Determinant criterion:* For a given diversity order target of q, maximize $(\prod_{n=1}^{q} \lambda_n)^{1/q}$ over all pairs of distinct codewords.

A similar set of design criteria can be stated for the quasi-static ISI fading channel using the PEP given in (28) and the corresponding error matrix given in (29). Therefore, if we need to construct codes satisfying these design criteria, we can guarantee performance in terms of diversity order. The main problem in practice is to construct such codes which do not have large decoding complexity. This sets up a familiar tension on the design in terms of satisfying the performance requirements and having low-complexity decoding. In Section IV, we outline constructions that explore this tension and are motivated by the theoretical considerations outlined above.

If coherent detection is difficult or too costly, one can employ noncoherent detection for the multiple-antenna channel [141], [142], [290]. Though it is demonstrated in [290] that a training-based technique achieves the same capacity–SNR slope as the optimal, there might be a situation where inexpensive receivers are needed because channel estimation cannot be accommodated. In such a case, differential techniques which satisfy the diversity order might be desirable. There has been significant recent work on differential transmission with noncoherent detection (see, for example, [140], [145], and references therein) and this is a topic we discuss in Section IV-B.2.

Link to Applications: Contrasting approaches of maximizing rate and maximizing reliability were brought together in Section III-C and this viewpoint becomes important in the context of which applications are needed on the wireless network. The outage probability is related to the diversity order of the system and we examine the effect of outage probability on the rate (see Figs. 8–10) for a flat power-delay profile on the CIR taps which are modeled as Rayleigh fading variables. Note that Fig. 8 illustrates the linear growth rate of achievable rate with respect to number of transmit and receive antennas (even for the frequency-selective case), similar to that shown in the flat-fading case in Theorems 3.1 and 3.2. Furthermore Fig. 8 quantifies the increase in spectral efficiency due to channel memory ν. From Theorems 3.9 and 3.10, diversity order plays a role in ensuring transmission reliability at the cost of reduced transmission rate. The question of which is

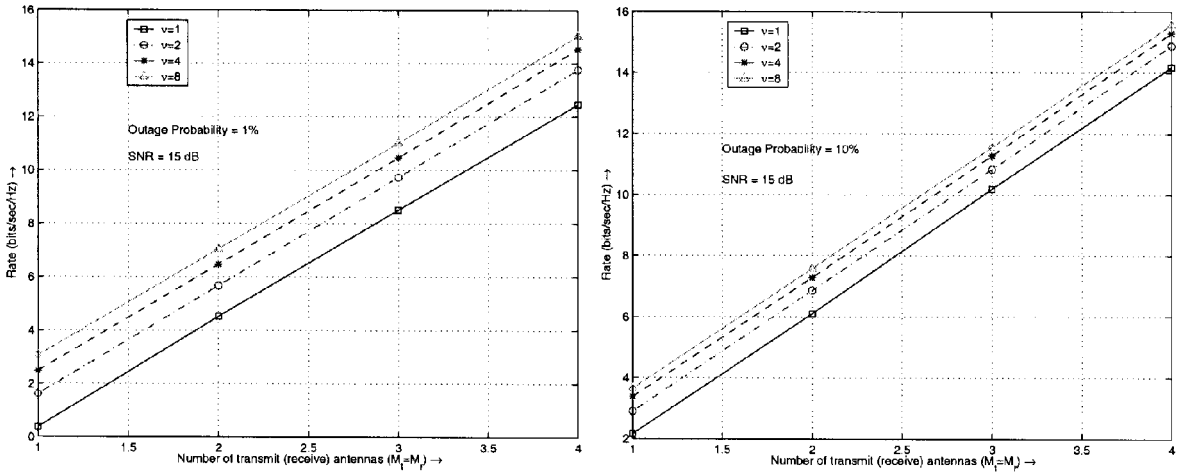

Fig. 8. Rates versus number of transmit (receive) antennas for outage of 1% and 10% and channel memory ν.

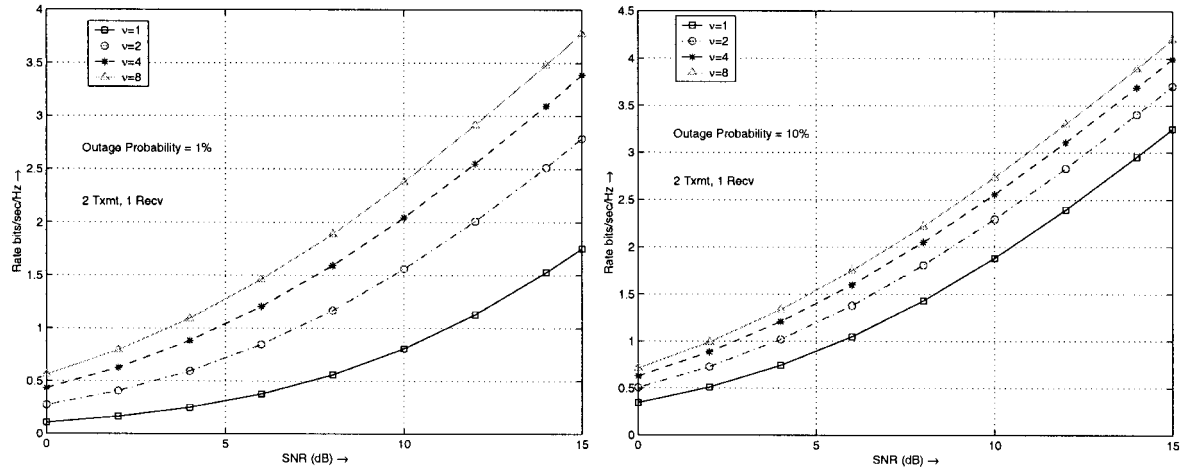

Fig. 9. Rates for outage probabilities of 1% and 10% for $M_t = 2$, $M_r = 1$, and various channel memory ν.

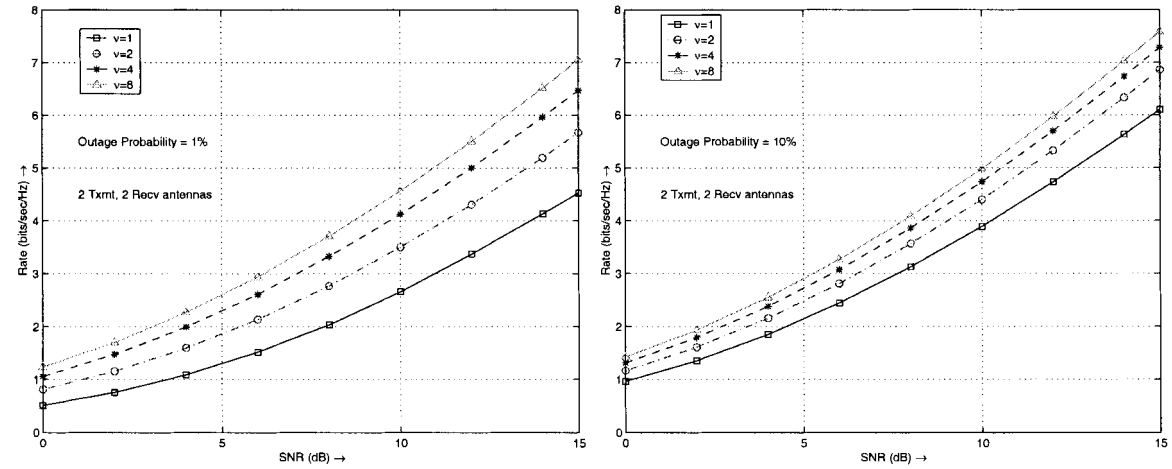

Fig. 10. Rates for outage probabilities of 1% and 10% for $M_t = 2$, $M_r = 2$, and various channel memory ν.

the more important design issue is dependent on the applications. In delay-constrained systems (for example, real-time traffic such as speech and video), it is more important to have larger diversity order (increased reliability) compared to rate. In delay-tolerant applications, larger rates might be a better operating point, and one might even be able to code across blocks (or retransmit lost packets) for improving performance at the cost of delay.

Connecting to the thread example introduced in Section II, we can now examine what the theoretical developments predict about these scenarios. For a SNR of 15 dB, Fig. 9 shows that a spectral efficiency of 3.78 b/s/Hz at 1% outage (and 4.2 b/s/Hz at 10% outage) is achievable for $\nu = 8$, as used in Scenario A. This implies that a single user in Scenario A can get (approximately) a data rate of 75 Mb/s at 1% outage (and 84Mb/s at 10% outage) using two transmit and one receive antenna. On the other hand, a single user in Scenario B with two transmit and one receive antenna can achieve a spectral efficiency of 2.7 b/s/Hz at 1% outage (and 3.4 b/s/Hz at 10% outage) for $\nu = 4$. This translates to (approximately) a data rate of 3.4 Mb/s at 1% outage (and 4.2 Mb/s at 10% outage) after taking into account a SNR penalty due to channel estimation errors (see Section IV-B). With two transmit and two receive antennas (see Fig. 10), for Scenario A we can get 7.1 b/s/Hz at 1% outage (and 7.6 b/s/Hz at 10% outage); similarly for Scenario B we can achieve 5.5 b/s/Hz at 1% outage (and 6.2 b/s/Hz at 10% outage) after the SNR penalty for channel estimation. With $M_t = M_r = 3$, we can achieve 11 b/s/Hz at a SNR of 15 dB and $\nu = 8$. Since for a given spectral efficiency, the rate achievable scales linearly with available transmission bandwidth, in this case, a rate of 1.1 Gb/s is possible for a 100-MHz transmission bandwith. In a multiuser environment, the spectrum has to be shared between users, and clearly the per-user rate will drop. Multiuser diversity increases the overall throughput but this needs to be divided by the number of users to reflect the throughput obtained per-user.

Multiuser Diversity: Another important idea that we discussed in Section III-B was a form of spatial diversity through multiple users, i.e., multiuser diversity. Here, the independent channel conditions among different users can be exploited by realizing that the wireless medium is a shared medium in contrast to a point-to-point channel as is the case in wireline communication. This idea motivates opportunistic resource allocation algorithms which attempt to maximize utilization of the wireless channel and, at the same time, attempt to guarantee some form of fairness between users. This leads us to an exploration of what are fairness criteria for sharing resources between multiple users, and what algorithms can be developed to optimize these criteria. This is a prelude to a more detailed discussion of such issues in Section V-A.3. Also, in ad hoc networks, multiuser diversity can be utilized by using communication nodes as relays for transmitting information from source to destination. From the results in Theorems 3.6 and 3.7 we have seen that the use of relays can substantially impact the long-term per-user throughput. We will revisit this again in Section V-B for ad hoc wireless networks.

The theoretical developments also indicate the strong interactions between the physical layer coding schemes and channel conditions with the networking issues of resource allocation and application design. This is another important insight we can draw in the design of wireless networks. Therefore, several problems which are traditionally considered as networking issues and are typically designed independent of the transmission techniques need to be reexamined in the context of wireless networks. As illustrated, spatial diversity needs to be taken into account while solving these problems. Such an integrated approach is another major lesson learnt from the theoretical considerations and we develop this topic in Section V.

Open Questions: Though recent research activity has shed considerable light on the theoretical underpinnings of modern wireless communication, this is an evolving research topic. We conclude this section by outlining some open questions whose answers would have significant impact on the understanding of next-generation multiuser wireless networks.

- **Single-User Shannon-theoretic Problems:** Though the capacity of the MIMO flat-fading channel with i.i.d. complex Gaussian fading channel response and channel side-information at the receiver has been established (see Theorem 3.1), the capacity of MIMO channels for various cases is unknown. This is true both for the coherent case, with correlated fading coefficients, and for the noncoherent case, where there is no receiver side information. There is some characterization of optimal codebook designs for the noncoherent case [141], [290], but the complete answer is still open.

  For frequency-selective channels, there is a larger number of unanswered questions. For example, the capacity of time-varying fading ISI channels is completely unknown. Some characterizations exist for the case when the channel is block time-invariant [48], [197], however, here too when the power-delay profile of the channel is not flat, the capacity is unknown. For the continuously varying fading ISI channel, there is no clear characterization of the capacity. Here some characterizations exist when the channel state information of the time-varying ISI response is known *both* at the transmitter and the receiver [183]. However, if we have fast time-varying channels, perhaps the assumption that the channel state information is fed back instantaneously to the transmitter might be a strong one. When the channel varies within a transmission block and when the transmitter has no channel state information available, even the optimal transmission basis functions are unknown. For example, the Fourier basis is not the optimal basis function for continuously varying channels and, therefore, OFDM is *not* asymptotically optimal, as was the case with time-invariant ISI channels. Some lower bounds on the capacity of such channels have been developed [84] but the capacity of this channel is still unknown.
- **Multiuser Problems:** As mentioned earlier, the characterization of rate regions for multiuser communication channels is full of open questions. The multiantenna broadcast channel problem is one that has generated significant interest and where some recent progress has been made [49], [269], [272], [289]. Here, an interesting coding technique comes into play which is loosely termed "writing on dirty paper." The technique was motivated by a Shannon-theoretic result

on capacity of Gaussian channels with (noncausal) transmitter side-information about the noise process [67]. The information streams for the different users in the broadcast channels act as "interference" to each other, but are known at the transmitter. Therefore, the coding technique is to use the result of [67] to code such that the interference is seemingly absent at the desired receiver. This is *not* done by cancelling the interference, but by riding on it [49], [67]. This idea is opening interesting new questions about coding for broadcast channels, though the general rate region for this problem is still unknown. The general rate region for the MIMO multiple-access channel is also not completely answered. When only the receiver has multiple antennas, the rate region is well understood [41], [271]. However, when the nodes also have multiple antennas, since there is no common basis that diagonalizes the channels (as is the case in ISI channels [56]), the capacity region has not been characterized completely. There is a vast number of open problems in many-to-many communication, and perhaps the most interesting is the relay channel mentioned in Section III-B. The relay channel is important in ad hoc wireless networks where relays are used to transmit information between source and destination.

These problems are a sampling of interesting open questions, and perhaps there are also important new problem formulations that have not been asked yet. There is also a large number of open questions related to communication under complexity and delay constraints. In summary, these topics promise to keep a generation of researchers busy.

IV. Signal Transmission Issues

The previous section dealt with information-theoretic issues related to spatial diversity including maximum achievable rates and space–time code design criteria. Several ideal assumptions were often made such as the availability of perfect CSI and the implementation of ML decoding at the receiver. This section deals with physical layer techniques for MIMO broadband wireless channels under practical conditions. The main focus of this section is on signal processing algorithms in space–time transceivers. We start in Section IV-A by examining transmitter techniques that improve the throughput and/or reliability of wireless communication over broadband MIMO channels. These techniques include spatial multiplexing (which improves throughput), transmit diversity (which improves reliability) including space–time trellis and block codes, and OFDM, which mitigates the channel's frequency selectivity. In Section IV-B, we describe in some detail two classes of receiver techniques for joint equalization and decoding of space–time-coded transmissions. The first class is coherent techniques that require channel estimation. We consider both quasi-static (as in Scenario A) and rapidly varying channels (as in Scenario B) for this class of techniques. The coherent class includes adaptive techniques that do not explicitly estimate the channel but rather learn and track its characteristics using training symbols and previous decisions. The second class of techniques are noncoherent where no channel estimation is required (hence, eliminating its overhead and complexity) at the expense of a 3 dB SNR loss from coherent techniques in the quasi-static case. These noncoherent techniques become especially attractive for rapidly varying channels (in Scenario B) where accurate channel estimation becomes very challenging.

We would like to emphasize the following subtle point about our CSI assumption at the transmitter. Recall from Section II that channel fading occurs at two time scales: fast (Rayleigh) fading due to constructive/destructive addition of multipath which results in rapid signal fluctuations and slow fading (log-normal shadowing) due to large structures in the signal transmission path. While knowledge of the fast fading component at the transmitter is not possible in practice (and, hence, is not assumed in this paper), it is possible to track and feedback the slow fading component to the transmitter to be used in scheduling transmissions (cf. Section V-A.3).

A. Transmitter Techniques

In this section, we describe transmitter techniques that enhance the performance of multiple-transmit-antenna systems either by increasing their bit rate through spatial multiplexing or by decreasing their average error rate through spatial diversity. We briefly discuss the fundamental tradeoff in achieving these two performance objectives. Other forms of transmit diversity are also described and compared. Finally, an effective modulation technique known as OFDM suitable for broadband channels is described and its pros and cons delineated.

1) Spatial Multiplexing [Bell Labs Layered Space–Time Architecture (BLAST)]: Communication theory suggests the following four guidelines for increasing the bit rate.

- *Increase the symbol rate* (or equivalently transmission bandwidth): the price paid is increased susceptibility to ISI and increased requirements of processing speed. Another limitation is the scarcity and high cost of radio frequency (RF) spectrum.
- *Increase signal constellation size* (or equivalently the transmission spectral efficiency): the price paid is reduced noise immunity (assuming fixed transmitted power) during bad channel conditions (low SNR due to fading) and increased susceptibility to synchronization errors (carrier frequency offsets and phase noise). Error correction coding and adaptive modulation techniques can be used to improve the link margin.
- *Increase transmitted power*: this is undesirable in wireless transmission because it could increase nonlinear distortion in power amplifiers, reduces battery life at the subscriber terminal, and increases cochannel interference.
- *Use of multiple transmit and/or receive antennas* to increase *throughput* (by using spatial multiplexing to create multiple parallel channels for transmission of independent information streams). The price paid is the

cost of implementing multiple RF chains and increased space requirements.

As an example, the 3G TDMA cellular standard known as *Enhanced Data Rates for Global Evolution* (EDGE) [110] achieves a significant bit rate increase over its 2G predecessors IS-136 and GSM by: 1) increasing the symbol rate from 24 Ksymbols/s in IS-136 to 271 ksymbols/s and 2) increasing the signal constellation size from DQPSK in IS-136 and (binary) GMSK in GSM to 8 PSK.

Using the information-theoretic results of Section III, we can see that by adding M_t transmit and M_r receive antennas, the achievable rate is further multiplied by $\min(M_t, M_r)$. A major challenge in realizing this significant additional throughput gain in practice is the development of cost-effective integrated multiple-chain RF implementation and of low-complexity MIMO receiver architectures. An example of the latter is BLAST [104], [105], where the multiple transmitted data streams are separated and detected successively using a combination of array processing (nulling) and multiuser detection (interference cancellation) techniques.

In its most basic form known as vertical (V)-BLAST, the receiver signal processing functions are similar to a decision feedback equalizer (DFE) operating in the spatial domain where the nulling operation is performed by the "feedforward filter" and the interference cancellation operation is performed by the "feedback filter" [59]. As in all feedback-based detection schemes, V-BLAST suffers from *error propagation* effects which are further exacerbated by the fact that the first detected stream (typically chosen to be the one with the highest detection SNR) enjoys the smallest diversity order of $(M_r - M_t + 1)$, since $(M_t - 1)$ spatial degrees of freedom are used to null interference from the other $(M_t - 1)$ data streams yet to be detected.

Several techniques have been proposed to enhance the performance of V-BLAST, such as using minimum mean square error (MMSE) interference cancellation instead of nulling, assigning variable data rates to the various streams where the first detected streams that pass through less reliable channels (smaller diversity order) are assigned a smaller data rate than streams passing through more reliable channels, and performing ML detection on the first few unreliable streams [59] or reduced-complexity ML detection for all streams using the sphere decoder algorithm [76],[24] (see Section IV-B.1 for more details), combining BLAST with soft-information-based iterative "turbo" detection algorithms [18], [224], and threaded approach to multiplexing [97].

We conclude with the following two remarks. First, the BLAST architecture has been extended to the broadband channel scenario using a MIMO generalization of the classical DFE [9], [176]. Second, the presence of antenna correlation and the lack of scattering richness in the propagation environment reduce the achievable rates of spatial multiplexing techniques from their theoretical projections under ideal assumptions [58], [117]. Nevertheless, recent experimental results show that a substantial fraction of these theoretical rates is still achievable under practical propagation scenarios [113].

2) Transmit Diversity Techniques: Fading is a major performance-limiting impairment on wireless channels that arises mainly from destructive addition of multipaths in the propagation medium. Diversity techniques mitigate fading by transmitting multiple correlated replicas of the same information signal through independently fading channel realizations that are much less likely to fade simultaneously than each individually. Diversity techniques can be classified according to the domain in which they are created into three main categories.

- *Temporal* diversity techniques that utilize channel coding and time interleaving to mitigate fading at the expense of added delay and bandwidth efficiency loss. This excludes their use on slow-fading channels and delay-sensitive applications.
- *Frequency* (or multipath) diversity, which is available for broadband signaling and can be realized using an ML equalizer to collect signal energy from multiple propagation paths. This form of diversity is not available in narrowband signaling.
- *Spatial* diversity, where multiple antennas are used at the transmitter and/or receiver to provide multiple independently fading paths[25] for the transmitted signals that carry the same information (spatial redundancy) [187]. Spatial diversity techniques provide significant performance gains without sacrificing precious bandwidth or transmit power resources.

Spatial diversity techniques fall under two main categories: transmit diversity and receive diversity. Receive diversity combines multiple independently received signals (corresponding to the same transmitted signal), can utilize CSI available at the receiver, and is more suitable for the uplink (since it is more cost effective to implement multiple antennas at the base station than at the terminal). Transmit diversity is more challenging to provision and realize because it involves the design of multiple correlated signals from a single information signal without utilizing CSI (typically not available accurately at the transmitter end). Furthermore, transmit diversity must be coupled with effective receiver signal processing techniques that can extract the desired information signal from the distorted and noisy received signal. Transmit diversity is more practical than receive diversity for enhancing the downlink (which is the bottleneck in broadband asymmetric applications such as Internet browsing and downloading) to preserve the small size and low power consumption features of the user terminal. A common attribute of transmit and receive diversity is that both experience "diminishing returns" (with respect to an error probability criterion) as the number of antennas increases [242], i.e., with respect to the SNR gain for a given probability of error. In more detail, since the slope (diversity) of the error probability curve increases with the number of antennas, for same increase in the number of antennas the SNR gain for a given error probability diminishes

[24]Using ML detection techniques also allows us to relax the assumption $M_t \geq M_r$ needed in the traditional BLAST algorithm [105].

[25]This can be ensured by placing the antennas sufficiently apart (more than the coherence distance).

quite rapidly. Therefore, from a performance-complexity tradeoff point of view, it is effective to use small numbers of antennas (typically less than four). This is in contrast with the point of view of increasing the rate as done through spatial multiplexing gains where the rate multiplexing gains continue to increase linearly with the number of antennas (assumed equal at both ends).

There are two main classes of multiple-antenna transmitter techniques: closed-loop and open-loop. The former uses a feedback channel to send CSI acquired at the receiver back to the transmitter to be used in signal design while the latter does not require CSI. Assuming availability of ideal (i.e., error free and instantaneous) CSI at the transmitter, closed-loop techniques have an SNR advantage of $10\log_{10}(M_t)$ dB over open-loop techniques due to the "array gain" factor [14]. However, several practical factors degrade the performance of closed-loop techniques including channel estimation errors at the receiver, errors in feedback link (due to noise, interference, and quantization effects), and feedback delay which causes a mismatch between available and actual CSI. All of these factors combined with the extra bandwidth and system complexity resources needed for the feedback link make open-loop techniques more attractive as a robust means for improving downlink performance for high-mobility applications (such as Scenario B) while closed-loop techniques (such as beamforming) become attractive under low-mobility conditions (such as Scenario A).[26] Our focus in this section will be exclusively on open-loop spatial transmit diversity techniques due to their applicability to both scenarios.[27] Beamforming techniques are discussed extensively in several tutorial papers, such as [120], [121], and [267].

The simplest example of open-loop spatial transmit diversity techniques is *delay diversity* [257], [282] where the signal transmitted at sampling instant k from the ith antenna is $x_i(k) = x(k - l_i)$ for $2 \leq i \leq M_t$ and $x_1(k) = x(k)$ where l_i denotes the time delay (in symbol periods) on the ith transmit antenna (with $l_i \neq l_j$ for $i \neq j$). Assuming a single receive antenna, the D-transform[28] of the received signal is given by

$$y(D) = x(D)\left(h_1(D) + \sum_{i=2}^{M_t} D^{l_i} h_i(D)\right) + z(D). \quad (36)$$

It is clear from (36) that delay diversity transforms spatial diversity into multipath diversity that can be realized through equalization [225]. For flat-fading channels, we can set $l_i = (i-1)$ and achieve full (i.e., order-M_t) spatial diversity using an ML equalizer with $(2^b)^{M_t-1}$ states where 2^b is the input alphabet size. However, for frequency-selective channels, a delay of at least $l_i = (i-1)(\nu+1)$ is needed to ensure that coefficients from the various spatial FIR channels do not interfere with each other causing a diversity loss. This, in turn, increases equalizer complexity to $(2^b)^{(M_t-1)(\nu+1)}$ states which is prohibitive even for moderate b, M_t, and ν. In the next subsection, we describe another family of open-loop spatial transmit diversity techniques known as space–time block codes that achieve full spatial diversity with practical complexity even for frequency-selective channels with long delay spread.

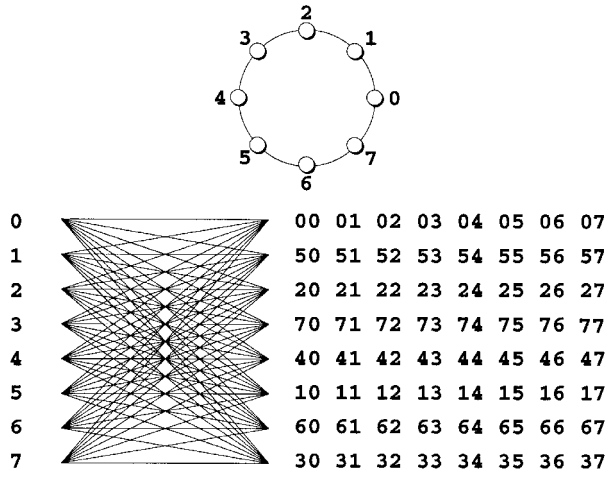

Fig. 11. Eight-state 8-PSK space–time trellis code with two transmit antennas and a spectral efficiency of 3 b/s/Hz.

3) Space–Time Coding: STC has received considerable attention in academic and industrial circles [11], [12] due to its many advantages. First, it improves the downlink performance without the need for multiple receive antennas at the terminals. For example, for WCDMA, STC techniques where shown in [199] to result in substantial capacity gains due to the resulting "smoother" fading which, in turn, makes power control more effective and reduces the transmitted power. Second, it can be easily combined with channel coding, as shown in [242], realizing a coding gain in addition to the spatial diversity gain. Third, they do not require CSI at the transmitter, i.e., operate in open-loop mode, thus eliminating the need for an expensive and, in case of rapid channel fading (such as Scenario B), unreliable reverse link. Finally, they have been shown to be robust against nonideal operating conditions such as antenna correlation, channel estimation errors, and Doppler effects [189], [241]. There has been extensive work on the design of space–time codes since its introduction in [242]. The combination of the turbo principle [35], [37] with space–time codes has been explored (see, for example, [31], [174] among several other references). Also the application of low-density parity check (LDPC) codes [111] to STC has been explored (see, for example, [177] and references therein). We focus our discussion on the basic principles of space–time codes and describe two main flavors: trellis and block codes, as described next.

Space–Time Trellis Codes: The space–time trellis encoder maps the information bit stream into M_t streams of symbols (each belonging in a size-2^b signal constellation) that are

[26]Channel state knowledge at the transmitter can be reasonable when the transmission and reception take place at the same carrier frequency as in time division duplexing (TDD) [213], [237]. However, even in frequency division duplex (FDD) for two-way communication systems, statistical knowledge of the channel could be used to enhance performance [211], [214].

[27]It is also possible to combine closed-loop and open-loop techniques as shown recently in [152], [232].

[28]The D-transform of a discrete-time sequence $\{x(k)\}_{k=0}^{N-1}$ is defined as $x(D) = \sum_{k=0}^{N-1} x(k)D^k$. It is derived from the Z-transform by replacing the unit delay Z^{-1} by D.

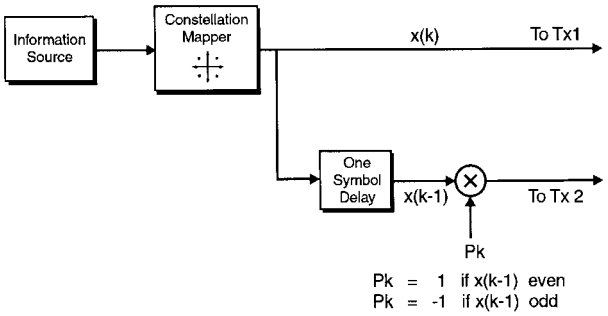

Fig. 12. Equivalent encoder model for eight-state 8-PSK space–time trellis code with two transmit antennas.

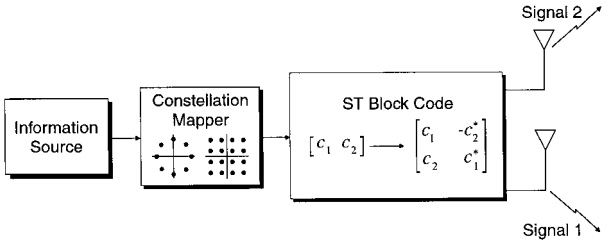

Fig. 13. Spatial transmit diversity with Alamouti's space–time block code.

transmitted simultaneously.[29] STTC design criteria are based on minimizing the PEP bound in Section III-C.

As an example, we consider the eight-state 8-PSK STTC for two transmit antennas introduced in [242] whose trellis description is given in Fig. 11, where the edge label c_1c_2 means that symbol c_1 is transmitted from the first antenna and symbol c_2 from the second antenna. The different symbol pairs in a given row label the transitions out of a given state, in order, from top to bottom. An equivalent and convenient (for reasons to become clear shortly) implementation of the eight-state 8-PSK STTC encoder is depicted in Fig. 12. This equivalent implementation clearly shows that the eight-state 8-PSK STTC is identical to classical delay diversity transmission [225] *except* that the delayed symbol from the second antenna is multiplied by -1 if it is an odd symbol, i.e., $\in \{1,3,5,7\}$. This slight modification results in additional coding gain over a flat-fading channel. We emphasize that this STTC does not achieve the maximum possible diversity gains (spatial and multipath) on a frequency-selective channel; however, its performance is near optimum for practical ranges of SNR on wireless links [107].[30] Furthermore, when implementing the eight-state 8-PSK STTC described above over a frequency-selective channel, its structure can be exploited to reduce the complexity of joint equalization and decoding. This is achieved by embedding the space–time encoder in Fig. 12 in the two channels $h_1(D)$ and $h_2(D)$ resulting in an equivalent single-input single-output (SISO) data-dependent CIR with memory $(\nu+1)$ whose D-transform is given by

$$h_{\text{eqv}}^{\text{STTC}}(k,D) = h_1(D) + p_k D h_2(D) \tag{37}$$

where $p_k = \pm 1$ is data dependent. Therefore, trellis-based joint space–time equalization and decoding with $8^{\nu+1}$ states can be performed on this equivalent channel. Without exploiting the STTC structure, trellis equalization requires $8^{2\nu}$ states and STTC decoding requires 8 states.

Several other full-rate full-diversity STTCs for different signal constellations and different numbers of antennas were introduced in [242]. This discussion just gives the flavor of the space–time trellis code designs, extensions of the constructions given in [242] can be found, for example, in [29] and [132].

Space–Time Block Codes: The decoding complexity of STTC (measured by the number of trellis states at the decoder) increases *exponentially* as a function of the diversity level and transmission rate [242]. In addressing the issue of decoding complexity, Alamouti [14] discovered an ingenious space–time block coding scheme for transmission with two antennas. According to this scheme, input symbols are grouped in pairs where symbols x_k and x_{k+1} are transmitted at time k from the first and second antennas, respectively. Then, at time $k+1$, symbol $-\bar{x}_{k+1}$ is transmitted from the first antenna and symbol $\bar{x}_k$ is transmitted from the second antenna, where $\bar{(.)}$ denotes complex conjugation (cf. Fig. 13). This imposes an orthogonal spatio-temporal structure on the transmitted symbols. Alamouti's STBC has been adopted in several wireless standards such as WCDMA [247] and CDMA2000 [248] due to its many attractive features including the following.

- It achieves full diversity at full transmission rate for any (real or complex) signal constellation.
- It does not require CSI at the transmitter (i.e., open loop).
- ML decoding involves only *linear* processing at the receiver (due to the orthogonal code structure).

The main drawbacks of Alamouti's STBC are the following.

- Unlike space–time trellis codes, it does not provide any coding gain.
- A rate-1 STBC cannot be constructed, in general, for any complex signal constellation with more than two transmit antennas [240].
- The simple decoding rule is valid only for a flat-fading channel where the channel gain is constant over two consecutive symbols.

The Alamouti STBC has been extended to the case of more than two transmit antennas [239] using the theory of orthogonal designs. There it was shown that, in general, full-rate orthogonal designs exist for all real constellations for two, four, or eight transmit antennas only while for all complex constellations they exist only for two transmit antennas (the Alamouti scheme). However, for particular constellations, it might be possible to construct orthogonal designs for other cases. Moreover, if a rate loss is acceptable, orthogonal designs exist for an arbitrary number of transmit antennas [239]. Further details of orthogonal designs are given in the Appendix.

Recently, STBCs have been extended to the frequency-selective channel case by implementing the Alamouti orthog-

[29]The total transmitted power is divided equally among the M_t transmit antennas.

[30]For examples of STTC designs for frequency-selective channels see, e.g., [173], among others.

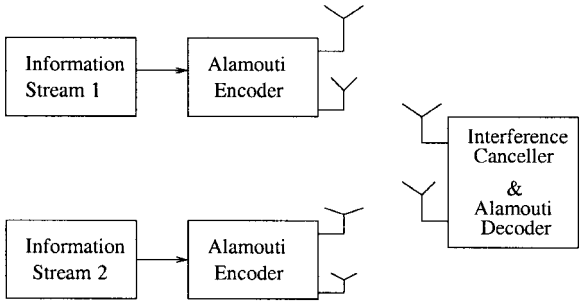

Fig. 14. An interference cancellation scheme that achieves rate 2 and diversity order 2 with 2 TX and 2 RX.

onal signaling scheme at a *block* level instead of *symbol* level. Depending on whether the implementation is done in the time or frequency domain, three STBC structures for frequency-selective channels have been proposed: TR-STBC [172], OFDM-STBC [175], and FDE-STBC [7]. As an illustration, we present next the space–time encoding scheme for FDE-STBC. Denote the nth symbol of the kth transmitted block from antenna i by $x_i^{(k)}$. At blocks $k = 0, 2, 4, \ldots$ pairs of length-N blocks $\mathbf{x}_1^{(k)}(n)$ and $\mathbf{x}_2^{(k)}(n)$ (for $0 \leq n \leq N-1$) are generated by the mobile user. Inspired by Alamouti STBC, we encode the information symbols as follows [7]:

$$\mathbf{x}_1^{(k+1)}(n) = -\mathbf{x}_2^{*(k)}((-n)_N) \text{ and } \mathbf{x}_2^{(k+1)}(n) = \mathbf{x}_1^{*(k)}((-n)_N) \quad \text{for } n = 0, 1, \ldots, N-1 \text{ and } k = 0, 2, 4, \ldots \tag{38}$$

where $(\cdot)_N$ denotes the modulo-N operation. In addition, a cyclic prefix (CP) of length ν (the maximum order of the FIR wireless channel) is added to each transmitted block to eliminate interblock interference (IBI) and make all channel matrices *circulant*. We refer the reader to [8] for a detailed description and comparison of these schemes. The main point we would like to stress here is that these three STBC schemes realize both spatial and multipath diversity gains at practical complexity levels. For channels with long delay spread, the frequency-domain implementation using the FFT either in a single-carrier or multicarrier fashion becomes more advantageous from a complexity point of view.

4) Diversity Versus Throughput Tradeoff: In the previous sections, we have seen two types of gains from MIMO transmitter techniques: spatial multiplexing gains and diversity gains. As discussed in Section III-C, Theorem 3.10 shows that one can simultaneously achieve diversity gains as well as multiplexing gains. In this context, V-BLAST achieves maximum spatial multiplexing gain but offers small diversity gain while space–time trellis and block codes are designed to achieve maximum diversity gains with no multiplexing gains. While there are many architectures that achieve both types of gains (an example is the system shown in Fig. 14), the design of MIMO systems that achieve the entire optimal diversity-multiplexing tradeoff curve in [291] remains a significant open research problem.

With a single receive antenna, the Alamouti scheme of Section IV-A.3 was shown in [291] to achieve the optimal rate-diversity tradeoff curve. However, it becomes suboptimal with two (or more) receive antennas and when extended to the case of more than two transmit antennas using orthogonal designs [239]. This suboptimality arises from the orthogonality constraint imposed in this class of space–time codes to achieve maximum diversity gains (at a given rate), while having linear decoding complexity.

A class of space–time codes known as *linear dispersion codes* (LDC) was introduced in [135] where the orthogonality constraint is relaxed to achieve higher rate while still enjoying (expected) polynomial decoding complexity for a wide SNR range by using the sphere decoder. This comes at the expense of signal constellation expansion and not guaranteeing maximum diversity gains (as in orthogonal designs). With M_t transmit antennas and a channel coherence time of T, the $T \times M_t$ transmitted signal space–time matrix $\mathbf{X}$ in LDC schemes has the form

$$\mathbf{X} = \sum_{q=1}^{Q} \alpha_q \mathbf{A}_q + j\beta_q \mathbf{B}_q \tag{39}$$

where the real scalars α_q and β_q are related to the Q information symbols x_q (that belong to a size-2^b complex signal constellation) by $x_q = \alpha_q + j\beta_q$ for $q = 1, 2, \ldots Q$. This LDC has a rate of $Q/T \log_2 M$. Several LDC designs were presented in [135] based on a judicious choice of the parameters T, Q and the so-called dispersion matrices $\mathbf{A}_q$ and $\mathbf{B}_q$ to maximize the mutual information between the transmitted and received signals.

An alternate way to attain diversity is to build in the diversity into the modulation. This idea is illustrated in Fig. 15. On the left is the usual QPSK modulation and on the right is a rotated QPSK modulation. Suppose we transmit information on the real and imaginary parts through independent fading channels. Then, we can view the transmission as being over two independent channels and the maximal diversity order attainable is 2. According to the criterion developed in Section III-C, the uncoded transmission would obtain diversity order dependent on whether the codewords (in this case we have four codewords) are different in each of the dimensions. Clearly, the unrotated modulation has a built-in diversity of 1 whereas the rotated modulation in Fig. 15 has a diversity of 2. This basic idea was proposed in [42], [158] and developed for higher dimensional lattices in [43] and references therein. Therefore, one can construct modulation schemes with built-in diversity, with the caveat that the constellation size is actually increasing. This is because the projections of the constellation points are now distinct and, therefore, the modulation scheme has to deal with a larger number of transmit levels. The main point to note here is that the Theorem 3.9 refers to rate versus diversity tradeoff for a given constellation size. Therefore, in order to consider the efficiency of coding schemes based on the rotated constellations, one needs to take into account the expansion in the constellation size. For example, transmit diversity using such rotated constellations is described in [77] and [91] and space–time codes designed using rotated constellations are given in [75].

Another point of view is explored in [86] where codes that achieve a high-rate, but have embedded within them a high-diversity (lower-rate) code are designed. Clearly the overall code will still be governed by the rate-diversity tradeoff, but

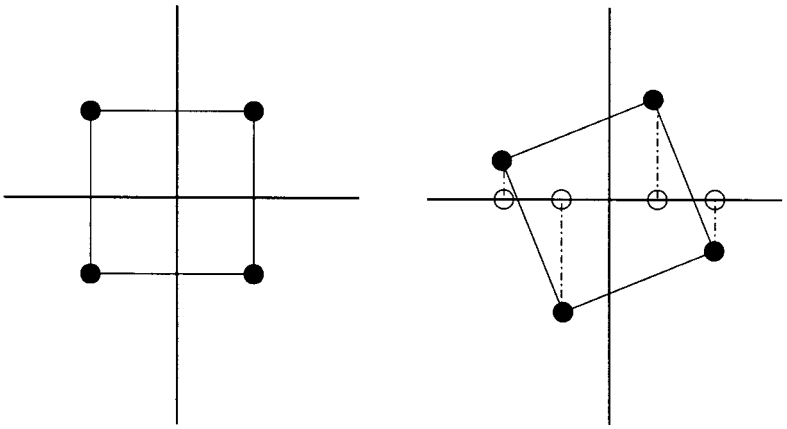

Fig. 15. Rotated constellations.

the idea is to ensure the reliability (diversity) of at least part of the total information.

5) OFDM: In OFDM, the high-rate input stream is demultiplexed and transmitted over N low-rate independent frequency subcarriers. This multicarrier transmission scheme is implemented digitally using the efficient FFT [280]. Since OFDM is a block transmission scheme, a guard sequence (of length at least equal to channel memory) is needed to eliminate IBI and ensure that individual subcarriers can be isolated at the receiver. In OFDM, the choice for guard sequence is a *cyclic prefix* which makes the channel matrix *circulant*, hence, diagonalizable by the FFT. If the FFT size is made large enough such that the width of each frequency bin is less than the *coherence bandwidth* of the channel, then no equalization is needed.[31] A large FFT size (compared to channel memory) also reduces the cyclic prefix guard sequence overhead at the expense of increased storage and processing requirements and increased delay which might not be acceptable for delay-sensitive applications.

OFDM is very attractive as modulation/equalization scheme for a channel with long delay spread (where trellis-based equalization is very complex) like Scenario A. For time-selective channels (as in Scenario B), the subcarriers in an OFDM signals lose their orthogonality (or equivalently the channel matrix is no longer circulant, hence, is no longer diagonalizable by the FFT) resulting in intercarrier interference (ICI). Successful application of OFDM to high-mobility scenarios is critically dependent on the implementation of effective channel estimation/tracking and ICI suppression schemes (see Section IV-B.1 and [235]).

OFDM offers great flexibility in that multiple signals with different rates and quality-of-service (QoS) requirements can be transmitted over the parallel frequency subchannels. Moreover, avoiding strong RF narrowband interference within the transmission bandwidth is easily accomplished by turning off the corresponding subchannels. OFDM was applied to MIMO broadband channels in [5] and [210]. OFDM has two main drawbacks, namely a high peak to average ratio (PAR), which results in larger backoff with nonlinear amplifiers [221], and high sensitivity to frequency errors and phase noise [206]. An alternative equalization scheme that overcomes these two drawbacks of OFDM while retaining its reduced implementation complexity advantage (but not its multirate capability) is the single-carrier frequency-domain equalizer SC FDE [222].

[31]Except for a simple one-tap complex equalizer for each subchannel, assuming negligible intercarrier interference due to Doppler effects or frequency offset errors.

B. Receiver Techniques

In this section, we present an overview of receiver signal processing algorithms suitable for the detection of space–time-coded signals. These algorithms can be classified under two main categories: coherent and noncoherent. Coherent detection requires CSI either explicitly by estimating the channel matrix and feeding this estimate to joint equalization/decoding algorithms (to be discussed in Section IV-B.1) or implicitly where the optimum settings of the joint equalizer/decoder are computed adaptively using training symbols (to be discussed in Section IV-B.1). Noncoherent techniques do not require CSI and, hence, are more suitable for rapidly time-varying channels (such as Scenario B) where it is more challenging to acquire accurate CSI or when cost and complexity constraints exclude the use of channel estimation modules. As seen in Section III-A, even when the channel is unknown at the receiver, [290] shows that the rate achievable using a training-based technique is only a constant factor (in mutual information) away from the optimal. This motivates our treatment of the training-based techniques in Section IV-B,1. The training sequences occupy a certain number of degrees of freedom in the transmitted sequence which could have been utilized to send additional information symbols. However, the result in [290] shows that it is an efficient way to utilize these degrees of freedom.

To explain the main ideas without unduly complicating the presentation, we assume (unless otherwise stated) a single receive antenna. When multiple independent receive antennas are available, additional *receive diversity gains* can be realized by *combining* the outputs of these antenna using spatio-temporal processing (see, e.g., [201], [202], and references therein).

1) Coherent Techniques: We start by considering channel estimation techniques for both quasi-static (or slowly time-varying) channels and for rapidly time-varying channels with appreciable channel variation within the block. This is followed by a discussion of joint equalization/decoding techniques for space–time trellis and block coded signals.

Channel Estimation for Quasi-Static Channels: For quasi-static channels (such as Scenario A), CSI can be estimated at the receiver using a training sequence embedded in each transmission block. For single-transmit-antenna signalling, the training sequence is only required to have "good" (i.e., impulse-like) auto-correlation properties. However, for the M_t transmit-antenna scenarios, the M_t training sequences should, in addition, have "low" (ideally zero) cross-correlation. In addition, it is desirable (in order to avoid amplifier nonlinear distortion) to use training sequences with constant amplitude. *Perfect root of unity sequences* (PRUS) [60] have these ideal correlation and constant amplitude properties. However, for a given training sequence length, PRUS do not always belong to standard signal constellations such as PSK. Additional challenges in channel estimation for multiple-transmit-antenna systems over the single-transmit-antenna case are the increased number of channel parameters to be estimated and the reduced transmit power (by a factor of M_t) for each transmit antenna.

In [108], it was proposed to encode a single training sequence by a space–time encoder to generate the M_t training sequences (cf. Fig. 16).[32] Strictly speaking, this approach is suboptimum, since the M_t transmitted training sequences are cross-correlated by the space–time encoder which imposes a constraint on the possible generated training sequences. However, it turns out that, with proper design, the performance loss from optimal PRUS training is negligible [108]. Furthermore, this approach reduces the training sequence search space from $(2^b)^{M_t N_t}$ to $(2^b)^{N_t}$ (assuming equal input and output alphabet size 2^b and length-N_t training sequences), making exhaustive search more practical and thus facilitating the identification of good training sequences from standard signal constellations such as PSK.

The search space can be further reduced by exploiting special characteristics of the particular STC. As an example, consider the eight-state 8-PSK STTC for two transmit and one receive antennas of Section IV-A.3 whose equivalent CIR is given by (37). For a given transmission block (over which the two channels $h_1(D)$ and $h_2(D)$ are constant), the input sequence determines the equivalent channel. By transmitting only "even" training symbols from the subconstellation $C_e = \{0, 2, 4, 6\}$, $p_k = +1$ and the equivalent channel is given by $h_e(D) = h_1(D) + Dh_2(D)$. On the other hand, transmitting only "odd" training symbols from the subconstellation $C_o = \{1, 3, 5, 7\}$, results in $p_k = -1$ and the equivalent channel $h_o(D) = h_1(D) - Dh_2(D)$. After estimating $h_e(D)$ and $h_o(D)$, we can compute

$$h_1(D) = \frac{h_e(D) + h_o(D)}{2} \quad \text{and} \quad h_2(D) = \frac{h_e(D) - h_o(D)}{2D}. \tag{40}$$

We propose to use a training sequence of the form $\mathbf{s} = [\mathbf{s}_e \quad \mathbf{s}_o]$ where $\mathbf{s}_e$ has length $N_t/2$ and takes values in the C_e subconstellation and $\mathbf{s}_o$ has length $N_t/2$ and takes values in the C_o subconstellation. Note that if $\mathbf{s}_e$ is a good sequence in terms of MMSE for the estimation of $h_e(D)$, the sequence $\mathbf{s}_o$ created as $\mathbf{s}_o = a\ \mathbf{s}_e$ where $a = \exp(i\pi k/4)$ and any $k = 1,3,5,7$ achieves the same MMSE for the estimation of $h_o(D)$. Thus, instead of searching over all possible 8^{N_t} sequences $\mathbf{s}$, we can further restrict the search space to the $4^{N_t/2}$ sequences $\mathbf{s}_e$. A reduced-size search can identify sequences $\mathbf{s}_e$ and $\mathbf{s}_o = a\ \mathbf{s}_e$ such that the channel estimation MMSE is achieved. We emphasize that similar reduced-complexity techniques can be developed for other STTCs by deriving their equivalent encoder models (as in Fig. 12).

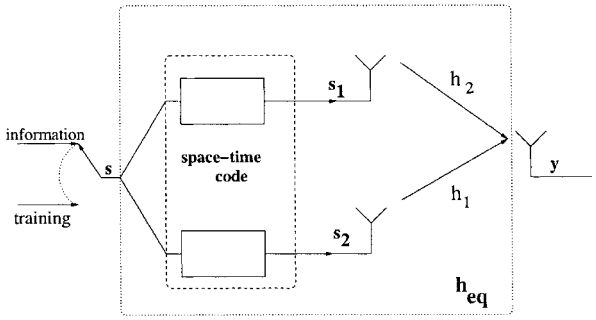

Fig. 16. Generation of two training sequences by space–time encoding a single training sequence.

In summary, the special STC structure can be utilized to simplify training sequence design for multiple-antenna transmissions without sacrificing performance.

Channel Estimation and Tracking for Rapidly Time-Varying Channels: In block transmissions over frequency-selective channels, time selectivity may have a negative impact on the system performance. Revisiting (3), we observe that equalization of the received block $\mathbf{y}$ and recovery of the transmitted symbols $\mathbf{x}$ depend upon the reliable estimation of the channel matrix $\mathbf{H}$. Estimation and equalization are facilitated by the fact that $\mathbf{H}$ possesses a structured-form in widely used block-transmission techniques when the channels remain constant during the transmission of a block. For example, $\mathbf{H}$ is a block Toeplitz matrix (when the guard sequence is all zeros) or is a block circulant matrix (when the guard sequence is the cyclic prefix). However, when the underlying channels vary significantly within a transmission block, $\mathbf{H}$ loses its special structured form. Consequently, both the estimation and the equalization of $\mathbf{H}$ become more challenging.

It is well documented that multicarrier transmissions (such as OFDM) are more sensitive to time variations than single-carrier transmissions. In OFDM, rapid time variation of the underlying channels within a transmission block result in ICI. Depending on the Doppler frequency and the block length chosen for transmission, ICI can potentially cause

[32]We assume, for simplicity, the same space–time encoder for the training and the information symbols. However, they could be different in general.

a severe deterioration of QoS. In SISO OFDM, previous studies have quantified the effects of ICI on the system performance [150], [167], [169], [216], [220]. For rapidly time-varying environments, and for cases where the channel obeys a parsimonious model, techniques in [118], [246], and [252] can be used for channel estimation and tracking. Here, we discuss *severe* Doppler cases, where the channel can no longer be assumed constant within a transmission block (as in Scenario B): for such rapidly time-varying channels, the challenges of channel estimation are addressed by the channel estimation and tracking scheme of [235].

To simplify the presentation, consider the case $M_r = M_t = 1$. In a time-varying environment, estimation of $\mathbf{H}$ amounts to estimating N channels $\mathbf{h}_n := [h(n;0), \ldots, h(n;\nu-1)]^T$, $0 \le n \le N-1$, that comprise the rows of $\mathbf{H}$. To reduce the number of parameters needed for channel estimation from $N\nu$ to less than N, [235] makes the assumption that some of the channels $\mathbf{h}_n$ can be obtained by linear interpolation. Such an assumption holds true if there is not significant variation between channels $\mathbf{h}_n$ and $\mathbf{h}_{n+1}$, $0 \le n \le (N-2)$. The matrix $\mathbf{H}$ is parametrized using a small number of its rows. Then, the entire matrix is expressed as a function of these rows which reduces the number of parameters to be estimated. Physically, this puts "markers" in time where the channel is estimated, and the estimates at other times are interpolated using these estimates.

The channel matrix $\tilde{\mathbf{H}}$ obtained by the interpolation of the channels $\mathbf{h}_m(1), \ldots, \mathbf{h}_m(M)$ (where $m(1), \ldots, m(M)$ are the channel rows used in the interpolation) is

$$\tilde{\mathbf{H}} = \sum_{1 \le i \le M} \mathbf{A}_{m(i)} H_C(\mathbf{h}_{m(i)}) \tag{41}$$

where $\mathbf{A}_{m(i)}$ is an $N \times N$ diagonal matrix with entries equal to the interpolation weights, and $H_c(\mathbf{h})$ is the OFDM circulant channel matrix that corresponds to an FIR channel with entries contained in the channel vector $\mathbf{h}$. Given the structured form of $\tilde{\mathbf{H}}$, its estimation amounts to estimating $M\nu$ parameters grouped in the $M\nu \times 1$ vector $\tilde{\mathbf{h}} := [\mathbf{h}^T_{m(1)}, \ldots, \mathbf{h}^T_{m(M)}]^T$. Using judiciously placed pilot tones, the latter can be obtained as the least-squares solution $\hat{\tilde{\mathbf{h}}} = \tilde{\mathbf{B}}^{\dagger}_{(P)} \mathbf{Y}_{(P)}$, where $\mathbf{Y}_{(P)}$ is the received vector when pilot tones are included in the transmission, and $\tilde{\mathbf{B}}_{(P)}$ is a matrix with entries that are function of the pilot tones and the interpolation weights (see [235] for more details).

The quality of channel estimates depends heavily on the placement of pilot tones on the FFT grid. In frequency-selective time-invariant channels, placing the pilot tones equispaced on the FFT grid is the optimal scheme [194], [196]. On the other hand, work on pilot symbol assisted modulation (PSAM) [52] has suggested that, for Rayleigh flat-fading time-varying channels, pilot symbols should be placed periodically in the time domain to produce channel estimates. Then, the coherent detection of the transmitted symbols is based on the interpolation of these channel estimates. The periodic transmission of pilot symbols in the time domain suggests a grouping of pilot tones in the frequency domain. We have found that our channel estimation method produces the best results when the *pilot tones are partitioned into equispaced groups on the FFT grid*, a result which is in contrast to the optimal pilot placement schemes of [194] and [196].

Furthermore, under relatively mild Doppler, the quality of the channel estimates (and, consequently, the SINR gains of the ICI mitigating methods in [235]) can be improved by channel tracking. A simple tracking scheme is the following: an initial channel estimate $\hat{\mathbf{H}}_{(0)}$ can be obtained by transmitting a full training block; subsequent blocks contain pilot tones which are used to acquire new estimates $\mathbf{H}'_{(n)}$. The channel estimate for the nth OFDM block is obtained using a forgetting factor α from the relation $\hat{\mathbf{H}}_{(n)} = \alpha \hat{\mathbf{H}}_{(n-1)} + (1-\alpha)\mathbf{H}'_{(n)}$. Frequent retraining can further improve the quality of the channel estimates at the expense of the overhead of the training symbols.

Joint Equalization/Decoding of Space–Time Trellis Codes: For broadband transmissions, equalization is indispensable for mitigating *ISI* [244]. STC makes equalization more challenging because it generates multiple *correlated* signals that are transmitted *simultaneously at equal power*. However, carefully designed joint equalization/decoding schemes can exploit this correlation to reduce implementation complexity while achieving significant performance gains over single-antenna transmissions (due to spatial and multipath diversity gains). In this section, we describe briefly examples of practical near-optimal joint equalization/decoding schemes for STTC and STBC. A detailed treatment of this subject is given in [8].

Both the STTC and the CIRs are finite-state machines described by a trellis. Hence, optimum performance is achieved by joint STTC equalization/decoding on the combined trellis. The complexity of full joint trellis equalization and decoding increases *exponentially* with the channel memory,[33] signal constellation size, and the number of transmit antennas. On the other hand, increasing these three parameters are effective means to achieve high bit rates. This motivates the need for reduced-complexity joint equalization and decoding techniques that achieve a practical performance-complexity tradeoff. An example of such schemes is described next.

When implementing the eight-state 8-PSK STTC over frequency-selective channels, its rich structure can be exploited to reduce equalization/decoding complexity. This is achieved by performing trellis-based joint equalization and space–time decoding with $8^{\nu+1}$ states on the equivalent channel given in (37). For channels with long memory, further complexity reductions (at some performance loss) are achieved by preceding the joint equalizer/decoder with a channel shortening FIR prefilter (cf. Fig. 17). The objective of the prefilter is to shorten and shape the effective CIR memory seen by the equalizer to reduce its complexity. Prefilter design algorithms suitable for space–time-coded signals are described in [106], [287].

Several reduced-state equalization/decoding algorithms have been proposed in the literature including DDFSE [93], RSSE [99], T-BCJR [109], and M-BCJR [106], [109]. We

[33]For a given channel delay spread, the channel memory increases linearly with the transmission bandwidth.

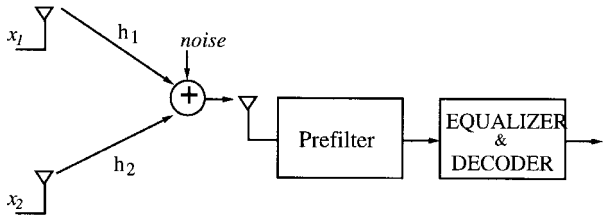

Fig. 17. Receiver structure for STTC prefiltered joint equalization/decoding with two transmit and one receive antennas.

have investigated and improved the M-BCJR algorithm and applied it to the problem of joint equalization and decoding of STTC, as described briefly next. The findings of our study are detailed in [106].

The M-BCJR algorithm is a reduced-complexity version of the famed BCJR algorithm [22] where at each trellis step, only the M active states associated with the highest metrics are retained. An improved version of the M-BCJR algorithm was proposed in [106] and applied to perform joint equalization and decoding of STTC. More specifically, it was shown in [106] that preceding the M-BCJR equalizer/decoder with a channel-shortening prefilter improves its performance, especially for small values of M. Even better performance is achieved when a different prefilter is used for the forward and backward recursions of the M-BCJR algorithm. The value of M and the number of prefilter taps can be jointly optimized to achieve the best performance-complexity tradeoffs. For channels with long delay spread, FFT-based techniques such as OFDM emerge as attractive alternative candidates for STTC equalization/decoding [5].

Fig. 18 plots the performance of the prefiltered M-BCJR equalizer/decoder as a function of the number of active states M for a quasi-static typical urban environment assuming two transmit antennas, one receive antenna, and the eight-state 8-PSK STTC. Both channels are shortened to three taps, hence, the maximum value of M with the prefilter is $8^3 = 512$ states. Also shown in the figure as a benchmark is the BER of a full BCJR-MAP equalizer with 4096 states[34] and no prefilter. In our simulations, the noise samples are generated as independent samples of a zero-mean complex circularly symmetric Gaussian random variable with a variance of $1/\text{SNR}$ per complex dimension. The reason for doubling the noise variance (compared to the single-transmit-antenna case) is that with two-antenna transmissions, we assume that the total transmitted power is the same as in the single-antenna case and is divided equally between the two antennas. The average energy of the symbol transmitted from each antenna is normalized to one so that the SNR ratio is SNR.[35] It can be seen that negligible performance improvement is achieved by increasing M from 16 to its maximum value of 512 due the effective action of the prefilter which concentrates most of the channel energy in its leading or last taps for the forward and backward recursions, respectively. The performance gap from the 4096–state BCJR-MAP is due to the prefilter loss but, in return, we achieve a significant reduction in the number of equalizer/decoder states (from 4096 to 16).

[34]The equivalent SISO channel of the eight-state 8-PSK STTC has memory of $\nu + 1 = 4$ (cf. (37)). Hence, the number of BCJR-MAP equalizers states is $8^4 = 4096$.

[35]Note that SNR is related to E_b/N_0 by the relation: $\text{SNR} = (E_b/N_0)b$, where 2^b is the signal constellation size.

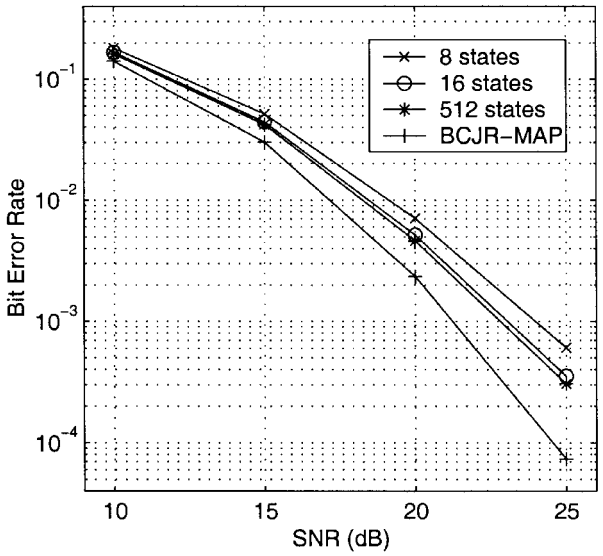

Fig. 18. BER performance of two-transmit one-receive eight-state 8-PSK STTC on a quasi-static typical urban channel with $\nu = 3$ with prefiltered M-BCJR equalizer/decoder as a function of M (the number of active states). The performance of a $8^{\nu+1} = 4096$-state full BCJR-MAP equalizer/decoder is shown as a benchmark.

Sphere Decoder: The M-BCJR algorithm described above is an example of a reduced-complexity MAP decoder algorithm for trellis-coded signals. It applies to both linear and nonlinear trellis codes, where in the latter the codewords (unlike the information symbols) do not necessarily form a lattice. However, when the input codewords belong to a lattice, the structure of the code can be utilized to reduce the computational complexity of the ML search. For specific lattices, efficient decoding algorithms have been developed [66]. Another technique which is applicable to any lattice is called the sphere decoder and was first proposed in [101] (a good exposition of this decoder for fading channels can be found in [276]). Note that the linear decoding techniques described in Section III-A and the decision feedback techniques of BLAST in Section IV-A.1 are not ML techniques. The main reason that these techniques are used is because ML decoding in general can be quite computationally expensive.

In order to illustrate the sphere decoder, let us look at an AWGN channel (i.e., in (3) set $\nu = 0$ and $\mathbf{H} = \mathbf{I}$). Fig. 19 depicts the lattice code $\mathcal{L}$. ML decoding uses the criterion $\min_{\mathbf{x}\in\mathcal{L}} \|\mathbf{y} - \mathbf{x}\|^2$ and, therefore, decodes to the closest lattice point in $\mathcal{L}$. Given that $\mathbf{x}$ comes from a lattice $\mathcal{L}$, its points can be written as $\mathbf{x} = \mathbf{M}\mathbf{u}$ where $\mathbf{M}$ is the generator matrix for the lattice [66] and $\mathbf{u}$ is an integer vector. Instead of performing an exhaustive search over the entire lattice, the sphere decoding algorithm (depicted in Fig. 19) searches over a suitably chosen radius R around the received point and finds the closest lattice point to it. For this given radius R, it is easy to write out the lattice points contained in it using the Gram matrix $\mathbf{M}\mathbf{M}^T$ of the lattice [276]. Since this contains a bounded number of points to search over, the complexity is

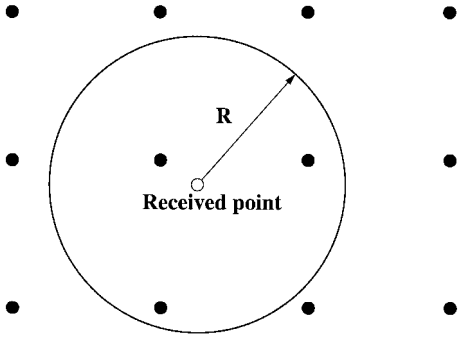

Fig. 19. Sphere decoding algorithm.

reduced considerably. In fact the complexity can be related to the minimum eigenvalue of the Gram matrix [101]. In an AWGN channel, the radius R can be chosen to be the covering radius of the lattice $\mathcal{L}$[66] and this ensures that there is at least one lattice point in the sphere. In fading channels, the matrix $\mathbf{H}$ in (4) can be seen to distort the lattice and, therefore, changes the generator matrix of the lattice. The channel parameters can, therefore, rotate, elongate, and compress the lattice along different directions. Therefore, the choice of the radius and the decoding complexity depend on the random channel. The same principle of the sphere decoder can be applied to the multiantenna channel [76] but a careful choice of the radius and repeated calculations of the channel-dependent Gram matrix is needed. Note that in this case, since the Gram matrix is random, the decoding complexity is also a random quantity (dependent on the minimum eigenvalue of the Gram matrix). This combined with the fact that the covering radius of the modified lattice is random causes the complexity to be random, and, hence, it is meaningful to analyze the expected complexity of the sphere decoder [137], [185].

Joint Equalization/Decoding of Space–Time Block Codes: Our focus will be on Alamouti-type STBC with two transmit antennas. The treatment can be extended to more than two antennas using orthogonal designs [239] at the expense of some rate loss for complex signal constellation.

The main attractive feature of STBC is the quaternionic[36] structure of the spatio-temporal channel matrix. This allows us to eliminate interantenna interference using a low-complexity linear combiner (which is a spatio-temporal matched filter and is also the ML detector in this case). Then, joint equalization and decoding for each antenna stream proceeds using any of well-known algorithms for the single-antenna case which can be implemented either in the time or frequency domains. For illustration purposes, we describe next a joint equalization and decoding algorithm for the single-carrier frequency-domain-equalizer (SC FDE)-STBC. A more detailed discussion and comparison is given in [8].

The SC FDE-STBC receiver block diagram is given in Fig. 20. After analog-to-digital (A/D) conversion, the CP part

[36]A 2×2 complex orthogonal matrix of the form

$$\begin{bmatrix} c_1 & c_2 \\ -\bar{c}_2 & \bar{c}_1 \end{bmatrix}$$

is isomorphic to the quaternion group (see Appendix).

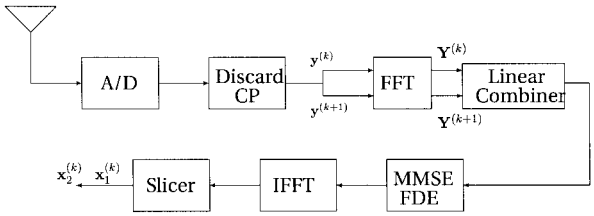

Fig. 20. FDE-STBC receiver block diagram.

of each received block is discarded. Mathematically, we can express the input-output relationship over the jth received block as follows:

$$\mathbf{y}^{(j)} = \mathbf{H}_1^{(j)}\mathbf{x}_1^{(j)} + \mathbf{H}_2^{(j)}\mathbf{x}_2^{(j)} + \mathbf{z}^{(j)} \tag{42}$$

where $\mathbf{H}_1^{(j)}$ and $\mathbf{H}_2^{(j)}$ are $N \times N$ circulant matrices whose first columns are equal to $\mathbf{h}_1^{(j)}$ and $\mathbf{h}_2^{(j)}$, respectively, appended by $(N - \nu - 1)$ zeros and $\mathbf{z}^{(j)}$ is the noise vector. Since $\mathbf{H}_1^{(j)}$ and $\mathbf{H}_2^{(j)}$ are circulant matrices, they admit the eigen-decompositions

$$\mathbf{H}_1^{(j)} = \mathbf{Q}^*\mathbf{\Lambda}_1^{(j)}\mathbf{Q} \quad ; \quad \mathbf{H}_2^{(j)} = \mathbf{Q}^*\mathbf{\Lambda}_2^{(j)}\mathbf{Q}\,,$$

where $\mathbf{Q}$ is the orthonormal FFT matrix and $\mathbf{\Lambda}_1^{(j)}$ (resp. $\mathbf{\Lambda}_2^{(j)}$) is a diagonal matrix whose (n, n) entry is equal to the nth FFT coefficient of $\mathbf{h}_1^{(j)}$ (resp. $\mathbf{h}_2^{(j)}$). Therefore, applying the FFT to $\mathbf{y}^{(j)}$, we get (for $j = k, k+1$)

$$\mathbf{Y}^{(j)} = \mathbf{Q}\mathbf{y}^{(j)} = \mathbf{\Lambda}_1^{(j)}\mathbf{X}_1^{(j)} + \mathbf{\Lambda}_2^{(j)}\mathbf{X}_2^{(j)} + \mathbf{Z}^{(j)}.$$

The SC FDE-STBC encoding rule is given by [7]

$$\mathbf{X}_1^{(k+1)}(m) = \bar{\mathbf{X}}_2^{(k)}(m) \quad \text{and} \quad \mathbf{X}_2^{(k+1)}(m) = -\bar{\mathbf{X}}_1^{(k)}(m) \tag{43}$$

for $m = 0, 1, \ldots, N-1$ and $k = 0, 2, 4, \ldots$. The length-N blocks at the FFT output are then processed in pairs resulting in the two blocks (we drop the time index from the channel matrices, since they are assumed fixed over the two blocks under consideration)

$$\underbrace{\begin{bmatrix} \mathbf{Y}^{(k)} \\ \bar{\mathbf{Y}}^{(k+1)} \end{bmatrix}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} \mathbf{\Lambda}_1 & \mathbf{\Lambda}_2 \\ -\bar{\mathbf{\Lambda}}_2 & \bar{\mathbf{\Lambda}}_1 \end{bmatrix}}_{\mathbf{\Lambda}} \underbrace{\begin{bmatrix} \mathbf{X}_1^{(k)} \\ \mathbf{X}_2^{(k)} \end{bmatrix}}_{\mathbf{X}} + \underbrace{\begin{bmatrix} \mathbf{Z}^{(k)} \\ \bar{\mathbf{Z}}^{(k+1)} \end{bmatrix}}_{\mathbf{Z}} \tag{44}$$

where $\mathbf{X}_1^{(k)}$ and $\mathbf{X}_2^{(k)}$ are the FFTs of the information blocks $\mathbf{x}_1^{(k)}$ and $\mathbf{x}_2^{(k)}$, respectively, and $\mathbf{Z}$ is the noise vector. We used the encoding rule in (43) to arrive at (44). To eliminate *interantenna interference*, the linear combiner $\mathbf{\Lambda}^*$ is applied to $\mathbf{Y}$. Due to the quaternionic structure of $\mathbf{\Lambda}$, a second-order diversity gain is achieved. Then, the two decoupled blocks at the output of the linear combiner are equalized separately using the MMSE FDE [222] which consists of N complex taps per block that mitigate *ISI*. Finally, the MMSE-FDE output is transformed back to the time domain using the inverse FFT where decisions are made.

Fig. 21 shows the diversity advantage achieved in SC MMSE FDE-STBC compared to single-antenna transmission. This figure assumes perfect channel knowledge at the

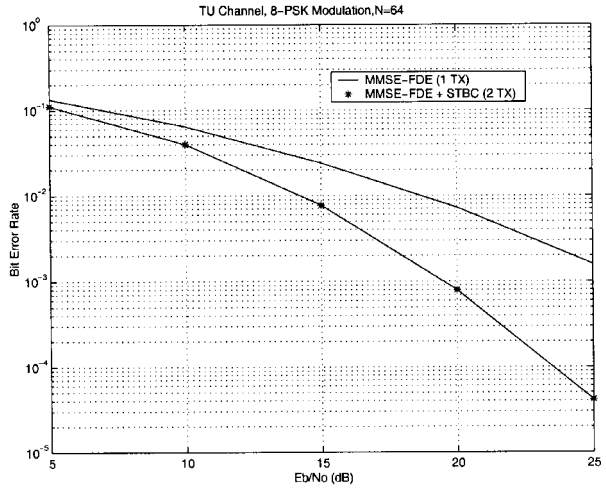

Fig. 21. BER of SC MMSE-FDE w/ and w/o STBC for typical urban channel ($\nu = 3$) with 8-PSK modulation and $N = 64$.

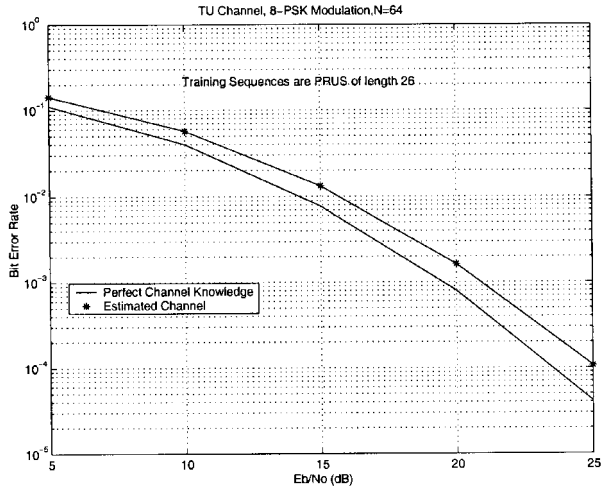

Fig. 22. Effect of channel estimation on performance of SC FDE-STBC.

receiver.[37] To investigate the effect of channel estimation errors on performance, we assumed that 26 training symbols per two blocks of information symbols (total length of 128 symbols). Two shifted PRUS [60] were transmitted from antennas 1 and 2 during training. The two channel estimates were jointly computed using a standard least squares algorithm [73]. From Fig. 22, it can be seen that the performance loss due to channel estimation errors is $\approx 1 - 1.5$ dB.

OFDM With Fast Channel Variations: As discussed in Section IV-A.5, OFDM is an effective equalization technique for broadband MIMO channels. The IFFT/FFT is used as a modulation/demodulation basis to partition the channel frequency response into a large number of orthogonal subcarriers each experiencing flat fading. This *diagonalization* of the channel matrix occurs only under quasi-static channel fading conditions. Therefore, if the channel is block fading, the Fourier basis is the eigenbasis, and one can estimate the channel from block to block (see, for example, [170] and references therein). However, in the presence of channel variations with the transmission block, the subcarrier are no longer orthogonal and ICI can result in significant performance degradation [63].

One approach, described in detail in [235], to mitigate ICI and restore (approximately) subcarrier orthogonality is to implement a time-domain MIMO prefilter $\mathbf{W}$ at the receiver front end that attempts to restore the *circulant* structure of the overall channel and, hence, make it diagonalizable by the FFT.

The matrix filter $\mathbf{W}$ can be designed (using channel knowledge acquired using the methods presented earlier for rapidly varying channels) to maximize signal to ICI plus noise ratio (SINR) which is defined at the mth frequency bin ($1 \leq m \leq N$) as follows:

$$\text{SINR}_m = \frac{E_{x,m}|\tilde{\mathbf{G}}(m,m)|^2}{\frac{1}{N}\text{tr}(\mathbf{R}_{\tilde{z}\tilde{z}}) + \sum_{n \neq m} E_{x,n}|\tilde{\mathbf{G}}(m,n)|^2} \tag{45}$$

where $E_{x,n}$ is the input energy allocated to the nth frequency bin, $\mathbf{R}_{\tilde{z}\tilde{z}}$ is the noise covariance (after the application of the $\mathbf{W}$ filter), and $\tilde{\mathbf{G}}$ is a matrix that represents the cascade of the DFT operation at the OFDM receiver, the $\mathbf{W}$ filter, the channel matrix $\mathbf{H}$, and the IDFT operation at the transmitter. It is shown in [235] that the design of $\mathbf{W}$ can be posed as a generalized eigenvalue problem, where $\mathbf{W}$ is calculated as a function of the estimated channel matrix.

Adaptive Techniques: The coherent receiver techniques described till now require CSI which is estimated and tracked using training sequences/pilot symbols inserted in each block and then used to compute the optimum joint equalizer/decoder settings. An alternative to this two-step channel-estimate-based approach is *adaptive* space–time equalization/decoding where CSI is not explicitly estimated at the receiver. Adaptive receivers still require training overhead to converge to their optimum settings which, in the presence of channel variations, are adapted using previous decisions to *track* these variations. Adaptive algorithms, such as the celebrated least mean square (LMS) algorithm [138], are widely used in single-antenna communication systems today due to its low implementation complexity. However, it has been shown to exhibit slow convergence and suffer significant performance degradation (relative to performance achieved with the optimum settings) when applied to broadband MIMO channels due to the large number of parameters that need to be simultaneously adapted and the wide eigenvalue spread problems encountered on those channels. Faster convergence can be achieved by implementing a more sophisticated family of algorithms known as recursive least squares (RLS). However, their high computational complexity compared to LMS and and their notorious behavior when implemented in finite precision limit their appeal in practice. It was shown in [288] that the orthogonal structure of STBC can be exploited to develop fast-converging RLS-type adaptive FDE-STBC at LMS-type complexity. A brief overview is given next.

Considering the structure of Fig. 20

$$\begin{bmatrix} \hat{\mathbf{X}}_1^{(k)} \\ \hat{\mathbf{X}}_2^{(k)} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 \\ \bar{\mathbf{A}}_2 & -\bar{\mathbf{A}}_1 \end{bmatrix} \mathbf{Y} \tag{46}$$

[37]The SC MMSE FDE-STBC performance can be further improved by adding a feedback section as discussed in [36], [100]

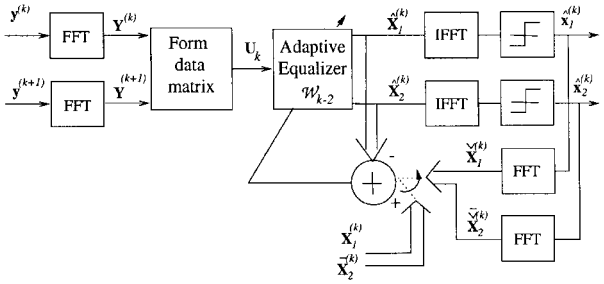

Fig. 23. Block diagram of adaptive FDE-STBC joint equalizer/decoder.

where $\mathbf{Y}$ was defined in (44) and the diagonal matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ are given by

$$\mathbf{A}_1 = \mathbf{\Lambda}_1^*.\text{diag}\left\{\frac{1}{\tilde{\mathbf{\Lambda}}(i,i)+\frac{1}{\text{SNR}}}\right\}_{i=0}^{N-1}$$
$$\mathbf{A}_2 = \mathbf{\Lambda}_2^*.\text{diag}\left\{\frac{1}{\tilde{\mathbf{\Lambda}}(i,i)+\frac{1}{\text{SNR}}}\right\}_{i=0}^{N-1} \tag{47}$$

with $\tilde{\mathbf{\Lambda}}(i,i) = |\mathbf{\Lambda}_1(i,i)|^2 + |\mathbf{\Lambda}_2(i,i)|^2$. Alternatively, we can write

$$\begin{bmatrix}\hat{\mathbf{X}}_1^{(k)}\\ \hat{\mathbf{X}}_2^{(k)}\end{bmatrix} = \begin{bmatrix}\text{diag}(\mathbf{Y}^{(k)}) & -\text{diag}(\bar{\mathbf{Y}}^{(k+1)})\\ \text{diag}(\mathbf{Y}^{(k+1)}) & \text{diag}(\bar{\mathbf{Y}}^{(k)})\end{bmatrix}\begin{bmatrix}\bar{\mathbf{W}}_1\\ \mathbf{W}_2\end{bmatrix} = \mathbf{U}_k\mathcal{W} \tag{48}$$

where $\bar{\mathbf{W}}_1$ and $\mathbf{W}_2$ are the vectors containing the diagonal elements of $\bar{\mathbf{A}}_1$ and $\mathbf{A}_2$, respectively, and $\mathcal{W}$ is a $2N\times 1$ vector containing the elements of $\bar{\mathbf{W}}_1$ and $\mathbf{W}_2$. The $2N\times 2N$ quaternionic matrix $\mathbf{U}_k$ contains the received symbols for blocks k and $k+1$. Equation (48) can be used to develop a frequency-domain block-adaptive RLS algorithm for $\mathbf{W}$ which, using the special quaternionic structure of the problem, be simplified to the following LMS-type recursions (see [288] for details of the derivation):

$$\mathcal{W}_{k+2} = \mathcal{W}_k + \begin{bmatrix}\mathbf{P}_{k+2} & \mathbf{0}\\ \mathbf{0} & \mathbf{P}_{k+2}\end{bmatrix}\mathbf{U}_{k+2}(\mathbf{D}_{k+2} - \mathbf{U}_{k+2}\mathcal{W}_k) \tag{49}$$

where $\mathbf{D}_{k+2} = \begin{bmatrix}\mathbf{X}_1^{(k+2)} & \bar{\mathbf{X}}_2^{(k+2)}\end{bmatrix}^T$ for the training mode and $\mathbf{D}_{k+2} = \begin{bmatrix}\hat{\mathbf{X}}_1^{(k+2)} & \bar{\hat{\mathbf{X}}}_2^{(k+2)}\end{bmatrix}^T$ for the decision-directed mode. The $N\times N$ diagonal matrix $\mathbf{P}_{k+2}$ is computed by the recursion

$$\mathbf{P}_{k+2} = \lambda^{-1}(\mathbf{P}_k - \lambda^{-1}\mathbf{P}_k\mathbf{\Gamma}_{k+2}\mathbf{P}_k) \tag{50}$$

where the diagonal matrices $\mathbf{\Gamma}_{k+2}$ and $\mathbf{\Delta}_{k+2}$ are computed from the recursions

$$\begin{aligned}\mathbf{\Gamma}_{k+2} &= \text{diag}(\mathbf{Y}^{(k)})\mathbf{\Delta}_{k+2}\text{diag}(\bar{\mathbf{Y}}^{(k)})\\ &\quad + \text{diag}(\mathbf{Y}^{(k+1)})\mathbf{\Delta}_{k+2}\text{diag}(\bar{\mathbf{Y}}^{(k+1)})\\ \mathbf{\Delta}_{k+2} &= (\mathbf{I}_N + \lambda^{-1}(\text{diag}(\mathbf{Y}^{(k)})\mathbf{P}_k\text{diag}(\bar{\mathbf{Y}}^{(k)})\\ &\quad + \text{diag}(\mathbf{Y}^{(k+1)})\mathbf{P}_k\text{diag}(\bar{\mathbf{Y}}^{(k+1)})))^{-1}.\end{aligned}$$

The initial conditions are $\mathcal{W}_0 = \mathbf{0}$, $\mathbf{P}_0 = \delta\mathbf{I}_N$ where δ is a large number, and the forgetting factor λ is chosen close to 1.

The block diagram of the adaptive FDE-STBC is shown in Fig. 23. Pairs of consecutive received blocks are transformed to the frequency domain using the FFT, then the data matrix in (48) is formed. The filter output is the product of the matrix $\mathbf{U}_k\mathcal{W}_{k-2}$ is transformed back to the time domain using IFFT and passed to a decision device to generate the input blocks estimates. The output of the adaptive equalizer is compared to the desired response to generate an error vector which is in turn used to update the equalizer coefficients according to the RLS recursions. The equalizer operates in a training mode until it converges, then it switches to a decision-directed mode where previous decisions are used for tracking. When operating over fast time-varying channels (as in Scenario B), retraining blocks can be transmitted periodically to prevent equalizer divergence (see [288]).

2) Noncoherent Techniques: Noncoherent transmission schemes do not require channel estimation, hence eliminating the need for bandwidth-consuming training sequences and reducing terminal complexity. This becomes more significant for rapidly fading channels (such as Scenario B) where frequent retraining is needed to track channel variations and for multiple-antenna broadband transmission scenarios where more channel parameters (several coefficients for each transmit-receive antenna pair) need to be estimated. Noncoherent techniques include blind identification and detection schemes. Here, the structure of the channel (finite impulse response), the input constellation (finite alphabet) and the output (cyclostationarity) are exploited to eliminate training symbols. Such techniques have a vast literature and we refer the interested reader to a good survey in [250]. An alternative technique is to have a generalized ML receiver which assumes statistics about channel state but not knowledge of the state itself [142], [258].

Several noncoherent space–time transmission schemes have been proposed for flat-fading channels including differential STBC schemes with two [238] or more [146] transmit antennas and group differential STC schemes (proposed in [140] and [145]). An extensive characterization and classification of group differential space–time codes has been given in [229]. In this section we describe a differential space–time transmission scheme for frequency-selective channels recently proposed in [80] that achieves full diversity (spatial and multipath) at rate one[38] with two transmit antennas. This scheme is a differential form for the OFDM-STBC structure described in [175] and, hence, is a nongroup code. A time-domain differential space–time scheme with single-carrier transmission is also presented in [80].

We consider two symbols $X_1(m)$ and $X_2(m)$ drawn from a PSK constellation which, in a conventional OFDM system, would be transmitted over two consecutive OFDM blocks on the same subcarrier m. Following the Alamouti encoding scheme described in Section IV-A.3, the two source symbols are mapped as

$$\mathbf{X}^{(1)}(m) = [X_1(m), X_2(m)]^T,\ \mathbf{X}^{(2)}(m) = [-\bar{X}_2(m), \bar{X}_1(m)]^T \tag{51}$$

[38]This does not include the rate penalty incurred by concatenating OFDM-STBC with an outer code and interleaving across frequency tones which is common to all OFDM systems (see, e.g., [222] for more discussion.)

where $\mathbf{X}^{(1)}$ represents the information-bearing vector for the first OFDM block and $\mathbf{X}^{(2)}$ corresponds to the second OFDM block.[39] Let N denote the FFT size, then $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are length-$2N$ vectors holding the symbols to be transmitted by the two transmit antennas. Consequently, after taking the FFT at the receiver, we have (at subcarrier m)

$$\begin{pmatrix} Y_1(m) & Y_2(m) \\ -\bar{Y}_2(m) & \bar{Y}_1(m) \end{pmatrix} = \begin{pmatrix} H_1(m) & H_2(m) \\ -\bar{H}_2(m) & \bar{H}_1(m) \end{pmatrix} \begin{pmatrix} X_1(m) & -\bar{X}_2(m) \\ X_2(m) & \bar{X}_1(m) \end{pmatrix} + \text{noise} \tag{52}$$

where $H_1(m)$ and $H_2(m)$ are the frequency responses of the two channels at subcarrier m.

For block k and subcarrier m, denote the source symbols as $\mathbf{u}_m^{(k)} = \begin{bmatrix} u_{1,m}^{(k)} & u_{2,m}^{(k)} \end{bmatrix}^T$, the transmitted matrix as $\mathbf{X}_m^{(k)}$, and the received matrix as $\mathbf{Y}_m^{(k)}$. Then, in the absence of noise, (52) is written as $\mathbf{Y}_m^{(k)} = \mathbf{H}_m \mathbf{X}_m^{(k)}$, where we assume that the channel is fixed over two consecutive blocks. Using the quaternionic structure of $\mathbf{H}_m$, it follows that

$$\bar{\mathbf{Y}}_m^{(k-1)} \, \mathbf{Y}_m^{(k)} = (|H_1(m)|^2 + |H_2(m)|^2) \, \bar{\mathbf{X}}_m^{(k-1)} \, \mathbf{X}_m^{(k)} .$$

Since we would like to estimate the source symbols contained in $\mathbf{U}_m^{(k)} \underset{=}{def} \begin{pmatrix} u_{1,m}^{(k)} & -\bar{u}_{2,m}^{(k)} \\ u_{2,m}^{(k)} & \bar{u}_{1,m}^{(k)} \end{pmatrix}$, we define the differential transmission rule

$$\mathbf{X}_m^{(k)} = (\bar{\mathbf{X}}_m^{(k-1)})^{-1} \, \mathbf{U}_m^{(k)} . \tag{53}$$

Note that no inverse computation is needed in computing $(\bar{\mathbf{X}}_m^{(k-1)})^{-1}$ due to the quaternionic structure of $\bar{\mathbf{X}}_m^{(k-1)}$.

Fig. 24 illustrates the 3-dB SNR loss of differential OFDM-STBC relative to its coherent counterpart[40] (with perfect CSI assumed) for an indoor environment (similar to Scenario A). In the presence of high mobility (as in Scenario B), the performance of the coherent schemes degrades significantly due to channel estimation errors (see [235]). Under such conditions, the assumption of a fixed channel over two transmission blocks made in differential OFDM-STBC will not hold causing some performance degradation, however, its lower implementation complexity still makes it a more attractive choice for high Doppler rates than coherent OFDM-STBC with channel estimation/tracking.

C. Summary, Design Issues, and Future Challenges

The main objective of the physical layer is delivery of the highest possible bit rates reliably over the wireless channel [39]. To achieve this objective, several techniques

[39]Intuitively, each OFDM subcarrier can be thought of as a flat-fading channel and the Alamouti code is applied to each of the OFDM subcarriers. As a result, the Alamouti code yields diversity gains at every subcarrier. However, in order to gather both spatial and multipath (frequency) diversity gains, one needs to appropriately code across the subcarriers and decode them jointly.

[40]This figure is for illustration of the 3-dB SNR difference, since it compares the coherent with the noncoherent. However, in these curves the full multipath diversity is not exploited through appropriate coding across subcarriers, and this would move both the curves to the left, equivalently.

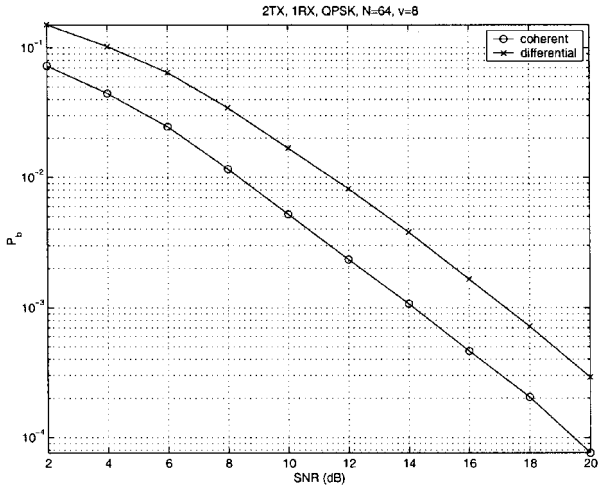

Fig. 24. Performance comparison between coherent and differential OFDM-STBC with 2 TX, 1RX, QPSK modulation, FFT size of 64, $\nu = 8$.

at the transmitter and receiver ends are needed to mitigate various performance impairments. This includes the use of diversity techniques to mitigate fading effects by exploiting space selectivity of the channel, the use of MIMO antenna techniques to realize spatial rate multiplexing gains, and the use of OFDM (or other equalization techniques such as M-BCJR or SC FDE) to mitigate channel frequency selectivity. In addition, three main techniques can be used to mitigate Doppler effects due to channel time selectivity: channel estimation and tracking using pilot symbols/tones, adaptive filtering, and differential transmission/detection. The first technique is more complex but achieves superior performance for quasi-static or slowly fading channels (as in Scenario A).

A discussion about the tradeoffs involved in system design issues is now in order. The system design depends on both the choice of key parameters such as block length, carrier frequency, and number of transmit/receive antennas as well as the operating environment conditions such as high versus low SNR, high versus low mobility, and strict versus relaxed delay constraints.

The length of the transmission block N (relative to the symbol period and the channel memory ν) is an important design parameter. Shorter blocks experience less channel time variation (which reduces the need for channel tracking within the block), incur smaller delay, and have smaller receiver complexity (typically, block-by-block signal processing algorithm complexity grows in a quadratic or cubic manner with the block size). On the other hand, smaller blocks could incur a significant throughput penalty due to overhead (needed for various functions including guard sequence, synchronization, training, etc.).

Concerning the carrier frequency f_c, the current trend is toward higher f_c where more RF bandwidth is available, the antenna size is smaller (at the same radiation efficiency), and the antenna spacing requirements (to ensure independent fading) are less stringent due to decreased wavelength. On the other hand, the main challenges in migrating toward higher f_c are the higher costs of manufacturing reliable

RF components, the increased propagation loss, and the increased sensitivity to Doppler effects.

When selecting the number of transmit/receive antennas, several practical considerations must be taken into account, as described next. Under strict delay constraints, achieving high diversity gains (i.e., high reliability) becomes critical in order to minimize the need for retransmissions. As discussed in Section IV-A.2, since transmit/receive diversity gains experience diminishing returns (with respect to an error probability criterion) as their numbers increase, complexity considerations dictate the use of small number of antennas (typically no more than four antennas at each end). Current technology limitations favor using more antennas at the base station than at the user terminal. Moreover as illustrated in Figs. 8–10), less than four transmit and receive antennas are required to achieve a spectral efficiency of 10 b/s/Hz.

For delay-tolerant applications (such as nonreal-time data transmission, e.g., data file transfers), achieving high throughput takes precedence over achieving high diversity and larger antenna arrays (of course still limited by cost and space constraints) can be used to achieve high spatial rate multiplexing gains. Likewise, high-mobility channel conditions substantially impact the choice of system parameters such as the use of shorter blocks, lower carrier frequencies, and noncoherent or adaptive receiver techniques.

Space–time trellis codes [242] use multiple transmit antennas to achieve diversity and coding gains. The first gain manifests itself as an increase in the slope of the BER versus SNR curve (on a log log scale) at high SNR, while the latter gain manifests itself as a horizontal shift in that same curve. At low SNR, it becomes more important to maximize the coding gain while at high SNR diversity gains dominate the performance. For SNR ranges typically encountered on broadband wireless terrestrial links, it might be wise to sacrifice some diversity gain in exchange for more coding gain. For example, using only two transmit and one receive antenna for Scenario A (with delay spread as high as 16 taps), the maximum (spatial and multipath) diversity gain possible is $16 \times 2 \times 1 = 32$! For typical SNR levels in the 10–25 dB range, it suffices to design STCs that achieve a much smaller diversity level (e.g., up to eight) to limit the receiver complexity and to use the extra degrees of freedom in code design to achieve a higher coding gain.

Many challenges still exist at the physical layer on the road to achieving high rate and reliability wireless transmission. We conclude this section by enumerating some of these challenges.

- **Code Design:** Since the introduction of space–time codes in [242], code-design has been an active research area. There are still many open questions in space–time code design especially for noncoherent receivers and fading ISI channels. Another significant open problem is the design of practical space–time codes that achieve the optimal diversity-rate tradeoff derived in [291] and have a practical decoding complexity.
- **Implementation Issues:** These include the development of low-cost integrated multiple RF chains and of low-power parallelizable implementations of the STC receiver signal processing algorithms suitable for DSP and ASIC implementations at the high sampling rates used for broadband transmission/reception. While it is desirable (from an implementation point of view) to use a single receive antenna to preserve the desirable small form factor and low power consumption of the user terminal, adding a second receive antenna multiplies the diversity gain by a factor two (hence doubling the decay rate of average probability of error with SNR at high SNR) and enables interference cancellation. Implementation and manufacturing innovations are needed to make multiple-antenna user terminals commercially viable. Antennas capable of realizing polarization diversity gains [20] further enhance communication reliability by utilizing another source of diversity.
- **Receiver Signal Processing:** While effective and practical joint equalization and decoding schemes that exploit the multipath diversity available in frequency-selective channels have been developed, the full exploitation of time diversity in fast time-varying channels remains elusive. The main challenge here is the development of practical adaptive algorithms that can track the rapid variations of the large number of taps in MIMO channels and/or equalizers. While some encouraging steps have been made in this direction [162], [288], the allowable Doppler rates (which depend on the mobile speed and carrier frequency) for high performance are still quite limited.
- **Standardization Activities:** Multiple-antenna technology has already found its way into several wireless standards including IS-136, W-CDMA [247], and CDMA-2000 [248] and is being discussed or contemplated for many of the broadband wireless systems currently being standardized. These include WLAN 802.11a, broadband wireless access (BWA) 802.16, and wireless personal area networks (WPANs) 802.15. Significant efforts are needed to evaluate the impact of multiple-antenna technology on the performance of these systems at the physical and networking layers for the specific parameters of these systems. One example of such a study for 802.11a is [233].

V. Networking Issues

Since the wireless channel is an inherently shared medium, it becomes important to take advantage of spatial diversity to efficiently and equitably share the resources, namely bandwidth and power. As Section III-B pointed out, use of spatial diversity significantly improves communication rates in multiuser channels. In this section, we investigate the impact of spatial diversity from the perspective of multiuser wireless networks (note that spatial diversity involves *both* multiple-antenna diversity, discussed extensively in Sections III-A and IV, and multiuser diversity, discussed briefly in Section III-B). We examine the effect of spatial diversity on both hierarchical and ad hoc communication topologies (see Sections III-B and V-A). As mentioned in Section I,

link variability in wireless networks makes it important to reexamine information flows and functional abstractions of the traditional IP network protocol stack. We study the impact of spatial diversity on possible interlayer interactions in network protocols.

Traditionally, the datalink layer deals with access techniques for multiple users and scheduling transmissions among users. In wireless networks, the access techniques include a partition of the available (shared) resources and how to resolve contention for the resources. When there is contention to share the medium, a natural question relates to problems associated with fair sharing of the wireless medium among multiple users. Sharing a medium entails specifying: 1) what resources are to be shared; 2) what are the fairness criteria used to distribute resources to the users; and 3) the complexity of implementing the algorithms optimizing the desired criteria. The discussion in Section V-A is centered around the role of spatial diversity in these design choices.

Mobility in the user nodes causes the transmission and interference conditions among the users to vary with time. This variation implies that the wireless network topology (the communication graph $\mathcal{G}_C$ defined in Section III-B) changes over time. Communication with a particular user depends on acquiring information about this topology change in order to route information to the user. In hierarchical networks (Section III-B), this requires keeping track of which access point would be able to reach the desired destination (a topic discussed in Section V-B). In ad hoc networks, spatial diversity can also be exploited by using other users as relays to convey information to the final destination. We have seen in Section III-B that use of relays is critical to enhancing the per-user throughput of the wireless network. In Section V-B, we discuss issues associated with spatial diversity and routing.

A wireless network could be the last hop in a path that has both wired and wireless components. In Section V-C, we briefly study how spatial diversity impacts wireless connectivity to a hybrid network. Ultimately, it is the applications that run over wireless networks that finally determine the impact of spatial diversity. In Section V-D, we examine how diversity can be used in conjunction with applications (such as speech, audio, or video) and what properties need to be designed into the applications themselves to make them suitable for the wireless medium. In Section V-E, we summarize some of the effects of spatial diversity on the functional abstractions of the networking layers. We briefly examine the "vertical" impact of energy-limited communication on all of the networking layer as an example of these effects.

A. Wireless Media Access Control

The problem of how to let users access and share the wireless medium is central to its efficient use. Factors that determine efficiency include the following.

- Organization of the communicating nodes. For example, nodes could be in a hierarchical set-up with access points and terminal nodes, or they could be in an ad hoc communication topology.
- Simultaneous access to the wireless channel. Should nodes be allowed to access it in a noninterfering manner, or is multiuser interfering communication allowed?
- Fair resource distribution given contention. When several users attempt to use the same channel resource, how do we equitably distribute this resource among the contending users?

We investigate these issues in Section V-A.1, Section V-A.2, and Section V-A.3, respectively.

1) Communication Graph: Following the communication graph structure defined in Section III-B, we broadly classify wireless node topologies into two categories: hierarchical and ad hoc. Note that in both cases, the nodes could be further organized into "cells" [237] where users geographically close to each other share the wireless spectrum and the frequencies can be reused in other distant geographical locations (see, for example, [213], [277]).

The most commonly deployed node topology is the hierarchical structure (it is the one implemented in today's cellular wireless communication networks). Here, the base station acts as a wireless access point with which all other nodes communicate, and it performs centralized functions. The other nodes, typically wireless end-units, have simpler distributed functionalities. The advantage of this node topology is that the base station can be built with much higher capabilities and perhaps be made aware of other parts of the network as well. This allows all the complexity of the network to be concentrated in the base station and the end-units can be much simpler. The main disadvantage of such a network is that the base station could be a bottleneck and, therefore, its capabilities would need to scale with network size.

An alternative is the ad hoc network where there is no clear distinction between infrastructure and users. This structure has been built into practical wireless data networks such as IEEE 802.11 and Bluetooth. The advantages of this structure is that bottlenecks can perhaps be avoided by intelligent routing through relays. However, given the distributed nature of communication, such a communication graph structure needs to be self-organizing and network protocols need to be redesigned. Throughout Section V, we discuss the role of spatial diversity for both hierarchical and ad hoc wireless network design.

2) Access Techniques: The access techniques refer to the methods by which users establish communication to each other through the shared medium. This can be viewed both in the context of hierarchical networks and ad hoc networks. We discuss only two forms of access techniques, many-to-one (or multiple access) and one-to-many (or broadcast). One motivating example of these access techniques is in hierarchical networks, where multiple access can be viewed as the uplink channel (mobile end-units to base station) and broadcast can be viewed as the downlink channel (base station to mobile end-units).

In data transmission systems, the times at which users might need to transmit is random and, therefore, their attempts to access the wireless medium might occur aperiodically. Access techniques that allow users to access the medium when they need it are broadly termed *random access* [159] techniques. The access method depends on whether we are allowed a centralized system which can control which users transmit or a distributed scheme where the users have to develop mechanisms to access the channel. Our focus will be on centralized techniques where the central station controls which users can be given access to the medium and, therefore, avoids "collisions" between intended transmissions. The distributed schemes will be briefly described while discussing user contention (for a detailed discussion of such techniques, see, for example, [159, Ch. 7]).

The wireless channel naturally gives rise to three basic degrees of freedom: time, frequency, and space. Almost all traditional access techniques involve partitioning the degrees of freedom of the channel between simultaneously communicating nodes. This can be done through assigning time slots as in TDMA, or through signature codes as in code-division multiple access (CDMA), or through frequency slots as in frequency-division multiple access (FDMA) and combinations of these techniques. In the case of multiple-antenna communication, another access technique is possible by realizing that the communication nodes may be at different geographical locations and could have different channel signatures at the receivers. This difference can be exploited to efficiently decode simultaneously transmitting users. From an information-theoretic point of view (see [70, Ch. 14]), certain points in the optimal rate region are achievable through partitioning the channel using the above techniques. However, the entire rate regions may not be achievable by just noninterfering channel partitioning and more sophisticated techniques might be needed [70].

Many of the 3G wireless systems are based on CDMA technology which offers inherent frequency diversity due to spreading and is also suitable for interference-limited environments. In CDMA, the users' signatures are not restricted to be orthogonal to each other and, therefore, the users may interfere with each other while accessing the wireless channel. Since different users may experience very disparate channel conditions, the power transmitted by each user is controlled in order to restrict the amount of interference they cause to other users. Power control schemes which allow for a desired level of interference in a distributed manner have been extensively investigated (see, for example, [28] and the references therein). However, the receiver still needs to deal with multiple interfering signals. One method is to treat the undesired users as noise and proceed with single-user decoding. This receiver has the advantage of simplicity, but does not perform well when power control is not accurate. More sophisticated multiuser detection techniques might be appropriate in these cases [268]. The issue of combining power control (and resource allocation) in CDMA multiple-access systems has also been explored in detail (see, for example, [156], [166]). The problems in CDMA access have been extensively covered in other literature (see, for example, [268], [275] and references therein) and, therefore, are outside the scope of this paper.

Techniques based on FDMA are now becoming important through the use of OFDM in wireless networks [62]. Since OFDM allows a flexible and dynamic allocation of frequency bins to different users, it is particularly attractive in wireless data networks over frequency-selective channels. The advantage of OFDM is that changing the frequency allocation just translates to changing the multicarrier tone assignments to the users [82], [217]. It can also utilize spatial diversity, since the fading could occur at different frequency bins for different users. Therefore, OFDM would allow such users to share the channel efficiently by accessing it with the right frequency bins.

For CDMA, the code signatures are assigned to the different users in order to reduce interference between users simultaneously accessing the channel. However, spatial (geographical) diversity also implies that different users would see almost independent channel conditions. One way to exploit this is to allow multiple users to simultaneously transmit information and have multiple antennas at the receiver utilize spatial diversity when decoding the information. Therefore, since nature gives disparate channel signatures to the different users, this could be used to suppress interference using multiple antennas (for example, see [87] and references therein). In the sequel, we focus our attention on the use of multiuser spatial diversity (i.e., different channel signatures for different users), combined with space–time codes and how this structure can be utilized (without use of spreading codes) for multipacket reception. One main idea is to utilize STC to introduce structure into the user transmissions in order to guarantee error performance in terms of diversity order (see Section III-C). We also discuss other multipacket reception strategies where the receiver is allowed to combine successive packet transmissions of the users.

In the downlink (broadcast) channel, an additional resource that needs to be partitioned is the transmitted power. Since there is a total power budget that needs to be shared among the different transmissions, we also need to allocate this resource equitably between the users. Aside from the access and resource allocation problems, another constraint that naturally arises is the amount of allowable cooperation amongst users. In multiple-access channels (uplink) the users would be distributed and, therefore, the transmission techniques should not allow user cooperation, since they are not colocated. However, the receiver can collate information from all communication nodes and do joint processing. The situation in the downlink (broadcast) channel could be different, since the information is to be disseminated from one source. We could allow cooperation between the transmissions to the users, but then the users have to decode in a distributed manner. Thus, topological and access questions form a complicated mosaic that forces design choices on the wireless system engineer.

Multipacket Reception: Spatial diversity implies that disparate channel conditions are realized for different users.

As mentioned earlier, if we use multiple antennas, this disparity can potentially allow multiple users to transmit simultaneously while still being able to decode the users information efficiently at the receiver. In such an environment, it has been shown that K cochannel users, where each user is equipped with M_t antennas, can be detected with M_t-order diversity gains, if the receiver is equipped with $M_r = M_t(K-1)+1$ antennas [281]. We emphasize that this result assumes that the interfering signals are not correlated. However, the number of receive antennas can be reduced if the rich structure of STBC (see Section IV-A.3) is exploited: only K receive antennas are needed to provide M_t-order diversity gains and suppress $K-1$ cochannel space–time coded users. Additionally, after interference cancellation, transmitted symbols can still be recovered with space–time diversity gains. In a sense, STC can be thought of as inducing space–time signatures in a manner similar to CDMA (but without bandwidth expansion) and, therefore, multiuser detection techniques [268] may be utilized.

We will illustrate this technique with the case where there are two cochannel users each equipped with two transmit antennas implementing an Alamouti-type space–time block code (see Section IV-A.3). For the flat-fading channel, [188] developed a simple interference cancellation (IC) scheme. Using two receive antennas, [188] shows that, under minor conditions, the diversity order is the same as that of the Alamouti code (i.e., as if only one user was transmitting, and the receiver was equipped with only one antenna). Effectively, [188] showed that it is possible to *double* the system capacity (in terms of number of users) by applying linear processing at the receiver *without* sacrificing space–time diversity gains. Furthermore, [234] showed that: 1) for any STBC based on a complex orthogonal design for more than two transmit antennas per user, IC of two cochannel users can be carried out with only two receive antennas; 2) $K > 2$ cochannel users that use the Alamouti code can be detected with space–time diversity gains with only K receive antennas; and 3) low-complexity IC schemes for frequency-selective channels can be designed for both OFDM and single-carrier transmissions. To illustrate how the rich structure of STBC facilitates IC, we provide a brief overview of IC for the Alamouti-type code in frequency-selective channels.

Multipacket Reception for STBC in ISI Channels: Here, we present a joint equalization and interfence cancellation scheme for STBC assuming the single-carrier FDE-STBC scheme given in (38). We start from the single-user case, where at the receiver end, the CP part of each received block is discarded to eliminate IBI. The resulting length-N blocks are then processed in pairs where they are first transformed to the frequency domain using the FFT, resulting in the two blocks

$$\underbrace{\begin{pmatrix}\mathbf{Y}^{(k)}\\ -\mathbf{Y}^{*(k+1)}\end{pmatrix}}_{\mathbf{Y}} = \underbrace{\begin{pmatrix}\mathbf{\Lambda}_1 & \mathbf{\Lambda}_2\\ -\mathbf{\Lambda}_2{}^* & \mathbf{\Lambda}_1{}^*\end{pmatrix}}_{\mathbf{\Lambda}} \underbrace{\begin{pmatrix}\mathbf{X}_1^{(k)}\\ \mathbf{X}_2^{(k)}\end{pmatrix}}_{\mathbf{X}} + \text{noise} \quad (54)$$

where $\mathbf{X}_1^{(k)}$ and $\mathbf{X}_2^{(k)}$ are the FFTs of the information blocks $\mathbf{x}_1^{(k)}$ and $\mathbf{x}_2^{(k)}$, respectively. Since circulant matrices are diagonalized by the FFT, intercarrier interference is eliminated and $\mathbf{\Lambda}_1$, $\mathbf{\Lambda}_2$ are diagonal matrices containing the FFT coefficients of the underlying wireless channels. To eliminate inter-antenna interference, the linear combiner $\mathbf{\Lambda}^*$ is applied to $\mathbf{Y}$. Due to the orthogonal structure of $\mathbf{\Lambda}$, a second-order diversity gain is achieved. Then, the two decoupled blocks at the output of the linear combiner are equalized separately using the MMSe frequency domain equalizer (FDE), which consists of N complex taps that mitigate intersymbol interference. Finally, the MMSE-FDE output is transformed back to the time domain using the inverse FFT where decisions are made.

Equation (54) constitutes the basis for IC of STBC in frequency-selective channels. The frequency-selective channels are transformed to flat-fading taps on the FFT grid. Consequently, the IC techniques of [188] can be readily applied in the frequency domain (a similar technique, but for only two cochannel users has appeared in [236] for OFDM transmissions).

Note that this technique can be easily extended to more than two cochannel users and other space–time coded orthogonal designs [234]. The form of detection performed above is a simple decorrelating receiver [268] and, hence, ignores the noise, making it suitable only for high SNRs. As in multiuser detection, an MMSE detector can be constructed for the detection in (54). Such an approach has been proposed for the flat-fading channel in [188]. In [85], the ISI multiple-access channel is studied when each of the K users apply a time-domain STBC (for $M_t = 2$) suitable for ISI channels. Using the algebraic properties of the TR-STBC, it is shown that an optimal receiver can achieve the maximal diversity order of $2M_r\nu$ when each of the K users transmit at "full rate" (i.e., a rate of b b/s/Hz, if they are using a fixed alphabet size of 2^b). This has two desirable properties. One is that the each of the users achieve an extremal point in the rate-diversity tradeoff. Another is that the users achieve a performance equivalent to not sharing the transmission medium with the other users, i.e., achieve single-user performance. Furthermore, linear multiuser detectors which utilize the structure of the particular STBC are also studied. In summary, all the sophisticated multiuser detection techniques, including multistage decoding [266], iterative detection [278] etc. can be brought to bear in detecting the symbols in (54), once we utilize the structure of the space–time coded signal.

Packet Combining Techniques: Most coding techniques described in Section IV-A code across a small number of transmission blocks due to delay constraints. However, when the channel realization is bad, such techniques could result in unrecoverable errors (i.e., errors that remain even after an outer code is concatenated with an inner space–time code [189]). In such cases, an error detecting code declares an outage and there needs to be a recovery process for the transmitted packet.

One practical method for such recovery is through an automatic repeat request (ARQ) mechanism [54], [134], [155] where the receiver (through a feedback mechanism) notifies the transmitter that the packet was in error. Therefore, there

can be several recovery techniques when a packet error is detected, depending on the possible actions that either the transmitter or the receiver may take (which basically pertain to hybrid ARQ protocols [171]). Upon receiving information that the packet was in error, the transmitter could retransmit, either using a more powerful channel code, or may contain coding information related to the previous packet (using, e.g., incremental redundancy [17], [131], [171], [264]). At the receiver side, all received packets may be combined using signal processing and joint decoding techniques (as opposed to systems without packet combining, where only the last received packet is used to retrieve the transmitted information symbols). Adaptive packet/code combining schemes (with variable channel coding in the subsequent transmissions of a specific packet) can be interpreted as a mechanism which improves throughput by tracking the channel conditions over a single-user link. Such hybrid ARQ techniques can also be combined with multiuser packet reception mechanisms such as space–time block codes with interference suppression as described above.

User Contention: As mentioned in the beginning of this section, our focus has been on centralized access techniques. Here, we briefly discuss distributed access techniques to the wireless medium where the users attempt to access the medium without a central station controlling when they can access it. There are several such standard protocols[41] for distributed access (see, for example, [159, Sec. 7.5]); however, there is always a possibility that the transmitted packets may "collide," i.e., transmissions could overlap causing errors. It is possible to use spatial diversity to allow multiple users to access the medium simultaneously, and still be able to decode the transmitted packets. However, when many different users contend for the resources, even multipacket reception techniques might be overwhelmed and the decoded packets might be in error. In such a situation a recovery mechanism needs to be established. For example, the users could back off their transmission rates and retransmit at a randomly chosen time. On the other hand, multiuser packet combining techniques attempt to improve throughput (see, e.g, [251], [253]) by a joint design of the physical layer and the link layer. The basic idea is that collided packets are not discarded at the receiver—instead, these collided packets are combined with newly transmitted packets (which have been encoded differently). Here the link layer is aware of the source separation capabilities of the physical layer. Consequently, the users do not back off to reduce the number of collisions. Instead, packet collisions are allowed, since the physical layer is capable of retrieving the transmitted packets. Stability of these protocols has also been studied (see, e.g, [89]). There are several protocols proposed for transmission on collision channels which are outside the scope of this paper (see, for example, [50], [179] and references therein).

If centralized control is available, the central station will be able to schedule transmissions depending on the user requirements and channel conditions and, therefore, utilize the wireless medium more efficiently. We illustrate this idea through downlink scheduling algorithms in Section V-A.3.

3) Resource Allocation: Resource allocation has a strong interaction with physical layer transmission techniques and, therefore, it is placed in the datalink layer [159]. The channel resources are distributed between the users in order to satisfy certain QoS criteria such as rate, delay, etc. Allocation techniques could be either fixed or dynamic. Fixed allocation techniques are suitable for circuit-switched networks, where each user is allocated a guaranteed fixed channel or "circuit" for their communication needs. Such an allocation might be inefficient when there is variability in the communication needs of the users, which might occur in data applications. For such applications, dynamic allocation strategies utilize the wireless medium more efficiently. The geographic location of the users and their mobility determine the *diverse* channel conditions of the nodes. This spatial diversity of the users implies that one user could obtain better performance (in terms of the QoS criterion) for the same allocation of resource as compared to another user. Given this diversity, the question becomes: how to distribute the resources equitably, i.e., what are the fairness criteria? what information is available to the resource allocation scheme? and what access technique is used by the users? It is important for any resource allocation scheme to take this spatial diversity into account and indeed exploit it if possible. The methods for sharing the resources critically depend on the QoS criteria as well as the information available to the link layer. We first discuss the possible control information that the resource allocation scheme could have access to. Then, we review the various QoS criteria that are of interest in wireless networks. We define a *scheduler* as a resource allocation algorithm that decides how and when to allocate the resources to users. The context of the scheduler is given in Fig. 25.

Input Information: There are two forms of control information that the resource allocation scheme could have access to. One would be about the QoS required by a request from a user and the other would be about the channel conditions experienced by the user. The former control information is commonly seen in wireline networks as well, but the latter is more particular to wireless networks. In a wireless environment, unlike traditional resource allocation scenarios, one must take into account the channel state in order to provide reasonable QoS. Therefore, existing wireline resource allocation algorithms can not be directly applied to manage wireless networks, since they have unique characteristics, such as location-dependent data rates and channel errors. As discussed in Section IV-B.1, channel estimation is an integral part of the physical layer and, therefore, this information can also be used for resource allocation. In addition, we could have access to information regarding the various requests in the system, including user request sizes and request arrival

[41]For example, in 802.11 networks, a multiple access collision avoidance (MACA) protocol is used for distributed access [159]. Here the transmitter sends a *request to send* (RTS) to the receiver and, therefore, silences its neighbors (they do not transmit for a period of time). If the receiver is in a position to receive the transmission, it sends a *clear to send* (CTS), which silences the neighborhood of the receiver. This allows the transmitter to communicate with the receiver with minimal interference.

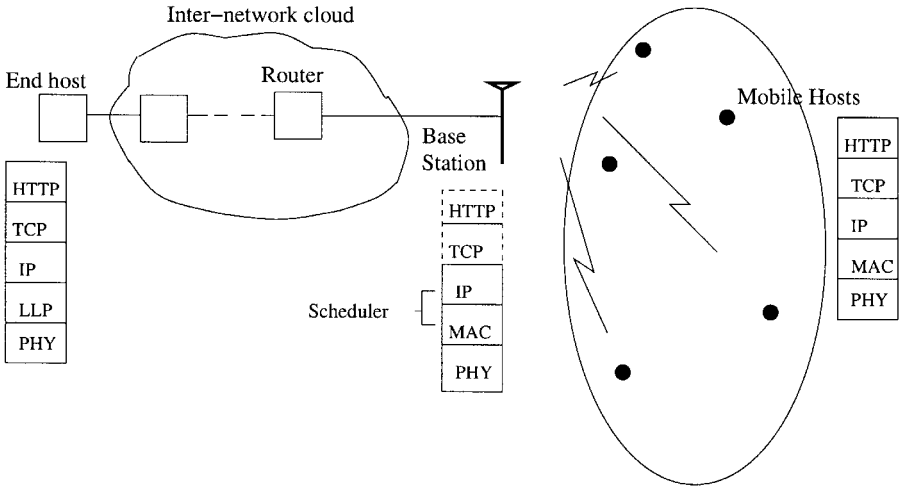

Fig. 25. Context of the downlink scheduling algorithms.

times. For example, this could occur by having a proxy which interprets packets up to the transport layer in order to deal with problems of hybrid networks (we delve into this issue in more detail in Section V-C). Therefore, the resource allocation scheme may have access to job request information such as size, type, request times, deadlines, etc. Note that we can take advantage of this diversity of user requirements in efficient scheduling schemes.

Allocation Criteria: The QoS criteria can be divided into two categories: one is rate-based and the other is job-based. In the rate-based criteria, the goal is to provide average data rates to users which satisfy certain properties. Therefore, the understanding here is that the flows are long-lived and thus assigning rates to different users is reasonable. In job-based criteria, the requests/flows can have very disparate ranges. Therefore, it would make sense to measure properties on individual flows and optimize criteria based on them.

- *Rate-Based Criteria:* Here, the resource allocation QoS criteria is based on the long-term average rate $\{T_k\}$ provided to the users. There could be several criteria based on the long-term rates. For example, a criterion based on proportional fairness maximizes $\sum_{k=1}^{K} \log(T_k)$ among the class of resource allocation algorithms. There are other criteria such as max-min fairness, utility fairness (which is determined by the utility function associated with the value of a particular rate to a user), etc. [74], [159].
- *Job-Based Criteria:* Here, the requests come in at certain arrival times a_i, and we have information about the size s_i (say in bytes). The goal is to schedule data delivery in order to optimize various QoS guarantees based on the job requests. These QoS criteria could be related to QoS guarantees per request, or to overall (average) QoS criteria.

Response time is defined to be $c_i - a_i$ where c_i is the time when a request was fully serviced and a_i is the arrival time of the request. This is a standard QoS criterion for a request. *Relative response* is defined as $(c_i - a_i)/s_i$[33]. Relative response was proposed in the context of heterogeneous workloads, such as the web. i.e., requests for data of different sizes (thus, different s_i). The response time metric is skewed toward large jobs, since jobs with large service times also tend to have large response time. On the other hand, the relative response metric is independent of size making it more fair to all job classes. Since data requests in the emerging data systems and applications would very likely be heterogeneous, relative response is an attractive metric to investigate. The above criteria relate to guarantees per request; we could also give guarantees only over all requests. For example, the overall performance criterion for a set of jobs could be the l_∞ norm, namely, $\max_i(c_i - a_i)$ (i.e., max response time)[42] or $\max_i (c_i - a_i)/s_i$ (i.e., max relative response). Another criterion is the l_1 norm where $\max_i$ is replaced by $\sum_i$. Since the l_1 norm for n jobs is equivalent to the average, the l_1 norm for the flow and the stretch are called average flow and average stretch, respectively.

Goal: For a given input, access schemeand desired QoS criterion, our overall goal is to produce a valid schedule which optimizes these metrics.

To have a comprehensive discussion of this goal itself is beyond the scope of this paper. Instead, we illustrate a small number of approaches to achieving this goal, focusing on the spatial diversity.

Typically, data traffic is asymmetric and we expect wireless terminal nodes to download larger amounts of data than uploading them to the access point. Therefore, the resource allocation problem becomes more crucial for transmission from the access point to the wireless terminal nodes making the downlink the bottleneck in wireless communication. This motivates our examination of downlink (broadcast) scheduling in a hierarchical network as given in the sequel. For the most part, we focus on a form of TDMA where all the resources are given to one user, but when they are scheduled depends on the QoS criterion. In this context, we discuss scheduling algorithms that critically use spatial diversity to optimize rate-based and job-based QoS criteria.

The main task of the scheduling algorithm is to determine which users access the channel and how to allocate power

[42]The term response time is sometimes called flow time in literature.

and bandwidth to them. We discuss the algorithms based on the category of the QoS criterion chosen.

Scheduling With Rate-Based QoS Criteria: For rate-based criteria, the goal is to give guarantees on the average rate obtained by the different users. Rate-based scheduling ideas such as generalized processor sharing and weighted fair queueing [78], [198], designed for wired networks, are not immediately applicable to the wireless case. All these algorithms assume that the rates achievable by the users are symmetric and constant over time and account for variations through bursty arrival processes. Therefore, when there is variation in user rates due to location-dependent channel errors arising from spatial diversity, the fair queueing algorithms need to be modified by taking this into account. In [178], a simple on–off rate model (with constant rate when transmission to a user is possible) is used for the rate variation in time for users. Therefore, by maintaining lead and lag accounts for the different users with respect to the "fair" rate predicted by [78], [198], the users are given less or more resources (time) for transmission. A discussion of related work about "fair" queueing techniques in wireless data networks can found in [192].

Wireless networks can support multiple rates for users [122], such as those proposed in the 3G wireless networks [34], [191]. Therefore, one can define rate optimization problems with criteria varying from maximizing overall throughput to providing minimum rate guarantees for all users. For example, in [15] a throughput (rate) maximizing algorithm for a multirate wireless network is studied. Recently, there has been interest in utilizing the multiuser diversity inherently available in wireless channels. This is motivated by the information-theoretic result, briefly discussed in Section III-B, that a form of TDMA, where only the user with the highest throughput transmits information, is throughput optimal. If the channel state for users in a wireless network varies with time, t, then the achievable throughput for a given outage level (see Section III-A) $R_k(t)$ to user k would also vary accordingly. Therefore, the idea is to schedule users when their channel state is close to the peak rate that it can support. There is inherent unfairness if only the users with the highest rates $R_k(t)$ are scheduled. In order to avoid this, the scheduling algorithm implemented in IS-856 [34] keeps track of the average throughput $T_k(t)$ of each user in a past window of length t_c. At time t, the scheduling algorithm transmits to the user with the largest $R_k(t)/T_k(t)$ among the active users. The average throughputs are then updated given the current allocation. This scheme is shown to have certain optimality properties related to proportional fairness [273]. This scheduling algorithm relies on the rates $R_k(t)$ to vary in time. But this assumption can be violated when the channels are constant or are very slowly time-varying. In order to artificially induce time variations, the authors of [273] propose to use multiple transmit antennas and introduce random phase rotations between the antennas to simulate fast fading. This idea of phase-sweeping for multiple antennas has been also proposed in [139], [279] in the context of creating time diversity in single-user systems. With such artificially induced fast channel variations, the same scheduling algorithm used in IS-856 (outlined above) inherently captures the multiuser spatial diversity of the network. In [273], this technique is shown to achieve the maximal diversity order (see Section III-C) for each user, asymptotically in number of (uniformly distributed) users.

In a heavily loaded system (large number of users) and where there is a uniform distribution of users, the scheduling algorithm proposed in [273] is attractive. However, for lightly loaded systems, *or* when delay is an important QoS criterion, its desirability is less clear. Given that the technique proposed in [273] is based on a rate-based QoS criterion, it cannot provide delay guarantees for the jobs of different users. This motivates the investigation of scheduling algorithms for job-based QoS criteria discussed below.

Scheduling Algorithms for Job-Based QoS Criteria: In job scheduling problems, there are two basic variants, namely *offline* or *online*. In the offline case, all the request arrivals are known ahead of time; in this case, the arrival times are also known as the *release times* since no request may be serviced before its "arrival" time. The offline case is of theoretical interest and is mainly useful to quantify the benefit to be accrued from scheduling. In the online case, requests arrive over time and the scheduling algorithms do not know the future requests when processing the current load. The relative performance of an online algorithm is measured in comparison to the offline case. In standard scheduling terminology, requests may be *preempted* (i.e., stop processing a request, process other requests, and resume the original request) or be *migrated* (i.e., requests may be served in one or more codes).

Job scheduling is very popular in the context of processor scheduling, and various algorithms have been proposed for different QoS metrics such as completion time, maximum response time, and weighted average response time [157]. In wireless networks, job scheduling has been addressed in the context of downlink *broadcast* scheduling [3].

As mentioned earlier, next-generation wireless networks could support multiple transmission rates depending on the channel conditions. For example, in the simplest case where the channel is quasi-static over a transmission block, for a user experiencing a channel $\mathbf{H}$, the rate achievable could be $\log(|\mathbf{I} + (P/M_t\sigma^2)\mathbf{H}\mathbf{H}^*|)$ (see Section III-A). In a recent work, downlink unicast scheduling in CDMA systems was studied [153]. Here, a linear rate model is assumed for the physical layer instead of a concave rate function with respect to power allocated. Assuming an accurate communication-theoretic model for the physical layer achievable rates (as described in Section III-A), job scheduling algorithms are proposed and analyzed for various QoS criteria in [32].

OFDM is an attractive transmission technique for broadband channels (see Section IV-A.5). In downlink transmission, multiple users can be served using an OFDM frame, which effectively allocates different users to separate subsets of OFDM subcarriers. Given the spatial diversity of the users, they experience disparate channel conditions on the

different OFDM subcarriers. Therefore, one can formulate the scheduling problem as one of allocating disjoint subsets of these parallel channels to the different users, and also distributing the total transmit power between these subsets. For different amounts of power allocated, the rates achievable would also vary depending on the channel conditions. In [32] such scheduling problems are formulated with respect to job-based criteria, such as minimizing the maximum delay. The main result in that paper shows that simple scheduling algorithms can be shown to be very efficient. Here the efficiency is with respect to the competitive ratio of the online algorithm to an offline optimal method. For example, if the QoS of interest is maximum delay, then simple algorithms can be shown to perform within a constant from the optimal, if resource augmentation (overprovisioning) is allowed. The analysis also bounds the amount of overprovisioning needed to guarantee performance of the online algorithms. The resource-augmented competitive analysis is a useful tool that is employed here to bound the worst case overprovisioning needed to the guaranteed QoS of choice.

Impact of Spatial Diversity: Both multiantenna and multiuser diversity give significant improvements in link-layer user throughputs. But the two effects seem to work in different ways: on the one hand, using multiple-antenna techniques, we can obtain greater reliability (and data rates) by "smoothening" the channel variations; on the other hand, the multiple-user diversity utilizes the channel variability across users for increased throughput. A natural question is, which form of diversity is more important in the context of multiuser wireless networks? Perhaps this question is ill-posed (or incompletely posed), as the answer might be, "It depends." STC is inherently an open loop technique, and increases reliability of delay-constrained transmission. In fact, if (partial) channel state information is available at the transmitter, then techniques like antenna selection diversity (which transmits all the information through the antenna with the "best" channel conditions) is the direct analog of the multiuser diversity used through scheduling best users. The channel variability of multiuser diversity techniques (e.g., [273]) rely on delay-tolerant applications and accurate low-delay feedback about channel conditions. So perhaps for low-delay applications requiring high reliability, space–time codes (perhaps combined with job scheduling algorithms) might be a good design choice. Moreover, in the presence of high mobility, the ability to schedule the "best user" (for multiuser diversity) when there might be scheduling (and channel feedback) delays, may require reliable data transmission, calling for a combination of multiuser diversity scheduling with space–time codes. Finally, multiple-antenna techniques can also be used for multipacket reception, as seen in Section V-A.2. The interaction between multiple antenna and multiuser diversity is still an active research topic.

B. Mobility and Routing

The main role of routing is to find an efficient path on the communication graph (see Section III-B) to deliver information to the end user. User mobility, which is a characteristic of wireless networks, changes the graph and, hence, affects routing. Even if sophisticated physical layer techniques provide HDRs over a single-hop link, the problem of providing an uninterrupted wireless connectivity is not completely solved. As the nodes move, the network needs to track their movement, route the data packets accordingly, and make arrangements so that QoS guarantees can still be provided. We also saw in Section III-B that mobility can be used to induce multiuser diversity (Theorem 3.7) in the context of ad hoc networks. In hierarchical networks as well, mobility allows nodes to move out of bad channel conditions and provides time-diversity in physical layer transmission. Moreover, mobile nodes could have better channel conditions to another base station and, therefore, might want to switch to that base station (called a "handoff" [213]). These handoffs from one base station to another should be tracked in order to reach the user. In this section we investigate how spatial diversity through mobility affects routing in wireless networks.

One abstraction of routing is obtained by defining the underlying communication graph (see Section III-B) and letting the routing protocol work on this graph. This abstraction works quite well in wired networks, where the time scale over which the graph structure changes is quite slow. In wireless networks, mobility can cause the node topology (and, hence, the graph structure) to change quite rapidly. Therefore, one way to envisage the "layer-interaction" is by conveying the changing graph to the network layer [159] for the routing to be done on this graph. In ad hoc networks, routing can also be helped by observing that this graph can be endowed with geometry. We describe how the protocols might use this observation in Section V-B.1.

However, there could be a deeper interaction in wireless networks. We demonstrate this using ad hoc wireless networks. Here the graph itself is malleable in that a node can in principle increase its transmit power to communicate with its destination directly. As seen in Section III-B, such a strategy is disruptive in terms of overall network throughput efficiency, and it was demonstrated that a better strategy (from a long-term per-user throughput point of view) is nearest neighbor multihop transmission [129]. In the presence of mobility, the user node topology changes and this can be utilized to change transmission strategies (and, therefore, the graph structures) based on routing (source-destination pairs). Hence, in these cases, there are strong connections between channel conditions and routing protocols which is based on channel fading and relative geographical locations of the nodes, i.e., the multiuser spatial diversity effect. Hence, power control is a "vertical" issue that causes a strong interaction between the layers. Another example is the case of energy-limited communication, where the battery on certain nodes could be close to being exhausted and this can define a further constraint on the communication graph (see, for example, [219] and references therein). In hierarchical schemes as well, the strong connection between routing and spatial diversity can be seen through the idea of *soft-handoff* (also called macrodiversity, for example, see [133] and references therein). Here the mobile node is in

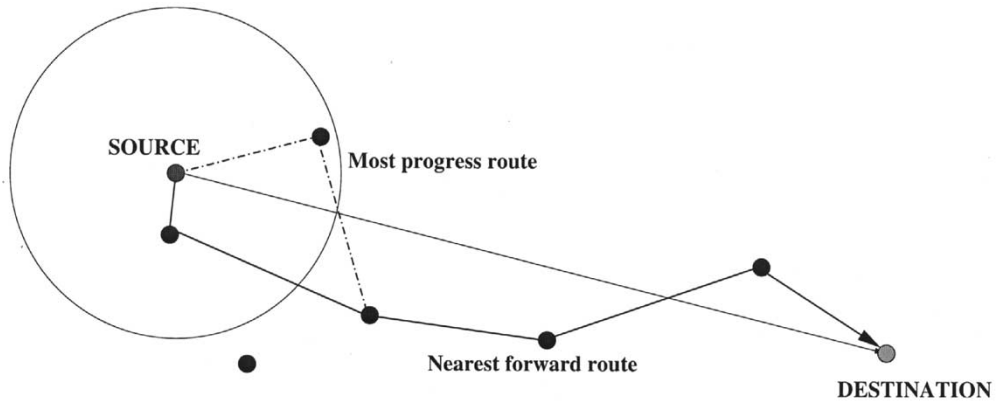

Fig. 26. Illustration of greedy position-based routing algorithms.

simultaneous contact with more than one base station and, therefore, utilizes spatial diversity to improve performance. Therefore, routing information to the end user can utilize this form of spatial diversity as well. An abstract way to think about this problem is by assuming that for a given node topology, the access technique implements different graphs which could be realized. There is a close interaction between the routing and construction of this graph in that we can choose between the "best" realization of the graph, where optimization depends on the criteria used for routing (minimum number of hops, maximal number of simultaneous transmissions, minimum energy, etc.). Mobility of the nodes induces different instantiations of the communication graph and routing can choose which instantiation to use. This is another way to think about the multiuser diversity through relays that was introduced in Section III-B.

The above discussion demonstrates that the division of functionalities in the network protocol stack becomes unclear in wireless networks. The simple abstraction of defining the communication graph and routing on it may not be a complete solution in wireless networks. Spatial diversity plays an important role in routing over both ad hoc (Section V-B.1) and hierarchical (Section V-B.2) wireless networks. By observing the impact of spatial diversity, perhaps the functional abstraction needs some rethinking and is a topic of ongoing research.

1) Ad Hoc Networks: The main classifications of routing algorithms in ad hoc wireless networks are *topology* based algorithms and *position* based algorithms [181]. In topology based algorithms, the communication graph is constructed based on the links that exist in the network. A variety of routing algorithms have been developed based on route quality and degree of mobility [181], [204], [243].

Position-based routing utilizes additional information about the node locations to assist the routing protocol. Here the observation is that the communication graph can be conveyed by utilizing the positional information of the nodes.[43] In order to describe and update the communication graph, instead of transferring the connectivity information, the users can utilize node locations to construct it. The main task of the position-based routing protocol is divided into two parts. One task is to maintain the locations of destinations, i.e., a location service. The other task is a routing protocol based on the locations. Both of these tasks are critical and clearly interdependent with several techniques proposed to efficiently perform them (see [181] for a survey of these techniques). When there is large-scale mobility, the positions of the nodes need to be updated and several of these techniques get overwhelmed by the rate of these updates. Spatial diversity through user mobility is an important consideration and, therefore, it is unclear if this task division is appropriate in high-mobility environments.

[43]This is accurate when only path loss is considered (see Section II).

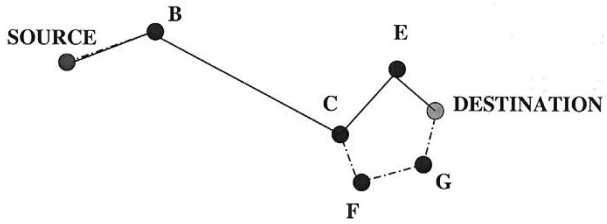

Fig. 27. Geographic routing algorithm.

A node can use GPS location service to determine its own location. In order to find the locations of the other users, one would need a location service which could be either centralized or distributed [181]. A centralized repository of location information could be maintained, which creates a hierarchy of nodes. Also, the nodes need to know the location of this centralized resource. In a distributed system, nodes carry positional information about some or all of the other nodes. This distributed system needs to be consistent and easily accessible. In [181], a survey of several such proposals is described. Due to mobility, the positional information needs to be updated. One proposal to deal with this problem is described in [30] using the "distance effect" where it is observed that one can make imprecise routing decisions far away from the destinations, because only when the packet is close to its ultimate destination does more precise routing (and, therefore, location information) become important.

In order to illustrate the ideas in position-based routing we describe two algorithms illustrated in Figs. 26 and 27. In Fig. 26, there are multiple paths from source to destination. Given the positional information of the nodes, one intuitive

strategy is to try to make the most progress toward the destination in each hop. Therefore, given a transmission range of the node (illustrated in Fig. 26 by a circle around the source), the next hop is chosen as the relay closest to the destination within the transmission range (called most progress route in Fig. 26 and illustrated with a broken line). Though this reduces the number of hops, it might cause large interference to other communicating nodes. An alternate strategy is to find the nearest neighbor (to the source) which makes progress toward the destination and, therefore, reduces the interference caused. This strategy is shown in Fig. 26 as nearest forward route with a solid line. Both of these strategies were proposed in the 1980s (see [181] for references) and modifications have been proposed to these basic ideas. One major problem that occurs with these greedy strategies is when the packet reaches a local extremum, i.e., there are no other nodes (within its communication range) closer to the destination than itself.[44] In this case, recovery mechanisms need to come into play such as transmitting the packet to a node which makes least negative progress toward destination, or to transmit to a randomly chosen intermediate node to allow the route to get out of a local extremum. These and several other recovery processes have been suggested in literature (see [181] for the survey).

A variation on the above positional routing algorithms has been suggested in [147] where the source might know the route to a node close to the destination rather than the destination itself. Therefore, as shown in Fig. 27, the source node might know the route to location F which is the closest node to the destination that it knows how to route to. Then, it sends the packet to the next hop B on this route. Along the way, the route can get modified if the relays know a better route to the destination. For example, in Fig. 27, when the packet reaches node C, it knows that the better route to destination is through node E rather than through node F (shown in broken line). Therefore, the route can get more accurate when the packet gets closer to the destination. A similar idea is proposed in the DREAM framework of [30], where not just the route but also the positional information about the destination becomes more accurate when the packet gets closer to the destination.

The problem of routing in the presence of significant node mobility has not been solved and would constitute an important step toward realizing the possible multiuser diversity effects that were seen in Section III-B. Here it might be interesting to examine algorithms that establish the communication graphs and routing jointly rather than in two steps. Some recent progress in utilizing mobility to distribute topology information has been reported in [127].

2) Hierarchical Networks: In hierarchical networks, the mobile user is handed off from one subnetwork to another subnetwork (served by base stations) depending on the link conditions to the base stations. In this section, we describe some methods to handle such node mobility (see also [6] for further issues on mobility management).

It is undesirable from a user perspective to have its wireless connectivity interrupted whenever the mobile user moves from one subnetwork to another. In order to connect to the network in IP protocols, a node is identified through an IP address [159]. The routing of packets from a correspondent node to the mobile node (where packets have as destination address the IP home address of the mobile node) is reminiscent of the way that the post office forward letters from an old address to a new address. Mobile IP [205] is the widely used approach for handling mobility at higher layers. It relies on the idea that a mobile node is associated with two IP addresses. The *home address* is an IP address used for identification of the node—this address is assigned to the node in the network location that the node originates from (this is the *home network*). The home address remains valid for long time intervals (i.e., time intervals longer than those the node spends in remote locations). The *care-of* address is an IP address used for routing, and is related to the new physical location the node has moved to; the care-of address remains valid as long as the node is in this remote location.

IP datagrams from the correspondent node to the mobile node are sent through the *home* and the *foreign* agents. The home agent is a router that resides in the home network, and is one of the routers through which IP datagrams would go through had the mobile node stayed in its home network. The home agent is aware of the new location of the mobile node, and forward packets to the new location (the forwarding procedure applies *tunneling*) based on the care-of address of the mobile node. When the packets reach the remote subnetwork, the foreign agent delivers the packets to the mobile node. The foreign agent is a router in the remote subnetwork. Mobile IP specifies the discovery of foreign agents, the authentication of mobile nodes, the registration of the mobile nodes to remote subnetworks, and the updating of information at the home agent [205].

Mobile IP was designed for macro-mobility scenarios, where the mobile node moves from one subnetwork to a remote subnetwork—such a transition entails computational and network overhead, as the home agent needs to be constantly updated about the whereabouts of the mobile node. When the mobile node moves between cells (in a cellular architecture) or between WLANs interconnected together (e.g., across buildings), such an overhead can be substantial, especially if the home agent is sufficiently far apart from the mobile node. To address the micro-mobility scenarios, attempts have been made to minimize the aforementioned overhead, by exploiting the hierarchical structure in the network (see, e.g., [212], [262] and references therein). The main idea is that the home agent does not need to be updated on every handoff if the mobile node moves across the same administrative domain (see also [47]). As a result, intradomain mobility does not reverberate signaling overhead to the home agent and the correspondent nodes, as the signaling overhead is restricted within the administrative domain.

Spatial diversity through soft handoff can improve performance. For example, in [212], with soft handoffs, the mobile

[44]This is of course a problem only when it is the communication range that has caused the local extremum and not when the destination is actually the nearest neighbor to the source.

node can receive packets from two base stations simultaneously. Consequently, packets which have been routed to the old location can still be delivered to the mobile node.

As a final note, we remark that node micro-mobility may also be addressed at the link layer. Learning bridges (see, e.g., [223],[21]) interconnect LANs by discovering routes between nodes (these routes are usually expressed in the form of spanning trees). Then, nodes may move from one network to the other, leaving it up to the bridges to forward data packets to appropriate locations (for scalability issues and methods to address them, see, e.g., [218]). Also, proxy servers may assist in mobility management (see, e.g., [151]).

C. Hybrid Networks

Hybrid networks have heterogeneous links, e.g., both wired and wireless links. Typically, wireless links which are part of a hybrid network, would be the last hop (or first hop) in such a network. For example, when connections are established from a server in the wired Internet to the mobile terminal, interoperability issues arise in maintaining an end-to-end semantic on this link. One of the central philosophies behind IP networks is to develop protocols that are distributed and have end-to-end functionality (i.e., controls do not need to intercept and interpret the protocols except at the source and destination). In this context, the main issues that arise in hybrid networks are: 1) the throughput mismatch between wired and wireless links 2) routing to mobile terminals 3) reacting to packet losses in different parts of hybrid links.

Distributed algorithms (congestion control) for sharing network resources in the wired networks have a rich history [65], [159]. These congestion control algorithms lie in the domain of the transport layer in the network protocol stack and the most widely implemented transport protocol is TCP. Hence, throughput mismatches can be handled through distributed end-to-end methods. The problem of routing for mobile terminals was addressed briefly in Section V-B.2. In this section we focus mainly on the impact of spatial diversity on packet losses in the wireless link.

The main issue is that TCP is designed for the wired environment where losses are mainly due to congestion and packet error rates are small. Therefore, the distributed TCP algorithm implements an aggressive back-off when it experiences packet losses, assuming that it arose due to congestion, and not link failure (as is common in wireless networks). There are two main approaches to increasing the efficiency of TCP in hybrid networks. One is to attempt to reduce the error rate in the wireless channel, using more sophisticated coding schemes, such as the space–time codes as described in Section IV-A. A complementary approach is to inform the sender that the packet loss occurred due to wireless link failure rather than congestion in the wired part of the network. Such a mechanism is called explicit loss notification (ELN) [25]. A combination of the two approaches would result in significant improvement in transport layer efficiency. Therefore, in this section we describe the complementary approaches with Section V-C.1 devoted to impact of STC on TCP efficiency and Section V-C.2 outlining implementations of scalable ELN mechanisms.

1) STC and TCP Throughput: In Section IV we showed that STC has a significant impact on the physical layer performance. Herein we focus on the impact of STC on the transport layer, and, as a case study, we discuss the impact of STBC on TCP throughput in IEEE 802.11a WLANs. At first sight, it appears to be obvious that a method (such as STBC) which improves physical layer performance eventually has a positive impact on the overall network performance. From a practical viewpoint, we want to quantify the gains of incorporating of STBC into wireless networks. It is important to keep in mind that transport layer performance is determined by a complex interaction between physical layer, link layer and TCP. Hence, it could be the case that the STC results in overall marginal network throughput improvements. In this section we trace the effects of STC on the link layer and TCP performance in an isolated 802.11a WLAN, where there are two antennas at the transmitter (and the Alamouti code is used), and one antenna at the receiver.

Several observations can be made based on the extensive simulation results in [233]. First, in all cases, STBC-enhanced 802.11a achieves a particular throughput value at a much lower SNR value than the standard 802.11a. Second, *STBC modify the SNR region under which a particular transmission rate should be chosen*. Third, *STBC increase the transmission range and improve robustness of WLANs*. In other words, with STBC, a particular throughput value can be achieved over a wider range of SNR. Consequently, not only the transmission range can be increased, but also the robustness of the network (in the events of sudden fading) can be improved.

Nevertheless, in an isolated random access wireless network, physical layer performance improvements do not translate to transport layer improvements in a straightforward way. Among the factors which determine the TCP throughput are the network load (i.e., how many nodes attempt to transmit at the same time), and the retransmission policy at the link layer. It is well-documented that for distributed random access protocols, packet collisions are the limiting performance factor. Therefore, in heavily loaded scenarios, the STC impact on TCP throughput is expected to be marginal. On the other hand, in mildly loaded networks (where frame collisions are a relatively rare event), TCP throughput is not only a function of frame-error-rate and physical layer transmission speed, but it is also a function of the correlation among frame errors (see, e.g., [16]), and the ARQ mechanism at the link layer. Errors at the wireless physical layer reverberate across layers and have a negative impact on TCP performance (see [25], and references therein). For bulk data transfers (i.e., sources which transmit data continuously for long time intervals) a rule of thumb is that TCP throughput scales as $1/\sqrt{(\text{packet loss rate})}$ (see, e.g., [180]). In mildly loaded scenarios, the (packet loss rate) is a function of the packet errors over the wireless channel (random collisions constitute a rare event). Assuming that there is no fragmentation of TCP packets into link-layer

frames (this assumption holds for small TCP packets, i.e., on the order of 1500 bytes, such as TCP packets in Ethernet), the (packet loss rate) is equal to the link layer loss rate, which is given by $(\text{frame error rate})^{N_{\text{ARQ}}}$, under the assumption of independent frame losses (N_{ARQ} is the maximum number of retransmissions).[45]

Consequently, when the link layer supports frame retransmissions, most of the catastrophic events (for TCP) could be eliminated. However, there are issues of timer interactions [79] (see Section V-C.2 for further discussion). Aside from this, STBC is expected to improve performance, because: 1) the number of frame retransmissions at the link layer is reduced and 2) TCP will attempt to transmit at a rate equal to the perceived "capacity" of the wireless link. On the other hand, whenever the link layer does not employ retransmissions and does not attempt to "hide" the frame errors from TCP, then STBC yield impressive performance improvements [233]. This is because when the frame error rate is higher than a certain threshold, the $1/\sqrt{(\text{packet loss rate})}$ rule of thumb does not hold; TCP throughput deteriorates fast, and essentially TCP breaks down [164].

To illustrate the TCP throughput gains, we consider a 802.11a network with only two active nodes (note that the IEEE 802.11 specifies a stop-and-go ARQ mechanism). In the first case, the transmit node has only one antenna and uses the 802.11a physical layer, whereas in the second case the transmitter has two antennas and uses the STBC-enhanced 802.11a. Fig. 28 shows the achievable average TCP throughput as a function of SNR for various transmission modes. There are some interesting observations to be made. First, for each transmission mode, the maximum achievable average TCP throughput is considerably less than the advertised physical layer throughput. This is due to the overhead introduced by the higher layers (see, e.g., [92] for more details). Second, as expected, when the channel quality is sufficiently good, the difference between STBC-802.11a and 802.11a becomes smaller. It is seen that at high SNR values (where channel coding takes care of the few bit errors that may occur), the TCP throughputs of STBC-802.11a and 802.11 are equal. However, with STBC-802.11a we may switch to faster transmission modes at much lower SNR values, as Fig. 28 indicates. It is at relatively low SNR values where it pays off to employ STBC. Even with ARQ retransmissions (which can compensate for residual errors after decoding at the receiver), TCP throughput performance is increased with STBC, because STBC improve the overall transmission medium quality. It also has an impact on the number of packet retransmissions N_{ARQ} and some of these details are given in [233].

2) Explicit Loss Notification: Despite the packet error rate improvement using STBC, there will be residual packet losses in the wireless transmissions resulting from bad channel realizations. Packet losses are more likely to be due to a wireless link failure rather than congestion in the hybrid

[45]For example, with up to seven retransmissions, and a 10^{-1} frame error rate, the probability that a frame will be received incorrectly even after the seventh retransmission is 10^{-8} with independent frame errors—with frame combining, the said probability becomes even smaller.

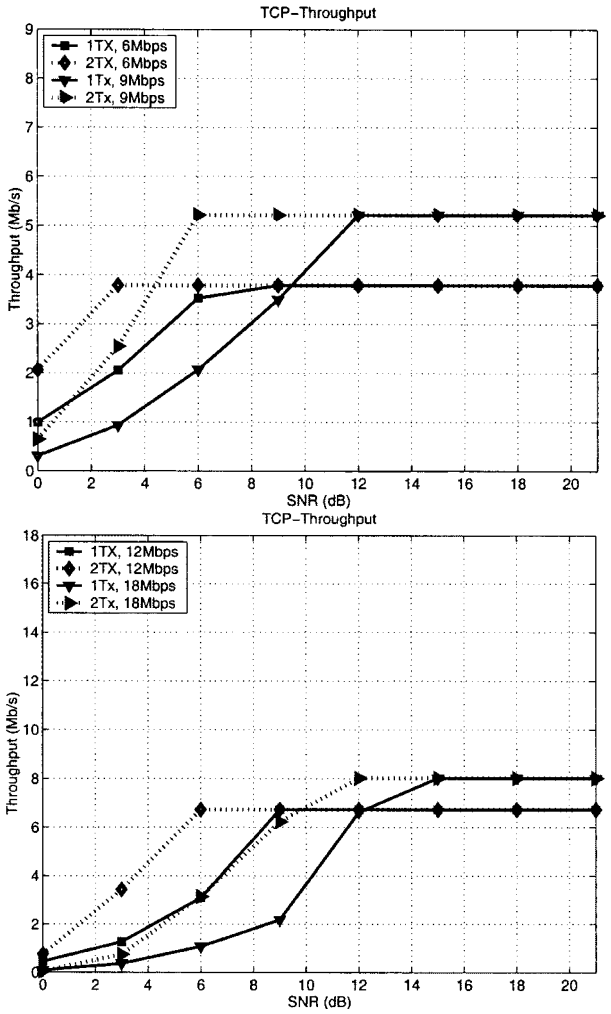

Fig. 28. TCP throughput for single session over HiperLan/2 channel "A."

network. As mentioned earlier, a complementary approach relies on being able to isolate the cause of packet loss to either link loss or congestion loss [19], [23], [26], [27], [46]. These schemes can be classified into three broad categories [25]: link layer (LL) protocols, split-connection protocols, and end-to-end protocols.

LL protocols use techniques such as ARQ, forward error correction (FEC), or their combination to provide the transport layer protocol with a dependable communication channel (see Section V-A.2). However, LL protocols lead to channel bandwidth and round trip time deviations in FEC and ARQ schemes, respectively. Also, without knowledge of upper-layer protocols, the LL protocols can possibly have adverse effects on TCP for two reasons [79]. First, competing retransmissions can be caused by an incompatible setting of timers at the two layers. Second, it cannot prevent TCP from going into an unnecessary fast retransmission mode because it cannot suppress the duplicate ACKs in the transport layer. The SNOOP protocol [25] fixes these problems by having an agent at the base station monitor the ACK packets and suppress unnecessary duplicate ACKs. However, this requires the base station to maintain a per-flow state.

Split-connection protocols, such as Indirect TCP [23] and M-TCP [46], also work to isolate the transport protocol from channel errors. They do so by splitting each TCP connection into two connections at the base station: ordinary TCP is used at the wireline link and protocols tuned specifically for wireless operations are used at the other end. Besides the problems of the violation of the end-to-end semantics and throughput mismatch between the two connections, the main concern of split-connection protocols is the maintenance of the per-flow state which is difficult to scale in large-scale wireless systems.

End-to-end protocols do not depend on the ability of the link layer to shield channel errors. Like TCP, they assume that the underlying network can only provide best effort, unreliable packet delivery service. So for end-to-end protocols to work well, they need mechanisms such as ELN to help the end hosts differentiate between congestion and link losses.

A simple way to implement an end-to-end ELN mechanism is proposed in [114]. The idea is to give extra *protection* to the TCP/IP header at the link layer. More specifically, a new end-to-end protocol is proposed to disambiguate between congestion losses and link failures. The protocol is based on a very simple principle: even if the IP packet is in error, if the header packet is received by the mobile terminal the receiver would know that there was a wireless link loss. This information can then be conveyed back to the sender using a special acknowledgment (ACK); this is an instance of ELN [25]. The extra protection for the header need not be done explicitly. By using an ARQ protocol and appropriate number of repeat attempts, one can get the required unequal error protection for the header quite naturally [114]. Using such an ELN mechanism improved transport efficiency significantly.

D. Real-Time Applications

It is clear from all the previous discussions that the inherent characteristic of wireless networks is the variability in rates, errors and, therefore, performance. This motivates the need to develop real-time applications that are robust to such variability. Design of robust applications is a vast topic and the techniques would vary significantly depending on the application of interest. Our focus in this short discussion is on problems related to coding of real-time applications such as voice, video, etc. A broad principle that might be required of any design is that of graceful degradation in performance with losses. The method for providing such a characteristic would depend on what the performance metric is for the given application.

Since in wireless networks there is location-dependent transmission rates and errors, having a *rate-scalable* application is desirable. For example, if a real-time source such as speech, audio or video is being transmitted, when the available rate varies (either because of transmission conditions or because of the scheduling scheme), the application would need to adapt to this variation. This can be done by layering the transmission source [215] so that when a lower rate is available, some layers of the source are dropped in order to provide a graceful degradation in performance. This adaptability is an important characteristic which needs to be built into real-time sources over wireless networks [53]. Also, if the relative importance of segments of the application frame is known, we may employ unequal error protection to provide greater reliability to higher priority segments (for example, see [72]).

In relaying strategies of ad hoc wireless networks described in Sections III-B and V-B.1, there could be multiple routes from source to destination. When there is a need to send delay-limited messages, one method is to transmit information through multiple routes in order to increase the probability of successfully receiving the information at the receiver despite route (path) failures. This is a form of *route diversity* and was first suggested in [182] in the context of wired networks. If the same information is transmitted over both routes, then this is a form of repetition coding. However, when both routes are successful, there is no performance advantage. Perhaps a more sophisticated technique would be to send correlated descriptions in the two routes such that each description is individually good, but they are different from one another so that if both routes are successful one gets a better approximation of the source. This is the basic idea behind multiple description (MD) source coding [96]. The idea is that the source is coded through several descriptions, where we require that performance (distortion) guarantees can be given to any subset of the descriptions and the descriptions mutually refine each other. Note, for example, that the layered coding schemes form a special case of such MD coding scheme, where guarantees of performance are not given for individual layers, but the layers refine the coarser description of the source. The information-theoretic characterization of MD source coding was provided in [96] and there have been several developments in this characterization since. Coding for MDs was proposed in [260]. Several recent lattice quantization schemes have been developed for MD coding [88], [124], [261].

As far as particular applications go, video could be an important application for future wireless networks. There has been significant research into robust video coding in the presence of errors [215]. The main problem that arises in video is that the compression schemes typically have motion compensation which introduces memory into the coded stream. Therefore, decoding the current video frame requires the previous video frames and depending on packetization errors, this could mean that either previous frames are corrupted or lost and, hence, decoding requires concealment of such errors. This is an active research topic especially in the context of wireless channels [119].

This discussion just scratches the surface of application requirements and design for wireless networks. The main point we wanted to illustrate was again the "interlayer" interactions that could occur by utilizing the characteristics of the wireless medium. Therefore, with availability of route diversity, variable transmission rates can significantly affect the design principles for the application layer.

E. Functional Abstractions

We conclude this section with some speculations on how spatial diversity induces rethinking on the functional abstractions of traditional IP protocol layers. Given the intricate interactions of the network layers in the wireless communication scenario, there is a temptation to redesign the entire protocol architecture. However, the success of the Internet has been largely due to the functional abstractions of the different layers and maintaining semantics for them to get just the right amount of information transferred between them. Given this successful model, perhaps the right question to ask in wireless networks is the kind of functional abstractions needed and maybe more importantly, what is the information to share between them. This enquiry is well motivated, given the resource constraints of the wireless medium.

We can view scheduling as an interaction between the physical layer and the link layer. Here, the resource allocation (power and bandwidth) depended on both the QoS criteria of interest and also the channel state of the users. This tight coupling was critical in the design of the scheduler. Routing in both ad hoc and hierarchical networks depended on the link layer conditions and, therefore, can be viewed as an "interlayer" interaction between the traditional routing and link layers. In hybrid networks, the efficiency of TCP, the preeminent transport protocol, was enhanced by using both physical layer (space–time codes) and link layer techniques (through ELN). Finally, MD source coding illustrated a strong interaction between the applications and routing.

One theme that can further illustrate these interactions is that of energy-limited communication arising out of the importance of conserving battery life of wireless mobile devices. In energy-aware scenarios, explicit interactions of the layers become important. For example, in scheduling transmissions, the strategies could be quite different from power-limited communication. One can tradeoff delay with energy consumption, as done in [95], [259], and references therein. Also in multihop routing, one can choose routes based on energy minimization as done in [219] and references therein. Finally, by utilizing spatial diversity, space–time codes can increase the reliability for a given SNR and, therefore, can be a tool for reducing energy consumption. We illustrated this in the context of how space–time block codes can increase reliability and, therefore, improve transport efficiency in Section V-C.1. We further examine how this affects energy consumption. Because STBC reduce the number of data packet retransmissions, compared to the single-antenna case, less energy is required for the transmission of a fixed-size data file when STBC are used. Consequently, battery life is prolonged. Interestingly enough, as STBC decrease the time it takes to successfully transmit a fixed-size data file, STBC may further prolong the battery life of wireless devices. This is because of the charge *recovery effect*: due to electrochemical properties of batteries (see, e.g, [4], [57], and references therein), whenever power is not drained from a battery, the battery can recover some of the lost charge. Though the maximum energy that can be drained from a battery is a fixed quantity, in practice, the wireless device may acquire different amounts of energy depending on how frequently the battery is used. Scheduling (so that the batteries are allowed to rest between transmissions) can lead to longer battery life, with small increase of the data packet delay [57]. STBC could prove beneficial to such scheduling strategies because STBC implicitly increase the rest period between successive transmissions.

In wireless networks, there is a very strong coupling of the physical and link layers. We believe this almost requires the removal of the functional separation of these two layers. The functional separation of routing is much more difficult to abstract. The simple abstraction of providing the underlying communication graph and routing on the graph may be inefficient in energy-limited communication. We do not believe there is yet a convincing answer to all these questions making it a rich research area.

All these topics illustrate the need to rethink several of the traditional approaches to networking in the context of wireless communication. This presents both a research challenge and an opportunity. The challenge is to identify the right design areas that would make the difference in overall system performance. The opportunity is to make a leap far beyond narrow problems of particular areas and an opportunity for new interdisciplinary problems.

VI. Summary

The goal of this paper was both to bring together major new developments in wireless communication (over the past decade) and also to investigate an emerging point of view. Sophisticated use of diversity in all its forms is the key tool that was the focus in this paper. The intricate interplay of different forms of diversity is the emerging view of reliable wireless networks.

It would be an oversimplification to suggest that it is straightforward to provide reliable QoS on broadband wireline networks, but it is usually possible to resolve many issues by overprovisioning [195]. That is not the case for wireless networks, where power and spectral bandwidth are limited. These constraints translate to limitations on signal processing at the mobile terminal, and require sophisticated resource allocation techniques due to significant variations in the capacity of the transmission link on short time scales. In fact, the difference is even more pronounced for ad hoc wireless networks, where radio resources are used both to maintain the network infrastructure and to transmit information.

Wireless networks present an opportunity to reexamine functional abstractions of traditional network layer protocols. Cross-layer interactions in wireless networks can optimize throughput by making additional performance information visible between layers in the IP protocol stack. Spatial diversity is critical in improving data rates and reliability of individual links and leads to innovations in scheduling that optimize global throughput. Space–time codes designed for small numbers of transmit and receive antennas have been

shown to significantly improve link capacity, and also system capacity through resource allocation. This coding technology can be integrated with sophisticated signal processing to provide a complete receiver that has computational complexity essentially implementable on current chip technology. This bounding of signal processing complexity is important given the energy constraints at the mobile terminal.

Fifty years ago the focus of communication and information theory was point-to-point communication in the presence of noise. Today, the challenge is packet data and real-time applications transmitted over heterogeneous networks where network management is distributed in space, and spatial diversity is the key to the great expectation of providing reliable end-to-end QoS for the wireless network as a whole.

Appendix
Orthogonal Designs and Algebraic Properties of STBC

This appendix describes how the mathematics of quadratic forms developed over 100 years ago is influencing the practice of wireless communication. The simplest example of a complex orthogonal design is the 2×2 code

$$(c_1, c_2) \rightarrow \begin{pmatrix} c_1 & -\overline{c}_2 \\ c_2 & \overline{c}_1 \end{pmatrix}$$

discovered by Alamouti [12]. The columns of this matrix represent different time slots, the rows represent different antennas, and the entries are the symbols to be transmitted. The signals r_1, r_2 received over two consecutive symbol periods are given by (assuming flat-fading channels)

$$\begin{pmatrix} r_1 \\ -\overline{r}_2 \end{pmatrix} = \begin{pmatrix} h_1 & h_2 \\ -\overline{h}_2 & \overline{h}_1 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$$

where h_1, h_2 are the *path gains* from the two transmit antennas to the mobile, and the noise samples z_1, z_2 are independent samples of a zero-mean complex Gaussian random variable with power spectral density $N_0/2$ per dimension. Decoding is remarkably simple, provided that the path gains are known at the mobile. We form

$$\begin{pmatrix} \overline{h}_1 & h_2 \\ \overline{h}_2 & -h_1 \end{pmatrix} \begin{pmatrix} r_1 \\ r_2^* \end{pmatrix} = (|h_1|^2 + |h_2|^2) \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + \begin{pmatrix} z_1' \\ z_2' \end{pmatrix}$$

and observe that the new noise samples remain independent of each other; the vector (z_1', z_2') has zero mean and covariance $(|h_1|^2 + |h_2|^2)\mathbf{I}_2$. We may then decode c_1, c_2 separately, rather than decoding the pair (c_1, c_2) jointly, which is far more complex. Decoding fails only if $h_1 = h_2 = 0$, when both transmission paths to the mobile antenna experience a deep fade simultaneously. This all works because the columns of the matrix

$$\mathbf{H} = \begin{pmatrix} h_1 & h_2 \\ -\overline{h}_2 & \overline{h}_1 \end{pmatrix}$$

are orthogonal, regardless of the actual values of the path gains h_1, h_2. If we were to view these path gains as indeterminates, then we have

$$\mathbf{H}\mathbf{H}^* = (|h_1|^2 + |h_2|^2)\mathbf{I}_2$$

which is the condition that defines a *complex orthogonal design*.

Formally, let $u_0, u_1, \ldots, u_{s-1}$ be positive integers, and let $x_0, x_1, \ldots, x_{s-1}$ be commuting indeterminates. A *real orthogonal design* of size N and type $(u_0, u_1, \ldots, u_{s-1})$ is an $N \times N$ matrix X with entries $0, \pm x_0, \pm x_1, \ldots, \pm x_{s-1}$ satisfying

$$\mathbf{X}\mathbf{X}^T = \left(\sum_{j=0}^{s-1} u_j x_j^2 \right) \mathbf{I}_N.$$

A *complex orthogonal design* of size N and type $(u_0, \ldots, u_{s-1}; v_1, \ldots, v_t)$ is a matrix $\mathbf{Z} = \mathbf{X} + i\mathbf{Y}$, where $\mathbf{X}$, $\mathbf{Y}$ are real orthogonal designs of type $(u_0, \ldots, u_{s-1})$ and $(v_1, \ldots, v_t)$ respectively, and where

$$\mathbf{Z}\mathbf{Z}^* = \left(\left(\sum_{j=0}^{s-1} u_j x_j^2 \right) + \left(\sum_{j=1}^{t} v_j y_j^2 \right) \right) \mathbf{I}_N.$$

Since

$$\begin{aligned} \mathbf{Z}\mathbf{Z}^* &= (\mathbf{X} + i\mathbf{Y})(\mathbf{X}^T - i\mathbf{Y}^T) \\ &= (\mathbf{X}\mathbf{X}^T + \mathbf{Y}\mathbf{Y}^T) + i(\mathbf{Y}\mathbf{X}^T - \mathbf{X}\mathbf{Y}^T) \end{aligned}$$

it follows that $\mathbf{X}\mathbf{Y}^T = \mathbf{Y}\mathbf{X}^T$. Note that if $t = s + 1$, then the entries of $X + iY$ are linear combinations of the complex indeterminates $z_j = x_j + iy_{j+1}$ and their conjugates $\overline{z}_j$, $j = 0, \ldots, s-1$.

Now consider the representation of the quaternions $\mathbb{Q}$ as a 4×4 matrix algebra over $\mathbb{R}$, where the quaternion $x_0 + ix_1 + jx_2 + kx_3$ corresponds to the matrix

$$\left[\begin{array}{cc|cc} x_0 & x_1 & x_2 & x_3 \\ -x_1 & x_0 & -x_3 & x_2 \\ \hline -x_2 & x_3 & x_0 & -x_1 \\ -x_3 & -x_2 & x_1 & x_0 \end{array}\right]$$

$$= x_0\mathbf{I} + x_1 \left[\begin{array}{cc|cc} & & 1 & \\ -1 & & & \\ \hline & & & -1 \\ & & 1 & \end{array}\right]$$

$$+ x_2 \left[\begin{array}{cc|cc} & & & 1 \\ & & 1 & \\ \hline -1 & & & \\ & -1 & & \end{array}\right] + x_3 \left[\begin{array}{cc|cc} & & & 1 \\ & & -1 & \\ \hline & 1 & & \\ -1 & & & \end{array}\right].$$

We may also view quaternions as pairs of complex numbers, where the product of quaternions (a, b) and (a', b') is given by

$$(a, b)(a', b') = (aa' - \overline{b}'b, ab' + \overline{a}'b).$$

These are Hamilton's Biquaternions. We may associate the pair (a, b) with the 2×2 complex matrix

$$\begin{pmatrix} a & b \\ -\overline{b} & \overline{a} \end{pmatrix}$$

and observe that matrix multiplication coincides with the rule for multiplying biquaternions. This correspondence between Alamouti matrices and quaternions means that the set

of Alamouti matrices is closed under addition, multiplication and taking inverses. This turns out to be critical to the signal processing algorithms that support interference suppression as we show next for the flat-fading channel case. The general case of frequency-selective channels is discussed in Section V-A.2 and treated in detail in [11] and [234].

We follow [188] in describing how a mobile terminal with two receive antennas uses the second antenna to separate two synchronous cochannel users, each employing the 2×2 block space–time code. If $\mathbf{c} = (c_1, c_2)$ and $\mathbf{s} = (s_1, s_2)$ are the codewords transmitted by the first and second users respectively, then

$$\mathbf{r}_1 = \mathbf{H}_1 \cdot \mathbf{c} + \mathbf{G}_1 \cdot \mathbf{s} + \boldsymbol{\eta}_1$$

$$\mathbf{r}_2 = \mathbf{H}_2 \cdot \mathbf{c} + \mathbf{G}_2 \cdot \mathbf{s} + \boldsymbol{\eta}_2,$$

where the vectors $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are complex Gaussian random variables with zero mean and covariance $N_0\mathbf{I}_2$. The matrices $\mathbf{H}_1$ and $\mathbf{H}_2$ capture the path gains from the first user to the first and second receive antennas respectively. The matrices $\mathbf{G}_1$ and $\mathbf{G}_2$ capture the path gains from the second user to the first and second receive antennas respectively.

Thus

$$\mathbf{r} = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{H}_1 & \mathbf{G}_1 \\ \mathbf{H}_2 & \mathbf{G}_2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{s} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{bmatrix}$$

and we set

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_1 & \mathbf{G}_1 \\ \mathbf{H}_2 & \mathbf{G}_2 \end{bmatrix}.$$

The zero-forcing decoder employs a linear combination of received symbols to remove interference between the two users without any regard for noise enhancement. It is further shown in [188] that if

$$\mathbf{W} = \begin{bmatrix} \mathbf{I}_2 & -\mathbf{G}_1\mathbf{G}_2^{-1} \\ -\mathbf{H}_2\mathbf{H}_1^{-1} & \mathbf{I}_2 \end{bmatrix}$$

then

$$\mathbf{W} \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{r}}_1 \\ \tilde{\mathbf{r}}_2 \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{H}} & 0 \\ 0 & \tilde{\mathbf{G}} \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{s} \end{bmatrix} + \begin{bmatrix} \tilde{\boldsymbol{\eta}}_1 \\ \tilde{\boldsymbol{\eta}}_2 \end{bmatrix}.$$

The algebraic structure of Alamouti matrices (closure under addition, multiplication, and taking inverses) implies that the matrices $\tilde{\mathbf{H}}$ and $\tilde{\mathbf{G}}$ have the same structure as the matrices $\mathbf{H}_1, \mathbf{H}_2, \mathbf{G}_1$ and $\mathbf{G}_2$. The matrix $\mathbf{W}$ transforms the problem of joint detection of two cochannel users into separate detection of two individual space–time users. Essentially it reduces the joint detection problem to two instances of a previously solved problem.

We refer the reader to Calderbank and Naguib [51] for more information, including the description of an MMSE decoder that avoids noise magnification associated with zero-forcing solutions, and more connections between space–time codes and exotic mathematical algebras.

We conclude with the following example of a diversity-4 rate-3/4 complex orthogonal design for four transmit antennas (see [239],[51], [249] and references therein for more examples and properties of orthogonal designs). It is based on the Cayley numbers or Octonions.

$$\begin{pmatrix} b_0 & b_1 & b_2 & 0 \\ -\overline{b}_1 & \overline{b}_0 & 0 & b_2 \\ -\overline{b}_2 & 0 & \overline{b}_0 & -b_1 \\ 0 & -\overline{b}_2 & \overline{b}_1 & b_0 \end{pmatrix}.$$

ACKNOWLEDGMENT

The authors would like to thank M. Grossglauser and V. Vaishampayan for their insightful comments on the paper, which helped to significantly improve it. The authors would also like to thank the following individuals (in alphabetical order) for many stimulating discussions on the topics of this paper: C. Fragouli, X. Gao, M. Grossglauser, S. Muthukrishnan, A. Naguib, N. Seshadri, V. Tarokh, D. Tse, V. Vaishampayan, and W. Younis.

REFERENCES

[1] "Waves of the future: Telecoms firms are betting that they will reap big rewards from the growth of mobile telephony," in *Economist*, 2000.

[2] A. Abidi, G. Pottie, and W. Kaiser, "Power-conscious design of wireless circuits and systems," *Proc. IEEE*, vol. 88, pp. 1528–1545, Oct. 2000.

[3] S. Acharya and S. Muthukrishnan, "Scheduling on-demand broadcasts for heterogenous: New metrics and algorithms," *ACM MobiCom*, pp. 43–54, 1998.

[4] M. Adamou and S. Sarkar, "A framework for optimal battery management for wireless nodes," *INFOCOM*, vol. 3, pp. 1783–1792, 2002.

[5] D. Agrawal, V. Tarokh, A. Naguib, and N. Seshadri, "Space-time coded OFDM for high data-rate wireless communication over wideband channels," *VTC*, pp. 2232–2236, May 1998.

[6] I. Akyildiz, J. McNair, J. Ho, H. Uzunalioglu, and W. Wang, "Mobility management in next-generation wireless systems," *Proc. IEEE*, vol. 87, pp. 1347–1384, Aug. 1999.

[7] N. Al-Dhahir, "Single-carrier frequency-domain equalization for space-time block-coded transmissions over frequency-selective fading channels," *IEEE Commun. Lett.*, vol. 7, pp. 304–306, July 2001.

[8] ——, "Overview and comparison of equalization schemes for space-time-coded signals with application to EDGE," *IEEE Trans. Signal Processing*, pp. 2477–2488, Oct. 2002.

[9] N. Al-Dhahir and A. H. Sayed, "The finite-length MIMO MMSE-DFE," *IEEE Trans. Signal Processing*, vol. 48, pp. 2921–2936, Oct. 2000.

[10] N. Al-Dhahir and S. Diggavi, "On the achievable rates of time-varying frequency-selective channels," in *CISS*, Princeton, NJ, Mar. 2002, pp. 860–865.

[11] N. Al-Dhahir, C. Fragouli, A. Stamoulis, Y. Younis, and A. Calderbank, "Space-time processing for broadband wireless access," *IEEE Commun. Mag.*, pp. 136–142, Sept. 2002.

[12] N. Al-Dhahir, G. Giannakis, B. Hochwald, B. Hughes, and T. Marzetta, "Editorial," *IEEE Trans. Signal Processing (Special Issue on Space-Time Coding)*, vol. 50, Oct. 2002.

[13] N. Al-Dhahir and S. N. Diggavi, "On the choice of guard sequence for block transmission over linear dispersive channels," *IEEE Trans. Commun.*, vol. 50, pp. 938–946, June 2002.

[14] S. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 1451–1458, Oct. 1998.

[15] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, vol. 39, pp. 150–154, Feb. 2001.

[16] F. Anjum and L. Tassiulas, "On the behavior of different TCP algorithms over a wireless channel with correlated packet losses," presented at the ACM SIGMETRICS'99, Atlanta, GA, 1999.

[17] A. Annamalai and V. K. Bhargava, "Analysis and optimization of adaptive multicopy transmission ARQ protocols for time-varying channels," *IEEE Trans. Commun.*, vol. 46, pp. 1356–1368, Oct. 1998.
[18] S. Ariyavisitakul, "Turbo space-time processing to improve wireless channel capacity," *IEEE Trans. Commun.*, vol. 48, pp. 1347–1359, Aug. 2000.
[19] E. Ayanoglu, S. Paul, T. F. LaPorta, K. K. Sabnani, and R. D. Gitlin, "Airmail: A link-layer protocol for wireless networks," *ACM Wireless Netw.*, vol. 1, pp. 47–60, Feb. 1995.
[20] L. Aydin, E. Esteves, and R. Padovani, "Reverse link capacity and coverage improvement for CDMA cellular systems using polarization and spatial diversity," in *Proc. ICC*, 2002, pp. 1887–1892.
[21] F. Backes, "Transparent bridges for interconnection of IEEE 802 LAN's," *IEEE Network*, vol. 2, pp. 5–9, Jan. 1988.
[22] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 284–287, Mar. 1974.
[23] A. Bakre and B. R. Badrinath, "I-tcp: Indirect tcp for mobile hosts," in *Proc. 15th Int. Conf. Distributed Computing System (ICDCS)*, May 1995.
[24] P. Balaban and J. Salz, "Optimum diversity combining and equalization in digital data transmission with applications to cellular mobile radio," *IEEE Trans. Commun.*, vol. 40, pp. 885–894, May 1992.
[25] H. Balakrishnan, V. N. Padmanabhan, S. Seshan, and R. H. Katz, "A comparison of mechanisms for improving TCP performance over wireless links," *IEEE/ACM Trans. Networking*, vol. 5, pp. 756–769, Dec. 1997.
[26] H. Balakrishnan, S. Seshan, and R. H. Katz, "Improving reliable transport and handoff performance in cellular wireless networks," *ACM Wireless Netw.*, vol. 1, no. 4, Dec. 1995.
[27] R. K. Balan, B. P. Lee, K. R. R. Kumar, L. Jacob, W. K. G. Seah, and A. L. Ananda, "TCP HACK: TCP header checksum option to improve performance over lossy links," presented at the IEEE INFOCOM'2001, 2001.
[28] N. Bambos, "Toward power-sensitive network architectures in wireless communications: Concepts, issues, and design aspects," *IEEE Pers. Commun.*, pp. 50–59, June 1998.
[29] S. Baro, G. Bauch, and A. Hansmann, "Improved codes for space-time trellis-coded modulation," *IEEE Commun. Lett.*, vol. 4, pp. 20–22, Jan. 2000.
[30] S. Basagni, I. Chlamtac, and V. R. Syrotuik, "A distance routing effect algorithm for mobility (DREAM)," in *Proc. ACM MOBICOM*, 1998.
[31] G. Bauch and N. Al-Dhahir, "Iterative equalization and decoding with channel shortening filters for space-time coded modulation," in *Proc. VTC*, 2000, pp. 1575–1582.
[32] L. Becchetti, S. Diggavi, S. Leonardi, A. Marchetti-Spaccamela, S. Muthukrishnan, T. Nandagopal, and A. Vitaletti, "Parallel scheduling problems in next generation wireless networks," in *Proc. ACM Symp. Parallel Algorithms and Architectures (SPAA)*, 2002, pp. 238–247.
[33] M. Bender, S. Chakrabarti, and S. Muthukrishnan, "Flow and stretch metrics for scheduling continuous job streams," in *Proc. Annu. Symp. Discrete Algorithms (SODA '98)*, 1998, pp. 270–279.
[34] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi, "CDMA/HDR; A bandwith-efficient high-speed wireless data service for nomadic users," *IEEE Commun. Mag.*, vol. 38, pp. 70–77, July 2000.
[35] S. Benedetto and G. Montorsi, "Unveiling turbo codes: Some results on parallel concatenated coding schemes," *IEEE Trans. Inform. Theory*, vol. 42, pp. 409–428, Mar. 1996.
[36] K. Berberidis and P. Karaivazoglou, "An efficient block adaptive decision feedback equalizer implemented in the frequency domain," *IEEE Trans. Signal Processing*, vol. 50, pp. 2273–2285, Sept. 2002.
[37] C. Berrou and A. Glavieux, "Near optimum error correcting coding and decoding: Turbo-codes," *IEEE Trans. Commun.*, vol. 44, pp. 1261–1271, Oct. 1996.
[38] S. Bhashyam, A. Sabharwal, and B. Aazhang, "Feedback gain in multiple antenna systems," *IEEE Trans. Commun.*, vol. 50, pp. 785–798, May 2002.
[39] E. Biglieri, "Digital transmission in the 21st century: Conflating modulation and coding," *IEEE Commun. Mag.*, pp. 128–137, May 2002.
[40] E. Biglieri, G. Caire, and G. Taricco, "Limiting performance of block-fading channels with multiple antennas," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1273–1289, May 2001.
[41] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: Information-theoretic and communications aspects," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2619–2692, Oct. 1998.
[42] K. Boulle and J. C. Belfiore, "Modulation schemes designed for the Rayleigh channel," in *Proc. Conf. Inform. Sci. Syst. (CISS '92)*, 1992, pp. 288–293.
[43] J. Boutros and E. Viterbo, "Signal space divesity: A power and bandwidth efficient diversity technique for the Rayleigh fading channel," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1453–1467, July 1998.
[44] L. H. Brandenburg and A. Wyner, "Capacity of the Gaussian channel with memory: The multivariate case," *Bell Syst. Tech. J.*, vol. 53, pp. 745–779, May-June 1974.
[45] D. Brennan, "Linear diversity combining techniques," *Proc. IEEE*, vol. 47, pp. 1075–1102, June 1959.
[46] K. Brown and S. Singh, "M-TCP: TCP for mobile cellular networks," *ACM Comput. Commun. Rev.*, vol. 27, no. 5, 1997.
[47] R. Caceres and V. Padmanabhan, "Fast and scalable wireless handoffs in supports of mobile Internet audio," *Mobile Netw. Appl.*, vol. 3, no. 4, pp. 351–363, 1998.
[48] G. Caire and S. Shamai, "On the capacity of some channels with channel state information," *IEEE Trans. Inform. Theory*, vol. 45, pp. 2007–2019, Sept. 1999.
[49] ——, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1691–1706, July 2003.
[50] G. Caire and D. Tuninetti, "ARQ protocols for the Gaussian collision channel," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1971–1988, July 2001.
[51] A. R. Calderbank and A. F. Naguib, "Orthogonal designs and third generation wireless communication," in *Surveys in Combinatorics 2001, London Math. Soc. Lecture Note Series #288*, J. W. P. Hirschfeld, Ed: Cambridge Univ. Press, 2001, pp. 75–107.
[52] J. Cavers, "An analysis of pilot symbol assisted modulation for Rayleigh fading channels (mobile radio)," *IEEE Trans. Veh. Technol.*, vol. 40, pp. 686–693, Nov. 1991.
[53] N. Chaddha and S. N. Diggavi, "A frame-work for joint source-channel coding of images over time-varying wireless channels," in *Proc. Int. Conf. Image Processing*, Sept. 1996, pp. 89–92.
[54] D. Chase, "Code combining—a maximum likelihood decoding approach for combining an arbitrary number of noisy packets," *IEEE Trans. Commun.*, vol. 33, pp. 385–393, May 1985.
[55] B. Chen and G. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embeddng," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1423–1443, May 2001.
[56] R. S. Cheng and S. Verdu, "Gaussian multiaccess channels with ISI: Capacity region and multiuser water-filling," *IEEE Trans. Inform. Theory*, vol. 39, pp. 773–785, May 1993.
[57] C. Chiasserini and R. Rao, "Improving battery performance by using traffic shaping techniques," *IEEE J. Select. Areas Commun.*, vol. 19, pp. 1385–1394, July 2001.
[58] D. Chizhik, G. Foschini, M. Gans, and R. Valenzuela, "Keyholes, correlations, and capacities of multielement transmit and receive antennas," *IEEE Trans. Wireless Commun.*, vol. 1, pp. 361–368, Apr. 2002.
[59] W. Choi, R. Negi, and J. M. Cioffi, "Combined ML and DFE decoding for the V-BLAST system," *Proc. ICC*, pp. 1243–1248, 2000.
[60] D. Chu, "Polyphase codes with good periodic correlation properties," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 531–532, July 1972.
[61] C. Chuah, D. Tse, J. Kahn, and R. Valenzuela, "Capacity scaling in MIMO wireless systems under correlated fading," *IEEE Trans. Inform. Theory*, vol. 48, pp. 637–650, Mar. 2002.
[62] J. Chuang and N. Sollenberger, "Beyond 3G: Wideband data access based on OFDM and dynamic packet assignment," *IEEE Commun. Mag.*, vol. 38, pp. 78–87, July 2000.
[63] L. J. Cimini, "Analysis and simulation of a digital mobile channel using orthogonal frequency division multiplexing," *IEEE Trans. Commun.*, vol. 33, pp. 665–675, July 1985.
[64] J. M. Cioffi, G. P. Dudevoir, M. V. Eyuboglu, and G. D. Forney, "MMSE decision-feedback equalizers and coding: Parts I & II," *IEEE Trans. Commun.*, vol. 43, pp. 2582–2604, Oct. 1995.
[65] D. Clark, "The design philosophy of the DARPA internet protocols," in *Proc. ACM SIGCOMM*, 1988, pp. 106–114.
[66] J. Conway and N. Sloane, *Sphere Packings, Lattices, and Groups*, 3rd ed. New York: Springer-Verlag, 1999.

[67] M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. Inform. Theory*, vol. 29, pp. 439–441, May 1983.

[68] T. M. Cover, "Some advances in broadcast channels," in *Advances in Communication Theory*, A. Viterbi, Ed. San Francisco, CA: Academic, 1975, vol. 4, Theory and Applications.

[69] T. M. Cover and A. E. Gamal, "Capacity theorems for the relay channel," *IEEE Trans. Inform. Theory*, vol. 25, pp. 572–584, Sept. 1979.

[70] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[71] D. Cox, "Portable digital radio communications—An approach to tetherless access," *IEEE Commun. Mag.*, pp. 30–40, July 1989.

[72] R. Cox, J. Hagenauer, N. Seshadri, and C.-E. Sundberg, "Subband speech coding and matched convolutional channel coding for mobile radio channels," *IEEE Trans. Signal Processing*, vol. 39, pp. 1717–1731, Aug. 1991.

[73] S. Crozier, D. Falconer, and S. Mahmoud, "Least sum of squared errors (LSSE) channel estimation," in *Radar and Signal Processing, IEE Proc. F*, 1991, pp. 371–378.

[74] R. L. Cruz, "Quality of service guarantees in virtual circuit switched networks," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1048–1056, Aug. 1995.

[75] M. O. Damen, K. Abed-Meraim, and J.-C. Belfiore, "Diagonal algebraic space-time block codes," *IEEE Trans. Inform. Theory*, vol. 48, pp. 628–636, Mar. 2002.

[76] M. O. Damen, A. Chkeif, and J.-C. Belfiore, "Lattice codes decoder for space-time codes," *IEEE Commun. Lett.*, vol. 4, pp. 161–163, May 2000.

[77] V. M. DaSilva and E. S. Sousa, "Fading resistant modulation using several transmitter antennas," *IEEE Trans. Commun.*, vol. 45, pp. 1236–1244, Oct. 1997.

[78] A. Demers, S. Keshav, and S. Shenker, "Analysis and simulation of a fair queueing algorithm," in *Proc. ACM Sigcomm'89*, pp. 1–12.

[79] A. DeSimone, M. C. Chuah, and O. C. Yue, "Throughput performance of transport-layer protocols over wireless LANs," in *Proc. Globecom*, 1993.

[80] S. Diggavi, N. Al-Dhahir, A. Stamoulis, and A. R. Calderbank, "Differential space-time coding for frequency-selective channels," *IEEE Commun. Lett.*, vol. 6, pp. 253–255, June 2002.

[81] S. Diggavi, M. Grossglauser, and D. N. C. Tse, "Even one dimensional mobility increases capacity of adhoc wireless networks," in *IEEE Symp. Information Theory*, 2002, p. 388.

[82] S. N. Diggavi, "Multiuser DMT: A multiple access modulation scheme," in *Proc. GLOBECOM*, 1996, pp. 1566–1570.

[83] ——, "Analysis of multicarrier transmission in time-varying channels," in *Proc. ICC*, 1997, pp. 1191–1195.

[84] ——, "On achievable performance of spatial diversity fading channels," *IEEE Trans. Inform. Theory*, vol. 47, pp. 308–325, Jan. 2001.

[85] S. N. Diggavi, N. Al-Dhahir, and A. R. Calderbank, "Algebraic properties of space-time block codes in inter-symbol interference multiple-access channels," *IEEE Trans. Inform. Theory*, vol. 49, pp. 2403–2414, Oct. 2003.

[86] ——, "Diversity embedded space-time codes," in *Proc. IEEE GLOBECOM*, 2003, pp. 1909–1914.

[87] S. N. Diggavi, B. C. Ng, and A. Paulraj, "An interference suppresion scheme with joint channel-data estimation," *IEEE J. Select. Areas Commun.*, vol. 17, pp. 1924–1939, Nov. 1999.

[88] S. N. Diggavi, N. Sloane, and V. A. Vaishampayan, "Asymmetric multiple description lattice vector quantizers," *IEEE Trans. Inform. Theory*, vol. 48, pp. 174–191, Jan. 2002.

[89] G. Dimic and N. Sidiropoulos, "Stability analysis of collision resolution protocols with retransmission diversity," in *Proc. ICASSP*, vol. 3, 2002, pp. 2133–2126.

[90] D. Divsalar and M. K. Simon, "The design of trellis coded MPSK for fading channels: Performance criteria," *IEEE Trans. Commun.*, vol. 36, pp. 1004–1012, Sept. 1988.

[91] X. Dong, N. C. Beaulieu, and P. H. Wittke, "Signaling constellations for fading channels," *IEEE Trans. Commun.*, vol. 47, pp. 703–714, May 1999.

[92] A. Doufexi, S. Armour, M. Butler, A. Nix, D. Bull, J. McGeehan, and P. Karlsson, "A comparison of the HIPERLAN/2 and IEEE 802.11a wireless LAN standards," *IEEE Commun. Mag.*, vol. 40, pp. 172–180, May 2002.

[93] A. Duel-Hallen and C. Heegard, "Delayed decision-feedback sequence estimation," *IEEE Trans. Commun.*, vol. 37, pp. 428–436, May 1989.

[94] A. Edelman, "Eigenvalues and condition numbers of random matrices," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, 1989.

[95] A. El-Gamal, C. Nair, B. Prabhakar, E. Uysal-Biyikoglu, and S. Zahedi, "Energy-efficient scheduling of packet transmissions over wireless networks," in *Proc. IEEE INFOCOM 2002*, pp. 1773–1782.

[96] A. A. El Gamal and T. M. Cover, "Achievable rates for multiple descriptions," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 851–857, Nov. 1982.

[97] H. El-Gamal and A. R. Hammons, "A new approach to layered space-time coding and signal processing," *IEEE Trans. Inform. Theory*, vol. 47, pp. 2321–2334, Sept. 2001.

[98] T. Ericson, "A Gaussian channel with slow fading," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 353–356, May 1970.

[99] M. Eyuboglu and S. Qureshi, "Reduced-state sequence estimation for coded modulation of intersymbol interference channels," *IEEE J. Select. Areas Commun.*, vol. 17, pp. 989–999, Aug. 1999.

[100] D. Falconer, S. Ariyavisitakul, A. Benyamin-Seeyar, and B. Eidson, "Frequency domain equalization for single-carrier broadband wireless systems," *IEEE Commun. Mag.*, vol. 40, pp. 58–66, Apr. 2002.

[101] U. Fincke and M. Pohst, "Improved methods for calculating vectors of shortest length in a lattice, including a complexity analysis," *Math. Comput.*, vol. 44, pp. 463–471, Apr. 1985.

[102] G. D. Forney, "Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 363–378, May 1972.

[103] G. Foschini, "Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas," *Bell Labs Tech. J.*, vol. 1, no. 2, pp. 41–59, Sept. 1996.

[104] ——, "Layered space-time architecture for wireless communication in a fading environment when using multiple antenna elements," *Bell Labs Tech. J.*, pp. 41–59, 1996.

[105] G. Foschini, G. Golden, R. Valenzuela, and P. Wolniansky, "Simplified processing for high spectral efficiency wireless communication employing multi-element arrays," *IEEE J. Select. Areas Commun.*, vol. 17, pp. 1841–1852, Nov. 1999.

[106] C. Fragouli, N. Al-Dhahir, S. Diggavi, and W. Turin, "Prefiltered space-time M-BCJR equalizer for frequency-selective channels," *IEEE Trans. Commun.*, vol. 50, pp. 742–753, May 2002.

[107] C. Fragouli, N. Al-Dhahir, and W. Turin, "Effect of spatio-temporal channel correlation on the performance of space-time codes," in *Proc. ICC*, vol. 2, 2002, pp. 826–830.

[108] ——, "Reduced-complexity training schemes for multiple-antenna broadband transmissions," in *Proc. WCNC*, vol. 1, 2002, pp. 78–83.

[109] V. Franz and J. Anderson, "Concatenated decoding with a reduced-search BCJR algorithm," *IEEE J. Select. Areas Commun.*, pp. 186–195, Feb. 1998.

[110] A. Furuskar, S. Mazur, F. Muller, and H. Olofsson, "EDGE: Enhanced data rates for GSM and TDMA/136 evolution," *IEEE Pers. Commun.*, pp. 56–66, June 1999.

[111] R. G. Gallager. (1963) *Low Density Parity Check Codes* [Online]. Available: http://justice.mit.edu/people/gallager.html

[112] ——, *Information Theory and Reliable Communications*. New York: Wiley, 1968.

[113] M. J. Gans, N. Amitay, Y. Yeh, H. Xu, T. Damen, R. Valenzuela, T. Sizer, R. Storz, D. Taylor, W. MacDonald, C. Tran, and A. Adamiecki, "Outdoor BLAST measurement system at 2.44 GHz: Calibration and initial results," *IEEE J. Select. Areas Commun.*, vol. 20, pp. 570–583, Apr. 2002.

[114] X. Gao, S. N. Diggavi, and S. Muthukrishnan, "LHP: An end-to-end reliable transport protocol over wireless data networks," in *Proc. IEEE Int. Conf. Communications (ICC)*, 2003, pp. 66–70.

[115] M. Gastpar and M. Vetterli, "On the capacity of wireless networks: The relay case," in *Proc. IEEE INFOCOM*, 2002, pp. 1577–1586.

[116] S. I. Gelfand and M. S. Pinsker, "Coding for channels with random parameters," *Probl. Contr. Inform. Theory*, vol. 9, no. 1, pp. 19–31, 1980.

[117] D. Gesbert, H. Bolcskei, D. A. Gore, and A. J. Paulraj, "Outdoor MIMO wireless channels: Models and performance prediction," *IEEE Trans. Commun.*, vol. 50, pp. 1926–1934, Dec. 2002.

[118] G. B. Giannakis and C. Tepedelenlioglu, "Basis expansion models and diversity techniques for blind equalization of time-varying channels," *Proc. IEEE*, vol. 86, pp. 1969–1986, Oct. 1998.

[119] B. Girod and N. Farber, "Wireless video," in *Compressed Video Over Networks*, A. Reibman and M.-T. Sun, Eds. New York: Marcel Dekker, 2000.

[120] L. Godara, "Applications of antenna arrays to mobile communications. Part I. Performance improvement, feasibility, and system considerations," *Proc. IEEE*, vol. 85, pp. 1031–1060, July 1997.
[121] ——, "Applications of antenna arrays to mobile communications. Part II. Beamforming and direction-of-arrival considerations," *Proc. IEEE*, vol. 85, pp. 1195–1245, Aug. 1997.
[122] A. Goldsmith and S. Chua, "Variable-rate variable-power MQAM for fading channels," *IEEE Trans. Commun.*, vol. 45, pp. 1218–1230, Oct. 1997.
[123] A. Goldsmith and P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1986–1992, Nov. 1997.
[124] V. K. Goyal and J. Kovacevic, "Generalized multiple description coding with correlating transforms," *IEEE Trans. Inform. Theory*, vol. 47, pp. 2199–2224, Sept. 2001.
[125] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series and Products*. San Diego, CA: Academic, 1994.
[126] M. Grossglauser and D. N. C. Tse, "Mobility increases the capacity of ad-hoc wireless networks," *IEEE/ACM Trans. Networking*, vol. 10, pp. 477–486, Aug. 2002.
[127] M. Grossglauser and M. Vetterli, "Locating nodes with ease: Last encounter routing in ad hoc networks through mobility diffusion," in *Proc. IEEE INFOCOM*, 2003, pp. 1954–1964.
[128] J.-C. Guey, M. P. Fitz, M. R. Bell, and W.-Y. Kuo, "Signal design for transmitter diversity wireless communication systems over Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 47, pp. 527–537, Apr. 1999.
[129] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. Inform. Theory*, vol. 46, pp. 388–404, Mar. 2000.
[130] ——, "Toward an information theory of large networks: An achievable rate region," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1877–1894, Aug. 2003.
[131] J. Hagenauer, "Rate compatable punctured convolutional codes (RCPC codes) and their applications," *IEEE Trans. Commun.*, vol. 36, pp. 389–400, Apr. 1988.
[132] A. R. Hammons and H. El-Gamal, "On the theory of space-time codes for PSK modulation," *IEEE Trans. Inform. Theory*, vol. 46, pp. 524–542, Mar. 2000.
[133] S. V. Hanly, "Capacity and power control in spread spectrum macrodiversity radio networks," *IEEE Trans. Commun.*, vol. 44, pp. 247–256, Feb. 1996.
[134] B. Harvey and S. Wicker, "Packet combining systems based on the Viterbi decoder," *IEEE Trans. Commun.*, vol. COM-26, pp. 540–547, Sept. 1980.
[135] B. Hassibi and B. Hochwald, "High-rate codes that are linear in space and time," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1804–1824, July 2002.
[136] B. Hassibi and T. Marzetta, "Multiple-antennas and isotropically random unitary inputs: The received signal density in closed form," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1473–1484, June 2002.
[137] B. Hassibi and H. Vikalo, "On the expected complexity of integer least-squares problems," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2002, pp. 1497–1500.
[138] S. Haykin, *Adaptive Filter Theory*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1991.
[139] A. Hiroike, F. Adachi, and N. Nakajima, "Combined effects of phase sweeping transmitter diversity and channel coding," *IEEE Trans. Veh. Technol.*, vol. 41, pp. 170–176, May 1992.
[140] B. Hochwald and W. Sweldens, "Differential unitary space-time modulation," *IEEE Trans. Commun.*, vol. 48, pp. 2041–2052, Dec. 2000.
[141] B. M. Hochwald and T. L. Marzetta, "Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading," *IEEE Trans. Inform. Theory*, vol. 45, pp. 139–157, Jan. 1999.
[142] ——, "Unitary space-time modulation for multiple-antenna communications in Rayleigh flat fading," *IEEE Trans. Inform. Theory*, vol. 46, pp. 543–564, Mar. 2000.
[143] M. Honig and M. K. Tsatsanis, "Adaptive techniques for multiuser CDMA receivers," *IEEE Signal Processing Mag.*, vol. 17, pp. 49–61, May 2000.
[144] T.-C. Hou and V. O. K. Li, "Transmission range control in multihop packet radio networks," *IEEE Trans. Commun.*, vol. COM-34, pp. 38–44, Jan. 1986.
[145] B. L. Hughes, "Differential space-time modulation," *IEEE Trans. Inform. Theory*, vol. 46, pp. 2567–2578, Nov. 2000.
[146] H. Jafarkhani and V. Tarokh, "Multiple transmit antenna differential detection from generalized orthogonal designs," *IEEE Trans. Inform. Theory*, vol. 47, pp. 2626–2631, Sept. 2001.
[147] R. Jain, A. Puri, and R. Sengupta, "Geographical routing using partial information for wireless adhoc networks," *IEEE Pers. Commun.*, vol. 8, pp. 48–57, Feb. 2001.
[148] W. Jakes, *Microwave Mobile Communications*. New York: IEEE Press, 1974.
[149] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency high data rate personal communication wireless system," in *Proc. Vehicular Technology Conf.*, 2000, pp. 1854–1858.
[150] W. Jeon, K. Chang, and Y. Cho, "An equalization technique for orthogonal frequency-division multiplexing systems in time-variant multipath channels," *IEEE Trans. Commun.*, vol. 47, pp. 27–32, Jan. 1999.
[151] Z. Jiang, K. Leung, B. Kim, and P. Henry, "Seamless mobility management based on proxy servers," in *Proc. WCNC*, vol. 2, 2002, pp. 563–568.
[152] G. Jongren, M. Skoglund, and B. Ottersten, "Combining beamforming and orthogonal space-time block coding," *IEEE Trans. Inform. Theory*, vol. 48, pp. 611–627, Mar. 2002.
[153] N. Joshi, S. R. Kadaba, S. Patel, and G. Sundaram, "Downlink scheduling in CDMA data networks," in *Proc. ACM MobiCom*, 2000, pp. 179–190.
[154] T. Kailath, "Channel characterization: Time-variant dispersive channels," in *Lectures on Communication System Theory*, E. J. Baghdady, Ed. New York: McGraw-Hill, 1961, pp. 95–123.
[155] S. Kallel and D. Haccoun, "Sequential decoding with ARQ and code combining: A robust hybrid FEC/ARQ system," *IEEE Trans. Commun.*, vol. 36, pp. 773–780, July 1988.
[156] S. Kandukuri and N. Bambos, "Power-controlled multiple access schemes for next-generation wireless packet networks," *IEEE Wireless Commun.*, vol. 9, pp. 58–64, June 2002.
[157] D. Karger, C. Stein, and J. Wein, "Scheduling algorithms," in *Algorithms and Theory of Computation Handbook*. Boca Raton, FL: CRC, 1999.
[158] K. J. Kerpez, "Constellations for good diversity performance," *IEEE Trans. Commun.*, vol. 41, pp. 1412–1421, Sept. 1993.
[159] S. Keshav, *An Engineering Approach to Computer Networking: ATM Networks, the Internet, and the Telephone Network*. Boston, MA: Addison-Wesley, 1995.
[160] R. Knopp and G. Caire, "Power control and beamforming for systems with multiple transmit and receive antennas," *IEEE Trans. Wireless Commun.*, vol. 1, pp. 638–648, Oct. 2002.
[161] R. Knopp and P. Humblet, "Information capacity and power control in single-cell multiuser communications," in *IEEE Int. Conf. Communications (ICC)*, 1995, pp. 331–335.
[162] C. Komninakis, C. Fragouli, A. Sayed, and R. Wesel, "Multi-input multi-output fading channel tracking and equalization using Kalman estimation," *IEEE Trans. Signal Processing*, vol. 50, pp. 1065–1076, May 2002.
[163] P. Kyritsi, D. Cox, R. A. Valenzuela, and P. Wolniansky, "Effect of antenna polarization on the capacity of a multiple element system in an indoor environment," *IEEE J. Select. Areas Commun.*, vol. 20, pp. 1227–1239, Aug. 2002.
[164] T. Lakshman and U. Madhow, "The performance of TCP/IP for networks with high bandwidth-delay products and random loss," *IEEE Trans. Networking*, vol. 5, pp. 336–350, June 1997.
[165] A. Lapidoth and S. Shamai, "Fading channels: How perfect need "perfect side information" be?," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1118–1134, May 2002.
[166] J. W. Lee, R. R. Mazumdar, and N. B. Shroff, "Downlink power allocation for multi-class CDMA wireless networks," in *Proc. IEEE INFOCOM 2002*, 2002, pp. 1480–1489.
[167] J. Li and M. Kavehrad, "Effects of time selective multipath fading on OFDM systems for broadband mobile applications," *IEEE Commun. Lett.*, vol. 3, pp. 332–334, Dec. 1999.
[168] L. Li and A. Goldsmith, "Optimal resource allocation for fading broadcast channels- Part I: Ergodic capacity," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1083–1102, Mar. 2001.
[169] Y. Li and L. Cimini, "Bounds on the interchannel interference of OFDM in time-varying impairments," *IEEE Trans. Commun.*, vol. 49, pp. 401–404, Mar. 2001.
[170] Y. Li, L. J. C. Jr, and N. R. Sollenberger, "Robust channel estimation for OFDM systems with rapid dispersive fading channels," *IEEE Trans. Commun.*, vol. 46, pp. 902–915, July 1998.

[171] S. Lin, D. Costello, and M. Miller, "Automatic-repeat-request error-control schemes," *IEEE Commun. Mag.*, vol. 22, pp. 5–17, Dec. 1984.
[172] E. Lindskog and A. Paulraj, "A transmit diversity scheme for delay spread channels," in *Proc. Int. Conf. Communications (ICC)*, 2000, pp. 307–311.
[173] Y. Liu, M. Fitz, and O. Takeshita, "Space-time codes performance criteria and design for frequency selective fading channels," in *Proc. ICC*, vol. 9, 2001, pp. 2800–2804.
[174] Y. Liu, M. P. Fitz, and O. Y. Takeshita, "Full rate space-time turbo codes," *IEEE J. Select. Areas Commun.*, vol. 19, pp. 969–980, May 2001.
[175] Z. Liu, G. Giannakis, A. Scaglione, and S. Barbarossa, "Decoding and equalization of unknown multipath channels based on block precoding and transmit-antenna diversity," in *Proc. Asilomar Conf. Signals, Systems, and Computers*, 1999, pp. 1557–1561.
[176] A. Lozano and C. Papadias, "Layered space-time receivers for frequency-selective wireless channels," *IEEE Trans. Commun.*, vol. 50, pp. 65–73, Jan. 2002.
[177] B. Lu, X. Wang, and K. Narayanan, "LDPC-based space-time coded OFDM systems over correlated fading channels: Performance analysis and receiver design," *IEEE Trans. Commun.*, vol. 50, pp. 74–88, Jan. 2002.
[178] S. Lu, V. Bharghavan, and R. Srikant, "Fair scheduling in wireless packet networks," *IEEE Trans. Networking*, vol. 7, pp. 473–489, Aug. 1999.
[179] J. L. Massey and P. Mathys, "The collision channel without feedback," *IEEE Trans. Inform. Theory*, vol. 31, pp. 192–204, 1985.
[180] M. Mathis, J. Semke, and J. Mahdavi, "The macroscopic behavior of the TCP congestion avoidance algorithm," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 27, no. 3, pp. 67–82, July 1997.
[181] M. Mauve, J. Widmer, and H. Hartenstein, "A survey on position-based routing in mobile ad hoc networks," *IEEE Network*, vol. 15, pp. 30–39, Nov. 2001.
[182] N. Maxemchuk, "Dispersity routing in store and forward networks," Ph.D. dissertation, Univ. Pennsylvania, Philadelphia, 1975.
[183] M. Medard, "Capacity of time-varying multiple user channels in wireless communications," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, 1995.
[184] ——, "The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel," *IEEE Trans. Inform. Theory*, vol. 46, pp. 933–946, May 2000.
[185] R. Motwani and P. Raghavan, *Randomized Algorithms*. New York: Cambridge Univ. Press, 1995.
[186] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*. New York: Wiley, 1982.
[187] R. Murch and K. Letaief, "Antenna systems for broadband wireless access," *IEEE Commun. Mag.*, pp. 76–83, Apr. 2002.
[188] A. Naguib, N. Seshadri, and A. Calderbank, "Applications of space-time block codes and interference suppression for high capacity and high data rate wireless systems," in *Proc. 32nd Annu. Asilomar Conf. Signals, Systems and Computers*, 1998, pp. 1803–1810.
[189] A. Naguib, V. Tarokh, N. Seshadri, and A. Calderbank, "A space-time coding modem for high-data-rate wireless communications," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 1459–1477, Oct. 1998.
[190] A. F. Naguib, "Combined interference suppression and frequency domain equalization for space-time block coded transmission," in *Proc. IEEE Int. Conf. Communications (ICC)*, 2003, pp. 3261–3266.
[191] S. Nanda, K. Balachandran, and S. Kumar, "Adaptation techniques in wireless packet data services," *IEEE Commun. Mag.*, vol. 38, pp. 54–65, Jan. 2000.
[192] T. Nandagopal, S. Lu, and V. Bharghavan, "A unified architecture for the design and evaluation of wireless fair queueing algorithms," in *Proc. ACM Mobicom*, 1999, pp. 132–142.
[193] F. Neeser and J. L. Massey, "Proper complex random processes with applications to information theory," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1293–1302, July 1993.
[194] R. Negi and J. Cioffi, "Pilot tone selection for channel estimation in a mobile OFDM system," *IEEE Trans. Consumer Electron.*, vol. 44, pp. 1122–1128, Aug. 1998.
[195] A. M. Odlyzko. The history of communications and its implications for the Internet. [Online]. Available: http://www.dtc.umn.edu/odlyzko/doc/networks.html
[196] S. Ohno and G. B. Giannakis, "Capacity maximizing pilots and precoders for wireless OFDM over rapidly fading channels," *IEEE Trans. Inform. Theory*, submitted for publication.
[197] L. H. Ozarow, S. Shamai, and A. D. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Trans. Veh. Technol.*, vol. 43, pp. 359–378, May 1994.
[198] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: The single-node case," *IEEE/ACM Trans. Networking*, vol. 1, pp. 344–357, June 1993.
[199] S. Parkvall, M. Karlsson, M. Samuelsson, L. Hedlund, and B. Goransson, "Transmit diversity in WCDMA: Link and system level results," in *Proc. Vehicular Technology Conf.*, 2000, pp. 864–868.
[200] P. Patel and J. Holtzman, "Analysis of a simple successive interference cancellation scheme in a DS/CDMA system," *IEEE J. Select. Areas Commun.*, vol. 12, pp. 796–807, June 1994.
[201] A. Paulraj and B. Ng, "Space-time modems for wireless personal communications," *IEEE Pers. Commun.*, pp. 36–48, Feb. 1998.
[202] A. Paulraj and C. Papadias, "Space-time processing for wireless communications," *IEEE Signal Processing Mag.*, pp. 49–83, Nov. 1997.
[203] C. Perkins, *Mobile IP: Design Principles and Practice*. Boston, MA: Addison-Wesley, 1998.
[204] ——, *Ad Hoc Networking*. Boston, MA: Addison-Wesley, 2001.
[205] ——, "Mobile IP," *IEEE Commun. Mag.*, vol. 40, pp. 66–82, May 2002.
[206] T. Pollet, M. V. Bladel, and M. Moeneclaey, "BER sensitivity of OFDM systems to carrier frequency offset and wiener phase noise," *IEEE Trans. Commun.*, pp. 191–193, Feb./Mar./Apr. 1995.
[207] V. Poor, "The wireless revolution: A signal processing perspective," presented at the ICASSP 2000 (Plenary Talk), Istanbul, Turkey, 2000.
[208] J. G. Proakis, *Digital Communications*, 3rd ed. New York: McGraw-Hill, 1995.
[209] G. Raleigh, S. N. Diggavi, A. F. Naguib, and A. Paulraj, "Characterization of fast fading vector channels for multi-antenna communication systems," in *Proc. 28th Asilomar Conf. Signals, Systems and Computers*, 1994, pp. 853–857.
[210] G. G. Raleigh and J. M. Cioffi, "Spatio-temporal coding for wireless communication," *IEEE Trans. Commun.*, vol. 46, pp. 357–366, Mar. 1998.
[211] G. G. Raleigh, S. N. Diggavi, V. K. Jones, and A. Paulraj, "Blind adaptive antenna algorithm for wireless communication," in *Proc. ICC*, 1995, pp. 1494–1499.
[212] R. Ramjee, K. Varadhan, L. Salgarelli, S. Thuel, S.-Y. Wang, and T. L. Porta, "HAWAII: A domain-based approach for supporting mobility in wide-area wireless networks," *IEEE/ACM Trans. Networking*, vol. 10, pp. 396–410, June 2002.
[213] T. Rappaport, *Wireless Communications*. New York: IEEE Press, 1996.
[214] F. Rashid-Farrokhi, K. Liu, and L. Tassiulas, "Transmit beamforming and power control for cellular wireless systems," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 1437–1450, Oct. 1998.
[215] A. Reibman and M.-T. Sun, *Compressed Video Over Networks*. New York: Marcel Dekker, 2000.
[216] P. Robertson and S. Kaiser, "Analysis of the loss of orthogonality through Doppler spread in OFDM systems," in *Proc. GLOBECOM*, 1999, pp. 701–706.
[217] J. R. Roche and A. D. Wyner, "Method and apparatus for transmitting signals in a multi-tone code division multiple access communication system," U.S. Patent 5 410 538, 1995.
[218] T. L. Rodeheffer, C. A. Thekkath, and D. C. Anderson, "SmartBridge: A scalable bridge architecture," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 30, no. 4, pp. 205–216, Oct. 2000.
[219] V. Rodoplu and T. H. Meng, "Minimum energy mobile wireless networks," *IEEE J. Select. Areas Commun.*, vol. 17, pp. 1333–1344, Aug. 1999.
[220] M. Russell and G. Stuber, "Interchannel interference analysis of OFDM in a mobile environment," in *Proc. VTC*, 1995, pp. 820–824.
[221] H. Sari, G. Karam, and I. Jeanclaud, "An analysis of orthogonal frequency division multiplexing for mobile radio applications," in *Proc. VTC*, 1994, pp. 1635–1639.
[222] H. Sari, G. Karam, and I. Jeanclaude, "Transmission techniques for digital terrestrial TV broadcasting," *IEEE Commun. Mag.*, pp. 100–109, Feb. 1995.
[223] W. Seifert, "Bridges and routers," *IEEE Network*, vol. 2, pp. 57–64, Jan. 1988.

[224] M. Sellathurai and S. Haykin, "TURBO-BLAST for wireless communications: Theory and experiments," *IEEE Trans. Signal Processing*, pp. 2538–2546, Oct. 2002.
[225] N. Seshadri and J. Winters, "Two signaling schemes for improving the error performance of frequency-division-duplex (FDD) transmission systems using transmitter antenna diversity," in *Proc. IEEE Vehicular Technology Conf. (VTC)*, 1993, pp. 508–511.
[226] S. Shamai and A. D. Wyner, "Information-theoretic considerations for symmetric, cellular, multiple access fading channels, parts I,II," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1877–1911, Nov. 1997.
[227] C. E. Shannon, "A mathematical theory of communications," *Bell Syst. Tech. J.*, vol. 27, pp. 623–656, 1948.
[228] ——, "Channels with side-information at the transmitter," *IBM J. Res. Develop.*, vol. 2, pp. 289–293, Oct. 1958.
[229] A. Shokrollahi, B. Hassibi, B. Hochwald, and W. Sweldens, "Representation theory for high-rate multiple-antenna code design," *IEEE Trans. Inform. Theory*, vol. 47, pp. 2335–2367, Sept. 2001.
[230] M. K. Simon and M. S. Alouini, *Digital Communication Over Generalized Fading Channels: A Unified Approach to the Performance Analysis*. New York: Wiley, 2000.
[231] S. Siwamogsatham, M. Fitz, and J. Grimm, "A new view of performance analysis of transmit diversity schemes in correlated Rayleigh fading," *IEEE Trans. Inform. Theory*, vol. 48, pp. 950–956, Apr. 2002.
[232] R. Soni, M. Buehrer, and R. Benning, "Intelligent antenna system for cdma2000," *IEEE Signal Processing Mag.*, pp. 54–67, July 2002.
[233] A. Stamoulis and N. Al-Dhahir, "Impact of space-time block codes on 802.11 network throughput," *IEEE Trans. Wireless Commun.*, vol. 2, pp. 1029–1039, Sept. 2003.
[234] A. Stamoulis, N. Al-Dhahir, and A. Calderbank, "Further results on interference cancellation and space-time block codes," in *Proc. 35th Asilomar Conf. Signals, Systems and Computers*, vol. 1, 2001, pp. 257–262.
[235] A. Stamoulis, S. Diggavi, and N. Al-Dhahir, "Intercarrier interference in MIMO OFDM," *IEEE Trans. Signal Processing*, vol. 50, pp. 2451–2464, Oct. 2002.
[236] A. Stamoulis, Z. Liu, and G. Giannakis, "Space-time coded generalized multicarrier CDMA with block-spreading for multirate services," in *Proc. Allerton Conf. Communication, Control, and Computing*, vol. 2, 2000, pp. 1076–1085.
[237] G. L. Stuber, *Principles of Mobile Communication*, 2nd ed. Norwell, MA: Kluwer Academic, 2000.
[238] V. Tarokh and H. Jafarkhani, "A differential detection scheme for transmit diversity," *IEEE J. Select. Areas Commun.*, vol. 18, pp. 1169–1174, July 2000.
[239] V. Tarokh, H. Jafarkhani, and A. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1456–1467, July 1999.
[240] ——, "Space-time block coding for wireless communications: Performance results," *IEEE J. Select. Areas Commun.*, vol. 17, pp. 451–460, Mar. 1999.
[241] V. Tarokh, A. Naguib, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communication: Performance criteria in the presence of channel estimation errors, mobility, and multiple paths," *IEEE Trans. Commun.*, vol. 47, pp. 199–207, Feb. 1999.
[242] V. Tarokh, N. Seshadri, and A. Calderbank, "Space-time codes for high data rate wireless communications: Performance criterion and code construction," *IEEE Trans. Inform. Theory*, vol. 44, pp. 744–765, Mar. 1998.
[243] L. Tassiulas and A. Ephremides, "Jointly optimal routing and scheduling in packet radio networks," *IEEE Trans. Inform. Theory*, vol. 38, pp. 165–168, Jan. 1992.
[244] D. Taylor, G. Vitetta, B. Hart, and Mammela, "Wireless channel equalization," *Eur. Trans. Telecommun.*, vol. 9, pp. 117–143, Mar./Apr. 1998.
[245] I. E. Telatar, "Capacity of multiple antenna Gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, no. 6, pp. 585–595, Nov./Dec. 1999.
[246] C. Tepedelenlioglu and G. B. Giannakis, "Transmitter redundancy for blind estimation and equalization of time- and frequency-selective channels," *IEEE Trans. Signal Processing*, vol. 48, pp. 2029–2043, July 2000.
[247] "Space-time block coded transmit antenna diversity for WCDMA," Texas Instruments Inc., Dallas, TX, SMG2 Doc. 581/98, 1998.
[248] TIA 45.5 Subcommittee, "The CDMA 2000 candidate submission (draft)," TIA, Arlington, VA, June 2, 1998.
[249] O. Tirkkonen and A. Hottinen, "Square-matrix embeddable space-time block codes for complex signal constellations," *IEEE Trans. Inform. Theory*, vol. 48, pp. 384–395, Feb. 2002.
[250] L. Tong and S. Perreau, "Mulichannel blind identification: From subspace to maximum likelihood methods," *Proc. IEEE*, vol. 86, pp. 1951–1968, Oct. 1998.
[251] L. Tong, Q. Zhao, and G. Mergen, "Multipacket reception in random access wireless networks: From signal processing to optimal medium access control," *IEEE Commun. Mag.*, pp. 108–112, Nov. 2001.
[252] M. Tsatsanis and Z. Xu, "Pilot symbol assisted modulation in frequency selective fading wireless channels," *IEEE Trans. Signal Processing*, vol. 48, pp. 2353–2365, Aug. 2000.
[253] M. Tsatsanis, R. Zhang, and S. Banerjee, "Network assisted diversity for random access wireless networks," *IEEE Trans. Signal Processing*, vol. 48, pp. 702–711, Mar. 2000.
[254] D. Tse and S. Hanly, "Linear multiuser receivers: Effective interference, effective bandwidth and user capacity," *IEEE Trans. Inform. Theory*, vol. 45, pp. 641–657, Mar. 1999.
[255] ——, "Spectral efficiency of CDMA with random spreading," *IEEE Trans. Inform. Theory*, vol. 45, pp. 622–640, Mar. 1999.
[256] D. N. C. Tse, "Optimal power allocation over parallel Gaussian broadcast channels," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, 1997, p. 27.
[257] J. Uddenfeldt and A. Raith, "Cellular digital mobile radio system and method of transmitting information in a digital cellular mobile radio system," U.S. Patent 5 088 108, Feb. 1992.
[258] M. Uysal, N. Al-Dhahir, and C. N. Georghiades, "A space-time block-coded OFDM scheme for unknown frequency-selective fading channels," *IEEE Commun. Lett.*, vol. 5, pp. 393–395, Oct. 2001.
[259] E. Uysal-Biyikoglu, B. Prabhakar, and A. El-Gamal, "Energy-efficient packet transmission over a wireless link," *IEEE/ACM Trans. Networking*, vol. 10, pp. 487–499, Aug. 2002.
[260] V. A. Vaishampayan, "Design of multiple description scalar quantizers," *IEEE Trans. Inform. Theory*, vol. 39, pp. 821–834, May 1993.
[261] V. A. Vaishampayan, N. Sloane, and S. Servetto, "Multiple description vector quantization with lattice codebooks: Design and analysis," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1718–1734, July 2001.
[262] A. Valko, "Cellular IP: A new approach to Internet host mobility," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 29, no. 1, pp. 50–65, Jan. 1999.
[263] E. C. van der Meulen, "A survey of multiway channels in information theory," *IEEE Trans. Inform. Theory*, vol. 23, pp. 1–37, Jan. 1977.
[264] R. van Nobelen, N. Seshadri, J. Whitehead, and S. Timiri, "An adaptive radio link protocol with enhanced data rates for GSM evolution," *IEEE Pers. Commun.*, vol. 6, pp. 54–64, Feb. 1999.
[265] M. Varanasi and T. Guess, "Optimum decision feedback multiuser equalization with successive decoding achieves the total capacity of the Gaussian multiple-access channel," in *Proc. Asilomar Conf. Signals, Systems and Computers*, 1997, pp. 1405–1409.
[266] M. K. Varanasi and B. Aazhang, "Multistage detection in asynchronous code-division multiple-access communications," *IEEE Trans. Commun.*, vol. 38, pp. 509–519, Apr. 1990.
[267] B. V. Veen and K. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, pp. 4–24, Apr. 1988.
[268] S. Verdu, *Multiuser Detection*. Cambridge, U.K.: Cambridge Univ. Press, 1998.
[269] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Trans. Inform. Theory*, vol. 49, pp. 2658–2668, Oct. 2003.
[270] E. Visotsky and U. Madhow, "Space-time transmit precoding with imperfect feedback," *IEEE Trans. Inform. Theory*, vol. 47, pp. 2632–2639, Sept. 2001.
[271] P. Viswanath, "Capacity of vector multiple access channel," Ph.D. dissertation, Univ. California, Berkeley, 2000.
[272] P. Viswanath and D. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, vol. 49, 2003, pp. 1912–1921.
[273] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1277–1294, June 2002.

[274] H. Viswanathan, "Capacity of Markov channels with receiver CSI and delayed feedback," *IEEE Trans. Inform. Theory*, vol. 45, pp. 761–770, Mar. 1999.
[275] A. J. Viterbi, *CDMA: Principles of Spread Spectrum Communications*. Reading, MA: Addison-Wesley, 1995.
[276] E. Viterbo and J. Boutros, "A universal lattice code decoder for fading channels," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1639–1642, July 1999.
[277] L.-C. Wang, G. L. Stuber, and C.-T. Lea, "Architecture design, frequency planning, and performance analysis for a microcell/macrocell overlaying system," *IEEE Trans. Veh. Technol.*, vol. 46, pp. 836–848, Apr. 1997.
[278] X. Wang and H. V. Poor, "Iterative (turbo) soft interference cancellation and decoding for coded CDMA," *IEEE Trans. Commun.*, vol. 47, pp. 1046–1061, July 1999.
[279] V. Weerackody, "Characteristics of a simulated fast fading indoor radio channel," in *Proc. IEEE Vehicular Technology Conf.*, 1993, pp. 231–235.
[280] S. Weinstein and P. Ebert, "Data transmission by frequency-division multiplexing using the discrete Fourier transform," *IEEE Trans. Commun.*, vol. COM-19, pp. 628–634, Oct. 1971.
[281] J. Winters, J. Salz, and R. Gitlin, "The impact of antenna diversity on the capacity of wireless communication systems," *IEEE Trans. Commun.*, vol. 42, pp. 1740–1751, Feb./Mar./Apr. 1994.
[282] A. Wittneben, "A new bandwidth efficient transmit antenna modulation diversity scheme for linear digital modulation," in *Proc. ICC*, 1993, pp. 1630–1634.
[283] G. Wornell and M. Trott, "Efficient signal processing techniques for exploiting transmit antenna diversity on fading channels," *IEEE Trans. Signal Processing*, vol. 45, pp. 191–205, Jan 1997.
[284] A. D. Wyner, "Recent results in the Shannon theory," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 2–10, Jan. 1974.
[285] L.-L. Xie and P. V. Kumar, "Network information theory for wireless communications," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, 2003, p. 288.
[286] J. Yang and S. Roy, "On joint transmitter and receiver optimization for multiple-input-multiple-output (MIMO) transmission systems," *IEEE Trans. Commun.*, vol. 42, pp. 3221–3231, Dec. 1994.
[287] W. Younis and N. Al-Dhahir, "Joint prefiltering and MLSE equalization of space-time-coded transmissions over frequency-selective channels," *IEEE Trans. Veh. Technol.*, vol. 51, pp. 144–154, Jan. 2002.
[288] W. Younis, N. Al-Dhahir, and A. Sayed, "Adaptive frequency-domain equalization of space-time block-coded transmissions," in *Proc. ICASSP*, 2002, pp. 2353–2356.
[289] W. Yu, "Competition and cooperation in multiuser communication environments," Ph.D. dissertation, Stanford Univ., Stanford, CA, 2002.
[290] L. Zheng and D. N. C. Tse, "Communication on the Grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel," *IEEE Trans. Inform. Theory*, vol. 48, pp. 359–383, Feb. 2002.
[291] ——, "Diversity and multiplexing: A fundamental tradeoff in multiple antenna channels," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1073–1096, May 2003.
[292] A. Zhou and G. B. Giannakis, "Space-time coding with maximum diversity gains over frequency-selective fading channels," *IEEE Signal Processing Lett.*, vol. 8, pp. 269–272, Oct. 2001.

Suhas N. Diggavi (Member, IEEE) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology (IIT), Delhi, India, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 1998.

From 1998 to 2003, he was a Principal Member of Technical Staff at the Information Sciences Center, AT&T Shannon Laboratories, Florham Park, NJ. Since 2003, he has been on the faculty of the School of Computer and Communication Services, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, where he is Head of the Laboratory of Information and Communication Systems (LICOS). His research interests include wireless networks, information theory, source coding, and signal processing.

Naofal Al-Dhahir (Senior Member, IEEE) received the M.S. and Ph.D. degrees in Electrical Engineering from Stanford University, Stanford, CA, in 1990 and 1994, respectively.

In 1992, he was an Instructor at Stanford University. From 1994 to 1999, he was a Member of the technical staff at General Electric's R&D center, where he worked on various aspects of satellite communication systems design and anti-jam GPS receivers. From 1999 to 2003, he was a Principal Member of technical staff at AT&T Shannon Laboratory, Florham Park, NJ, where he worked on space–time coding and signal processing. He is currently an Associate Professor at the University of Texas, Dallas. He has authored 50 journal papers and holds 11 U.S. patents. He is coauthor of *Doppler Applications for LEO Satellite Systems* (Norwell, MA: Kluwer, 2001). His current research interests include space–time coding and signal processing, OFDM, wireless networks, and digital subscriber line technology.

Dr. Al-Dhahir is a Member of the IEEE SP4COM technical committee. He served as Editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING and IEEE COMMUNICATIONS LETTERS. He is currently an Editor for IEEE TRANSACTIONS ON COMMUNICATIONS and will serve as cochair of the Communication Theory Symposium at Globecom'04.

A. Stamoulis, photograph and biography not available at the time of publication.

A. R. Calderbank (Fellow, IEEE) received the B.Sc. degree in mathematics from the University of Warwick, Coventry, U.K., in 1975, the M.Sc. degree in mathematics from Oxford University, Oxford, U.K., in 1976 and the Ph.D. degree in mathematics from the California Institute of Technology, Pasadena, in 1980.

He joined Bell Telephone Laboratories as a Member of Technical Staff in 1980, and retired from AT&T in 2003 as Vice President of Research. He is currently a Visiting Scholar in the Program for Applied and Computational Mathematics at Princeton University, Princeton, NJ. He was Associate Editor for *Coding Techniques* from 1986 to 1989. His research interests that range from algebraic coding theory and quantum computing to the design of wireless systems.

Dr. Calderbank served as Editor in Chief of the IEEE TRANSACTIONS ON INFORMATION THEORY from 1995 to 1998. He was a member of the Board of Governors of the IEEE Information Theory Society from 1991 to 1996. Dr. Calderbank was honored by the IEEE Information Theory Prize Paper Award in 1995 for his work on the Z_4 linearity of Kerdock and Preparata Codes (joint with A. R. Hammons, Jr., P. V. Kumar, N. J. A. Sloane, and P. Sole), and again in 1999 for the invention of space–time codes (joint with V. Tarokh and N. Seshadri). He is a recipient of the IEEE Millennium Medal, and was honored by AT&T's highest technical award in 2000 when he was appointed an AT&T Fellow.