◆ Physical-Layer Design for Next-Generation Cellular Wireless Systems

Gerard J. Foschini, Howard C. Huang, Sape J. Mullender, Sivarama Venkatesan, and Harish Viswanathan

The conventional cellular architecture will remain an integral part of next-generation wireless systems, providing high-speed packet data services directly to mobile users and also backhaul service for local area networks. In this paper, we present several proposals addressing the challenges associated with the critical design issues for the next-generation cellular wireless physical layer. These proposals include an orthogonal frequency division multiplexed (OFDM) or multicarrier code division multiple access (MC-CDMA) air interface for wide channel bandwidths, considerations for resource allocation, multiple antenna techniques, a hybrid duplexing technique combining the benefits of time division and frequency division duplexing, and coordinated networks for reducing intercell interference and increasing overall spectral efficiency.

Introduction

Third-generation (3G) wireless cellular networks efficiently provide circuit voice services and packet data services in bandwidths up to 5 MHz. Now that 3G networks are being successfully deployed, the focus and context of wireless research is shifting to the next generation of wireless networks. These next-generation wireless systems are expected to deliver a wide variety of high data rate services with packet-based access and variable data rates, as opposed to constant rate, circuit-based services, for 3G voice users. While alternative architectures such as local area, ad-hoc, or relay networks are expected to be components of this system, the conventional cellular architecture will remain an integral part, providing high-speed packet data services directly to mobile users and also backhaul service for local area networks. These cellular networks must support an order-of-magnitude increase compared to 3G in peak rates and throughput using a combination of wider channel bandwidths and increased spectral efficiency. These higher data rates will be required to support new applications such as video unicast and broadcast. Minimizing physical layer latency is also an important requirement, from the point of view of optimizing performance at higher layers (as is well known, Transmission Control Protocol [TCP] performance is adversely affected by large round-trip delays). In this paper, we present several proposals for the physical layer design of a next-generation wireless network, including the air interface design, resource allocation, multiple antenna techniques, duplexing method, and network coordination.

While this paper is focused on the physical layer design, we briefly describe the higher layer design to provide context. We envision that the access network of next-generation systems will be based on a

Bell Labs Technical Journal 10(2), 157–172 (2005) © 2005 Lucent Technologies Inc. Published by Wiley Periodicals, Inc. Published online in Wiley InterScience (www.interscience.wiley.com). • DOI: 10.1002/bltj.20099

distributed radio network architecture in which radio resource management and mobility management functions are performed at the base stations without any centralized node with radio-related functionality. This is in contrast to 3G network architectures where a single, central, radio network controller performs most radio resource management functions for a large number of base stations. Furthermore, the architecture would be all-IP in the sense that IP traffic will be terminated directly at the base stations that have IP routing and mobile-IP foreign agent functionality. Quality of service requirements in the network will be based on IP, for example, using Differentiated Services (DiffServ). Data transport and control signaling will be separated within the network with control signaling implemented through an IP-based call state control function defined as a part of the IP Multimedia Subsystem (IMS) architecture. With IP traffic terminated at the base stations, the physical and link layers of the backhaul link could be arbitrary, so with sufficient bandwidth, Asynchronous Transfer Mode (ATM), T1/E1, Ethernet, Optical, or Gigabit Ethernet could carry the traffic.

Current 3G cellular system air interfaces are based on code division multiple access (CDMA). Compared to narrowband multiple access techniques based on frequency division multiple access (FDMA) or time division multiple access (TDMA), CDMA provides high system spectral efficiency for voice services due to the benefits of interference averaging and frequency diversity. In the downlink of 3G packet data systems such as CDMA2000* 1xEV-DO, signals are modulated using CDMA. However, each base station transmits to only a single data user at a time, and service to multiple users is time multiplexed. The 1xEV-DO system operates in 1.25 MHz bandwidth and can provide peak data rates in excess of 2 Mb/s. At this bandwidth, the frequency selectivity of the channel causes minimal to moderate self-interference among the CDMA codes due to time-dispersive multipath fading. Furthermore, the intersymbol interference (ISI) is likewise manageable. The receiver can effectively mitigate both effects using a rake receiver or a relatively simple equalizer. Next-generation systems will provide higher peak rates, obtained by higher bandwidths and spectral efficiency. In higher bandwidths, the self-interference can be quite severe, and the sophisticated CDMA equalizers needed to mitigate it could become prohibitively complex.

In the "Air Interface" section, we present two air interface options that address these complexity issues while still providing high spectral efficiency and

Panel 1. Abbreviations, Acronyms, and Terms

1xEV-DO—CDMA2000* evolution–data optimized
3G—Third generation
802.16—The IEEE standard for wireless metropolitan networks, also known as WiMAX, WirelessMAN, and WMAN
ATM—Asynchronous transfer mode
BLAST—Bell Labs Layered Space-Time
CCT—Coherently coordinated transmission
CDMA—Code division multiple access
CSI—Channel state information
DiffServ—Differentiated Services
DPC—Dirty paper coding
FDD—Frequency division duplexing
FDMA—Frequency division multiple access
FFT—Fast Fourier transform
GPS—Global Positioning System
HSDPA—High-speed downlink packet access
IEEE—Institute of Electrical and Electronics Engineers
IFFT—Inverse FFT
IMS—IP Multimedia Subsystem
IP—Internet Protocol
ISI—Intersymbol interference
MC-CDMA—Multicarrier CDMA
MRC—Maximal ratio combining
OFDM—Orthogonal frequency division multiplexed/multiplexing
SBT—Single-base transmission
SDMA—Spatial division multiple access
SINR—Signal-to-interference-plus-noise ratio
SNR—Signal-to-noise ratio
TCP—Transmission Control Protocol
TDD—Time division duplexing
TDMA—Time division multiple access
UMTS*—Universal Mobile Telecommunications System
WiMAX—An implementation of the IEEE 802.16 standard
ZF—Zero forcing

robustness against channel impairments in very high bandwidths. These options are multicarrier CDMA (MC-CDMA) and orthogonal frequency division multiplexing (OFDM). We discuss the benefits and tradeoffs of these two options and highlight key design parameters. Both of these air interfaces divide the total available bandwidth into multiple subchannel resources that can be transmitted in parallel within each cell. Resource allocation of these subchannels to users and cells becomes an important issue when trying to improve spectral efficiency and mitigate intercell interference. In the "Resource Allocation" section, we introduce a taxonomy of options for resource allocation and discuss the benefits and tradeoffs associated with each. Multiple antennas will be an essential component of next-generation systems for improving spectral efficiency. In the "Multiple Antenna Techniques" section, we present a survey of various multiple antenna techniques and discuss the most suitable options. A fundamental design choice of any two-way wireless communication system is the duplexing technique for separating the uplink and downlink transmissions. The "Duplexing" section describes a novel technique for duplexing called band switching, which aims to combine the benefits of both frequency and time division duplexing while avoiding their respective drawbacks. This technique allows base stations to measure channels on the uplink and to use this channel state information on the downlink for beamforming, scheduling, or more sophisticated transmission via network coordination. Motivated by the demand for more spectrally efficient networks beyond what can be attained by air interface choices and conventional link-level multiple antenna techniques, there has been interest in understanding and enhancing the performance of multicell networks by allowing cooperation and coordination among base stations across the network. The "Network Coordination" section discusses two options for network coordination. In the final section of the paper, we conclude by synthesizing the options presented in the previous sections and by offering possible system configurations for near-, medium-, and long-term deployment scenarios.

Our goal in this paper is to discuss underlying system-level design issues and to present a qualitative overview of potential system evolution options. Quantitative comparisons are outside the scope of this paper. While some detailed performance results can be found in references, further quantitative work should be done to provide a full understanding of the relative merits and tradeoffs. Furthermore, we are not interested in comparing merits of specific implementations. Nevertheless, we hope the high-level insights and enhancements presented here can be applied to present CDMA-based systems such as 1xEV-DO and UMTS* HSDPA and to near-term OFDM proposals such as WiMax and 802.16.

Air Interface

An air interface for next-generation cellular systems is expected to meet requirements supporting data rates and system throughputs that are an order of magnitude higher than those in existing 3G systems. Burst rates in the 10–100 Mb/s range for cellular deployments are envisioned. Support for low-latency services and applications is also important. The range of desired data rates will probably require a channel bandwidth in the 10–20 MHz range. As mentioned earlier, the dominant impairment on such a channel is likely to be the self-interference caused by resolvable multipath, at least in pedestrian or moderate-speed vehicular environments. Arguably, these are the environments where the higher data rates are likely to be used.

The two promising options for the next-generation air interface are both based on the principle of dividing the channel bandwidth into multiple parallel subchannels. In the first option, MC-CDMA, each subchannel is modulated using CDMA spreading sequences in the time domain. (We note that the MC-CDMA system described here should not be confused with an alternative system with frequency-domain spreading, which sometimes goes by the same name.) The second option, OFDM, divides the bandwidth into much narrower subchannels (sometimes called subcarriers or tones), and each is modulated directly by the information-bearing symbols. In the following subsections, we describe the motivation, parameter design, benefits, and tradeoffs of these two air interface options.

MC-CDMA

The current 3G packet data system based on the 1xEV-DO standard uses CDMA modulation and operates in a 1.25 MHz with a chip timing of 1.2288×10^6 chips per second. We propose that the MC-CDMA air interface consists of multiple parallel 1.25 MHz subchannels using the same chip and symbol timing as the 1xEV-DO system. By basing the MC-CDMA system on the legacy system, migration to the next generation for the product developer is greatly simplified since much of the hardware, including radios and chip-level baseband processing, can be reused with relatively minor modifications. In 1xEV-DO, multipath interference on the downlink due to intercode interference is most effectively mitigated using a relatively simple equalizer. Similar robustness to multipath fading could be retained in an MC-CDMA system by using receivers of similar complexity in each subchannel. In general, the system complexity would scale approximately linearly with the number of subchannels.

In a 20 MHz total channel bandwidth, after accounting for necessary guard bands, it would be possible to accommodate 12 or more subchannels of 1.25 MHz bandwidth. The bandwidth of the guard bands can be reduced to the point where intercarrier interference is about the same level as the worst-case intercell interference. Under the assumption of universal frequency reuse (see Section 3), the intercell interference could be quite high so that the guard band can be quite narrow. Therefore the subchannel packing can be correspondingly aggressive.

Depending on the resource allocation technique (see the "Resource Allocation" section), one could transmit on any number of subchannels, and we would want to perform channel coding across the multiple subchannels in order to benefit from the frequency diversity offered in such a wide bandwidth. Coding across subchannels has the additional benefit of reducing the frame period on the order of the number of subchannels coded over, allowing for much lower latencies. Alternatively, one could encode independently within each subchannel. In this case, the baseband processing would be greatly simplified because one could simply implement multiple parallel 1xEV-DO receivers followed by a multiplexer, but this option does not take advantage of frequency diversity as effectively. For channel estimation at the receiver, pilot signals should be multiplexed on each subchannel.

An MC-CDMA approach could also be applied to a UMTS CDMA system by concatenating multiple 5 MHz channels. Because of the wider bandwidth, there is more frequency diversity in each subchannel; however, the tradeoff is a more complex equalizer to mitigate the multipath interference. In general, an MC-CDMA air interface could be implemented in a similar fashion for the uplink. For a given subchannel, spreading codes could be assigned to the same user or to different users. In the latter case, even when there is no delay spread, because of the timing requirements, it would be practically impossible to maintain orthogonality among codes received by the base.

OFDM

In OFDM, a large number of orthogonal narrowband subchannels (also known as subcarriers in the context of OFDM) are transmitted in parallel. In order to achieve very compact spectral utilization, there is some overlap between spectra of adjacent subchannels, but signals are modulated so that the spectrum nulls and peaks are properly aligned to minimize interference [12]. The structure of OFDM effectively addresses both the frequency selectivity and ISI caused by multipath fading in high bandwidths. For the former, because each subchannel operates over a narrow bandwidth, frequency selectivity in any subchannel is essentially nonexistent. For the latter, modulating at a very low symbol rate on each subchannel makes the symbols much longer than the channel impulse response, thereby diminishing the ISI. Using powerful error-correcting codes together with time and frequency interleaving yields even more robustness against frequency selective fading, and the insertion of an extra guard interval between consecutive OFDM symbols can further reduce the effects of ISI. Therefore, OFDM is well-suited to handle multiple impairments with reasonable complexity. An additional advantage of OFDM over CDMA is that

on the uplink, it allows the possibility of orthogonalization of users without requiring time division multiplexing. Furthermore, the modulation and demodulation for the multiple subchannels can be efficiently implemented using inverse fast Fourier and fast Fourier transforms (IFFT and FFT), respectively. Previously, a major problem of OFDM was in designing amplifiers to address the peak-to-average power problem. Many proposals have been made for addressing this problem with minimal loss of efficiency (see, e.g., [14]).

Here, for the purpose of illustration, we propose a skeletal OFDM design for the downlink of a next-generation cellular system that is designed for high-speed packet data services in a 20 MHz channel (the parameters can be scaled appropriately for use with other bandwidths). This design utilizes 810 OFDM subcarriers spaced 20 kHz apart, for an approximate channel bandwidth of 16.2 MHz, thus permitting a carrier spacing of 20 MHz. A 1024-point IFFT, with a sampling rate of 20.48 MHz, is used at the transmitter to modulate the subcarriers efficiently. A cyclic prefix of 176 samples is added to each 1024-sample IFFT output, and the resulting OFDM symbol is subjected to a 22-sample raised-cosine windowing operation for spectral shaping. These parameters provide immunity to delay spreads of up to 154 samples, or about 7.5 microseconds, with less than 1 dB overhead. With a subchannel spacing of 20 kHz, Doppler spreads up to 200 Hz should be tolerable (this corresponds to a vehicular speed of about 100 km/h at 2 GHz). Channel estimation at the receiver is facilitated by pilot symbols spread across the band.

On the uplink, a similar design can be used, except that a larger guard interval may be needed in each OFDM symbol to accommodate the largest differential path delay expected between users in the same cell. This results in a larger overhead in order to maintain user orthogonality. Alternatively, each user could adjust the timing of its uplink transmissions to compensate for the path delay to the intended base, thereby reducing the required guard interval duration (this would require the transmission of a reference timing signal on the downlink). Naturally, larger cells would require longer guard intervals (on both uplink and downlink), so it would be useful to make the duration of the guard interval flexible in order to accommodate a wide range of deployments.

In comparing MC-CDMA with OFDM, the granularity of the subchannels for OFDM is much finer (20 kHz versus 1.25 MHz). So, in general, OFDM is more robust than MC-CDMA to delay spread because the symbol period in OFDM is much longer than the chip period for MC-CDMA. The dual statement would be that MC-CDMA is more robust than OFDM to doppler spread because the subchannel spacing in MC-CDMA is much wider than that of OFDM. However, this dual statement would apply if there was only a single spreading code within a MC-CDMA channel. Because the MC-CDMA channel in general consists of multiple spreading codes, the code orthogonality is not maintained when there is substantial doppler spread. Consequently, if MC-CDMA uses multiple spreading codes per channel, then MC-CDMA and OFDM would have comparable robustness against doppler spread. In terms of receiver complexity, the CDMA equalizer would typically be more complex than the OFDM receiver for a given bandwidth.

Resource Allocation

Downlink transmission in 3G systems is typically done in unicast mode where each base station transmits independent information to separate users either to one user at a time or to multiple users simultaneously over multiplexed subchannels. We expect that unicast will remain an important transmission mode for next-generation systems, but we also expect broadcasting to become an important mode where all bases synchronously transmit identical information to all users in the network. For both transmission modes, in a multicell environment the challenge is to allocate subchannel resources in an efficient manner so that intercell interference effects are minimized. In this section we discuss the different resource allocation options shown in **Figure 1** and their relative merits, first for unicast and then for broadcast. In general, the discussion can be applied to either OFDM or MC-CDMA since, in the context of resource allocation, the main difference in the subchannel resources

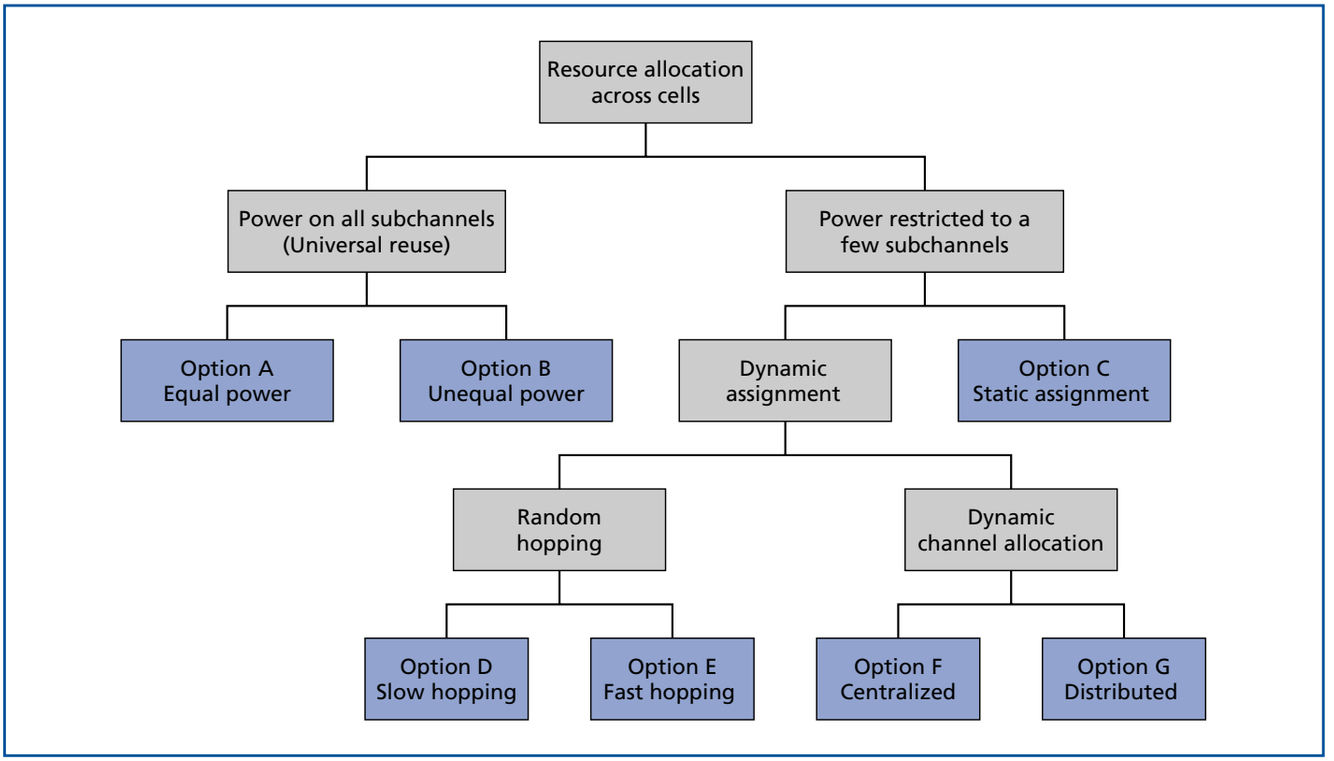

Figure 1.
Taxonomy of resource allocation options for multicell OFDM.

is the subchannel bandwidth and the number of subchannels per total transmit bandwidth.

In the first class of options, all base stations transmit power on all subchannels, meaning that universal frequency reuse is employed. Frequency reuse will not be required with proper design of link layer rates, and thus no frequency planning at the time of deployment is required. This approach is similar to the approach in the 1xEV-DO system where all CDMA codes are always used in every cell. Interference power from surrounding cells is received by mobiles without significant fluctuation across code blocks. The base station can optimally assign different mobiles to different subchannels depending on the channel conditions. Traditional opportunistic scheduling can be extended to the frequency dimension as well for this class of options. Variable rate transmission is achieved using adaptive coding and modulation; for very low coding rates, symbol repetition can be used. Since all subchannels are used to signal to mobiles, significant frequency diversity gains can be obtained. Within this class, the two options A and B correspond respectively to equal and unequal power assignment on the subchannels. Option B may be more difficult to implement since it requires knowledge of the subchannel conditions at the transmitter. However, by matching the powers to subchannel conditions, optimal power allocation can be achieved through waterfilling techniques. In low signal-to-noise ratio (SNR) conditions, the performance benefits of waterfilling are significant compared to equal power allocation. However, the performance difference diminishes with increasing SNR.

In contrast to the above two options, the remaining approaches involve using only some of the subchannels in a given cell. In Option C, subchannels of a given carrier are statically divided among different adjacent cells. This method is essentially equivalent to traditional fixed frequency reuse. Thus, it involves frequency planning and the interference is somewhat fixed. Because of its rigid nature of resource allocation across cells, this method does not exploit variations in

traffic patterns across cells and thus can be no better than the schemes that use dynamic (time-varying) assignments. The dynamic subchannel assignment can be accomplished through either random frequency hopping or coordinated dynamic allocation [4].

Under random hopping, subchannel assignment is done independently in each cell and is changed over time using random sequences. Thus, no frequency planning is required in this case. Frequency hopping provides frequency diversity and interference averaging. In slow frequency hopping (Option D), the subchannel assignment is fixed over the duration of an entire channel code block that spans multiple symbols, while in fast hopping (Option E) frequency hopping is done within a code block from symbol to symbol. The main difference between slow hopping and fast hopping occurs when the bandwidth occupied by a channel code block is too small to achieve frequency diversity. In this case, with slow hopping, there may be insufficient interference averaging within the code block. Consequently, there may be significant interference fluctuation across code blocks making it difficult to do adaptive modulation. As a result, fast hopping would be preferred when the bandwidth is not large (e.g., in a low data rate OFDM transmission). We note that the performance of random hopping can be improved through retransmission (e.g., in the form of incremental redundancy).

As a more sophisticated alternative to random hopping, tones can be dynamically assigned across cells in a coordinated manner by taking into account the specific terminals and their path loss from the different base stations from which transmission is scheduled (Option F). This results in optimum management of out-of-cell interference and maximizes throughput. However, the implementation requires exchange of messages between base stations and the radio network controller at a very fine time scale and thus may not be practical in the near term. An alternative is to achieve partial coordination in a distributed manner (Option G). This could be achieved, for example, through feedback from the mobile, providing information on its measured signal strength of signals from base stations in its local vicinity. Another approach is a hybrid of fixed frequency planning and hopping, in which base stations start using tones that are disjoint and as traffic grows they start using additional tones that are reserved for neighboring base stations. Such a scheme avoids interference at low traffic conditions and will require frequency planning.

When evaluating the options presented above in the design of next-generation systems, one of the main objectives is deployment flexibility. In fact, the base stations should be able to autonomously configure themselves. Thus, it is logical to ignore options that involve frequency planning, and we reject Options C and G from further consideration. Next, consider that Option F involves intercell coordination. It is likely that intercell coordination does not become practical in the near future because of the complexity involved in the coordination process in a mobile environment. Furthermore, we lose scalability or have to restrict the amount of coordination. Thus, we exclude Option F from further consideration in the remainder of this section. However, in the "Network Coordination" section, we assume a longer term perspective and consider network coordination as an option. We are thus left with the options of random tone hopping (Options D and E) or spreading the power across all the tones (Options A and B).

On the uplink, we choose tone hopping over transmitting across all tones for two reasons. First, since mobiles have limited power, it is more spectrally efficient to transmit over a narrower bandwidth for a given power. Second, the effects of interference are less when tone hopping is used to average the interference. On the downlink, the option where each base transmits power across all the tones is preferred. This can potentially exploit frequency diversity better than frequency hopping because of coherent combining across carriers (as opposed to achieving frequency diversity through coding in the frequency hopping case) for low rate transmissions or through optimum assignment of users to tones. Furthermore, with frequency reuse, the relative gains obtained from improved signal-to-interference-plus-noise ratio (SINR) characteristics on the downlink do not overcome its inherent spectral inefficiency [1]. This observation provides additional motivation for universal frequency reuse.

The preceding discussion in this section was in the context of unicast transmission. For broadcast transmissions, the mobile would combine the signals from all bases within range to maximize diversity. One could restrict transmission to only a subset of subchannels per base, but since interference is not an issue in broadcasting, any modest advantage obtained from frequency diversity is outweighed by its inefficient spectral untilization. Therefore, each base should transmit power on all subchannels, and because the subchannel characteristics to each mobile will be different, the logical choice for broadcast applications is Option A and transmitting with equal power on all subchannels. Because a mobile combines signals from multiple bases, timing differences among them will lead to intercell interference, which degrades the SINR. Interference can be reduced by increasing guard times between symbols, but doing so increases the signaling overhead. One can optimize the signaling overhead in order to maximize the resulting SINR for an OFDM system or an MC-CDMA system with receiver equalizers.

Multiple Antenna Techniques

Multiple antennas at the transmitter and/or at the receiver can be used in a variety of ways to improve the link and system performance of wireless systems. The various techniques, also known as multiple-input multiple-output, or MIMO, techniques, can be broadly classified based on the number of antennas at the transmitter and receiver, and they can be applied in most cases to either uplink or downlink transmission. As we did for resource allocation, we present a brief survey of various multiple antenna techniques before deciding on the best options for our next-generation cellular system.

As a baseline, we consider a single point-to-point link with a single transmit antenna and single receive antenna where the channel is known at the receiver. In contrast to the baseline, suppose we use $N > 1$ receive antennas and we assume that the channels to each antenna are independently fading and known at the receiver. Then maximal ratio combining (MRC) of the received signal improves the link performance and provides diversity gain N and an average combining gain "N. By diversity gain N," assuming an ergodic channel, we mean that, for fixed N, the probability of error averaged over the fading at high SNR is proportional to the SNR raised to the power $-N$. This is because the transmitted signal arrives at the receiver through N independently fading channels. By "combining gain N," we mean that the average SNR after combining is increased by a factor of N compared to the baseline. For indoor receivers, multiple antennas can be used to counteract in-building penetration losses. In a system context where multiple transmitters are present, more sophisticated optimum combining [19] can be used to mitigate the interfering transmissions.

We now consider options when there is a single receive antenna but multiple ($M > 1$) transmit antennas. In this case, there are two broad options. The first is transmit diversity whereby the same information is transmitted from different antennas using space-time coding. If the transmit power is fixed with respect to the single antenna baseline, then transmit diversity via space-time coding potentially provides diversity gain M when the channel is not known at the transmitter [2]. In wideband transmissions that already experience frequency diversity, the marginal benefit due to transmit diversity may provide only minimal link performance improvement. If power information is known at the transmitter (in other words, if the received powers of the transmit antennas are known by the transmitter), then selection transmit diversity can be used to transmit a single stream using the antenna that is received with the highest instantaneous SNR. Otherwise, if coherent channel realizations are known at the transmitter (e.g., through the duplexing technique described in the "Duplexing" section), ideal coherent weighting can be performed at the transmitter, and an average combining gain of M and diversity gain M can be achieved, similar to the MRC case.

Whereas transmit diversity improves the link performance for a single user, the second broad option, spatial division multiple access (SDMA) [13], improves the overall system throughput by transmitting simultaneously to multiple users over separate spatial channels. Note that SDMA provides capacity

gains even with single-antenna mobiles. With M antennas, up to M separate spatial channels (with varying degrees of mutual interference) can be formed to M distinct receivers. These spatial channels can be fixed (in the form of beamforming or through physical antenna backplanes) or can be dynamically adapted to individual users. Dynamically adapted beams can be formed using implicit channel estimates obtained from uplink transmissions [15] or using explicit channel knowledge obtained through feedback or appropriate duplexing (see the "Duplexing" section). For dynamically adapted beams, it is also possible to place spatial nulls in the directions of interferers. Sectorization can also be considered as a form of SDMA, and it is probably the simplest way to get significant system throughput improvements with multiple transmit antennas since it does not require coherent channel knowledge at the transmitter. Recently, a sophisticated SDMA technique using beamforming and a sphere detection algorithm has been proposed in [8], which reports impressive capacity performance for a system with a multiple antenna transmitter and many single-antenna receivers. This technique requires coherent channel knowledge at the transmitter and offers a potentially practical way to implement the closely related coherent coordinated transmission techniques discussed in the "Network Coordination" section.

While higher data rates can be achieved by increasing bandwidth, the need for higher spectral efficiency necessitates exploring other techniques. When multiple antennas are available at both the transmitter and receiver, a general technique, known as Bell Labs Layered Space-Time (BLAST) transmission [6], can be an effective means of spatial multiplexing for improving the spectral efficiency, and thereby increasing the peak rate, of point-to-point links. It is well known, for example, that, with M transmit antennas and M receive antennas, the average gain in capacity compared to the single-antenna baseline is proportional to M. In practical implementations with finite signaling constellations, it is most efficient to use single-stream transmit diversity techniques at lower SNRs and to reserve BLAST transmission for higher SNRs [9]. Therefore, the techniques discussed in this paper for increasing the SNR (e.g., multiple receive antennas, resource allocation, and network coordination) can be thought of as increasing overall spectral efficiency by enabling the use of BLAST techniques.

The three classes of multiple antenna transmission techniques (transmit diversity, SDMA, and spatial multiplexing) in general can be used together in any combination. For example, with four transmit antennas, one could achieve both diversity and spatial multiplexing gain by transmitting to a single user two independent data streams, each of which is encoded with Alamouti space-time block coding across two antennas [11]. A rigorous treatment of the tradeoffs between diversity and spatial multiplexing gains is developed in [16]. If these four antennas had physical backplanes to limit their transmission within a sector, then 4 additional antennas transmitting in a similar manner could be used to serve users in another sector. In this way SDMA can be achieved on top of transmit diversity and spatial multiplexing using a total of eight antennas.

Based on the discussions in previous sections, the characteristics of a next-generation cellular system include wide channel bandwidth, downlink universal frequency reuse, and uplink frequency hopping. These design choices will impact the multiple antenna design choices. Because of the frequency diversity achieved from the wide channel bandwidth, transmit diversity in general will not provide much link performance improvement. For near-term deployments where mobile terminals will most likely have only one or two antennas, the BLAST transmission opportunity should be carefully evaluated because universal frequency reuse would preclude high SNRs over much of the cell area. If each mobile had more antennas, BLAST transmission could be used over a larger fraction of the cell. On the uplink, it is possible for BLAST to be used especially if the mobile is close enough to the base, if enough power is concentrated in a sufficiently small subchannel, and if the hopping patterns create favorable interference averaging.

In general, SDMA is a very efficient way to increase the overall system throughput, since it can be implemented with single-antenna terminals. Assuming a

uniform distribution of users throughout a cell, sectorization is a simple and effective SDMA technique. We recommend that the multiple antenna strategy to be used first on both the uplink and downlink is sectorization and that the cell be divided into the maximum number of sectors determined by the following two factors. First, the angular width of the sector must be larger than the angle spread of the channel; otherwise, there would be unacceptable interference from adjacent sectors. Second, because the physical size of the antenna increases with decreasing sector size, there are physical and aesthetic constraints on the number of sectors per cell. Accounting for both these issues, it is reasonable to expect support for six sectors per cell. If multiple antennas are available within each of these sectors, dynamic beamforming could be used for both uplink and downlink. If the mobile has only one or two antennas, we recommend MRC as the downlink receiver technique and single-stream transmission for the uplink. When the mobile has many more antennas and also a sufficiently high SNR, BLAST is a strong candidate for both uplink and downlink.

Duplexing

Traditionally, the decision on how to partition the communication between uplink and downlink is between two options: either frequency division duplexing (FDD) or time division duplexing (TDD). Each contending option has accepted advantages as well as clear drawbacks. Spectrum allocation, performed by national and international agencies and mostly beyond the control of equipment manufacturers, is a highly political issue. However, the allocation of spectrum strongly conditions the duplexing choice because, if unpaired spectrum is allocated, FDD cannot be used. Although unpaired spectrum is easier to find, the historical tendency is to assign paired spectrum for wide-area systems, in which case both options are viable.

In FDD, uplink and downlink are orthogonal in frequency, provided there is sufficient separation between the corresponding blocks. In TDD, temporal orthogonality requires synchronized uplink and downlink switch patterns plus guard times to account for propagation delays. Since such orthogonality is essential in wide-area systems to prevent catastrophic interfer, given the availability of low-cost Global Positioning System (GPS) technology. In wide-area systems, the additional guard times are on the order of 100–200 μs, which imposes a lower value of about 2 ms for the duplex time in order to keep the overhead below 10%. Furthermore, this guard time should be large enough so that the power transients from transmissions do not affect the receiver processing on the following frame.

For low to moderate Doppler speeds, link reciprocity is usually regarded as the most attractive feature of TDD. As a result of reciprocity, sophisticated transmit processing schemes that necessitate instantaneous channel information become feasible. The lack of reciprocity in FDD, in turn, makes these schemes dependent on relaying channel state information (CSI) through feedback, which tends to incur unacceptable delays if conventional transmission techniques are employed.

A drawback of TDD comes from the periodic interruptions in the links, which are active only for half of the duty cycle. One of the central goals in the design of future-generation systems is to achieve a large reduction in latency. With discontinuous links, no message (not even a 1-bit acknowledgement) can be relayed back with a delay inferior to the duplex time. This implies that the time taken by a basic roundtrip at the physical layer level cannot go below a few milliseconds and thus the aggregate delay experienced by a packet running through a scheduler and subject to ARQ can easily be on the order of 10 ms. This latency propagates through the protocol stack posing serious problems to the upper layers and causing bottlenecks.

We propose to go beyond the paradigm of using either FDD or TDD by proposing a new duplexing scheme called band switching that blends the best characteristics of each [1]. Given paired spectrum blocks, instead of reserving a block for uplink and the other for downlink, we alternate their use every T sec, as depicted in **Figure 2**. With this scheme, reciprocity is achieved as the channel can be estimated in each band when it is used for uplink and then exploited when it is used for downlink. However, both links are

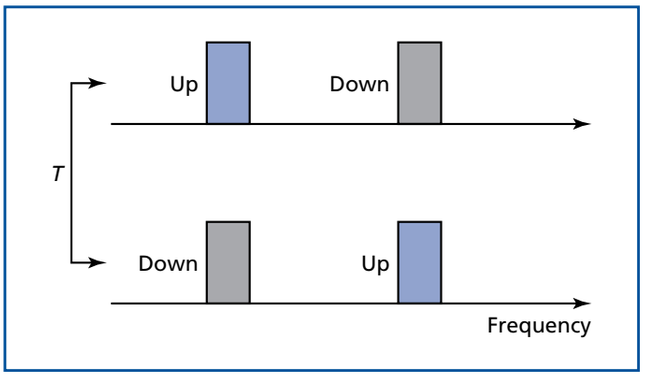

Figure 2.
Band-switching duplexing.

always active except during guard times, which are still required to account for propagation delays and power transients. Therefore, the chief features of both FDD and TDD are seized. Synchronicity and guard times are still needed, as in TDD. Note that band switching is both TDD and FDD. It is TDD because every unit of bandwidth is used, alternatively, half of the time for uplink and half of the time for downlink; it is FDD because, at every point in time, half the spectrum is used for uplink and half for downlink.

In either TDD or band switching, synchronization is required to prevent interference, for example, when a base station is listening for uplink transmissions but receives an unsynchronized downlink transmission from a nearby base. Mobiles can synchronize to base stations by tracking pilot signals transmitted by the base stations. Furthermore, base stations can synchronize to each other either over the backhaul network or by monitoring transmissions from mobiles (or base stations) that are in range but transmitting to another base station. When such mobiles (or base stations) are not present, synchronization is not required. Preliminary experiments with synchronization look promising, but further detailed study is still required.

To summarize the previous sections, we have proposed a system architecture with the following components:

- Either an MC-CDMA or an OFDM air interface,
- Centralized packet scheduling within each cell on both uplink and downlink with random hopping on the uplink to mitigate out-of-cell interference and full bandwidth transmission on the downlink,
- Multiple antenna techniques for SDMA and/or spatial multiplexing, and
- Band-switching duplexing.

These techniques target the data rate and latency demands expected in the near term. However, when further improvements are required, one promising avenue for investigation is network coordination, which we address in the following section.

Network Coordination

While we rejected the possibility of network coordination for resource allocation in the "Multiple Antenna Techniques" section, we did so out of concern that the complexity required to communicate information among base stations could be prohibitive for the short term. However, if we take a longer term perspective in which such communication becomes possible, then so does network coordination. We consider two forms of network coordination in this section. First, under the assumption that fast cell selection is used instead of soft handoff and that channel state information is not known at base stations, intercell interference can be reduced by coordinating the transmission times of the bases. For example, for a user at the cell edge, the intercell interference experienced can be reduced by restricting transmission from an adjacent base, resulting in a higher throughput. Second, if coherent channel state information is known at the base via TDD or band switching, then the antennas across multiple bases can act as a network antenna array. The signal for a given user would be transmitted from multiple antennas, and each would be weighted in such a way as to mitigate interference caused to other users. This significantly improves the SINR distribution, bringing us into the realm where BLAST techniques provide large gains. These concepts are advanced techniques, and they are independent of the air interface (OFDM or MC-CDMA).

Network Coordination for Interference Avoidance

By coordinating the channel allocation or, equivalently, the base station transmission times across the network for a given channel, interference can be reduced and the overall network throughput can be increased. For a user near its desired base, the received signal is often noise limited (i.e., the signal is limited by

thermal noise as opposed to intercell interference); therefore, turning off transmissions from adjacent bases will not affect the performance significantly. On the other hand, for a user near the cell edge, the received signal from its desired base may not be much stronger than the signal from an adjacent base. (Due to fast fading, the desired base's signal may in fact be weaker.) In this case, the user's throughput could be improved significantly by turning off the adjacent base. From a network perspective, this strategy is desirable if the improved performance compensates for the missed transmission opportunity by the adjacent base.

Ideally, coordination would include all interfering cells. However, the complexity of doing so could be prohibitive since this would require each base (or some centralized entity) to know the received power by each mobile from all bases. As an alternative, coordination could occur locally among a group of two or three bases, reducing the complexity but still providing the bulk of the gains since the one or two most strongly interfering bases are typically much stronger than the remaining interfering bases.

Preliminary results have shown that interference avoidance through network coordination can improve the system throughput significantly for interference limited systems—i.e., for a network of densely packed, small (less than 1 km radius) cells [5]. In addition, for systems with constraints on minimum rate and maximum delay, there will be pressure to serve the user at the cell edge more often in order to meet these constraints, and interference avoidance can again potentially provide significant performance improvements.

Network Coordination Using Coherent Channel State Information

Instead of just knowledge of power levels received by the users, if the bases had knowledge of the coherent channel state information to each user from all bases that reach it, then the antennas across multiple bases could act as a single network antenna array. The signal for a given user would be transmitted from multiple antennas, and each would be weighted in such a way as to mitigate interference caused to other users. This technique of transmitting simultaneously to users with all base stations in a network is motivated by fundamental downlink performance limits of multicell networks. In the context of information theory, the system of interest can be modeled as a multiple antenna Gaussian broadcast channel whose capacity region can be achieved using coherently coordinated transmission combined with a sophisticated coding technique known as dirty paper coding (DPC) [18]. The value of coherently coordinated transmission (CCT) over conventional transmission techniques has been investigated using asymptotic analysis in [10] and numerical simulations in [3 and 7]. We will see that these techniques offer substantial capacity gains even with single-antenna mobiles.

We consider a system model similar to the one in [7] and consisting of a hexagonal network of cells with a single base antenna in the center of each cell with single-antenna users placed randomly in the network. In the baseline system, which is called single-base transmission (SBT), each active base transmits to a uniquely assigned user. The assignment is made so that, for each user, its SINR is maximized over all other choices of bases, under the assumption that all bases transmit with full power. Pathloss, Rayleigh fading, and shadowing are considered when measuring the SINR, and the users with the worst 10% SINR are considered outages and discarded from the system. The bases for these discarded users are inactive, and the remaining active bases transmit with full power. The rate achieved by each user is given by its Shannon capacity as a function of its SINR. Apart from the assignment of 10% of the users to outage, nothing is done to mitigate intercell interference for those users that are served. The operation of the SBT baseline is similar to exisiting packet data systems such as 1xEV-DO.

Two CCT techniques are considered using the same placement of users obtained from SBT. The first is based on zero forcing (ZF) where the transmissions for the users are coherently weighted such that each user experiences no interference from any other user. The second uses a combination of ZF and DPC. In the interest of fairness, the proposed CCT techniques transmit to all users not in outage with equal rate, and power control is used to maximize this common

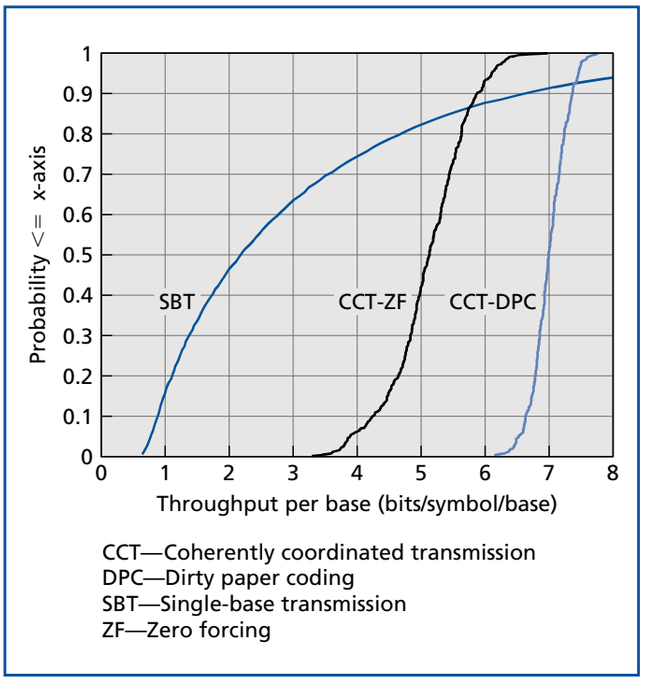

Figure 3.
Cumulative distribution function of the throughput base for SBT, CCT-ZF, and CCT-DPC systems.

rate. (In contrast to SBT where 10% of the bases are inactive, all bases for the CCT techniques are active.) We assume perfect channel knowledge at all bases for both CCT techniques.

Figure 3 shows the cumulative distribution function of the throughput per base for SBT, CCT-ZF, and CCT-DPC systems. (The CCT-DPC curve is a lower bound on the optimal CCT-DPC performance because, for example, not all possible encoding orders for the users were considered.) The mean throughputs of these systems are 3.23, 5.11, and 7.01 bits per second per base for SBT, CCT-ZF, and CCT-DPC, respectively. The gains of CCT-ZF and CCT-DPC over SBT are 1.58 and 2.17, respectively. The performance gains are even more impressive when considering outage capacity. The minimum value of each curve corresponds to a 10% user outage, and the throughputs at these points are 0.637, 3.30, and 6.14 bits per second per base for SBT, CCT-ZF, and CCT-DPC, respectively. The gains of CCT-ZF and CCT-DPC over SBT for outage capacity are 5.18 and 9.64, respectively. Note that these numerical results have assumed a single antenna per receiver. If instead multiple antennas are available at the receivers (assuming there are sufficient transmit antennas in the network), then we could use spatial multiplexing techniques, taking advantage of the improved SINR distribution from network coordination and providing much greater spectral efficiency.

On the uplink for a single-cell system, the capacity-achieving technique is for all users (assumed to have a single antenna) to transmit over the entire bandwidth and to rely on multiuser detection consisting of a minimum mean-squared error linear front end followed by successive interference cancellation [17]. Generalizing to the multicell system, the optimum receiver would require coordinated detection among all the bases in the network using the same minimum mean-squared error detector with successive interference cancellation. This technique is therefore the dual of DPC used for the downlink.

Network coordination on either the uplink or downlink requires that all bases have knowledge of all channel realizations between all users and all bases. Furthermore, the processing required for DPC and multiuser detection needs to be done at a centralized location. On the downlink, one implementation would be to convey all channel estimates via a high-speed backhaul network to a central processor. The coding for DPC and antenna weights could be computed and then conveyed to all the bases. Likewise on the uplink, the baseband received signals from all bases could be conveyed to a central processor where the detection is performed. To simplify the implementation, coordination could be performed over overlapping clusters of cells (as determined by the pathloss exponent) instead of over the entire network. We emphasize that the delay and bandwidth requirements on the backhaul for implementing network coordination are considerable and need to be quantified.

Conclusions

We have presented an overview of technology options for the physical layer design of next-generation cellular networks to provide high-speed packet-based services. We summarize the findings in **Table I** by listing three different system configurations that depend on time of deployment and complexity

Table I. Summary of physical-layer proposals.

Design elements	Near-term	Medium-term	Long-term
Air interface	MC-CDMA	OFDM	OFDM
Resource allocation	Random hopping (uplink) Universal reuse (downlink)	Random hopping (uplink) Universal reuse (downlink)	Network coordination
Duplexing	TDD or FDD	Band switching	Band switching
Multiple antenna technique	SDMA and MRC	SDMA, MRC, and BLAST	Network coordination

BLAST—Bell Labs Layered Space-Time
FDD—Frequency division duplexing
MC-CDMA—Multicarrier code division multiple access
MRC—Maximal ratio combining
OFDM—Orthogonal frequency division multiplexing
SDMA—Spatial division multiple access
TDD—Time division duplexing

constraints. For near-term deployment and with minimal complexity, an MC-CDMA air interface could be used with random hopping on the uplink and universal reuse on the downlink. The duplexing could be based on the legacy system (either TDD or FDD). Assuming insufficient antennas for BLAST, multiple antennas should be used for SDMA and MRC. A more ambitious mid-term deployment requiring more complexity would use an OFDM air interface and band-switching duplexing. Coherent channel estimates at the base station transmitter could be used for more sophisticated multiple antenna transmissions (e.g., dynamic beamforming), and BLAST could also be used if there are sufficient receive antennas to significantly improve spectral efficiency. Finally, as part of a long-term proposal, coherent channel estimates could be used for network coordination on the downlink, and network-wide successive interference cancellation could be used on the uplink.

We conclude by noting that this paper has focused on physical layer design, but the design is by no means complete because it has not addressed higher layer issues to a large extent. In order to design next-generation systems for optimizing end-to-end user experiences, it is essential that a more thorough design consider cross-layer interactions from the physical layer all the way up to the application layer. Topics to consider include channel coding, medium access control design, scheduling, dynamic network optimization, backhaul design, interaction of the cellular network with other networks, and base station and radio access network architectures.

Acknowledgments

The authors would like to acknowledge the following people who made contributions to other documents upon which parts of this paper are based: Angela Alexiou, Peter Bosch, Dmitry Chizhik, Suman Das, Kemal Karakayali, Jonathan Ling, Angel Lozano, Constantinos Papadias, Dragan Samardzija, and Reinaldo Valenzuela.

References

[1] A. Alexiou, D. Avidor, P. Bosch, S. Das, P. Gupta, B. Hochwald, T. E. Klein, J. Ling, A. Lozano, T. Marzetta, S. Mukherjee, S. Mullender, C. Papadias, R. Valenzuela, and H. Viswanathan, "Duplexing, Resource Allocation and Intercell Coordination-Design Recommendations for Next-Generation Wireless Systems," Wireless Commun. and Mobile Computing J., forthcoming 2005.

[2] G. Caire, M. Damen, and H. El Gamal, "Principles of Space-Time Coding," Cambridge Univ. Press, Cambridge, UK, 2005.

[3] C. Chow, B. I. Shraiman, A. M. Sengupta, and M. R. Andrews, "Using Phase and Amplitude

Control Across Networks to Increase Capacity up to Fourfold," Proc. of the URSI National Radio Science Meeting (Maastricht, Neth., 2002).
[4] S. Das and H. Viswanathan, "Dynamic Frequency Assignment in a Multi-user OFDM System," Proc. IEEE Veh. Technol. Conf. (Los Angeles, CA, 2004).
[5] S. Das, H. Viswanathan, and G. Rittenhouse, "Dynamic Load Balancing Through Coordinated Scheduling in Packet Data Networks," Proc. IEEE INFOCOM (San Francisco, CA, 2003), 786–796.
[6] G. J. Foschini, D. Chizhik, M. J. Gans, C. Papadias, and R. A. Valenzuela, "Analysis and Performance of Some Basic Space-Time Architectures," IEEE J. Select. Areas Commun, 21:3 (2003) 303–320.
[7] G. J. Foschini, H. Huang, K. Karakayali, R. A. Valenzuela, and S. Venkatesan, "The Value of Coherent Base Station Coordination," Proc. Conf. Inform. Sci. and Syst. (Baltimore, MD), forthcoming 2005.
[8] B. Hochwald, C. Peel, and L. Swindlehurst, "Achieving Near-Capacity in a Multi-Antenna Multi-User System," Proc. 41st Allerton Conference (Allerton, IL, 2003), <http://mars.bell-labs.com>)
[9] H. Huang, S. Venkatesan, A. Kogiantis, and N. Sharma, "Increasing the Peak Data Rate of 3G Downlink Packet Data Systems Using Multiple Antennas," Proc. IEEE Veh. Technol. Conf., (Jeju, Korea, 2003), 311–315.
[10] H. Huang and S. Venkatesan, "Asymptotic Downlink Capacity of Coordinated Cellular Networks," Proc. Asilomar Conf. Signals, Syst., and Comput. (Pacific Grove, CA, 2004).
[11] H. Huang and H. Viswanathan, "Multiple Antennas and Multiuser Detection in High Data Rate CDMA Systems," Proc. Veh. Technol. Conf. (Tokyo, Japan, 2000), 556–560.
[12] R. van Nee and R. Prasad, OFDM for Wireless Multimedia Applications, Artech House, Boston, MA, 2000.
[13] H. Viswanathan and K. Kumaran, "Rate Scheduling in Multiple Antenna Downlink Wireless Systems," Proc. 39th Allerton Conference (Monticello, IL, 2001), pp. 747–756.
[14] A. A. Salvekar, C. Aldana, J. Tellado, and J. Cioffi, "Peak-to-Average Power Ratio Reduction for Block Transmission Systems in the Presence of Transmit Filtering," Proc. IEEE Internat. Conf. Commun. (Helsinki, Fin., 2001), 175–178.
[15] R. A. Soni, R. M. Buehrer, and R. D. Benning, "Intelligent Antenna System for cdma2000," IEEE Signal Processing Mag., 19:2 (2002) 54–67.
[16] D. Tse, P. Viswanath, and L. Zheng, "Diversity-Multiplexing Tradeoff in Multiple Access Channels," IEEE Trans. Inform. Theory, 50:9 (2004), 1859–74.
[17] M. K. Varanasi and T. Guess, "Optimum Decision Feedback Multiuser Equalization with Successive Decoding Achieves the Total Capacity of the Gaussian Multiple-Access Channel," Proc. Asilomar Conf. Signals, Syst., and Comput. (Pacific Grove, CA, 1997), 1405–1409.
[18] H. Weingarten, Y. Steinberg, and S. Shamai, "The Capacity Region of the Gaussian MIMO Broadcast Channel," Proc. Conf. Inform. Sci. and Syst. (Princeton, NJ, 2004).
[19] J. Winters, "Optimum Combining in Digital Mobile Radio with Co-Channel Interference," IEEE Trans. Veh. Technol., 33:3 (1985), 144–155.

(Manuscript approved March 2005)

GERARD J. FOSCHINI is a distinguished member of technical staff in the Wireless Research Lab at Bell Labs in Holmdel, New Jersey. He holds a B.S.E.E. from the New Jersey Institute of Technology in Newark, an M.E.E. from New York University in New York City, and a Ph.D. in mathematics from Stevens Institute of Technology in Hoboken, New Jersey. A Bell Labs Fellow and an IEEE Fellow as well, he is experienced in a wide range of communication areas, including wireless and optical communications at both point-to-point and network levels. Dr. Foschini has taught at Princeton University and lectured at numerous other universities throughout the United States and elsewhere. He has published over 100 papers, and he holds many patents, one of which won the 2001 Lucent Inventor Award and another, the 2002 Thomas Alva Edison Patent Award. For his innovation and his outstanding contributions to communications theory, he received the 2004 IEEE Eric E. Sumner Award.

HOWARD C. HUANG is a distinguished member of technical staff in the Wireless Research Lab at Bell Labs in Holmdel, New Jersey. He holds a B.S.E.E degree from Rice University in Houston, Texas, and M.S. and Ph.D. degrees in electrical engineering from Princeton University in New Jersey. Dr. Huang's research interests include communication theory and signal processing for wireless communication systems.

SAPE J. MULLENDER is a member of technical staff at the Computing Sciences Department at Bell Labs in Murray Hill, New Jersey. He has worked extensively in operating systems, multimedia systems, and, in recent years, wireless systems research. He was a principal designer of the Amoeba distributed system; he led the European Union's Pegasus project, which resulted in the design of the Nemesis multimedia operating system; and he made valuable contributions to work on the Plan 9® and Inferno® operating systems. He received a Ph.D. from the Vrije Universiteit in Amsterdam, The Netherlands, where he also was formerly a faculty member. He currently holds a chair part time in the Computer Science Department at the University of Twente. Dr. Mullender has published papers on file systems, high-performance RPC protocols, migratable object location in computer networks, and protection mechanisms, and he was involved in the organization of a series of advanced courses on distributed systems—Arctic'88, Fingerlakes'89, Bologna'90, Karuizawa'91, and Lisboa'92.

SIVARAMA VENKATESAN is a member of technical staff in the Wireless Research Lab at Bell Labs in Holmdel, New Jersey. He received a B.Tech. degree in electrical engineering from the Indian Institute of Technology in Chennai, India, and M.S. and Ph.D. degrees, in electrical engineering as well, from Cornell University in Ithaca, New York. Dr. Venkatesan's research interests include information theory and multiple-antenna wireless communication systems.

HARISH VISWANATHAN is a distinguished member of technical staff in the Wireless Research Lab at Bell Labs in Murray Hill, New Jersey. He holds a B. Tech. degree in electrical engineering from the Indian Institute of Technology in Chennai, India, and M.S. and Ph.D. degrees, also in electrical engineering, from Cornell University in Ithaca, New York. Dr. Viswanathan's research interests include information theory, communication theory, wireless networks, and signal processing. ◆