Foundations and Trends® in
Networking
Vol. 1, Nos. 3-4 (2006) 271–425

DOI: 10.1561/1300000004

now
the essence of knowledge

Cooperative Communications

Gerhard Kramer[1], Ivana Marić[2] and Roy D. Yates[3]

[1] *Bell Laboratories, Alcatel-Lucent, Murray Hill, New Jersey 07974, USA, gkr@bell-labs.com*
[2] *Dept. of Electrical Engineering, Stanford University, Stanford, California 94305, USA, ivanam@systems.stanford.edu*
[3] *WINLAB, Rutgers University, North Brunswick, New Jersey 08902, USA, ryates@winlab.rutgers.edu*

Abstract

This article reviews progress in cooperative communication networks. Our survey is by no means exhaustive. Instead, we assemble a representative sample of recent results to serve as a roadmap for the area. Our emphasis is on wireless networks, but many of the results apply to cooperation in wireline networks and mixed wireless/wireline networks. We intend our presentation to be a tutorial for the reader who is familiar with information theory concepts but has not actively followed the field. For the active researcher, this contribution should serve as a useful digest of significant results. This article is meant to encourage readers to find new ways to apply the ideas of network cooperation and should make the area sufficiently accessible to network designers to contribute to the advancement of networking practice.

1

Overview

1.1 Introduction

The classic representation of a communication network is a graph, as in Figure 1.1, with a set of nodes and edges. The nodes usually represent *devices* such as a router, a wireless access point, or a mobile telephone. The edges usually represent communication *links* or *channels*, for example a fiber-optic cable or a wireless link. Both the devices and the channels may have constraints on their operation. For example, a router might have limited processing power, or perhaps it can accept data from only a few of its ports simultaneously. A fiber-optic link has a limited bandwidth (which can be quite large!). A wireless phone,

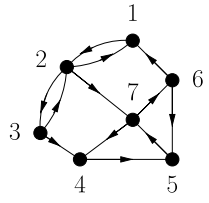


Fig. 1.1 A network graph.

Table 1.1 Device and channel properties.

	Devices (Nodes)	Channels (Edges)
Wireline	Constraints:	Independent channels
	– processing speed/energy	No interference
	– input/output (ports)	Limited bandwidth
	– delay	Slow changes
	Limited network knowledge	Packet erasures
Wireless	Constraints:	Broadcasting
	– transmit energy	Interference
	– processing speed/energy	Noise
	– half-duplex	Limited bandwidth
	– delay	Slow or rapid changes
	Limited network knowledge	
	Limited channel knowledge	

on the other hand, has limited battery resources and likely wishes to conserve energy. A wireless link can have rapid time variations arising from mobility and multipath propagation of signals. Some of these properties are collected in Table 1.1 and are described in more detail in this text.

The purpose of a communication network is to enable the exchange of *messages* between its nodes. These messages, as generated by an application, are organized into data packets. In the traditional model of a network, the nodes operate as store-and-forward packet routers that transmit packets over point-to-point links. However, this model is unnecessarily restrictive as it ignores two important possibilities:

- **Node Coding:** Nodes can combine, or encode, any of their received information and symbol streams.
- **Broadcasting:** Nodes overhear the transmissions of other nodes from which they are not required to receive messages.

Node coding is possible in any network, while the ability to overhear transmissions is a property of the physical communication channel. In particular, wireless devices inherently broadcast information in that a signal to a particular receiving node can be overheard by other nodes. Typically, the wireless nodes treat these overheard signals as interference and the system provides mechanisms to mitigate this interference. For example, many second generation cellular phones employ code division multiple access (CDMA) to permit signal decoding in

the presence of interference [185]. As a second example, 802.11× wireless LANs employ a media access control (MAC) protocol that enables nearby nodes to share the broadcast channel by taking turns transmitting. The goal of these multiple access strategies is to ensure that communication over a link is reliable. A consequence is that the paradigm of a network as a set of point-to-point communication links is reinforced.

On the other hand, it has been understood in the information theory community for over three decades that wireless communication from a source to destination can benefit from the cooperation of nodes that overhear the transmission. As these intermediate nodes may generate transmissions based on *processing* of the overheard signals, the signaling strategies go well beyond the inherent cooperation offered by nodes that act as store-and-forward packet routers.

The relay channel [181] models this communication situation. Fundamental coding strategies for the relay channel were developed in [33]. The key idea was that the relay can use the overheard transmission from the source to form its own transmission that aids in decoding at the receiver. These pioneering efforts focused on the simplest case of one source, one destination, and one relay.

More recently, the seminal work of [4] has spurred progress in network coding. The connection between relaying and network coding is that both approaches represent cooperative techniques. A node processes either received signals (in the case of relaying) or messages and encoded packets (in the case of network coding) based on the transmissions of multiple nodes. Although final capacity results remain elusive, it has become apparent that cooperative techniques using node coding and broadcasting offer significant improvements over traditional store-and-forward wireless networks with point-to-point links.

1.2 Networking Protocols

Advances in cooperative communication have focused on two goals:

- increasing the communication rates across the network, and
- increasing the communication reliability in networks with time varying channels.

The advances are often demonstrated by mathematical (information theory) analyses of signaling strategies over models of communication channels. The signaling strategies typically involve coordinated transmissions by multiple network nodes; but despite recent progress, there remain significant barriers to applying these results to the development of practical network protocols.

For example, many information-theory results are based on asymptotically large block lengths and usually ignore the overhead needed to set up and maintain coordinated transmissions. In practice, however, and in particular for mobile wireless networks, the time variations in the communication channels necessitate variations in the signaling strategies. The setup and maintenance signaling therefore becomes a recurring cost. We use the term *network protocol* for a set of distributed algorithms executed by the network nodes that configure, execute, and reconfigure signaling strategies in response to temporal variations. A cooperative network protocol must incorporate methods so that recurring setup costs do not dominate the efficiency gains of a cooperative signaling strategy.

Much of the value of networking is derived from the interconnection of heterogeneous devices and systems. It is worthwhile to keep in mind that the Internet is not so much a network as a network of networks. These interconnections of networks are made possible by designing protocols based on abstract models, as in the layers of the OSI or TCP/IP network protocol stacks [18]. In particular, a network protocol is more valuable by providing a certain capability over a more generic network abstraction, even if this is at the expense of reduced performance. In this context, the development of practical cooperative protocols will be facilitated by identifying the simplest possible abstractions of cooperative signaling strategies.

1.3 Objectives and Outline

This text reviews progress in cooperative communication networks. Our survey is by no means exhaustive. Instead, we assemble a representative sample of recent results to serve as a roadmap for the area. Our emphasis is on wireless networks, but many of the results apply to cooperation

in wireline networks and mixed wireless/wireline networks. We intend our presentation to be a tutorial for the reader who is familiar with information theory concepts but has not actively followed the field. For the active researcher, we would like this contribution to serve as a useful digest of significant results.

We hope this text will encourage readers to find new ways to apply the fundamental ideas of network cooperation. We also hope we have made the area sufficiently accessible to network designers to contribute to the advancement of networking practice.

Chapter 2 provides an overview of conventional communication and networking. This includes information-theoretic capacity definitions for the point-to-point links that are the traditional basis for networks, as well as key elements of the link and network layer protocols. While much or all of this material will be familiar to the reader, we include it to provide contrast to the cooperative networks that follow. Chapters 3, 4, and 5 emphasize an information-theoretic treatment of the capacity of cooperative networks. In particular, Chapter 3 provides basic definitions, and a sampling of fundamental network models and corresponding capacity results. Chapter 4 develops the key ideas in information-theoretic relaying and makes apparent how the methods go well beyond store-and-forward routing. Chapter 5 focuses on time-varying wireless networks and how they can be made more reliable through the use of cooperative diversity mechanisms. Finally, Chapter 6 investigates implementation aspects of these cooperative strategies, concluding with a review of recent proposals in the literature for network cooperation protocols.

2

Conventional Networks: A Review

Before describing cooperative protocols and networks, we start with a review of networking models and common practices. Our discussion introduces the conventional decomposition of a network into protocol layers. We then employ these layers to organize the subsequent review of the physical, link, and network layer protocols. We recognize that future cooperative networks will also demand a re-examination of transport layer protocols; however, we omit a review of the higher layers as they are beyond the scope of this text. While much or all of this material is likely to be familiar, this review will serve as a baseline for comparisons with cooperative networks.

2.1 Network Layering

Layering as embodied in the protocol stack of Figure 2.1 is a key idea in the development of networks. The stack of boxes (modules) arranged as layers represents a network node. Each module operates at a particular layer. The horizontal dashed arrows between modules in different nodes signify that a module may exchange messages with its peer modules in other network nodes. Messages that are sent through lower layer

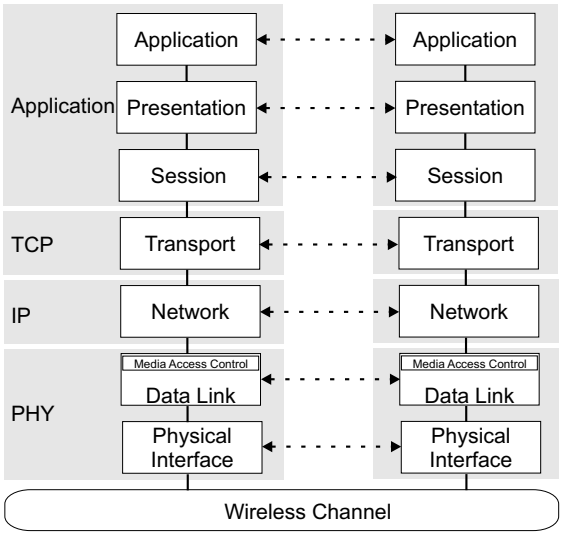


Fig. 2.1 Network protocol stack.

modules to peer modules in other network nodes are the basis for distributed network algorithms.

Data packets are used to communicate from a source node to a destination node via a *path* with intermediate nodes. At the source, data packets generated by an application are passed from module to module down through successive layers of the protocol stack. Each module typically appends its own header to each such data packet. A module may *repackage* the data packets, by dividing packets into smaller packets. The packet headers serve as protocol signaling for peer modules, either at intermediate nodes or at the destination node. A module also may inject its own control packets in order to communicate with peer modules.

When a packet proceeds through a multihop route to a destination, a packet climbs no higher than necessary in the protocol stack. That is, a packet passing through an intermediate node will reach the network layer where the routing algorithm will decide to what node the packet should be forwarded. Thus a packet reaches the transport layer and application layer only at the destination. At the destination, each module is responsible for undoing the repackaging of its source node

peer by stripping the additional headers and control packets injected by its peer. That is, the higher layer packets passed down to a module at the source should be passed back up to the higher layer module at the destination.

In traditional wireline or wireless networks, these modules are well defined. Packets are buffered and sequenced by the transport layer, typically TCP, that implements both a reliable end-to-end connection as well as end-to-end flow control. Finding routes (via a sequence of links) to a destination is the job of the network layer. Maintaining these routes and forwarding packets along these routes is also a network layer task. The link layer ensures reliable packet communication on a single link. As shown in Figure 2.1, this may include a MAC sublayer that regulates channel access. The physical (PHY) layer represents the hardware that performs transmission and reception.

In an IP network, the full stack has the simplified representation shown in the gray boxes. In a wireless setting, the MAC sublayer, the link layer, and the PHY layer are lumped together as a PHY layer. This combined PHY layer is just an interface queue that accepts IP packets.

For our purposes, we start with a source node s running an application layer process that wishes to transmit messages to an application layer process at a destination node t. The messages are encoded as data packets with appropriate headers that identify the application process, the source node s and the destination node t. When these packets are passed to a TCP transport layer, sequence numbers are appended and the release of packets to the network layer is controlled by reverse stream of TCP ACKs from the receiver TCP process. The network layer examines the destination address (an IP address) and determines where to send the packet using a routing table. For example, the routing table for a source node attached to an Ethernet might specify only two rules: direct transmission to destination nodes on the same Ethernet and forwarding to a gateway node for all other packets.

At the data link layer, it is common practice to append a cyclic redundancy check (CRC) to each packet. The CRC allows the data link layer at the receiver to detect packet reception errors. Sequence numbers may also be added to facilitate automatic repeat request (ARQ)

retransmission protocols at the link layer. The PHY layer is responsible for the transmission of bits to a receiver of a link. The coding and modulation employed at the PHY layer for a single point-to-point link may be quite complex. When forward error correction (FEC) and hybrid ARQ protocols are employed, the line between the physical and link layers is blurred. However, above the link layer, one can assume that the interfaces between layers are based on binary data packets.

In the following sections of this chapter, we climb the protocol stack in describing the traditional functions of the physical, link and network layers.

2.2 Physical Layer: Communication Theory

We review basic concepts of communication theory. This theory is useful for all network layers, but in particular for the PHY and application layers where source and/or channel coding are used. The following sections also introduce notation that we use throughout the document.

2.2.1 Information Theory

Information theory began as the mathematics of communications for one source and one channel [164]. Consider Figure 2.2 and suppose the source puts out a message W that consists of B independent and uniformly distributed bits. Let X and Y be random variables representing the channel input and output with alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively. A discrete memoryless channel (DMC) is a conditional probability distribution $P_{Y|X}(\cdot)$, where $\mathcal{X}$ and $\mathcal{Y}$ are discrete and finite.[1] The encoder transmits a string $X^n = X_1, X_2, \ldots, X_n$ that is a function of W.

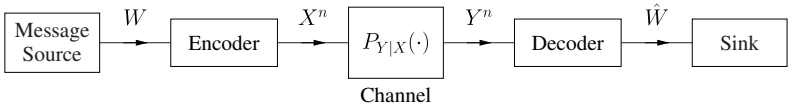


Fig. 2.2 Communications model.

[1] The theory generalizes to continuous random variables in natural ways.

The decoder sees $Y^n = Y_1, Y_2, \ldots, Y_n$ and puts out a message estimate $\hat{W}$ that is a function of Y^n.

The *capacity* is the maximum rate $R = B/n$ bits per channel use for which, for sufficiently large n, there exists a W-to-X^n mapping (an encoder) and a Y^n-to-$\hat{W}$ mapping (a decoder) so that the error probability $\Pr[\hat{W} \neq W]$ can be made as close to 0 as desired (but not necessarily exactly 0). It is a remarkable fact that such a capacity exists. Perhaps just as remarkably, the capacity is given by the compact formula

$$C = \max_{P_X(\cdot)} I(X;Y) \quad \text{bits/use}, \tag{2.1}$$

where

$$I(X;Y) = \sum_{a \in \mathcal{X}, b \in \mathcal{Y}, P_{XY}(a,b) > 0} P_{XY}(a,b) \log_2 \frac{P_{XY}(a,b)}{P_X(a) P_Y(b)} \tag{2.2}$$

is the *mutual information* between X and Y. The formula (2.1) is universal in the sense that it gives the capacity for any DMC. Furthermore, the capacity proof is based on a random code construction method that is useful for many other communication problems [34, 35].

Similarly, consider the additive white Gaussian noise (AWGN) channel

$$Y = \frac{h}{d^{\alpha/2}} X + Z \tag{2.3}$$

with power constraint

$$\sum_{i=1}^{n} E[|X_i|^2]/n \leq P, \tag{2.4}$$

where X and Z are real random variables; Z has variance N; h, d, and α are real numbers; and $E[\cdot]$ denotes expectation. The variables h and d represent channel gains and distances between the encoder and decoder, respectively, and α is an attenuation exponent (e.g., $\alpha = 2$ for free-space propagation). The model (2.3) is not accurate for small d. For example, as d approaches zero, we have the physically untenable result that the received power exceeds the transmitted power. However, we

will be mainly interested in "long-range" communication where (2.3) is a reasonable model for distance-dependent attenuation. The capacity formula (2.1) generalizes in a straightforward way to AWGN channels, namely

$$C = \max_{p_X(\cdot)} \int p_{XY}(a,b)\ \log_2 \frac{p_{XY}(a,b)}{p_X(a)p_Y(b)}\, da\, db\ \text{ bits/use}, \tag{2.5}$$

where $p_{XY}(a,b)$ is the joint probability density of X and Y (if $\mathcal{X}$ is discrete, then replace $p_X(a)$ and $p_{XY}(a,b)$ by $P_X(a)$ and $P_X(a)\,p_{Y|X}(b|a)$, respectively, and replace the integral over $\mathcal{X}$ by a sum). The maximization in (2.5) is interpreted as constraining $p_X(\cdot)$ to satisfy (2.4). The maximum entropy theorem [34, p. 234] establishes that the best X is Gaussian with zero mean and variance P, and the capacity is

$$C = \frac{1}{2}\log_2(1+\gamma)\ \text{ bits/use}, \tag{2.6}$$

where

$$\gamma = \left(\frac{P}{N}\right)\frac{|h|^2}{d^\alpha} \tag{2.7}$$

is the signal-to-noise ratio (SNR) [34, p. 241]. Suppose next that X, h, and $Z = Z_R + jZ_I$ are complex with $j = \sqrt{-1}$, and Z_R and Z_I are independent, real, Gaussian random variables with variance $N/2$. The capacity is now

$$C = \log_2(1+\gamma)\ \text{ bits/use}. \tag{2.8}$$

2.2.2 Modulation

Consider the AWGN channel (2.3) with complex symbol alphabets. Let the SNR in decibels be

$$\gamma_{dB} = 10\log_{10}\gamma\ \text{ dB} \tag{2.9}$$

so that we can write the capacity (2.8) as

$$C = \log_2\left(1 + 10^{\gamma_{dB}/10}\right)\ \text{ bits/use}. \tag{2.10}$$

We plot (2.10) in Figure 2.3 as the curve labeled "Gaussian."

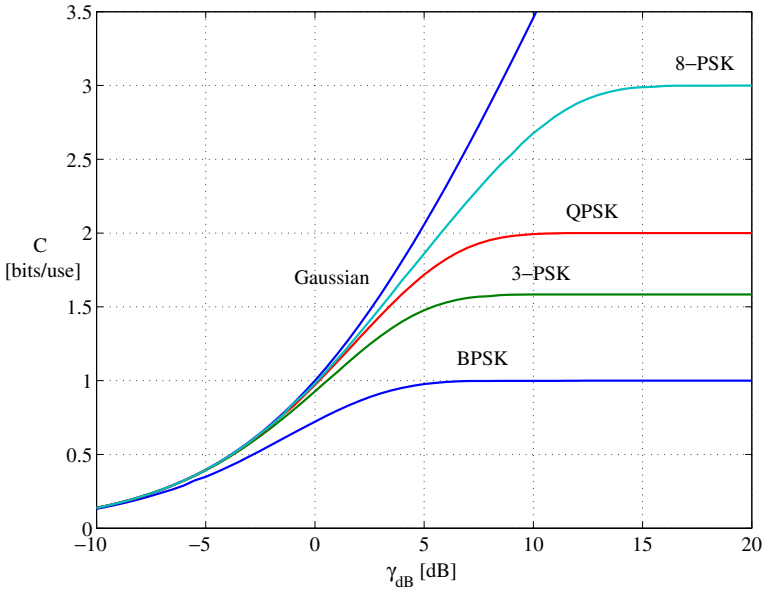


Fig. 2.3 Capacity of M-PSK signal sets.

The formula (2.10) is based on using $X = X_R + jX_I$, where X_R and X_I are independent, real, zero-mean, Gaussian random variables with variance $P/2$. However, for practical reasons such as amplifier constraints, receiver synchronization, and detector complexity, one usually restricts attention to X with a limited number of values. For example, for wireless communication one might restrict attention to *M-ary phase-shift keying* (M-PSK) where X is uniform over the alphabet

$$\mathcal{X} = \left\{\sqrt{E_s}, \sqrt{E_s}e^{j\,2\pi/M}, \ldots, \sqrt{E_s}e^{j\,2\pi(M-1)/M}\right\}, \tag{2.11}$$

where $E_s = P$ is the (average) per-symbol-energy. The set $\mathcal{X}$ is called the *signal set* or the *modulation set* (cf. [148, Ch. 4]).

The capacity of the complex AWGN channel with M-PSK is (see (2.5))

$$C = -\log_2(\pi e) - \int p_Y(b)\, \log_2 p_Y(b)\, db \ \text{ bits/use}, \tag{2.12}$$

where

$$p_Y(b) = \sum_{m=0}^{M-1} \frac{1}{M}\, p_{Y|X}\left(b|\sqrt{E_s}e^{j\,2\pi m/M}\right) \tag{2.13}$$

and

$$p_{Y|X}(b|a) = p_{Z_R}\left(\Re\{b\} - \Re\{a\}\right) p_{Z_I}\left(\Im\{b\} - \Im\{a\}\right), \tag{2.14}$$

where Z_R and Z_I are independent, Gaussian, and have variance $N/2$, and $\Re\{x\}$ and $\Im\{x\}$ are the respective real and imaginary parts of x. Suppose we use $M = 2, 3, 4, 8$, whose signal sets are called binary PSK (BPSK), 3-PSK, quaternary PSK (QPSK) and 8-PSK, respectively. The capacities are plotted in Figure 2.3 as a function of γ_{dB}. Observe how the M-PSK capacities saturate at $\log_2(M)$ bits/use at large γ. Clearly, at high SNRs one must use large signal sets, while at low SNRs it seems that BPSK suffices (however, see the next sections on spectral efficiency).

Finally, we remark that the capacity calculations in this chapter assume that the receiver can accurately estimate the channel SNR. If this is not the case, then many interesting issues arise [184]. For instance, an important practical issue is the choice of signal set for the best capacity scaling behavior at low and high SNRs [115, 117].

2.2.3 Spectral Efficiency

We next include the effect of *bandwidth*. Loosely speaking, the *spectral efficiency* of a modulation set is the number of bits per second per Hertz that the set can support. To define spectral efficiency more precisely, we follow an approach similar to [165, Sec. VIII] and [188, Sec. 7.2] and make the following idealizations:

(i) The channel passes frequencies f in the range

$$f_0 - W/2 < |f| < f_0 + W/2, \tag{2.15}$$

where the center frequency f_0 is much larger than W.[2] The equivalent baseband representations of the channel input and

[2] Note that W is here the bandwidth rather than a message.

output signals are therefore complex and band-limited to $|f| < W/2$. We say that the channel bandwidth is W.

(ii) The channel is used for a period of T seconds. We describe the channel input and output signals by $n = TW$ complex samples spaced $T_s = 1/W$ seconds apart, i.e., we can reconstruct the baseband input signal $X(t)$ as

$$X(t) = \sum_{i=1}^{n} X_i \cdot \frac{\sin(\pi W t - \pi i)}{\pi W t - \pi i}, \tag{2.16}$$

where X_i is a complex number whose squared amplitude and phase are the respective energy and phase of the ith input sample. The average energy per sample is

$$E_s = \sum_{i=1}^{n} E\left[|X_i|^2\right]/n. \tag{2.17}$$

We consider primarily Gaussian signaling for which X_i is a complex Gaussian random variable, or PSK for which $|X_i|^2 = E_s$ for all i.

(iii) The channel adds complex Gaussian noise with power N to each input sample, i.e., the ith channel output sample is $Y_i = X_i + Z_i$ where $Z_i = Z_{R,i} + j\,Z_{I,i}$ and $Z_{R,i}$ and $Z_{I,i}$ are independent, Gaussian random variables each having variance $(NT_s)/2$. Usually N increases proportionally with W, and we write $N = WN_0$ where $N_0/2$ is the noise power per Hertz.

The modulation rate is $R_{\text{mod}} = \log_2(M)$ bits, where M is the number of values that X takes on. For instance, QPSK has $R_{\text{mod}} = 2$ bits. Suppose the encoder maps blocks of k_c bits to blocks of n_c bits at the rate $R_c = k_c/n_c$. The overall coded modulation rate is therefore $R = R_c R_{\text{mod}}$ bits. The energy consumed per information bit is $E_b = E_s/R$ and we define the *information bit* SNR ratio as E_b/N_0. Let P_b be the decoder bit-error probability.

It is now natural to define the *spectral efficiency* of the modulation at the bit-error probability P_b as

$$\eta(E_b/N_0, P_b) = R_c^* R_{\text{mod}} \ \text{ bits/s/Hz}, \tag{2.18}$$

where R_c^* is the *maximum code rate* for which one can achieve a bit-error probability of P_b when the information bit SNR ratio is E_b/N_0. For wireless transmission P_b can be 10^{-6}, while for optical transmission P_b can be required to be 10^{-15} or less. However, one usually considers only the case $P_b \to 0$ because the spectral efficiency hardly changes as long as $P_b \leq 10^{-3}$. We denote the resulting spectral efficiency by $\eta(E_b/N_0)$.

Note that (2.18) defines spectral efficiency without taking into account the spectral guard bands that are necessary in practice. The reason for doing this is that the amount of guard band needed, in proportion to the bandwidth, will depend primarily on the *pulse shape*, and less on the number of points in the signal set. We will not consider pulse shaping.

2.2.4 Computing Spectral Efficiency

Recall that we use the channel $n = TW$ times over T seconds, and that the channel is memoryless. The capacity normalized by the time and bandwidth is therefore (see (2.1))

$$C = \frac{\max_{P_{X^n}(\cdot)} I(X^n;Y^n)}{TW} = \frac{\max_{P_X(\cdot)} I(X;Y)}{(T/n)W} \text{ bits/s/Hz.} \tag{2.19}$$

Let $T_s = T/n$ be the symbol time. We are using $T_s = 1/W$ so the denominator of (2.19) is simply one. We further define $E_s = PT_s$ and $N_s = NT_s$ to be the transmit-symbol and noise-symbol energies, respectively. The mutual information in (2.19) is in general some function of $P/N = E_s/N_s$, i.e., we have

$$C = f(E_s/N_s) = f\left(R \cdot E_b/N_0\right), \tag{2.20}$$

where $f(\cdot)$ is some non-decreasing function, and where we have used $E_s = RE_b$ and $N_s = N_0(WT_s)$. Note further that $R \leq C$, so that $C \leq f(C \cdot E_b/N_0)$ and, given E_b/N_0, the "best" C is the largest C^* satisfying

$$C^* = f\left(C^* \cdot E_b/N_0\right). \tag{2.21}$$

Equation (2.21) gives the ultimate E_b/N_0 in terms of C^*, or the ultimate C^* in terms of E_b/N_0. We thus have

$$\eta(E_b/N_0) = C^* \quad \text{bits/s/Hz}. \tag{2.22}$$

For example, suppose we wish to compute $\eta(E_b/N_0)$ when arbitrary complex X are permitted. Since R_{mod} is in principle infinite here, one now maximizes the overall rate R without separating it into R_{mod} and R_c. We thus have

$$C = \log_2(1 + E_s/N_s) \quad \text{bits/s/Hz}. \tag{2.23}$$

We set $E_s = RE_b$, $N_s = N_0$, and use (2.23) and $R \leq C$ to obtain

$$\frac{2^\eta - 1}{\eta} \leq \frac{E_b}{N_0}. \tag{2.24}$$

The η which satisfy (2.24) with equality are given by the curve labeled "Gaussian" in Figure 2.4, where $(E_b/N_0)_{dB} = 10\log_{10} E_b/N_0$. Observe that as $\eta \to 0$ we have $E_b/N_0 \to \ln(2) \approx 0.6931$ which is -1.59 dB.

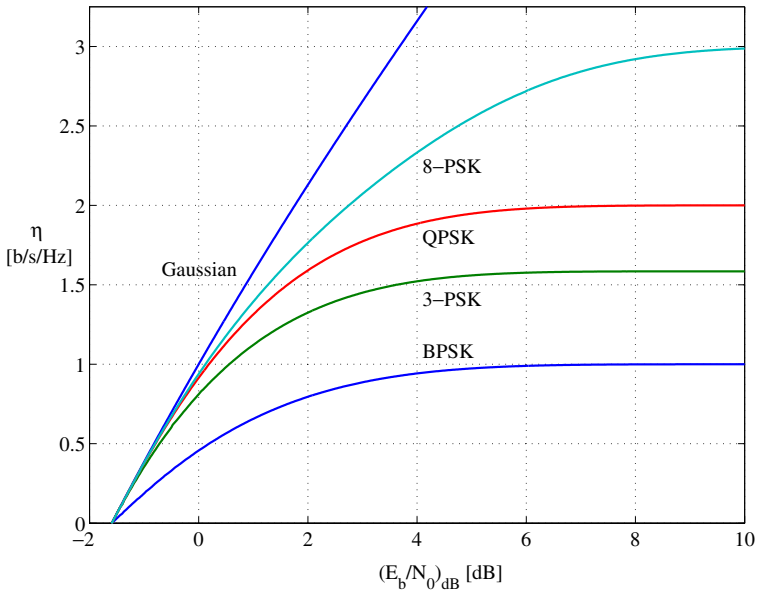


Fig. 2.4 Spectral efficiencies for M-PSK.

The other curves in Figure 2.4 are those for M-PSK with $M = 2, 3, 4, 8$. Note the striking differences between Figures 2.4 and 2.3 at low SNRs. Note also that at $(E_b/N_0)_{dB} \approx -1.59\,\text{dB}$ we have that 3-PSK, QPSK, and 8-PSK all outperform BPSK by a factor of two in rate, but they perform just as well as Gaussian signaling. Thus, BPSK is inefficient at low SNRs, but 3-PSK and QPSK are practically optimal (cf. [183]).

2.2.5 Multi-Antenna Communication

The information theory developed above applies as well to multi-antenna communication. We model such problems by making the channel inputs and outputs *vectors*, i.e., we have

$$\underline{Y} = \frac{H}{d^{\alpha/2}}\,\underline{X} + \underline{Z}, \tag{2.25}$$

where $\underline{X}$, $\underline{Y}$, and $\underline{Z}$ are complex column vectors of length n_X, n_Y, and n_Y, respectively, and H is a complex $n_Y \times n_X$ matrix. The additive noise vector $\underline{Z}$ has independent, Gaussian, entries with independent real and imaginary parts with variance $N/2$ each. The block power constraint is now

$$\sum_{i=1}^{n} E[\|\underline{X}_i\|^2]/n \le P, \tag{2.26}$$

where $\|\underline{X}\|^2 = \underline{X}^\dagger \underline{X}$ and $\underline{X}^\dagger$ is the complex-conjugate transpose of $\underline{X}$.

One can show that the best $\underline{X}$ are zero-mean and Gaussian by the maximum entropy theorem [34, p. 234]; we write the covariance matrix of $\underline{X}$ as $Q_{\underline{X}} = E[\underline{X}\,\underline{X}^\dagger]$. The capacity is (see (2.1))

$$C = \max_{p_{\underline{X}}(\cdot)} I(\underline{X};\underline{Y}) \tag{2.27}$$

$$= \max_{Q_{\underline{X}}} \log \left| I + \frac{1}{d^\alpha N}\, H Q_{\underline{X}} H^\dagger \right|, \tag{2.28}$$

where (2.28) follows by [34, Thm. 9.4.1], $|A|$ is the determinant of A, and where the maximizations in (2.27) and (2.28) are interpreted as constraining $p_{\underline{X}}(\cdot)$ to satisfy (2.26). Consider the matrix H and recall that we can use the singular value decomposition to write $H = V\Sigma U^\dagger$ where U and V are unitary and Σ is a $n_Y \times n_X$ matrix that is zero except

for the (non-negative) singular values on the diagonal [75, p. 414]. We thus have

$$C = \max_{Q_{\underline{X}}} \log \left| I + \frac{1}{d^{\alpha} N} \Sigma Q_{\underline{X}} \Sigma^{\dagger} \right|, \tag{2.29}$$

where we have used $|V| = 1$ and $V^{\dagger} V = I$ to remove the V [75, p. 414], and where we have included the U in the optimization of $Q_{\underline{X}}$.

The resulting optimization problem is the same as for parallel Gaussian channels [34, p. 250]. The optimal $Q_{\underline{X}}$ for this problem is known to be diagonal with entries $P_i = E[|X_i|^2]$, $i = 1, 2, \ldots, n_X$, chosen according to a "waterfilling" solution. More precisely, if we let σ_i be the ith diagonal entry of Σ, we choose P_i to satisfy

$$P_i = \begin{cases} \max\left(Q - \frac{N}{\sigma_i^2}, 0\right) & \text{if } \sigma_i \neq 0, \\ 0 & \text{otherwise} \end{cases} \tag{2.30}$$

where the "water level" Q is chosen so that

$$\sum_{i=1}^{n_X} P_i = P. \tag{2.31}$$

2.3 Link Layer: ARQ and HARQ Protocols

For the PHY layer, error control is usually based on both Automatic Retransmission ReQuest (ARQ) and FEC schemes ("forward" refers to the non-feedback aspect of error control where a code automatically corrects errors detected at the receiver). ARQ is a link layer protocol that provides reliability based on error-detecting codes and retransmissions. A parity-bit or CRC check of a received packet triggers retransmission requests. If the receiver determines that the packet is in error, it sends a negative acknowledgment (NAK) to the transmitter, otherwise it sends an acknowledgment (ACK). In the former case, the packet is retransmitted. From the perspective of the application layer, such retransmissions are *Automatic* (the acronym ARQ is based on the Morse designation for retransmission request [52]). Some flavors of this protocol include stop-and-wait [186], go-back-N [17], and selective-repeat [121], and they each provide different tradeoffs in throughput and buffering at the transmitter and/or receiver.

The throughput of ARQ protocols can be improved by combining FEC with ARQ in the form of *Hybrid-ARQ* [31, 120]. Hybrid-ARQ lets erroneous received packets be collected and combined in various ways before decoding. One can exploit this idea for cooperative communication as well, and we address such protocols in detail in Chapter 6. Packet combining can be based on hard decisions or soft channel outputs. In the latter case, noisy versions of the same packet are combined by maximal-ratio, equal gain, or selection combining diversity techniques. The transmitted packets can thus be viewed as symbols of a repetition code, and this idea can be extended to more general classes of codes. As another approach, *incremental redundancy* ARQ can be realized with rate compatible punctured convolutional codes (RCPC) by first using the highest rate code from the RCPC code family and sending additional bits as needed [67]. Alternatively, with turbo codes one punctures parity bits [141].

2.4 Network Layer: Wireless Routing Protocols

The role of the network layer is to route packets to their intended destinations. In a traditional network, a node implements a routing table that maps the destination address of a packet to a rule for forwarding that packet. In a wired network, the routing table specifies the outgoing link on which to forward a packet. In a wireless network in which the physical medium is a broadcast channel, the routing table may specify the intended next recipient of a packet.

In wired networks, link topology changes are infrequent and routing tables tend to be fairly static. This is reinforced by the use of policy-based routing in which the routing of packets depends on negotiated agreements among the operators of autonomous systems [51].

By contrast, wireless networks enable node mobility which can induce a stochastic process of link connection and disconnection. In response, mobile ad hoc networks implement specialized routing algorithms in the network layer [88, 147] that adapt to these frequent topology changes. These routing algorithms have been the starting point for mobile networking research. Changes in wireless link topology result in frequent routing table updates and necessitate distributed algorithms

for route discovery and route repair, as well as appropriate metrics for choosing efficient routes.

In the mobile networking community, *Dynamic Source Routing* (DSR) [88] and *Ad hoc On-demand Distance Vector* (AODV) [147] routing have emerged as *de facto* standards for mobile network routing. Each has open-source implementations for linux [30, 42, 129, 155] and the ns-2 simulator [8, 21, 130, 177]. Both protocols implement route discovery, route caching, and route maintenance in response to time variations in link connectivity. The protocols differ in how the routing information is stored. Here we describe these protocols in detail to convey the complexity of the task, even if the network is limited to store-and-forward operations. We start with DSR and then describe how AODV differs.

2.4.1 DSR

DSR is a source routing protocol that adds to each data packet a header specifying the complete route (a sequence of nodes) that the packet must follow to the destination. Each mobile node maintains a cache that holds source routes that it has learned. When a node wants to send a packet to another node, it first checks its route cache for a source route to the destination. If no route is found, the source will invoke the *route discovery* procedure.

A node initiating route discovery (called the initiator) broadcasts a route request (RREQ) packet which may be received by those nodes within range. The RREQ packet identifies the intended destination, referred to as the *target node* and contains a unique request id set by the initiator from a locally maintained sequence number. In addition to the addresses of the initiator and target of the request, each RREQ packet contains a *route record*, the accumulated sequence of hops taken by the RREQ packet as it is forwarded.

When a node receives a RREQ packet, it processes the request according to the following steps:

- If the pair (initiator address, request id) of the RREQ packet is found in this node's list of recently seen requests, the request

is discarded. This removes later copies of the request that arrive at this node by alternate routes.

- If this node's address is already in the route record, the RREQ packet is discarded. This guarantees that no single copy of the request can propagate around a loop.
- If the target of the request matches this node's own address, then the route record contains the route from the initiator by which the request reached this node. In this case, this node returns a copy of this route in a route reply (RREP) packet to the initiator.
- If this node has a route cache entry for the target, it appends this cached route to the accumulated route record in the packet, and returns this route in a RREP packet to the initiator without re-broadcasting the RREQ.
- Otherwise, this node appends its own address to the route record in the RREQ packet, and re-broadcasts the request.

The RREQ thus propagates through the *ad hoc* network until it reaches the target host, which then replies to the initiator. Effectively, route discovery is implemented by controlled flooding of the RREQ packet. If the route discovery is successful, the initiator receives a RREP packet listing a sequence of network hops through which it may reach the target.

In conventional routing protocols, nodes exchange periodic routing updates. If the status of a link or a node changes, the periodic updates eventually reflect the changes to all other nodes, resulting in the computation of new routes. However DSR does not have periodic messages of any kind from any of the mobile nodes. Instead, while a route is in use, the route maintenance procedure monitors the operation of the route and informs the source of any routing errors.

Since wireless links are generally less reliable than wired links, many wireless networks, including 802.11, utilize a hop-by-hop acknowledgment at the data link layer in order to provide early detection and retransmission of lost or corrupted packets. In these networks, route maintenance can be easily provided, since over each link, the transmit-

ting node can determine if the link is up. If the data link layer reports an unrecoverable transmission problem, the transmitting node sends a route error (RRER) packet to the original source of the packet. The RRER packet contains the addresses of the nodes at both ends of the link in error. When a RRER packet is received, the link in error is removed from this node's route cache, and all routes that contain this link are truncated.

We note that a node can add entries to its route cache whenever it learns a new route. In particular, when a node forwards a data packet as an intermediate link on the route in that packet, the forwarding node is able to observe the entire route in the packet. If a node forwards an RREP packet, it can also add the route information from the route record being returned in that route reply, to its own route cache. Finally, since all wireless network transmissions are inherently broadcast, a node may be able to configure its network interface into promiscuous receive mode, and can then add to its route cache the route information from any overheard data or RREP packet.

In order to return the RREP packet to the initiator of the route discovery, the target reverses the route in the route record from the RREQ packet, and use this route to send the RREP packet. This, however, requires the wireless network communication between each of these pairs of hosts to work equally well in both directions, which may not be true in some environments or with some MAC-layer protocols.

2.4.2 AODV

AODV uses a broadcast route discovery mechanism, as is also used by DSR. Instead of source routing however, AODV relies on dynamically establishing route table entries at intermediate nodes. This difference pays off in networks with many nodes, where a larger overhead is incurred by carrying source routes in each data packet. AODV uses destination sequence numbers, as in destination-sequenced distance-vector (DSDV) routing [146], to maintain the most recent routing information. Each node maintains a monotonically increasing sequence number which is used to supersede stale cached routes. AODV also features

timer-based states in each node. A routing entry is deleted if not used during a specific amount of time.

Route discovery in AODV is very similar to DSR. The main difference is the use of sequence numbers. In addition to the similar fields in the DSR RREQ, the AODV RREQ contains the pair (source sequence number, last destination sequence number known to the source). The source sequence number is used to maintain freshness information about the reverse route to the source and the destination sequence number specifies how fresh a route to the destination must be in order to be accepted by the source.

As the RREQ is flooded, it automatically sets up the reverse path from all nodes back to the source. To set up a reverse path, a node records the address of the neighbor from which it received the first copy of the RREQ. These reverse path route entries are maintained long enough for the RREQ to traverse the network and produce a reply to the source.

When an RREQ arrives at a node (possibly the destination itself) that has a routing table entry for the destination, it checks the freshness of that entry by comparing the destination sequence number of the route entry with that in the RREQ. If the sequence number of the RREQ is strictly greater, the routing table entry cannot be used to respond to the RREQ; instead, the node rebroadcasts the RREQ. Otherwise, if the RREQ has not been processed previously, the node then unicasts a RREP packet back to the sender of the RREQ. As the RREP travels back to the source, each node along the path sets up a forward pointer to the node from which the RREP came, updates its timeout information for route entries to the source and destination, and records the latest destination sequence number for the requested destination.

A node receiving an RREP propagates the first RREP for a given source node toward that source. If further RREPs are received, the node updates its routing information and propagates the RREP only if the RREP contains either a greater destination sequence number than the previous RREP or the same destination sequence number with a smaller hop count.

When a link fails, the node upstream of the break sends a RRER packet with a fresh sequence number (i.e., a sequence number that is one greater than the previously known sequence number) and a hop count of infinity to all active upstream neighbors. Then these nodes repeat the same process and so on. This process continues until all active source nodes are notified and then terminates because AODV maintains only loop free routes and there are only a finite number of nodes in the network. We note that this is unlike DSR in which broken link information may not be propagated to all route caches with that link.

2.4.3 Properties of *Ad Hoc* Routing Algorithms

It is important to note that the DSR and AODV algorithms are based on IP packets. Logically, mobile *ad hoc* IP networks differ from ordinary IP networks only in the network layer. In particular, the model abstraction of point-to-point links is preserved; only link state (up/down) changes become more frequent. The cooperation between nodes is limited strictly to packet store-and-forward. Despite this relative simplicity, the deployment of *ad hoc* routing at all nodes in a mobile wireless network results in a network with a very large state space. Whether in the form of a DSR cache or an AODV routing table, every node maintains its best guesses regarding available routes. As mobility causes link topology changes, these guesses will be outdated and likely have conflicts from node to node. As a result, the analysis of routing algorithms in mobile environments can do little more than verify correctness in quasistatic environments.

Instead, a large literature of experimental evaluation has emerged. Much of this is simulation-based [21, 38, 74, 130, 155, 194] but there has also been testing with small networks of nodes and 802.11 radio interfaces [95, 131, 142]. Much of this mobile *ad hoc* routing literature has focused on certain fundamental problems. In particular, route discovery and route maintenance generate a flurry of packets and unnecessary route discovery should be avoided [21, 38, 155]. On the other hand, when mobility induces link failures, fast route maintenance and discovery are vital. In addition, there are non-trivial cross-layer inter-

actions between the MAC, the routing protocol and the transport layer [140, 142]. For example, excess consumption of bandwidth resources by TCP can cause congestion in the 802.11 CSMA protocol, which can cause route maintenance to be invoked [140].

Additional issues arise with interactions with transport layer protocols. Consider an 802.11 *ad hoc* wireless network with stationary nodes such that the PHY layer radio connectivity is adequate. Suppose that these nodes are supporting a TCP file transfer over a multihop radio path. In this case, data packets in the forward direction (from sender to receiver) contend for the channel with RTS/CTS messages at the PHY layer as well as with TCP ACK messages in the reverse direction. These contending data packets cause self interference to the multihop route and can disrupt timely control message exchanges. This condition can be perceived as a link failure, triggering inappropriate route repair or route discovery mechanisms, ultimately resulting in transport layer timeouts and dramatic reductions in throughput [64]. This deficiency is in addition to the problems caused by PHY layer outages induced by fading on a single link, for which transport layer solutions [15, 16] have been developed.

Finally, we recall that DSR and AODV were designed to use only IP packets. As a result, simple hop metrics are the primary basis for routing. Frequently, the fastest route reply is equated with the best route. When links have variable rates, this is not necessarily correct. Since the RREQ and RREP packets are small, they travel faster across long distance hops at lower communication rates. Moreover, in an 802.11 environment which employs uncoded packets, these short control packets can be received on marginal links that are useless for longer data packets. Thus, longer data packets generally benefit from routes with shorter-distance higher-data-rate links. MAC-layer and network-layer enhanced routing algorithms have emerged to address such issues [13, 39]. We will see that these approaches have analogies to the cooperative approaches described in Chapter 6.

3

Network Models

The aim of this chapter is to develop a common framework for analyzing capacities of wireline and wireless networks. Our focus will, in fact, be on physical- and link-layer issues for wireless problems. However, there are close relations between wireline and wireless networks that we wish to highlight, and that we hope will lead to a better understanding of both types of networks.

As is customary, we will represent a network by a graph, i.e., a set $\mathcal{N}$ of nodes and a set $\mathcal{E}$ of edges that are pairs of nodes (see Figure 1.1). If an edge (u, v) is directed, then the ordering tells us that the edge goes from node u to node v. For example, the directed network in Figure 3.1 below has $\mathcal{N} = \{1, 2, 3\}$ and $\mathcal{E} = \{(1, 2), (2, 3)\}$.

3.1 Wireline Network Models

The channels (or edges) of a wireline network are usually modeled as being independent in the sense that the signals carried by different channels do not interfere with each other (see Table 1.1). Moreover, one often assumes that any noise has been removed by ARQ and FEC at the PHY layer. The resulting network is hence noise free, but the edges have

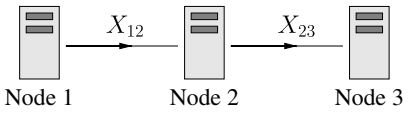


Fig. 3.1 A wireline network.

capacity constraints due to bandwidth limitations. The bandwidths of different channels in a wireline network can differ greatly, e.g., if the channels represent copper wires or fiber-optic cables. The channels usually change slowly so that nodes can learn their channels during an initialization phase. However, the nodes may be aware of only the local topology of the network. A common approach is to organize data into packets that are sometimes lost, e.g., due to congestion, in which case one encounters packet erasures.

A simple such network with three nodes and two edges is shown in Figure 3.1 where the signal carried on the edge between nodes 1 and 2 is labeled X_{12}, and similarly for X_{23}. Suppose the capacities of the first and second edges are C_{12} and C_{23} bits per channel use, respectively. As an abstraction, we choose the alphabets of X_{12} and X_{23} to have sizes $2^{C_{12}}$ and $2^{C_{23}}$ (assumed to be integers), respectively. One can, therefore, transmit at most C_e bits through edge e every time one uses this edge.

A wireline network often has delay, processing, and input/output constraints on the nodes. For instance, suppose node 2 in Figure 3.1 has limited processing power. We could model this as shown in Figure 3.2

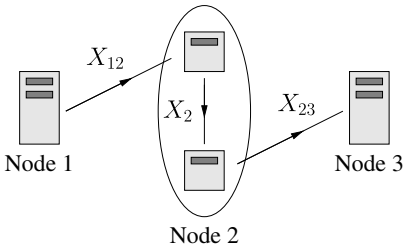


Fig. 3.2 A wireline network with a node constraint.

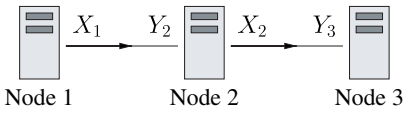


Fig. 3.3 A wireline network with another type of node constraint.

where, as compared to Figure 3.1, node 2 is split into two nodes and an edge carrying X_2 that has an alphabet of size 2^{C_2}.

Consider next a second type of node constraint where node 2 in Figure 3.1 has only one input/output port on which it can either transmit or receive. We model this as shown in Figure 3.3: we introduce input and output variables X_u and Y_u, respectively, for every node u and write

$$Y_2 = \begin{cases} X_1 & \text{if } X_2 = 0, \\ 0 & \text{if } X_2 \neq 0 \end{cases} \tag{3.1}$$

while for node 3 we have $Y_3 = X_2$. The symbol 0 in (3.1) might represent a "silence" symbol. The abstraction in (3.1) will later help to define information-theoretic models for wireless networks as well.

As a somewhat more complex example, consider the network shown in Figure 3.4 where both nodes 1 and 2 have port constraints as in (3.1). We write the resulting network equations as $Y_{13} = X_{13}$, $Y_{23} = X_{23}$, and

$$Y_{12} = \begin{cases} X_{12} & \text{if } X_{21} = 0 \text{ and } X_{23} = 0, \\ 0 & \text{if } X_{21} \neq 0 \text{ or } X_{23} \neq 0 \end{cases} \tag{3.2a}$$

$$Y_{21} = \begin{cases} X_{21} & \text{if } X_{12} = 0 \text{ and } X_{13} = 0, \\ 0 & \text{if } X_{12} \neq 0 \text{ or } X_{13} \neq 0. \end{cases} \tag{3.2b}$$

Note that the network in Figure 3.4 has two paths to node 3 from nodes 1 and 2, rather than just one as in Figure 3.1. We can thus use more sophisticated communication strategies for the network of Figure 3.4 than for the network in Figure 3.1. Note further that, for wireline networks, it is often useful to give the edge variables two indices, one index for the "start" node and one for the "end" node.

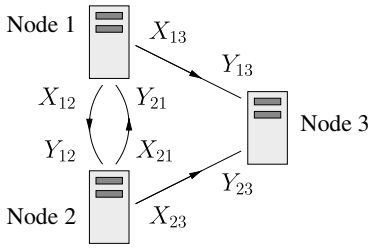


Fig. 3.4 Another wireline network.

3.2 Wireless Channel Models

Wireless channels have been the subject of a large body of research, including both analytic and empirical modeling; see [61, 151, 178] and references therein. Consider, for instance, the wireless network depicted on the left in Figure 3.5. The signals transmitted by the devices are *bandlimited* and can, by Nyquist sampling theory, be represented by sequences of discrete-time symbols. A commonly studied class of problems is based on an AWGN model where every output sample at nodes 2 and 3 can be written as

$$Y_2 = \frac{h_{12}}{d_{12}^{\alpha/2}} X_1 + Z_2, \tag{3.3a}$$

$$Y_3 = \frac{h_{13}}{d_{13}^{\alpha/2}} X_1 + \frac{h_{23}}{d_{23}^{\alpha/2}} X_2 + Z_3, \tag{3.3b}$$

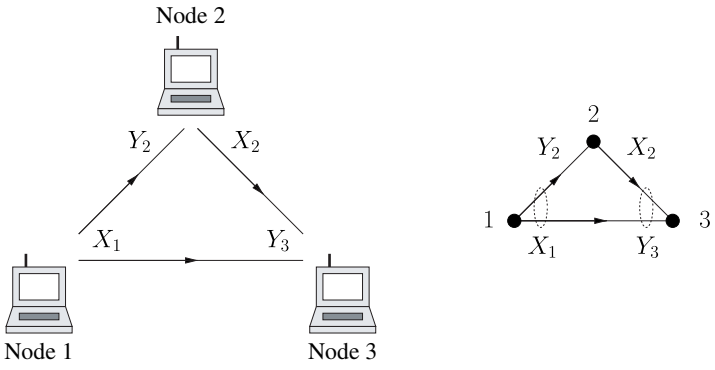


Fig. 3.5 A wireless network and its graph.

respectively, where $X_1, X_2, Y_2, Y_3, Z_2, Z_3$ are complex random variables, h_{uv} and d_{uv} are the respective channel gains and distances between nodes u and v, and α is an attenuation exponent (see Section 2.2.1). The Z_2 and Z_3 are defined as $Z_2 = Z_{2R} + jZ_{2I}$ and $Z_3 = Z_{3R} + jZ_{3I}$, where $j = \sqrt{-1}$ and $Z_{2R}, Z_{2I}, Z_{3R}, Z_{3I}$ are independent, Gaussian random variables with variance $N/2$. The Z_2 and Z_3 are independent of X_1 and X_2.

This wireless network is naturally represented by the graph shown on the right in Figure 3.5. The idea of this graph is that every node u has one channel input X_u and one channel output Y_u. Of course, if a node only receives or transmits then one can ignore the former or latter variable. We draw a directed edge from node u to node v if X_u contributes to Y_v. The graph now permits *broadcasting* (X_1 contributes to both Y_2 and Y_3) but this causes *interference* (X_1 and X_2 interfere at node 3).

Thus wireless channels differ in many ways from wireline channels (see Table 1.1). Additionally, there are significant *time variations* due to node mobility and the propagation environment. The channel changes are considered to be either slow or fast depending on many factors, for example the device velocity, bandwidth, delay constraints, and electromagnetic wave scattering and absorption. Slow channel variations might occur when the device has low velocity (laptops), the bandwidth is large so that individual symbols are short, there are tight delay constraints (voice traffic), or there is line-of-sight communication. In fact, signal strength can fluctuate rapidly even at low device velocity due to multipath propagation; this spatial phenomenon is called *fading*. Rapid channel variations occur when the device has high velocity (planes, trains), the bandwidth is small so that individual symbols are long, there are relaxed delay constraints (data traffic), or there is rich scattering causing multipath.

Stated more succinctly, a channel (or edge) is called *slow fading* if all the encoded symbols of each data packet traversing this channel encounter only one channel realization (often represented by a multiplicative gain, see (3.3) below). The channel is called *fast fading* if the encoded symbols of each data packet encounter many channel realizations. Of course, there are intermediate situations where the channel

is neither slow nor fast fading. However, for simplicity we will consider only the two extreme cases.

We next model channel time variations in more detail. Suppose we use the channel n times, and we index the edge variables at time i, $i = 1, 2, \ldots, n$, as $X_{1,i}$, $Y_{2,i}$, $h_{12,i}$, and so forth (see (3.3)). We might wish to choose the sequences $\{h_{uv,i}\}_{i=1}^{n}$ by using electromagnetic wave propagation equations for specified geographies, and for specified device trajectories and velocities. This approach is, however, sensitive to the choice of system variables and is too complex to be useful for many scattering environments encountered in practice. Instead, we admit uncertainty and model the sequences $\{h_{uv,i}\}_{i=1}^{n}$ as realizations of integer time stochastic processes $\{H_{uv,i}\}_{i=1}^{n}$. Several types of processes have been considered in the literature, and each has its own peculiarities. We opt for simple classes of models since our focus will be on cooperative strategies rather than channel characterization. These models are characterized by the *marginal distributions* of the H_{uv} and the *temporal correlation* of the $H_{uv,i}$.

With respect to the distribution of the H_{uv}, we consider two models. First, the *Rayleigh fading* channel model assumes the signal at an antenna consists of a large number of independent randomly phased unresolvable multipath components. In this case, the H_{uv} are independent, complex, Gaussian, zero-mean, unit variance random variables with independent real and imaginary parts with variance $1/2$. We further assume that all the H_{uv} are independent of X_1, X_2, Z_2, Z_3. In this case, the channel capacity of the individual link (u, v) does not depend on the channel phase, and it is sometimes convenient to follow [22] and use the simplified real-valued signals

$$Y_{v,i} = \frac{\sqrt{H_{uv,i}}}{d_{uv}^{\alpha/2}} X_{u,i} + Z_i \tag{3.4}$$

in which we ignore the quadrature signal components (but one should keep the underlying complex model in mind). In this case, the channel gain $H_{uv,i}$ is a real-valued exponential random variable with expected value $E[H_{uv,i}] = 1$. The actual data rates in the complex baseband channel will be double those derived under the real-valued model. This same real-valued model may also be applied to the OFDM channel

model as the capacity of each subchannel is insensitive to the subchannel phase.

An instructive second model is the *uniform-phase fading* channel in which $H_{uv} = e^{j\Phi_{uv}}$ and the Φ_{uv} are independent and uniform over the interval $[0, 2\pi)$. This model is unrealistic for wireless environments, and may seem peculiar since the capacity of the individual link (u, v) does not depend on the phase of H_{uv}. However, as we shall see in Chapter 4, this model has the didactic advantages of giving simple capacity expressions and important insight for wireless networks in which multiple transmitters can derive advantages by phase aligning signals at a receiver. We remark that the name "fading" is perhaps inappropriate here because there are no signal strength variations. However, we feel that the suggestive label compensates adequately for the loss in precision.

With respect to temporal correlation, as mentioned above we focus on two classes of models that correspond to extreme forms of fast and slow fading. More precisely, in the fast fading model, the channel gains $h_{uv,i}$, $i = 1, 2, \ldots, n$, are chosen independently with the distributions $P_{H_{uv}}(\cdot)$. Our second model class describing slow fading has the $h_{uv,1}$ drawn randomly before transmission for all (u, v), and we set $h_{uv,i} = h_{uv,1}$ for $i = 2, \ldots, n$. In this slow fading case, the reliable communication rates one can achieve are therefore random variables, and one is usually interested in characterizing their probability distributions. This model is often referred to as *block-fading* or *quasistatic*, and it is a realistic model for shadow fading, or for systems employing time-division multiaccess (TDMA) or orthogonal frequency division multiplexing (OFDM) with TDMA symbols.

An appropriate definition of capacity depends on both the marginal distributions and the temporal correlations of the link gains. In addition, the capacity depends on how the transmitter and receiver observe the channel state. In the next section, we interpret the incomplete knowledge of the wireless channel state process as a constraint on a wireless device. We then return in Section 3.4 to examine capacity metrics in terms of wireless channel models and constraints on wireless devices.

3.3 Wireless Device Models

Wireless devices usually have several types of constraints. A commonly studied constraint is that the powers (or energies) of the inputs satisfy (see (2.4))

$$\sum_{i=1}^{n} E[|X_{t,i}|^2]/n \le P_t, \quad t = 1, 2. \tag{3.5}$$

A more severe type of constraint is

$$|X_{t,i}|^2 \le E_t, \quad t = 1, 2, \tag{3.6}$$

where the per-symbol energy is constrained.

A second constraint is similar to the wireline port constraint (3.1). Note that the model defined by (3.3a) and (3.3b) implicitly lets node 2 transmit and receive at the same time in the same frequency band. However, this is often not possible due to the large differences in transmit and receive energies at the antenna of node 2. Most practical wireless devices operate under a *half-duplex* constraint that we can model as

$$Y_2 = \begin{cases} \dfrac{h_{12}}{d_{12}^{\alpha/2}} X_1 + Z_2 & \text{if } X_2 = 0, \\ 0 & \text{if } X_2 \neq 0. \end{cases} \tag{3.7}$$

Note the similarity between (3.1) and (3.7).

A third constraint is that, because the channel changes over time, a device usually does not know the internode distances d_{uv} or channel gains h_{uv} ahead of time. A simple approach is to assume that each node knows the gain h_{uv}/d_{uv}^{α} between itself and the nodes with which it cooperates. However, the validity of this assumption depends strongly on the speed of channel variations. Learning the channel and network parameters is usually a challenging task even for nodes near each other. In general, different models must be considered for different scenarios. For example, there might be two tiers of nodes: one tier with high capacity to neighboring nodes and therefore sufficient channel and network knowledge, and a second tier that lacks such knowledge. The communications theory for the latter case is fascinating in its own right, and we refer to [115, 117, 135, 184] for further information.

3.4 Wireless Capacity and Channel State Information

The direct application of Shannon's theory, as early as Shannon's own work [166], provided the capacity measure now often called *ergodic capacity* [19, 24, 62]. The ergodic capacity is the maximum achievable time-average rate of reliable communication, and it is appropriate when the receiver observes a typical sequence of channel states during the reception of a codeword. As the codeword length is proportional to the receiver's decoding delay, one can also say that ergodic capacity is an appropriate metric when the channel variation is fast relative to the delay an application can tolerate.

The quasistatic channel model that is the limiting case of *very* slow fading has given rise to two additional capacity formulations, namely *delay limited capacity* [69] and *capacity versus outage probability* [24, 144] (see [19] for a comprehensive survey). Delay limited capacity is the maximum rate at which the instantaneous mutual information can be kept constant over all states of the fading process. Capacity versus outage probability was introduced for block interference fading channels, and an outage occurs if the instantaneous mutual information is less than a fixed transmission rate. Slow block-fading channel models are the subject of Chapter 5, where one of the important figures of merit is *diversity*.

We next turn to device channel state information (CSI). A receiver may deliberately introduce significant delay to form a more accurate estimate of the channel state from both past and future received symbols. It is common to assume that the receiver has accurate CSI as a byproduct of the symbol detection process. On the other hand, transmitter CSI is less likely to be accurate because the timely delivery of channel measurements from the receiver to the transmitter becomes increasingly difficult as the channel varies more quickly. In time-division duplex (TDD) systems, a transmitter and receiver may alternate their transmitter and receiver roles, permitting each to garner accurate CSI as the receiver. Nevertheless, in employing this CSI at the transmitter, measurement delay may degrade accuracy.

In general, CSI issues are complex and methods are often motivated by detailed considerations such as, e.g., how quantized CSI is embedded

in feedback packets. However, for a point-to-point link, we consider four models corresponding to whether there is CSI at the transmitter (CSIT) and whether there is CSI at the receiver (CSIR). We summarize properties of ergodic capacity when the CSI is either perfect or completely unavailable; more information can be found in [22]. We define a channel with input X, output Y, and channel state H by $P_{Y|XH}(\cdot)$ (in the following, we use uppercase "P" for both distributions and densities for simplicity).

- **No CSIT, No CSIR**
  Neither device observes H, so the channel is

$$P_{Y|X}(y|x) = \sum_h P_{Y|XH}(y|x,h)\, P_H(h) \tag{3.8}$$

  and has capacity $\max_{P_X(\cdot)} I(X;Y)$. The simplicity in the capacity formula is perhaps deceiving since the optimization over $P_X(\cdot)$ can be tricky. For instance, the capacity-achieving $P_X(\cdot)$ for power-constrained AWGN fading channels are discrete point mass distributions [1, 66, 81].
- **CSIR, No CSIT**
  The communication system can be viewed as a channel with the usual input X but with the output pair (see [22, 174])

$$(Y,H) = \left(\frac{H}{d^{\alpha/2}}\, X + Z, H\right). \tag{3.9}$$

  The capacity is thus

$$\max_{P_X(\cdot)} I(X;YH) = \max_{P_X(\cdot)} I(X;Y|H). \tag{3.10}$$

- **CSIT, No CSIR**
  Shannon studied this problem with causal CSI [166]. For some applications, for example information storage, the channel state sequence $\{H_i\}_{i=1}^n$ may even be known noncausally [60, 71]. In the latter case, the capacity is $\max[I(U;Y) - I(U;H)]$ where the maximization is over all $P_{UX|H}$ for which $U - [X,H] - Y$ forms a Markov chain [60].

- **CSIT, CSIR**
  In this case, the instantaneous capacity for state h is

$$C(h) = \max_{P_{X|H}(\cdot)} I(X;Y|H=h) \tag{3.11}$$

  and the ergodic capacity is $E[C(H)]$ (see [22, Prop. 2]). Let $P_i = E[|X_i|^2]$ be chosen as a function $P(H_i)$ of H_i. To identify the transmission policy that maximizes $E[C(H)]$ subject to the average power constraint

$$E[P(H)] \leq \overline{P} \tag{3.12}$$

  we observe that a sequence of channel uses in time provides the same mutual information as would those same channel uses in parallel Gaussian channels (see Section 2.2.5). The optimal $P_{X|H}(\cdot|h)$ thus corresponds to a "waterfilling" power allocation analogous to (2.30), i.e., we choose

$$P(h) = \begin{cases} \max\left(Q - \frac{N}{|h|^2}, 0\right) & \text{if } |h| \neq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.13}$$

  The waterfilling level Q is chosen to satisfy (3.12) with equality; see [22, 62].

3.5 Mixed Networks

A mixed wireline/wireless network and its graph is shown in Figure 3.6. Note that the graphs for the networks of Figures 3.4 and 3.6 are the same, but they are interpreted differently. In general, we will say that every node u has one channel input X_u and one channel output Y_u. We draw a directed edge from node u to node v if X_u contributes to Y_v. For example, the wireline network of Figure 3.4 has X_1 being a vector $[X_{12}\ X_{13}]$, and similarly for X_2 and Y_3. The X_{12} and X_{13} here represent symbols sent on non-interfering (or parallel) channels. The same is true for the mixed network of Figure 3.6. However, instead of

$$Y_3 = [Y_{13}\ Y_{23}] = [X_{13}\ X_{23}] \tag{3.14}$$

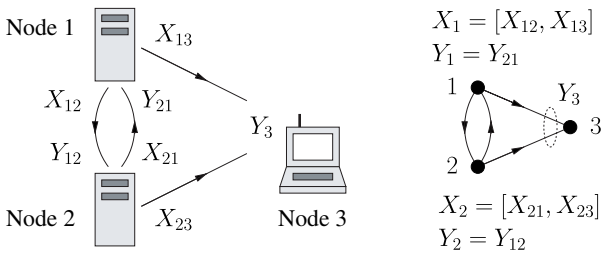


Fig. 3.6 A mixed wireline/wireless network and its graph.

as in Figure 3.4, we might now have

$$Y_3 = \frac{h_{13}}{d_{13}^{\alpha/2}} X_{13} + \frac{h_{23}}{d_{23}^{\alpha/2}} X_{23} + Z_3, \tag{3.15}$$

where X_{13} and X_{23} are complex random variables. Hence, the signal X_{13} interferes with X_{23}. Furthermore, if there are "port" constraints in the network of Figure 3.6 then one might wish to add (3.2a) and (3.2b) to the model.

3.6 Source and Destination Models

The previous sections describe device (node) and channel (edge) models. A communication network also has sources and sinks, and we will consider bit sources where every bit is uniformly distributed and is independent of all other message bits. We thereby implicitly assume that multimedia signals, such as voice or video, are compressed to bit streams, and any intra- or inter-source correlations are ignored. In other words, we assume there is a separation (layering) between the application layer and the rest of the network protocol stack. This chapter further assumes that the sources are not bursty, i.e., any source time variations are smoothed by large buffers, and delay is not critical.

It remains to specify where sources and sinks are located. Once this choice is made, one has a network that is often referred to as a "channel" in the information theory literature. For example, Figures 3.7–3.9 show several such "channels." In these figures, the source emits a message W_m (represented by a hollow node) and a sink accepts an estimate $\hat{W}_m$

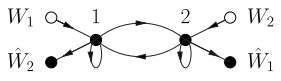
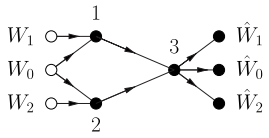
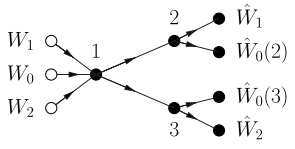
Network	Device Nodes	Channel Edges	Sources	Sinks	Graph
Point–to–Point Channel (DMC & AWGN Channel)	2	1	1	1	
Two–way Channel (2WC)	2	4	2	2	
Multiaccess Channel (MAC)	3	2	3	3	
Broadcast Channel (BC)	3	2	3	4	

Fig. 3.7 Basic networks.

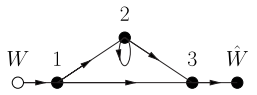
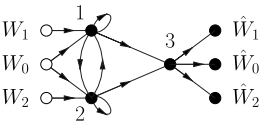
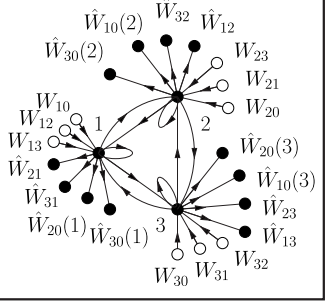
Network	Device Nodes	Channel Edges	Sources	Sinks	Graph
Relay Channel (RC)	3	4	1	1	
MAC with Generalized Feedback (MAC–GF)	3	6	3	3	
Three–way Channel (3WC)	3	9	9	12	

Fig. 3.8 Cooperative networks with three device nodes.

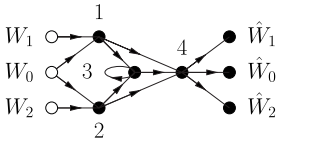
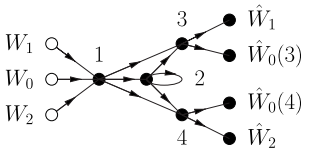
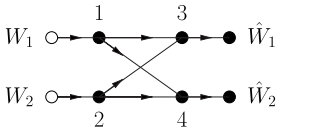
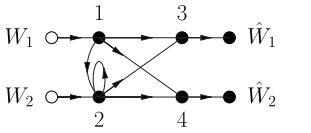
Network	Device Nodes	Channel Edges	Sources	Sinks	Graph
MAC with a Dedicated Relay (MAC–DR)	4	6	3	3	
BC with a Dedicated Relay (BC–DR)	4	6	3	4	
Interference Channel (IC)	4	4	2	2	
Cognitive Radio Channel	4	6	2	2	

Fig. 3.9 Networks with four device nodes.

(represented by a solid node) of W_m. If W_m is decoded at more than one node, we write $\hat{W}_m(u)$ to represent an estimate of W_m at node u. We begin by discussing some basic networks.

(1) A *point-to-point channel* (see Section 2.2.1) has two device nodes, one channel edge, one source, and one sink (see Figure 3.7). Device node 1 has a channel input X and device node 2 has a channel output Y. The channel is often modeled as being memoryless in the sense that an output Y_i at time i is affected only by the input X_i at time i, i.e., Y_i is statistically independent of all other inputs and outputs given X_i. One can therefore define a noisy channel by a single-sample (or "single-letter") conditional probability distribution $P_{Y|X}(\cdot)$. If the alphabets $\mathcal{X}$ and $\mathcal{Y}$ of the respective X and Y are discrete and finite, the channel is called a *discrete memoryless*

channel or DMC [34, p. 193], [35, p. 100], [56, p. 73] (see Section 2.2.1). If the channel is noise-free, as in Figure 3.1, then we have

$$P_{Y|X}(b|a) = 1(b = a), \tag{3.16}$$

where $1(\cdot)$ is the indicator function that is 1 if its argument is true and is 0 otherwise. If the alphabets are continuous, e.g., for AWGN channels such as (3.3a), then one focuses on a conditional probability density $p_{Y|X}(\cdot)$. Alternatively, one simply considers a sum as in (3.3a).

(2) A *two-way channel* (2WC) has two sources and sinks [167]. Device node u, $u = 1, 2$, has one channel input X_u and one channel output Y_u. The self-loops in Figure 3.7 specify that X_u contributes to Y_u for both u. This would occur, e.g., if the graph represents a wireless network where both nodes have the half-duplex constraint (3.7). The channel is often modeled as being memoryless in the sense that an output $Y_{u,i}$ at time i is affected only by the inputs $X_{1,i}$ and $X_{2,i}$ at time i, and is statistically independent of all other *past* inputs and outputs (note that $Y_{u,i}$ generally contributes to future $X_{u,k}$, $k > i$; $Y_{u,i}$ can thus be statistically dependent on future $X_{u,k}$). One can define the memoryless channel by a conditional probability distribution $P_{Y_1Y_2|X_1X_2}(\cdot)$. Note that a 2WC includes a DMC as a special case, both without or with feedback from node 2 to node 1.

(3) A *multiaccess channel* (MAC) has three device nodes, three sources and sinks, and all sinks are located at device node 3 [2, 119], [35, p. 270], [34, p. 388]. Note that the common message W_0 is given to both nodes 1 and 2. The memoryless MAC is defined by a conditional probability distribution $P_{Y_3|X_1X_2}(\cdot)$. For simplicity, one often ignores the subscript 3 and replaces Y_3 by Y.

(4) A *broadcast channel* (BC) has three device nodes, three sources, and four sinks, and all sources are located at node 1 [35, p. 359], [34, p. 418]. The common message W_0 is decoded at both receive nodes 2 and 3. The memoryless BC is

defined by a conditional probability distribution $P_{Y_2Y_3|X_1}(\cdot)$. Again, for simplicity one often writes X in place of X_1, and Y_1, Y_2 in place of Y_2, Y_3.

For the networks in Figure 3.7, only the 2WC has nodes that actively cooperate; the other networks have nodes that cooperate only in the sense that they can, say, agree on a common time reference. Figure 3.8 shows some *cooperative* networks with three nodes.

(1) A *relay channel* (RC) has one source and sink [33], [34, p. 428], [180], [181]. More generally, a RC has many device nodes, but only one source and sink. The device nodes without sources and sinks are called *relays* and aid communication, perhaps generously, or through incentives, or competitively. Observe that the relay in Figure 3.8 has a self-loop, which is needed if there is a half-duplex constraint (3.7) for example. The graph in Figure 3.5 thus permits full-duplex transmission since X_2 does not affect Y_2. A memoryless RC can be defined by $P_{Y_2Y_3|X_1X_2}(\cdot)$. Note that, even if the RC is memoryless, the network does have memory in the sense that there is a processing delay at the relay.

(2) A *MAC with generalized feedback* (MAC-GF) has three sources and sinks like a MAC, but now the source device nodes can cooperate [28], [99], [187]. The MAC-GF includes the RC as a special case by setting W_0 and W_2 to be constants and by letting node 1 transmit but not receive. In this case, node 2 transmits only as a relay for node 1. A memoryless MAC-GF can be defined by $P_{Y_1Y_2Y_3|X_1X_2}(\cdot)$.

(3) A *three-way channel* (3WC) is the most general network with three device nodes. The 3WC has 9 sources and 12 sinks: there is a common message and two "private" messages at each node. The 3WC includes all of the previous networks as special cases (DMC, 2WC, MAC, BC, RC, MAC-GF). The memoryless 3WC channel is defined by $P_{Y_1Y_2Y_3|X_1X_2X_3}(\cdot)$.

The last set of networks we consider have four device nodes, as shown in Figure 3.9. These networks will serve as examples describing communication strategies later on.

(1) A multi-access relay channel (MARC) or *MAC with a dedicated relay* (MAC-DR) is a MAC with one extra device node that acts as a relay [110, 156]. For example, such a situation might occur in a wireless cellular network where a relay station (node 3) helps to transmit data to a base station (node 4). A memoryless MAC-DR can be defined by $P_{Y_3Y_4|X_1X_2X_3}(\cdot)$.
(2) A *BC with a dedicated relay* (BC-DR) is a BC with an extra relay node [106, 118]. A memoryless BC-DR can be defined by $P_{Y_2Y_3Y_4|X_1X_2}(\cdot)$.
(3) An *interference channel* (IC) is a non-cooperative model where two device nodes interfere with each other's transmissions [3, 68]. A memoryless IC can be defined by $P_{Y_3Y_4|X_1X_2}(\cdot)$.
(4) Alternatively, suppose node 2 attempts to cooperate. One might add a directed edge from node 1 to node 2, say, as well as a self-loop at node 2 to include a half-duplex constraint. This channel has been dubbed a *cognitive radio channel* because of its relation to cognitive radio applications [40].

3.7 Network Capacity

In the context of multi-node networks, we would like to define capacity in a similar manner as in Section 2.2.1 for a DMC and AWGN channel. Suppose there are M sources and source m puts out message W_m with B_m bits, $m = 1, 2, \ldots, M$. The messages are assumed to be independent. For convenience, we introduce a common network *clock* that governs the transmissions of channel inputs $X_{u,i}$ and outputs $Y_{u,i}$. Basically, the clock ticks n times and node u can transmit $X_{u,i}$ after clock tick $i-1$ and before clock tick i. Node u receives $Y_{u,i}$ at clock tick i. Hence, $X_{u,i}$ can be any function of its own messages and its past channel outputs $Y_u^{i-1} = Y_{u,1}, Y_{u,2}, \ldots, Y_{u,i-1}$ [34, p. 444], [27, 102, 167, 180].

The main advantage of introducing a network clock is that it helps to clarify exposition, notation, and concepts such as encoding, decoding, causality, cooperation, relaying, and so forth. An obvious criticism is that a network clock seems artificial and restrictive because all nodes must operate synchronously. However, we remark that node asynchronism can often be dealt with by introducing appropriate device constraints on channel knowledge.

A network clock makes it easy to define the capacity of a network. Suppose the network clock ticks n times so that the rate of source message W_m is $R_m = B_m/n$ bits per clock tick. Let $\mathcal{D}_m$ be the set of nodes that decode W_m. The *capacity region* is the closure of the set of rate-tuples $(R_1, R_2, \ldots, R_M)$ for which, for sufficiently large n, there exist encoders and decoders so that the error probability

$$\Pr\left[\bigcup_{m=1}^{M} \bigcup_{u\in\mathcal{D}_m} \left\{\hat{W}_m(u) \neq W_m\right\}\right] \tag{3.17}$$

can be made as close to 0 as desired (but not necessarily exactly 0).

An interesting fact is that the capacity region is not yet known for *any* of the memoryless networks in Figures 3.7–3.9 *except* for the DMC and the MAC. This perhaps surprising situation suggests that finding the information-theoretic capacity regions for networks is a difficult problem indeed. Fortunately, for certain networks with AWGN one can say more. For example, the capacity regions of the 2WC and BC with AWGN are known.[1] The capacity regions of the RC, MAC-GF, MAC-DR, and BC-DR have recently been resolved for some cases.

As an example, consider the AWGN MAC with (see Figure 3.7)

$$Y = \frac{h_1}{d_1^{\alpha/2}} X_1 + \frac{h_2}{d_2^{\alpha/2}} X_2 + Z \tag{3.18a}$$

$$\sum_{i=1}^{n} E[|X_{u,i}|^2]/n \leq P_u, \quad u = 1, 2, \tag{3.18b}$$

where X_u, h_u, and Z are real, Z has variance N, and where the expectation is over the codewords. Consider the case $R_0 = 0$. The capacity

[1] The capacity region of the AWGN BC with vector symbols is not yet resolved if $R_0 > 0$, $R_1 > 0$, and $R_2 > 0$.

region is the set of non-negative rate pairs (R_1, R_2) in the pentagon defined by the bounds (see [34, p. 405])

$$R_1 \leq \frac{1}{2}\log_2(1+\gamma_1) \tag{3.19a}$$

$$R_2 \leq \frac{1}{2}\log_2(1+\gamma_2) \tag{3.19b}$$

$$R_1 + R_2 \leq \frac{1}{2}\log_2(1+\gamma_1+\gamma_2), \tag{3.19c}$$

where

$$\gamma_u = \left(\frac{P_u}{N}\right)\frac{|h_u|^2}{d_u^\alpha}, \quad u = 1, 2, \tag{3.20}$$

and where the rate units are bits/clock tick. This region is shown in Figure 3.10.

We remark that a simple communication strategy for the MAC with block power constraints (3.18b) is to use time-division multiplexing (TDM) or frequency-division multiplexing (FDM). For example, suppose nodes 1 and 2 use the fractions α and $1-\alpha$ of the available bandwidth, respectively. The noise powers for nodes 1 and 2 thus reduce to αN and $(1-\alpha)N$, respectively (see Section 2.2.3), and the rates are

$$R_1 = \frac{\alpha}{2}\log_2\left(1+\frac{\gamma_1}{\alpha}\right) \tag{3.21a}$$

$$R_2 = \frac{1-\alpha}{2}\log_2\left(1+\frac{\gamma_2}{1-\alpha}\right). \tag{3.21b}$$

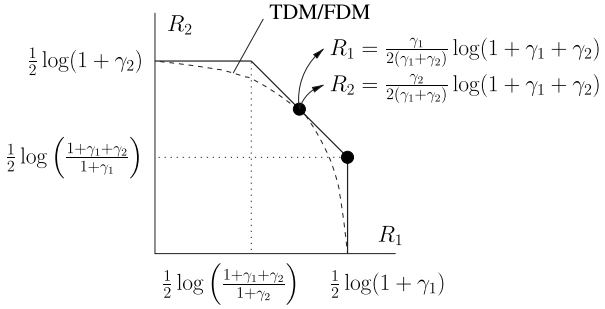


Fig. 3.10 AWGN MAC channel capacity region.

Similarly, if nodes 1 and 2 use the fractions α and $1-\alpha$ of the available time, then they can boost their powers while transmitting by the factors α^{-1} and $(1-\alpha)^{-1}$, respectively. The resulting rate pairs (3.21a) and (3.21b) are plotted in Figure 3.10. In particular, by choosing $\alpha = \gamma_1/(\gamma_1+\gamma_2)$ one achieves a boundary point with

$$R_1 + R_2 = \frac{1}{2}\log_2(1+\gamma_1+\gamma_2). \tag{3.22}$$

TDM and FDM are thus effective techniques for the MAC with AWGN.

3.7.1 Cut-Set Bound on Capacity

The capacity of networks is useful to know because it provides a benchmark for what is possible and guides the design of protocols and codes. However, as we have seen, the capacity is often difficult to determine. As a compromise, one is left with trying to find capacity *bounds*. *Inner* bounds on the capacity region are based on creatively designing protocols and codes to show that certain rate-tuples are achievable. *Outer* bounds on the capacity region are, however, often more difficult and tricky to develop as they must hold for all possible protocols and codes. A very useful outer bound for large networks is known as a *cut-set* bound [34, p. 445], [9, 43, 167].

Suppose that $\mathcal{U}$ and $\mathcal{V}$ are disjoint subsets of the set $\mathcal{N}$ of network device nodes, i.e., $\mathcal{N}$ does not include message nodes and message-estimate nodes. Let $(\mathcal{U},\mathcal{V})$ denote the set of edges that lead from $\mathcal{U}$ to $\mathcal{V}$. Consider a set $\mathcal{S} \subseteq \mathcal{N}$ and let $\overline{\mathcal{S}}$ be the complement of $\mathcal{S}$ in $\mathcal{N}$. A *cut* separating the message W_m from one of its estimates $\hat{W}_m(u)$ is a pair $(\mathcal{S},\overline{\mathcal{S}})$ where the W_m message node is connected to a node in $\mathcal{S}$ but not in $\overline{\mathcal{S}}$, and where the $\hat{W}_m(u)$ message-estimate node is connected to a node in $\overline{\mathcal{S}}$.

The above definition of a cut is the same as the usual notion of a cut in wireline networks. Furthermore, there is a cut-set bound that we develop below that applies to all the networks considered in this Chapter. For directed wireline networks, this bound reduces to the usual cut-set bound, but we caution that for undirected wireline networks the following bound can be better than standard bounds [108].

Let $X_{\mathcal{S}} = \{X_u : \ u \in \mathcal{S}\}$ and similarly for $Y_{\mathcal{S}}$. Consider any choice of encoders and decoders, and suppose we compute the per-clock-tick average joint distribution

$$P_{X_{\mathcal{N}}Y_{\mathcal{N}}}(a,b) = \left[\frac{1}{n}\sum_{i=1}^{n} P_{X_{\mathcal{N},i}}(a)\right] P_{Y_{\mathcal{N}}|X_{\mathcal{N}}}(b|a), \tag{3.23}$$

where $P_{X_{\mathcal{N},i}}(\cdot)$ is the marginal distribution of the channel inputs at time i. Let $\mathcal{M}(\mathcal{S})$ be the set of messages separated from one of their sinks by the cut $(\mathcal{S},\overline{\mathcal{S}})$. Then any rate-tuple in the capacity region must satisfy

$$\sum_{m\in\mathcal{M}(\mathcal{S})} R_m \le I\left(X_{\mathcal{S}}; Y_{\overline{\mathcal{S}}}|X_{\overline{\mathcal{S}}}\right), \tag{3.24}$$

where for discrete alphabets

$$I(X;Y|Z) = \sum_{a,b,c} P_{XYZ}(a,b,c)\ \log_2 \frac{P_{XY|Z}(a,b|c)}{P_{X|Z}(a|c)P_{Y|Z}(b|c)} \tag{3.25}$$

is the mutual information between X and Y conditioned on Z. The expression (3.25) generalizes to continuous alphabets in natural ways. Furthermore, if node u has the power constraint (3.18b), then we require

$$E\left[|X_u|^2\right] \le P_u. \tag{3.26}$$

Next, for a given input distribution $P_{X_{\mathcal{N}}}(\cdot)$ satisfying (3.26), let $\overline{\mathcal{R}}(\mathcal{S}, P_{X_{\mathcal{N}}})$ be the set of non-negative rate-tuples satisfying (3.24). We further define

$$\overline{\mathcal{R}}(P_{X_{\mathcal{N}}}) = \bigcap_{\mathcal{S}\subset\mathcal{N}} \overline{\mathcal{R}}(\mathcal{S}, P_{X_{\mathcal{N}}}). \tag{3.27}$$

The *cut-set bound* is the union of (3.27) over all input distributions, i.e., the cut-set region is

$$\overline{\mathcal{R}} = \bigcup_{P_{X_{\mathcal{N}}}(\cdot)} \overline{\mathcal{R}}(P_{X_{\mathcal{N}}}) = \bigcup_{P_{X_{\mathcal{N}}}(\cdot)} \bigcap_{\mathcal{S}\subset\mathcal{N}} \overline{\mathcal{R}}(\mathcal{S}, P_{X_{\mathcal{N}}}) \tag{3.28}$$

and $\overline{\mathcal{R}}$ is an outer bound on the capacity region.

A few remarks are in order. First, we emphasize that if $P_{X_{\mathcal{N}}}(\cdot)$ is fixed then (3.27) is itself an upper bound on the achievable rates. Second, one can show that $I(X_{\mathcal{S}};Y_{\overline{\mathcal{S}}}|X_{\overline{\mathcal{S}}})$ in (3.24) is concave in $P_{X_{\mathcal{N}}}(\cdot)$. Finding boundary points of (3.28) in terms of $P_{X_{\mathcal{N}}}(\cdot)$ is therefore a concave optimization problem because the intersection of concave functions is concave. Finally, we remark that $\overline{\mathcal{R}}$ is the best known capacity outer bound even for many small networks.

For example, consider the wireline network in Figure 3.4 but without the port constraints (3.2). Suppose the capacity of edge (u,v) is C_{uv} and nodes 1 and 2 have messages W_1 and W_2, respectively, destined for node 3. The cut-set bound is

$$\overline{\mathcal{R}} = \left\{ (R_1,R_2): \begin{array}{c} 0 \le R_1 \le C_{12}+C_{13} \\ 0 \le R_2 \le C_{21}+C_{23} \\ R_1+R_2 \le C_{13}+C_{23} \end{array} \right\}, \tag{3.29}$$

where the optimal input distribution has the X_{uv} being independent and uniform. It turns out that the cut-set bound is the capacity region in this case, and it is also the usual networking cut-set bound.

Consider next some networks with noise. For the DMC we have

$$\overline{\mathcal{R}} = \bigcup_{P_X(\cdot)} \{R: 0 \le R \le I(X;Y)\}, \tag{3.30}$$

which is consistent with (2.1). For the 2WC, we have

$$\overline{\mathcal{R}} = \bigcup_{P_{X_1X_2}(\cdot)} \left\{ (R_1,R_2): \begin{array}{c} 0 \le R_1 \le I(X_1;Y_2|X_2) \\ 0 \le R_2 \le I(X_2;Y_1|X_1) \end{array} \right\} \tag{3.31}$$

and it remains to optimize over the joint distribution $P_{X_1X_2}(\cdot)$. For the MAC with output Y, we obtain

$$\overline{\mathcal{R}} = \bigcup_{P_{X_1X_2}(\cdot)} \left\{ (R_1,R_2): \begin{array}{c} 0 \le R_1 \le I(X_1;Y|X_2) \\ 0 \le R_2 \le I(X_2;Y|X_1) \\ R_1+R_2 \le I(X_1X_2;Y) \end{array} \right\}. \tag{3.32}$$

We again must optimize over $P_{X_1X_2}(\cdot)$. For the RC, we have

$$\overline{\mathcal{R}} = \bigcup_{P_{X_1X_2}(\cdot)} \{R:\ 0 \le R \le \min\{I(X_1;Y_2Y_3|X_2),\ I(X_1X_2;Y_3)\}\}. \tag{3.33}$$

In fact, the cut-set bounds (3.31)–(3.33) are loose in general because optimizing over all $P_{X_1X_2}(\cdot)$ is too optimistic (although anticipated, the bound (3.33) was only recently shown to be loose for a class of RCs [6]). Nevertheless, the cut-set bound often provides a useful benchmark, and it sometimes even gives the capacity region. As a final remark, one can improve cut-set bounds by using an approach that progressively removes edges from the network graph [107, 109].

4

Cooperative Strategies and Rates

This chapter introduces several cooperative strategies. We will consider primarily wireless networks, and we further consider either no fading or fast fading. Slow fading channels will be treated in Chapter 5. For all cases, we consider only "CSIR, No CSIT" models where each node knows the channel gains between itself and the nodes with which it cooperates (see Section 3.3). However, before continuing with wireless issues, we first consider basics of wireline networks.

4.1 Wireline Strategies

Cooperative strategies for wireline networks exist on all layers of the protocol stack, primarily to coordinate packet flows. Our focus will be on strategies that use *coding* to combine bits, symbols, or packets to form new bits or packets. We further focus on *rate* rather than reliability or delay, not because rate is the most important parameter, but because the theory for rates is simpler and currently more developed. We consider three types of strategies: routing, network coding, and mode coding [4, 105], [5, p. 4].

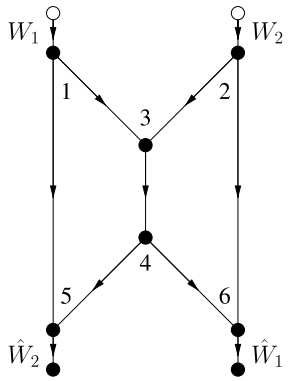


Fig. 4.1 A butterfly network.

4.1.1 Routing

Routing treats communication as path-based flows of bits or packets. Of course, routing does not use coding to combine bits or packets, although one might expand the definition to include copying (see, e.g., [26]). We review routing for the sake of comparing it with network coding in the next section.

Consider the "butterfly" network shown in Figure 4.1, where all edges are non-interfering point-to-point channels with unit capacity. Note that there are no self-loops so that there are no node constraints. Suppose we have two *unicast sessions*, i.e., two source-sink pairs, as shown in the figure. Clearly, there is exactly one *path* from node 1 to node 6, namely the sequence of nodes $(1,3,4,6)$ corresponding to a sequence of transmissions on the links $(1,3)$, $(3,4)$, and $(4,6)$. Routing assigns *flow* (rate) to every path so that no *coding* (combining of bits, symbols, or packets) is done. A bit received as an input to an intermediate node on a path is merely copied to an output link. One can easily check that routing achieves any rate pair $(R_1, R_2) = (1-\beta, \beta)$ for any $0 \leq \beta \leq 1$.

Consider next the *undirected* network shown on the left in Figure 4.2. The idea is that every edge can be used in *either* direction as long as the sum of rates (flow) in both directions is at most the capacity of this edge. In other words, every edge is a 2WC that has the

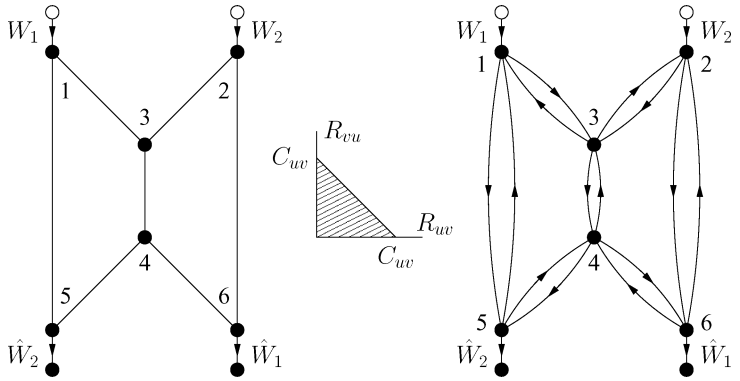


Fig. 4.2 An undirected butterfly network as a network of 2WCs.

capacity region shown in the middle of Figure 4.2. We thus redraw this graph as shown on the right in Figure 4.2. We further model every pair of edges between nodes u and v as Shannon's push-to-talk 2WC [167, Sec. 1] defined by

$$(Y_{uv}, Y_{vu}) = \begin{cases} (X_{uv}, Z_{vu}) & \text{if } X_{uv} \neq 0, X_{vu} = 0, \\ (Z_{uv}, X_{vu}) & \text{if } X_{uv} = 0, X_{vu} \neq 0, \\ (Z_{uv}, Z_{vu}) & \text{if } X_{uv} = X_{vu} = 0, \end{cases} \tag{4.1}$$

where (Z_{uv}, Z_{vu}) is uniform over $\mathcal{X}_{uv} \times \mathcal{X}_{vu}$, and $|\mathcal{X}_{uv}| = 2^{C_{uv}} + 1$ for all edges (u, v). The channel (4.1) has the capacity region shown in the middle of Figure 4.2.

Suppose that $C_{uv} = 1$ for all edges (u, v). Routing performs better than in Figure 4.1 because there are more paths available from each source to its sink. In fact, there are four paths from node 1 to node 6: path $(1, 3, 4, 6)$, path $(1, 3, 2, 6)$, path $(1, 5, 4, 6)$, and path $(1, 5, 4, 3, 2, 6)$. We can thus perform routing as shown in Figure 4.3 and achieve the rate pair $(R_1, R_2) = (1, 1)$. This pair defines the capacity region of this network, since the cut-set bound of Section 3.7.1 gives $R_1 \leq 1$, $R_2 \leq 1$. In fact, the maximum flow for two unicast sessions on any undirected graph is given by the minimum cut [80]. Note also that one can *separate* channel coding (PHY layer coding) and routing (network layer coding) to achieve the capacity for this special case. However, such separation of channel and network coding is not always optimal [152].

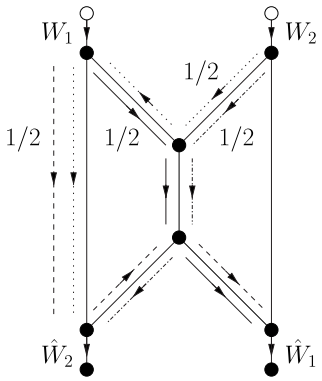


Fig. 4.3 An optimal routing for the undirected butterfly network.

4.1.2 Network Coding

We found that routing achieves a smaller rate region for the directed network of Figure 4.1 than for the undirected network in Figure 4.2. However, suppose that node 3 *combines* the packets arriving from nodes 1 and 2 by XORing them bit-wise, shown as $X_1 \oplus X_2$ in Figure 4.4. This combining of raw bits or packets is called *network coding* [4]. More specifically, it is called *linear* network coding if the combining operations are done over a (finite) field [101, 116]. Node 5 collects X_1 and $X_1 \oplus X_2$ and computes X_2. Similarly, node 6 collects X_2 and $X_1 \oplus X_2$ and computes X_1. Thus, network coding achieves the capacity point

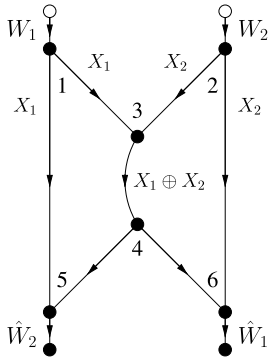


Fig. 4.4 Network coding for the directed butterfly network.

$(R_1, R_2) = (1, 1)$ even for the directed network of Figure 4.1, and thus outperforms routing by a factor of two in rate. Furthermore, network coding gives an additional method for achieving capacity on the undirected network of Figure 4.2.

We remark that the rate gains of network coding over routing are the largest for a *multicast* session, i.e., there is one source and many sinks. For example, the graph in Figure 4.5, which is a simple modification of the graph in Figure 4.1, shows that $R = 2$ is feasible with network coding while only $R = 1$ is feasible with routing.

We further remark that the codes in Figures 4.4 and 4.5 were specially chosen by using our knowledge of the network graph. In general, however, the individual nodes must operate without this information (see Table 1.1). An attractive coding method is then to use *random* network coding [4] or random linear network coding [73]. For the latter approach, consider node u and let $\mathcal{E}_{\text{in}}(u)$ and $\mathcal{E}_{\text{out}}(u)$ be its sets of incoming and outgoing edges, respectively. Suppose that every packet has $L = nb$ data bits for some integers n and b. Coding at node u proceeds as follows:

(a) for every pair $(e, f) \in \mathcal{E}_{\text{in}}(u) \times \mathcal{E}_{\text{out}}(u)$ choose a $w_{e,f}$ randomly from the Galois field $\text{GF}(2^b)$;

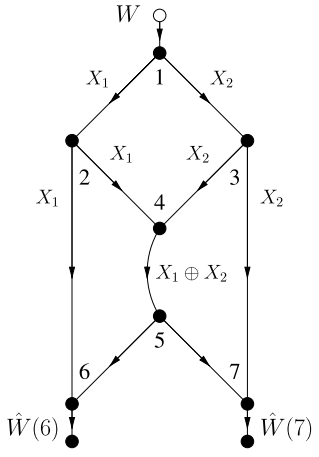


Fig. 4.5 Network coding for a multicast network.

(b) collect one packet from each incoming edge e and strip off its headers to form the data string x_e^L;
(c) parse x_e^L into n pieces $x_e^b(i)$, $i = 1, 2, \ldots, n$, having b bits each;
(d) for every outgoing edge f compute

$$\sum_{e \in \mathcal{E}_{\text{in}}(u)} w_{e,f} \cdot x_e^b(i) \tag{4.2}$$

for $i = 1, 2, \ldots, n$, where the multiplications and additions are over the Galois field GF(2^b);
(e) collect the n pieces corresponding to outgoing edge f into x_f^L;
(f) augment x_f^L with a header that includes the $w_{e,f}$, $e \in \mathcal{E}_{\text{in}}(u)$, to inform the decoders of how the packets were combined;
(g) transmit the resulting packet on edge f.

Random linear network coding is known to work well if the size b of the pieces is sufficiently large and if the weights $w_{e,f}$ are chosen independently and uniformly over GF(2^b). For example, for practical reasons we would like to use $b = 1$ so that the weights $w_{e,f}$ are in GF(2) but then one cannot guarantee that random network coding will work. In this case we must add redundancy, e.g., by using coding at the sources in the form of LT-codes or Raptor codes [128, 168]. The combination of coding at the sources and network coding can provide good rate versus reliability tradeoffs. Many other results on network coding can be found in a recent special issue of the IEEE Transactions on Information Theory devoted in part to this topic [171].

4.1.3 Mode Coding

There are further rate gains possible in wireline networks beyond network coding [105]. To show this, we consider the wireline network in Figure 3.3 with the node constraint (3.1) and $Y_3 = X_2$. The graph for this network is shown in Figure 4.6. Suppose for simplicity that the alphabets $\mathcal{X}_1, \mathcal{X}_2, \mathcal{Y}_2, \mathcal{Y}_3$ of the respective random variables X_1, X_2, Y_2, Y_3 are all the set $\{0, 1\}$. One might guess that the capacity is 1/2 bit/clock tick because the relay can either receive or transmit, but not both.

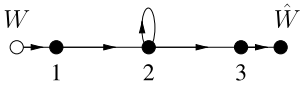


Fig. 4.6 A network graph with a relay device constraint.

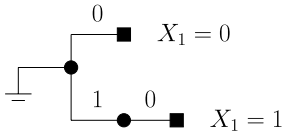


Fig. 4.7 A code tree for the relay in Figure 4.6.

Consider, however, the code tree depicted in Figure 4.7 that is labeled with the symbols X_2 that the relay transmits after having decoded an X_1. The idea is that after decoding $X_1 = 0$ the relay sends $X_2 = 0$, while after decoding $X_1 = 1$ the relay uses the channel twice to send the pair of symbols $X_{2,1} = 1$ and $X_{2,2} = 0$. Note that every codeword in the code tree has exactly one 0, so that the source can transmit exactly one new message bit for every relay codeword. Note further that the tree is labeled so that the code is *prefix-free* or *instantaneously decodable* [34, Ch. 5]. This means that the sink can correctly parse its received sequence to extract the message bits.

For example, suppose the message w has 8 bits $0,1,0,0,1,1,1,0$. The transmit and receive sequences are then

$$\begin{array}{rccccccccc} i = & 1, & 2, & 3,4, & 5, & 6, & 7,8, & 9,10, & 11,12, & 13 \\ X_1^n = & 0, & 1, & x,0, & 0, & 1, & x,1, & x,1, & x,0, & x \\ Y_2^n = & 0, & 1, & 0,0, & 0, & 1, & 0,1, & 0,1, & 0,0, & x \\ X_2^n = & 0, & 0, & 1,0, & 0, & 0, & 1,0, & 1,0, & 1,0, & 0, \end{array} \tag{4.3}$$

where x denotes a "do not care" symbol. Note that, for this particular message, we have transmitted 8 bits in $n = 13$ clock ticks which gives a rate $R = 8/13$ that is larger than $1/2$.

We would like to determine the rate of the above code when the source transmits b bits for large b. The codewords have variable lengths, so we define a random variable L_2 to be the length of the codeword of any of the message bits. If the source bits are coin-tossing, we compute

$E[L_2] = \frac{1}{2}(1) + \frac{1}{2}(2) = \frac{3}{2}$ and

$$R = \frac{1}{E[L_2]} = 2/3 \text{ bits/clock tick} \tag{4.4}$$

implying that we can substantially increase the rate beyond 1/2 bit/clock tick. In fact, it turns out that better compression codes such as Huffman codes or arithmetic source codes (see [34, Ch. 5]) can achieve $R = 0.773$ bits/clock tick. Furthermore, this rate is the capacity of the network, as one can check by computing the cut-set bound.

We call the above method *mode coding* because one is effectively transmitting information through the choice of the relay's operating mode (listen or talk).[1] We will later consider this type of coding in more detail for wireless relays. Note that the relative gains of mode coding are small when the alphabets $\mathcal{X}_1, \mathcal{X}_2, \mathcal{Y}_2, \mathcal{Y}_3$ are large. In fact, mode coding boosts the rate by at most 1 bit/packet over an edge time-sharing strategy. But packets often have several thousands of bits. Furthermore, mode coding requires rapid on/off switching by the relay. On the other hand, this coding might be useful for *covert* communication where, e.g., a router time-jitters transmissions according to a signature pattern.

4.2 Wireless Strategies

The previous section shows how network coding combines raw bits or packets at the network layer to improve rates. More generally, *cooperative coding* combines *symbols* at the physical and higher layers to produce new symbols. We consider several types of cooperative coding strategies, including

(1) amplify-and-forward (AF)
(2) classic multi-hop
(3) compress-and-forward (CF)
(4) decode-and-forward (DF)
(5) multipath decode-and-forward (MDF)
(6) DF or MDF with network coding.

[1] Mode coding seems related to sending "bits through queues" [7].

The above strategies can be used for both wireline and wireless networks and they require progressively more coordination. For example, consider a RC. AF and classic multi-hop do not necessarily require changes at the source or destination nodes, e.g., for multi-hop the relay can behave as if it is the destination or the source. CF does not necessarily require changes at the source but it does require some extra knowledge about the link capacities. DF requires changes at both the source and destination, and MDF requires additional changes at higher layers of the protocol stack. DF or MDF with network coding require even more changes at higher layers. Some of these issues are discussed in the sections below. However, we begin by specifying the models used in this section.

4.2.1 Idealized Wireless Models

In this section, we will describe and compare wireless strategies by using two idealizations:

- we consider full-duplex radios, i.e., devices that can transmit and listen at the same time in the same frequency band;
- if the channels are time-varying, we assume per-link channel state information is available at the receiver but not at the transmitter (CSIR, No CSIT, see Section 3.4).

We defer consideration of duplexing constraints to Section 4.3. Note that full-duplex operation is often not possible due to large difference in transmit and receive power[2] (see Section 3.3). However, full-duplex models have didactic utility in that their capacities can sometimes be written in closed-form. Furthermore, the information theory developed for the full-duplex models applies directly to wireline problems and, perhaps surprisingly, to half-duplex models as well [103, 106].

The basic cooperative model is the relay channel shown in Figure 3.5 [33, 106]. In our derivations, we consider a general geometry shown in Figure 4.8(a) with nodes u and v at distance d_{uv}. For numerical evaluations, we employ the linear geometry shown in Figure 4.8(b),

[2] Some devices can operate in full-duplex mode with good echo cancellation.

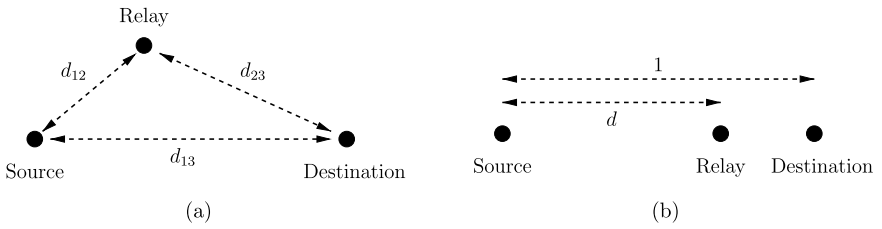


Fig. 4.8 Relay channel geometries: (a) In general, nodes u and v are separated by distance d_{uv}; (b) In the linear case, the source-destination distance is $d_{13} = 1$, the source-relay distance is $d_{12} = |d|$ and the relay-destination distance is $d_{23} = |1 - d|$.

where the source and destination nodes are a distance $d_{13} = 1$ apart, and the relay is a distance $d_{12} = |d|$ to the right of the source, and a distance $d_{23} = |1 - d|$ to the left of the destination. A negative d means that the relay is to the left of the source. We remark that we are here considering distances d_{uv} that are less than one, in seeming disregard of our claim in Section. 2.2.1 that we are interested in long-range communication where the model (2.3) is accurate. However, we are in effect normalizing the source–destination distance to be unity, and including long-range attenuation in the power constraints. For example, when we later find in Section 4.2.6 that DF achieves capacity when the relay is near the source, the distance need not be less than one but rather shows where the relay must be relative to the source and destination.

We consider channels that exhibit one of three kinds of fading: (1) no fading; (2) fast uniform-phase fading; and (3) fast Rayleigh fading (see Section 3.2). That is, for every clock tick we have the complex channels (see (3.3a) and (3.3b))

$$Y_2 = \frac{H_{12}}{|d|^{\alpha/2}} X_1 + Z_2, \tag{4.5a}$$

$$Y_3 = H_{13} X_1 + \frac{H_{23}}{|1-d|^{\alpha/2}} X_2 + Z_3. \tag{4.5b}$$

For no fading, the H_{uv} are constants. For uniform-phase fading, the H_{uv} are independent and uniform over $\{e^{j\phi} : \phi \in [0, 2\pi)\}$. For Rayleigh fading, the H_{uv} are independent and Gaussian with zero mean and unit variance.

For CSIR, we assume that all nodes know the distances d_{uv}, but for uniform-phase and Rayleigh fading H_{uv} is known only to node v, i.e., only the relay knows H_{12}, and only the destination knows H_{13} and H_{23}. As in Section 3.4, we model this by augmenting Y_2 and Y_3 to become the vectors $[Y_2\ H_{12}]$ and $[Y_3\ H_{13}\ H_{23}]$, respectively. As we will see, channels whose phase varies rapidly exhibit a special property that we exploit to derive capacity theorems. Furthermore, we shall later see that making the relay half-duplex brings both complications and simplifications to our theory and strategies.

4.2.2 Amplify-and-Forward

Amplify-and-forward has the relay transmit

$$X_{2,i} = a_{2,i} Y_{2,i-1}, \tag{4.6}$$

where the $a_{2,i}$, $i = 1, 2, \ldots, n$, are chosen to satisfy the relay's power constraint [59, 111, 161]. More generally, the relay might transmit some function of a small number of the past received symbols, e.g.,

$$X_{2,i} = \underline{a}_{2,i} \cdot [Y_{2,i-1}\ Y_{2,i-2}\ \ldots\ Y_{2,i-D}]^T, \tag{4.7}$$

where the $\underline{a}_{2,i}$, $i = 1, 2, \ldots, n$, are row vectors chosen to satisfy the relay's power constraint. Consider (4.6) where $a_{2,i} = a$ for all i so that we have (see (4.5b))

$$\begin{aligned} Y_{3,i} &= \frac{H_{13,i}}{d_{13}^{\alpha/2}} X_{1,i} + \frac{H_{23,i}}{d_{23}^{\alpha/2}} X_{2,i} + Z_{3,i}, && (4.8)\\ &= \frac{H_{13,i}}{d_{13}^{\alpha/2}} X_{1,i} + a \frac{H_{12,i-1} H_{23,i}}{d_{12}^{\alpha/2} d_{23}^{\alpha/2}} X_{1,i-1} + a \frac{H_{23,i}}{d_{23}^{\alpha/2}} Z_{2,i-1} + Z_{3,i}. && (4.9) \end{aligned}$$

To satisfy the power constraint, we require

$$|a|^2 \le \frac{P_2}{N + P_1 E[|H_{12}|^2]/d_{12}^{\alpha}}. \tag{4.10}$$

Observe that, without fading, (4.9) is an AWGN channel with unit-memory intersymbol interference (ISI). In general, one should therefore perform a waterfilling optimization of the spectrum of X_1^n (cf. Section 2.2.5, Section 3.4, and [106, Sec. VII.B]). The result is shown in

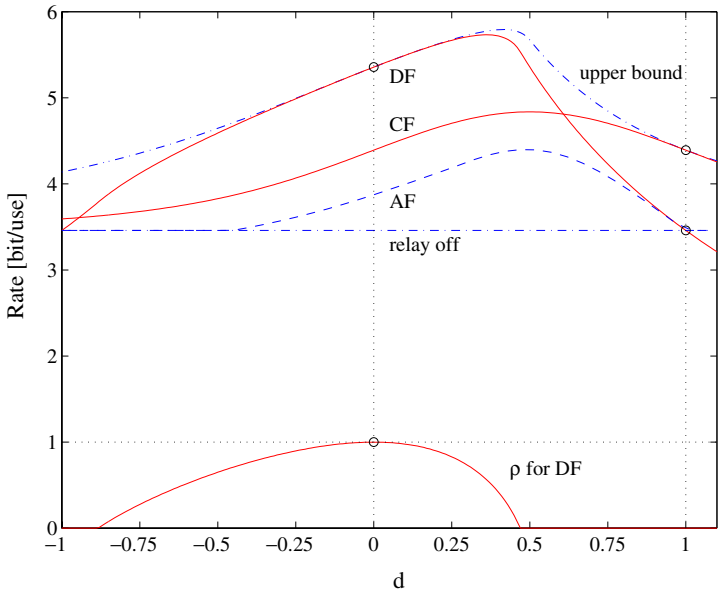


Fig. 4.9 Rates for a full-duplex relay, $P_1/N = P_2/N = 10$, $H_{uv} = 1$ for all (u, v), and $\alpha = 2$.

Figure 4.9 for the linear geometry of Figure 4.8(b) as the curve labeled "AF," where $P_1/N = P_2/N = 10$, $H_{uv} = 1$ for all (u, v) and $\alpha = 2$. The relay-off rate is simply $\log_2(11)$. The curve labeled "upper bound" is the cut-set bound (3.33). The other curves are based on the CF and DF strategies developed further below. Note that AF always performs as well as using no relay by choosing $a = 0$. Note also that increasing a from 0 increases the destination's ISI and noise power. The relay should thus not always transmit with maximum power.

The AF rates in Figure 4.9 are often significantly worse than the rates achieved by other strategies. However, AF can sometimes perform very well [59]. For example, consider the RC with T relays[3] shown in Figure 4.10. Suppose every relay u, $u = 2, 3, \ldots, T + 1$, has

$$\sum_{i=1}^{n} E\left[|X_{u,i}|^2\right]/n \leq P \tag{4.11}$$

[3] Note that T is here the number of relays rather than time.

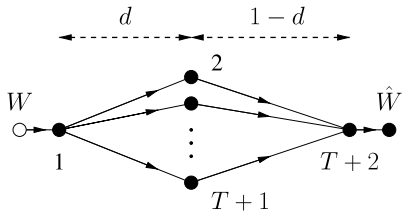


Fig. 4.10 A multi-relay channel and geometry.

and $E[|Z_u|^2] = N$. If $0 < d < 1$, $H_{uv} = 1$ for all (u, v), and every relay uses AF as in (4.6) with $a_{u,i} = a$, then we have

$$X_{u,i} = aY_{u,i-1} \tag{4.12}$$

and the T-relay version of (4.9) is

$$Y_{T+2,i} = \frac{Ta}{|d(1-d)|^{\alpha/2}} X_{1,i-1} + Z_{T+2,i} + \sum_{u=2}^{T+1} \frac{a}{|1-d|^{\alpha/2}} Z_{u,i-1}. \tag{4.13}$$

Note that there is no channel from node 1 to node $T + 2$. Equation (4.13) thus defines an AWGN channel with SNR

$$\gamma = \frac{T^2|a|^2 P_1}{N\left[|d(1-d)|^\alpha + T|a|^2|d|^\alpha\right]}. \tag{4.14}$$

But the cut bound (3.24) in Section 3.7.1 with $\mathcal{S} = \{1, 2, \ldots, T+1\}$ gives

$$C \le \log_2\left(1 + \frac{T^2 P}{N}\right). \tag{4.15}$$

We thus have

$$\begin{aligned} &\log_2(T) + \log_2\left(\frac{P_1}{N} \cdot \frac{|a|^2}{1+|a|^2}\right) \\ &\qquad < C \le 2\log_2(T) + \log_2\left(\frac{1}{T^2} + \frac{P}{N}\right) \end{aligned} \tag{4.16}$$

so that C grows as $\kappa \log_2(T)$ with T, where $1 \le \kappa \le 2$. Furthermore, AF achieves this scaling law up to a constant factor.

The above scenario permits the total system power to grow linearly with T. Instead, one might require $P = P_{\text{sum}}/T$ for some constant P_{sum}. We choose (see (4.10))

$$|a|^2 = \frac{P_{\text{sum}}/T}{N + P_1/|d|^\alpha}. \tag{4.17}$$

Inserting (4.17) into (4.14) and (4.15), we find that

$$\begin{aligned} \log_2(T) + \log_2\left(\frac{P_{\text{sum}}}{N} \cdot \frac{P_1}{N + P_1/|d|^\alpha + P_{\text{sum}}}\right) \\ < C \leq \log_2(T) + \log_2\left(\frac{1}{T} + \frac{P_{\text{sum}}}{N}\right). \end{aligned} \tag{4.18}$$

Now C grows as $\log_2(T)$ with T and AF achieves this scaling law up to a rate offset. We remark that both the scaling laws (4.16) and (4.18) require coherent combining of signals at the destination.

4.2.3 Classic Multi-Hop

Classic multi-hop has the source transmitting its message W to the relay in one-time slot, and then the relay forwarding W to the destination in a second-time slot. This scheme can be used with both full-duplex and half-duplex relays. Suppose we assign the time fractions τ to the first hop and $\bar{\tau} = 1 - \tau$ to the second hop. For constant H_{12} and H_{23}, we achieve

$$R = \min\left[\tau \log_2\left(1 + \frac{P_1|H_{12}|^2}{\tau d_{12}^\alpha N}\right), \bar{\tau}\log_2\left(1 + \frac{P_2|H_{23}|^2}{\bar{\tau} d_{23}^\alpha N}\right)\right]. \tag{4.19}$$

However, even after optimizing τ one always performs worse than using no relay for any d in Figure 4.9. This happens because $\alpha = 2$ is too small to make multi-hopping useful. We will see later that multi-hop can work well for half-duplex relays and larger α.

4.2.4 Compress-and-Forward

We continue to consider the no-fading case for simplicity. CF is a block-transmission strategy with the block structure shown in Figure 4.11 [33]. There are three codebooks: $\underline{x}_1(\cdot)$, $\underline{x}_2(\cdot)$, and a quantization codebook $\hat{\underline{y}}_2(\cdot)$. The source transmits $\underline{x}_1(w_b)$ indexed by a new

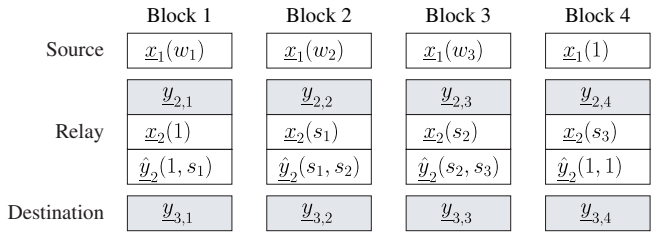


Fig. 4.11 A CF strategy for a full-duplex relay.

message w_b in every block b. The relay observes $\underline{y}_{2,b}$ in block b and chooses its next codeword as follows: $\underline{y}_{2,b}$ is quantized to $\hat{\underline{y}}_2(s_{b-1}, s_b)$ and the index s_b chooses the codeword $\underline{x}_2(s_b)$. The destination first decodes s_b using $\underline{y}_{3,b+1}$ and then decodes w_b by using $\hat{\underline{y}}_2(s_{b-1}, s_b)$ and $\underline{y}_{3,b}$. For the last block, the source transmits a default codeword $\underline{x}_1(1)$.

Consider, for example, the following relay encoding in block b. After canceling the effect of $\underline{x}_2(s_{b-1})$ on $\underline{y}_{2,b}$, the relay uses a vector quantizer to map $\underline{y}_{2,b}$ to $\hat{\underline{y}}_2(s_{b-1}, s_b)$ so that the average distortion (using some per-letter distortion function) between these two vectors is at most D. Shannon's rate-distortion theory [34, Ch. 13] tells us that the rate of the codebook $\hat{\underline{y}}_2(\cdot)$ need to be at most

$$R_Q(D) = I(Y_2; \hat{Y}_2 | X_2). \tag{4.20}$$

The relay can transmit s_b reliably to the destination in block $b+1$ via $\underline{x}_2(s_b)$ as long as

$$R_Q(D) \le I(X_2; Y_3). \tag{4.21}$$

The destination decodes s_b and now knows $\underline{x}_2(s_b)$ and $\hat{\underline{y}}_2(s_{b-1}, s_b)$. It further knows $\underline{x}_2(s_{b-1})$ from its decoding in block b and can cancel the effect of this vector on $\hat{\underline{y}}_2(s_{b-1}, s_b)$ and $\underline{y}_{3,b}$. Finally, the destination decodes w_b using $\hat{\underline{y}}_2(s_{b-1}, s_b)$ and $\underline{y}_{3,b}$ at the rate

$$R = I(X_1; \hat{Y}_2 Y_3 | X_2). \tag{4.22}$$

In fact, one can improve the rate specified by (4.20)–(4.22) by using a more sophisticated vector quantizer and destination decoder [33].

The idea is that the relay transmits a hash $h(s_{b-1})$ of the string s_{b-1} in block b, i.e., the relay sends $\underline{x}_2(h(s_{b-1}))$ and finds a quantization vector $\hat{\underline{y}}_2(h(s_{b-1}), s_b)$ in block b. The hashing is also known as *binning* and $h(s_{b-1})$ is called a bin index. The result is that any rate R satisfying

$$R = I(X_1; \hat{Y}_2 Y_3 | X_2) \tag{4.23a}$$

subject to

$$I(\hat{Y}_2; Y_2 | X_2 Y_3) \leq I(X_2; Y_3) \tag{4.23b}$$

is achievable where the joint probability distribution of the random variables factors as

$$P(x_1, x_2, y_2, y_3, \hat{y}_2) = P(x_1)P(x_2)P(y_2, y_3|x_1, x_2)P(\hat{y}_2|x_2, y_2). \tag{4.24}$$

As done in (4.24), we will write $P_X(x)$ as $P(x)$ if the argument of $P(\cdot)$ is a lower-case version of the random variable symbol. Note that the only change from (4.20)–(4.22) is that the quantization codebook has a lower rate than in (4.20), namely $I(\hat{Y}_2; Y_2|X_2Y_3)$. The factorization (4.24) reflects that the $\underline{x}_1(w_b)$ and $\underline{x}_2(h(s_{b-1}))$ codewords in block b are chosen independently, and that $\hat{\underline{y}}_2(h(s_{b-1}), s_b)$ is chosen as a function of $\underline{x}_2(h(s_{b-1}))$ and $\underline{y}_{2,b}$.

A few remarks are in order. First, note that the rate (4.23) is the rate of a point-to-point channel where the receiver has a vector output $[\hat{Y}_2\ Y_3]$. The CF strategy is thus closely related to multi-antenna reception (see [57, 106]). In fact, if the relay-to-destination channel capacity is very large (say, if the relay is close to the destination) then the quantization code book can be very large and $\hat{Y}_2$ is "almost" Y_2. The CF strategy will then work well. Second, it is interesting that the maximization of R in (4.23) turns out to be equivalent to the maximization of (see [45])

$$R = \min\Big\{I(X_1; \hat{Y}_2 Y_3|X_2), I(X_1X_2; Y_3) - I(\hat{Y}_2; Y_2|X_1X_2)\Big\} \tag{4.25}$$

subject to (4.24). Third, one can time-share several modes of operation in (4.23) and (4.24), i.e., the source and relay use the respective distributions $P_{X_1|Q}(\cdot|q)$ and $P_{X_2|Q}(\cdot|q)$ for a fraction $P(q)$ of the time, where q represents one of a finite number of modes. Suppose that the source

and relay cycle through these modes several times. The relay can then collect its quantization bits over all modes, and transmit them at the average rate $I(X_2;Y_3|Q)$ to the destination. The result is that any rate satisfying

$$R = I(X_1;\hat{Y}_2 Y_3|X_2 Q) \tag{4.26a}$$

subject to

$$I(\hat{Y}_2;Y_2|X_2 Y_3 Q) \leq I(X_2;Y_3|Q) \tag{4.26b}$$

is achievable where the joint probability distribution of the random variables factors as

$$P(q)P(x_1|q)P(x_2|q)P(y_2,y_3|x_1,x_2)P(\hat{y}_2|x_2,y_2,q). \tag{4.27}$$

The above rate sometimes improves the basic CF rate in (4.23) [45]. In fact, one can similarly add a *time-sharing* random variable Q to all the strategies in this chapter, but we will not consider this explicitly. Finally, note that the relay needs to know the statistics of Y_3 to compute its compression and channel coding rates in (4.23b). The relay might thus need to know $\{h_{13,i}\}_{i=1}^n$ and $\{h_{23,i}\}_{i=1}^n$.

As a numerical example, suppose there is no fading. We choose X_1 and X_2 to be the usual Gaussian distributions, and $\hat{Y}_2 = Y_2 + \hat{Z}_2$ where $\hat{Z}_2 = \hat{Z}_{2R} + j\hat{Z}_{2I}$ and $\hat{Z}_{2R}$ and $\hat{Z}_{2I}$ are independent, Gaussian random variables with variance $\hat{N}_2/2$ (this choice of $\hat{Y}_2$ is not necessarily optimal). The resulting information rates are

$$R = \log_2\left(1 + \frac{P_1|H_{12}|^2}{d_{12}^{\alpha}(N+\hat{N}_2)} + \frac{P_1|H_{13}|^2}{d_{13}^{\alpha}N}\right), \tag{4.28a}$$

$$I(\hat{Y}_2;Y_2|X_2Y_3) = \log_2\left(1 + \frac{N\left(\frac{P_1|H_{12}|^2}{d_{12}^{\alpha}} + \frac{P_1|H_{13}|^2}{d_{13}^{\alpha}} + N\right)}{\hat{N}_2\left(\frac{P_1|H_{13}|^2}{d_{13}^{\alpha}} + N\right)}\right), \tag{4.28b}$$

$$I(X_2;Y_3) = \log_2\left(1 + \frac{\frac{P_2|H_{23}|^2}{d_{23}^{\alpha}}}{\frac{P_1|H_{13}|^2}{d_{13}^{\alpha}} + N}\right). \tag{4.28c}$$

The choice

$$\hat{N}_2 = N \cdot \frac{P_1|H_{12}|^2/d_{12}^{\alpha} + P_1|H_{13}|^2/d_{13}^{\alpha} + N}{P_2|H_{23}|^2/d_{23}^{\alpha}} \tag{4.29}$$

thus satisfies (4.23b) with equality (see also [57] and [79, Sec. 3.2]). Consider again the case $P_1/N = P_2/N = 10$, $H_{uv} = 1$ for all (u, v), $\alpha = 2$, and the linear geometry of Figure 4.8(b). The resulting CF rates are shown in Figure 4.9 as the curve labeled "CF." Observe that CF performs well when the relay is close to the destination ($d \approx 1$) and even meets the cut-set bound (3.33) when $d = 1$. Furthermore, CF significantly outperforms AF everywhere.

4.2.5 Decode-and-Forward

Decode-and-forward is a block-transmission strategy that has the block structure shown in Figure 4.12 [33, 106]. There are two code books: $\underline{x}_1(\cdot)$ and $\underline{x}_2(\cdot)$. The source uses *block Markov* encoding, i.e., the encoding has one block memory in that the source codeword in block b is $\underline{x}_1(w_{b-1}, w_b)$. The idea is that the relay knows w_{b-1} and decodes w_b after having removed the effect of $\underline{x}_2(w_{b-1})$. The resulting rate bound is

$$R \le I(X_1; Y_2|X_2). \tag{4.30}$$

The relay can now transmit $\underline{x}_2(w_b)$ in block $b+1$. The destination decodes w_b with a sliding-window decoder that uses $\underline{y}_{3,b}$ and $\underline{y}_{3,b+1}$ (cf. [28, 191, 106]). If the codebooks of every block in Figure 4.12 were chosen independently, the resulting destination rate bound is

$$R \le I(X_1; Y_3|X_2) + I(X_2; Y_3) = I(X_1X_2; Y_3). \tag{4.31}$$

The term $I(X_1; Y_3|X_2)$ in (4.31) comes from $\underline{y}_{3,b}$ (note that the destination already knows w_{b-1} and thus $\underline{x}_2(w_{b-1})$) while the term $I(X_2; Y_3)$ comes from $\underline{y}_{3,b+1}$ ($\underline{x}_1(w_b, w_{b+1})$ is treated as interference).

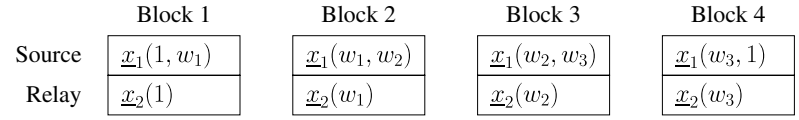


Fig. 4.12 A DF strategy for a full-duplex relay.

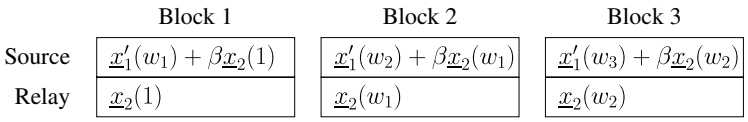


Fig. 4.13 A DF strategy for a full-duplex relay on an AWGN relay channel.

Summarizing, DF achieves the rate

$$R = \max_{P_{X_1X_2}(\cdot)} \min\left\{I(X_1;Y_2|X_2),\, I(X_1X_2;Y_3)\right\}. \tag{4.32}$$

The reader might wonder why one can optimize over all *joint* distributions $P_{X_1X_2}(\cdot)$. One can permit this by constructing the codebook $\underline{x}_1(\cdot)$ from $\underline{x}_2(\cdot)$ via *superposition* coding [32]. To explain this method, consider a full-duplex Gaussian relay channel with the DF block structure shown in Figure 4.13. The idea is that the code book $\underline{x}_1(\cdot)$ is constructed by adding (or superposing) codewords from a Gaussian codebook $\underline{x}'_1(\cdot)$ to codewords from a Gaussian codebook $\underline{x}_2(\cdot)$ scaled by β. The codewords in $\underline{x}_2(\cdot)$ use power P_2. The codewords in $\underline{x}'_1(\cdot)$ use power P'_1, where $P'_1 \le P_1$, and the scaled codewords from $\underline{x}_2(\cdot)$ use power $P_1 - P'_1$, i.e., we have $\beta = \sqrt{(P_1 - P'_1)/P_2}$. Furthermore, the codeword $\underline{x}'_1(w_b)$ in block b is independent of the codeword $\beta\underline{x}_2(w_{b-1})$, so the source power constraint is satisfied.

The relay decodes w_b at the rate $\log_2(1 + P'_1|H_{12}|^2/(d_{12}^\alpha N))$. The destination decodes w_b by using $\underline{y}_{3,b}$, $\underline{y}_{3,b+1}$, and its past estimate of w_{b-1}. The codeword $\underline{x}'_1(w_{b+1})$ is treated as interference. The resulting DF rate is

$$\begin{aligned} R = \max_{\rho} \min\Bigg\{ & \log_2\left(1 + \frac{P_1|H_{12}|^2(1-|\rho|^2)}{d_{12}^\alpha N}\right), \\ & \log_2\left(1 + \frac{P_1|H_{13}|^2}{d_{13}^\alpha N} + \frac{P_2|H_{23}|^2}{d_{23}^\alpha N} + \frac{2\sqrt{P_1P_2}\Re\{\rho H_{13}H_{23}^*\}}{d_{13}^{\alpha/2}d_{23}^{\alpha/2}N}\right)\Bigg\}, \end{aligned} \tag{4.33}$$

where we recall that $\Re\{x\}$ is the real part of x, and where $\rho = E[X_1X_2^*]/\sqrt{P_1P_2}$ is the *correlation coefficient* of the zero-mean X_1 and X_2, which in this case is $\rho = \sqrt{1 - P'_1/P_1}$. Note that, by phase-rotating the relay signal appropriately, the source and relay can *coherently*

combine their signals to give the destination an SNR boost. The price paid is a loss in the information rate from the source to the relay.

We make a few remarks. First, DF differs from classic multi-hop in several important ways:

(1) the source and relay transmit simultaneously
(2) the source and relay signals can be made to coherently combine at the destination if h_{13} and h_{23} are known to the transmitter or relay
(3) the destination decodes using several or all of its available output blocks.

These differences remain for half-duplex relays. However, the second difference disappears for fast uniform-phase fading or fast Rayleigh fading, as we shall see. Next, note that the last term in (4.32) is the rate of a point-to-point channel where the transmitter has a vector input $[X_1\ X_2]$. The DF strategy is thus closely related to multi-antenna transmission. In fact, if the source-to-relay channel capacity is large (say, if the relay is close to the source) then the limiting term in (4.32) is $I(X_1X_2;Y_3)$.

Consider again the case $P_1/N = P_2/N = 10$, $H_{uv} = 1$ for all (u, v), $\alpha = 2$, and the linear geometry of Figure 4.8(b). The resulting DF rates are shown in Figure 4.9 as the curve labeled "DF." The curve labeled "ρ for DF" shows the optimal correlation coefficient. Observe that DF performs well when the relay is close to the source ($d \approx 0$) and even meets the cut-set bound (3.33) when $d = 0$. Based on Figure 4.9, one would choose DF if the relay is close to the source, and CF if the relay is close to the destination. This insight applies to other networks as well. For example, suppose we have two relays. If both relays are close to the source or the destination, then they should both use DF or CF, respectively. On the other hand, if one relay is close to the source and the other is close to the destination, then the former relay should use DF while the latter should use CF.

4.2.6 CF and DF for Fast Uniform-Phase Fading

Consider the geometry of Figure 4.8(a) and where the H_{uv} are fast fading. As noted in Section 4.2.1, we augment Y_2 and Y_3 to become

$[Y_2\ H_{12}]$ and $[Y_3\ H_{13}\ H_{23}]$, respectively. From (4.23), the CF rate is thus

$$R = I(X_1; \hat{Y}_2 Y_3 | X_2 H_{13} H_{23}) \tag{4.34a}$$

subject to

$$I(\hat{Y}_2; Y_2 H_{12} | X_2 Y_3 H_{13} H_{23}) \leq I(X_2; Y_3 | H_{13} H_{23}), \tag{4.34b}$$

where

$$\begin{aligned} &P(\underline{h}, x_1, x_2, y_2, y_3, \hat{y}_2) \\ &\quad = P(\underline{h}) P(x_1) P(x_2) P(y_2, y_3 | x_1, x_2, \underline{h}) P(\hat{y}_2 | x_2, y_2, h_{12}) \end{aligned} \tag{4.35}$$

and where $\underline{h} = [h_{12}\ h_{13}\ h_{23}]$. For uniform-phase fading, we have $H_{uv} = e^{j\Phi_{uv}}$ and choose $\hat{Y}_2 = Y_2 e^{-j\Phi_{12}} + \hat{Z}_2$, where $\hat{Z}_2$, is the same as in Section 4.2.4. The result is that we recover the same rate as in (4.28a)–(4.29) (cf. [106]).

On the other hand, the DF rate is now

$$R = \max_{P_{X_1 X_2}(\cdot)} \min \{ I(X_1; Y_2 | X_2 H_{12}),\ I(X_1 X_2; Y_3 | H_{13} H_{23}) \}. \tag{4.36}$$

For the AWGN channel, this rate is

$$\begin{aligned} R = \max_{\rho} \min \Bigg\{ &\log_2 \left(1 + \frac{P_1 (1 - \rho^2)}{d_{12}^{\alpha} N} \right), \\ &E \left[\log_2 \left(1 + \frac{P_1}{d_{13}^{\alpha} N} + \frac{P_2}{d_{23}^{\alpha} N} + \frac{2 \rho e^{j(\Phi_{12} - \Phi_{13})} \sqrt{P_1 P_2}}{d_{12}^{\alpha/2} d_{13}^{\alpha/2} N} \right) \right] \Bigg\}, \end{aligned} \tag{4.37}$$

where the expectation is over the Φ_{12} and Φ_{13} that are independent and uniform over $[0, 2\pi)$. But Jensen's inequality gives

$$E[\log_2(X)] \leq \log_2(E[X]) \tag{4.38}$$

and we have $E\left[e^{j(\Phi_{12} - \Phi_{13})}\right] = 0$. The best ρ in (4.37) is therefore zero [79, 106]. This result is intuitive: without phase knowledge the source and relay cannot coherently combine their signals at the destination. The DF block structure thus simplifies to that shown in Figure 4.14.

	Block 1	Block 2	Block 3	Block 4
Source	$\underline{x}_1(w_1)$	$\underline{x}_1(w_2)$	$\underline{x}_1(w_3)$	$\underline{x}_1(w_4)$
Relay	$\underline{x}_2(1)$	$\underline{x}_2(w_1)$	$\underline{x}_2(w_2)$	$\underline{x}_2(w_3)$

Fig. 4.14 A DF strategy for a full-duplex relay and $\rho = 0$.

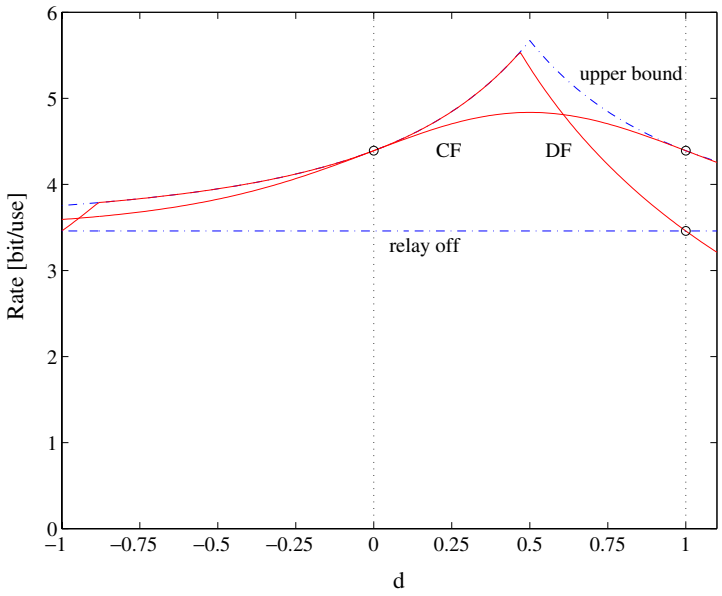


Fig. 4.15 Rates for a full-duplex relay, uniform-phase fading, $P_1/N = P_2/N = 10$, and $\alpha = 2$.

Figure 4.15 plots the CF, DF, and cut-set rates for the linear geometry shown in Figure 4.8(b) and the uniform-phase fading channel with $P_1/N = P_2/N = 10$ and $\alpha = 2$. Perhaps surprisingly, DF achieves capacity when the relay is in a region around the source [106]. Figure 4.16 plots the two-dimensional positions of the relay where DF achieves capacity when $P_1 = P_2$. Note that the region grows as α increases. The results in Figures 4.15 and 4.16 generalize to other fast fading channels, e.g., Rayleigh fading channels, and it reinforces the insight from Section 4.2.5 that one should use DF when the relay is near the source. We remind the reader that the distances in

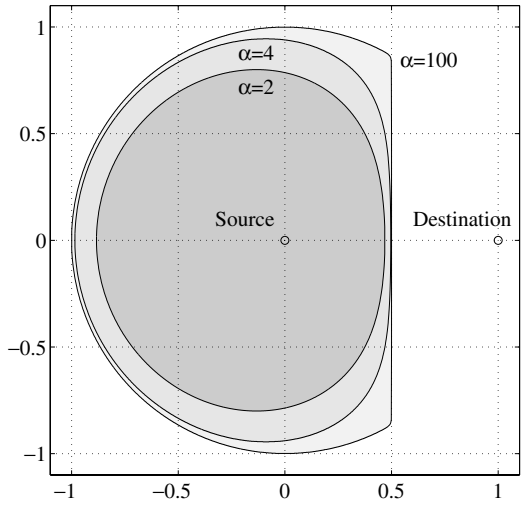


Fig. 4.16 Positions of the relay where DF achieves capacity with uniform-phase fading and $P_1 = P_2$.

Figures 4.15 and 4.16 should be interpreted as relative distances and not as absolute ones.

4.2.7 Multipath Decode-and-Forward

Multipath decode-and-forward is a variation of DF where the source node first performs *rate-splitting*, i.e., the message W with rate R is split into W' and W'' having the respective rates R' and R'' such that $R' + R'' = R$ [33, 44, 106]. The same type of rate-splitting is done for *multipath routing* in Figure 4.3. In fact, MDF generalizes multipath routing from wireline networks to wireless networks. Alternatively, one could say that MDF combines multipath routing and DF.

The MDF block structure is shown in Figure 4.17. The idea is that one replaces X_1 in DF with an *auxiliary* random variable U, i.e., one performs DF with U playing the role of X_1. The U represents codewords $\underline{u}(w'_{b-1}, w'_b)$ that transmit W', as shown in Figure 4.17. The "path" for W' can thus be considered to be the entire RC [65]. On the other hand, W'' is transmitted by the codewords $\underline{x}_1(w'_{b-1}, w'_b, w''_b)$ that are superposed on $\underline{u}(w'_{b-1}, w'_b)$ and $x_2(w'_{b-1})$. The "path" for W'' can thus be considered to be the direct link from the source to the relay.

	Block 1	Block 2	Block 3	Block 4
Source	$\underline{u}(1, w_1')$	$\underline{u}(w_1', w_2')$	$\underline{u}(w_2', w_3')$	$\underline{u}(w_3', 1)$
Relay	$\underline{x}_2(1)$	$\underline{x}_2(w_1')$	$\underline{x}_2(w_2')$	$\underline{x}_2(w_3')$
Source	$\underline{x}_1(1, w_1' \circ w_1'')$	$\underline{x}_1(w_1', w_2', w_2'')$	$\underline{x}_1(w_2', w_3', w_3'')$	$\underline{x}_1(w_3', 1, w_4'')$

Fig. 4.17 An MDF strategy for a full-duplex relay.

Suppose the destination first decodes the messages w_b' and then the w_b''. Based on the above discussion, the rate of MDF is a sum of a DF rate and a point-to-point rate:

$$R' = \min\{I(U;Y_2|X_2),\, I(UX_2;Y_3)\}, \tag{4.39a}$$

$$R'' = I(X_1;Y_3|UX_2), \tag{4.39b}$$

where we can choose any $P(u, x_1, x_2)$ and where we consider no fading for simplicity. The best MDF rate $R = R' + R''$ is therefore

$$R = \max_{P_{UX_1X_2}(\cdot)} \min\{I(U;Y_2|X_2) + I(X_1;Y_3|UX_2), I(X_1X_2;Y_3)\}. \tag{4.40}$$

The MDF superposition method is shown for AWGN channels in Figure 4.18 that generalizes Figure 4.13. The codebook $\underline{x}_1(\cdot)$ is constructed by adding codewords from the Gaussian codebooks $\underline{u}'(\cdot)$, $\beta\underline{x}_2(\cdot)$, and $\underline{u}''(\cdot)$. The codewords in $\underline{x}_2(\cdot)$ use power P_2. The codewords in $\underline{u}'(\cdot)$ and $\underline{u}''(\cdot)$ have powers P_1' and P_1'', respectively, where $P_1' + P_1'' \le P_1$. We must therefore choose $\beta = \sqrt{(P_1 - P_1' - P_1'')/P_2}$. The decoding procedure is as follows. The relay decodes w_b' after receiving $\underline{y}_{2,b}$. The destination decodes w_b' after receiving $\underline{y}_{3,b}$ and $\underline{y}_{3,b+1}$ by assuming that its past estimate of w_{b-1}' is correct, and by treating

	Block 1	Block 2
Source	$\underline{u}'(w_1') + \beta\underline{x}_2(1) + \underline{u}''(w_1'')$	$\underline{u}'(w_2') + \beta\underline{x}_2(w_1') + \underline{u}''(w_2'')$
Relay	$\underline{x}_2(1)$	$\underline{x}_2(w_1')$

Fig. 4.18 An MDF strategy for a full-duplex relay on an AWGN relay channel.

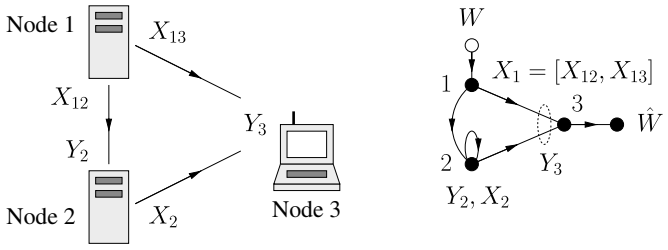


Fig. 4.19 A mixed wireline/wireless relay channel and its graph.

$\underline{u}'(w'_{b+1})$ and $\underline{u}''(w''_{b+1})$ as interference. Finally, the destination removes the effect of $\underline{u}'(w'_b)$ from $\underline{y}_{3,b}$ and decodes w''_b.

Unfortunately, for full-duplex AWGN channels the MDF method does not improve the DF method because the optimal choice of $P_{UX_1X_2}(\cdot)$ gives either DF or a point-to-point strategy. However, for other AWGN channels the MDF strategy performs better than DF. For example, we shall later see that MDF improves on DF for half-duplex channels. MDF even achieves capacity for an important class of channels discussed next.

Suppose we have the mixed network shown in Figure 4.19 where the source-to-relay channel is a wireline channel. We model this by using $X_1 = [X_{12}\ X_{13}]$ and the channel

$$P(y_2, y_3|x_1, x_2) = P(y_2|x_{12}, x_2)P(y_3|x_{13}, x_2), \tag{4.41}$$

where we permit X_2 to contribute to Y_2, i.e., the channel (4.41) includes port constraints such as (3.1). For the MDF rate (4.40), we choose $U = X_{12}$ and

$$P(x_{12}, x_{13}, x_2) = P(x_2)P(x_{12}|x_2)P(x_{13}|x_2). \tag{4.42}$$

We compute

$$R = \max \min \{I(X_{12}; Y_2|X_2) + I(X_{13}; Y_3|X_2), I(X_{13}X_2; Y_3)\}, \tag{4.43}$$

where the maximization is over all distributions of the form (4.42). One can check that the cut-set bound (3.33) also gives (4.43) because the best input distributions factor as (4.42). The MDF rate (4.43) is therefore the capacity of the mixed relay channel [47].

We remark that this channel also models cases where nodes 1 and 2 are wireless and the source-to-relay channel is "orthogonal" to the destination channel. For example, this happens if the source-to-relay channel uses a different frequency band than the destination channel.

4.2.8 Decode-and-Forward with Network Coding

Decode-and-forward can be combined with network coding in several ways, and we develop such strategies for two classes of wireless channels in this and the next section. Consider first the network shown in Figure 4.20 and suppose the channels are defined by

$$Y_1 = \frac{H_{31}}{d_{13}^{\alpha/2}} X_3 + Z_1, \tag{4.44a}$$

$$Y_2 = \frac{H_{32}}{d_{23}^{\alpha/2}} X_3 + Z_2, \tag{4.44b}$$

$$Y_3 = \frac{H_{13}}{d_{13}^{\alpha/2}} X_1 + \frac{H_{23}}{d_{23}^{\alpha/2}} X_2 + Z_3, \tag{4.44c}$$

where the H_{uv} and Z_u are the usual complex, Gaussian, random variables. Note that this model orthogonalizes the MAC from nodes 1 and 2 to node 3, and the BC from node 3 to nodes 1 and 2. For example, this could happen if nodes 1 and 2 are mobile stations and node 3 is a base station, and if one uses FDM for the uplink (MAC) and downlink (BC). AF, CF, and DF strategies for such channels have been analyzed in [100, 149, 150].

Suppose there is no fading with $H_{uv} = d_{uv} = 1$ for all (u, v), the noise variances are zero, and the alphabets of the X_u, $u = 1, 2, 3$, are all now binary. It is then easy to see that the best coding strategy is

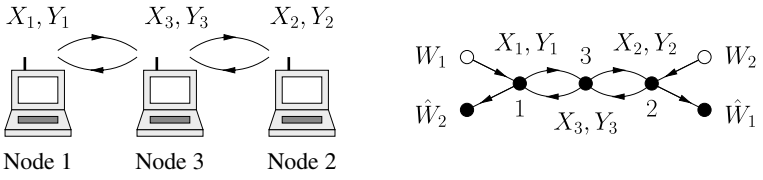


Fig. 4.20 A 3-node wireless network and its graph.

to use $X_{3,i} = X_{1,i-1} \oplus X_{2,i-1}$ for all i. The resulting capacity is the set of (R_1, R_2) satisfying $0 \le R_1 \le 1$ and $0 \le R_2 \le 1$ (see, e.g., [190, Figure 1]).

Suppose next that there is fast uniform-phase fading with noise and complex alphabets as in (4.44a)–(4.44c). There are two natural coding approaches. The first is similar to the binary case just described above: node 3 superposes codeword sequences for W_1 and W_2 over the modulo-lattice additive channels (MLACs) described in [48, 49] to satisfy its power constraint.[4] The second approach is to simply encode the pair (W_1, W_2) using a codebook with $2^{n(R_1+R_2)}$ codewords $\underline{x}_3(w_1, w_2)$ generated by $P_{X_3}(\cdot)$. The latter method is directly related to broadcasting when receivers have side information [179] and should prove useful when there are non-uniform rates.[5]

For example, suppose we choose X_1 and X_2 as independent Gaussian random variables. The MAC bounds are

$$R_1 \le \log_2\left(1 + \frac{P_1}{d_{13}^{\alpha} N}\right), \tag{4.45a}$$

$$R_2 \le \log_2\left(1 + \frac{P_2}{d_{23}^{\alpha} N}\right), \tag{4.45b}$$

$$R_1 + R_2 \le \log_2\left(1 + \frac{P_1}{d_{13}^{\alpha} N} + \frac{P_2}{d_{23}^{\alpha} N}\right). \tag{4.45c}$$

The BC rate bounds when using an MLAC (or a common codebook $\underline{x}_3(\cdot)$ as described above) to encode at node 3 are

$$R_1 \le \log_2\left(1 + \frac{P_3}{d_{23}^{\alpha} N}\right), \tag{4.46a}$$

$$R_2 \le \log_2\left(1 + \frac{P_3}{d_{13}^{\alpha} N}\right). \tag{4.46b}$$

In fact, the bounds (4.45a)–(4.46b) give the capacity region of our network if the sum of (4.46a) and (4.46b) implies (4.45c). One can verify this by using the cut-set bound: the bound (3.24) with

[4] We are grateful to S. Shamai for suggesting this approach.

[5] This approach was developed together with S. Shamai. The approach was also suggested to us by E. Telatar, and independently appeared in [143] while this text was being completed. The idea was further developed independently by L.-L. Xie.

$\mathcal{S} = \{1\},\{2\},\{1,3\},\{2,3\}$ gives (4.45a), (4.45b), (4.46a), and (4.46b), respectively.

4.2.9 Multipath Decode-and-Forward with Network Coding

Consider next the MAC-DR shown in Figure 4.21, where the X_{13} and X_{23} links are noise-free. Suppose X_{13} and X_{23} are binary while

$$Y_4 = \frac{H_{14}}{d_{14}^{\alpha/2}} X_{14} + \frac{H_{24}}{d_{24}^{\alpha/2}} X_{24} + \frac{H_{34}}{d_{34}^{\alpha/2}} X_3 + Z_4, \tag{4.47}$$

where H_{14}, H_{24}, H_{34}, and Z_4 are the usual complex, Gaussian, random variables. This model orthogonalizes the MAC from nodes 1 and 2 to node 3, and the MAC from nodes 1, 2, and 3 to node 4.

One approach to coding is to apply existing MAC-DR strategies such as AF, CF, and DF [156, 157, 158, 159, 160]. Instead, for variety we here wish to combine MDF strategies with network coding. Suppose nodes 1 and 2 split their messages as $W_u = [W'_u\, W''_u]$ so that $R_u = R'_u + R''_u$, $u = 1,2$. We further choose $R'_1 = R'_2 = 1$ so that node 3 can XOR the W'_1 and W'_2 bits it decodes and map them to a codeword in the codebook $\underline{x}_3(\cdot)$.

Node 1 associates its W'_1 bits with an auxiliary random variable U_1, and then superposes its W''_1 bits onto U_1 via X_1 as in Section 4.2.7. Node 2 uses similar steps. Node 4 first decodes W'_1 and W'_2 jointly, strips off their effect on Y_4, and then decodes the W''_1 and W''_2. The MAC rate

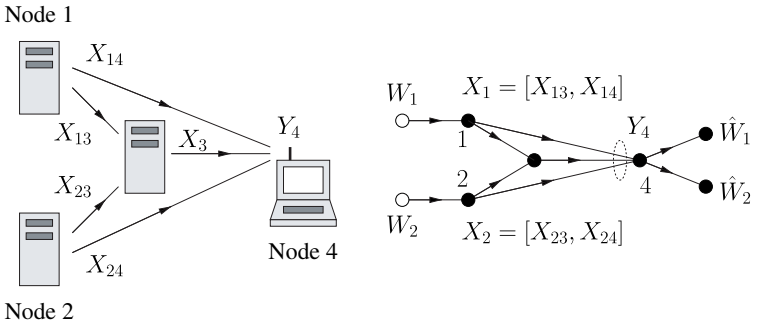


Fig. 4.21 A mixed wireline/wireless MAC-DR and its graph.

bounds for decoding W_1' and W_2' at node 4 are

$$R_1' \leq I(U_1 X_3; Y_4 | U_2 H), \tag{4.48a}$$

$$R_2' \leq I(U_2 X_3; Y_4 | U_1 H), \tag{4.48b}$$

$$R_1' + R_2' \leq I(U_1 U_2; Y_4 | X_3 H), \tag{4.48c}$$

where $[U_1\ X_{14}]$, $[U_2\ X_{24}]$, and X_3 are independent, and where $H = [H_{14}\ H_{24}\ H_{34}]$. The bounds (4.48) are computed by considering the different error events that can occur at node 4. The MAC rate bounds for decoding W_1'' and W_2'' at node 4 are

$$R_1'' \leq I(X_{14}; Y_4 | U_1 U_2 X_3 H), \tag{4.49a}$$

$$R_2'' \leq I(X_{24}; Y_4 | U_1 U_2 X_3 H), \tag{4.49b}$$

$$R_1'' + R_2'' \leq I(X_{14} X_{24}; Y_4 | U_1 U_2 X_3 H). \tag{4.49c}$$

Combining the above rate bounds, we get an achievable region. However, it is not clear that this region gives capacity points.

4.3 Half-Duplex Strategies and Mode Modulation

The information theory developed in the previous sections applies to both full-duplex and half-duplex devices [106]. In this section, we consider the theory in more detail for half-duplex devices. For example, we demonstrate that half-duplex nodes can transmit data by modulating their "listen" to "talk" modes [103]. However, despite their fundamental nature, one might not want to harness such *mode modulation* gains for reasons outlined below.

Consider a half-duplex RC with no fading. A DF strategy is depicted in Figure 4.22 where the relay listens for some fraction of the time and talks for the other fraction of the time.[6] Furthermore, all nodes know ahead of time when the relay listens and talks. We include the relay's operating modes explicitly in the model as follows [103]. Let M_2 be a *mode* random variable that takes on the values $M_2 = L$ and $M_2 = T$ if the relay listens and talks, respectively. One can check that the DF

[6] Observe that the relay decodes all message blocks w_b.

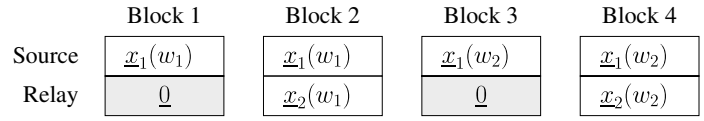


Fig. 4.22 A DF strategy for a half-duplex relay channel.

strategy in Figure 4.22 achieves the rate

$$R = \max_{P_{X_1X_2M_2}(\cdot)} \min\{I(X_1;Y_2|X_2M_2),\ I(X_1X_2;Y_3|M_2)\}. \tag{4.50}$$

However, suppose we re-define the channel as $P(y_2, y_3|x_1, x_2, m_2)$, where the relay's channel input is $X_2' = [X_2\ M_2]$. We use X_2' in place of X_2 in (4.32) and find that the DF rate is

$$R = \max_{P_{X_1X_2M_2}(\cdot)} \min\{I(X_1;Y_2|X_2M_2),\ I(X_1X_2M_2;Y_3)\}. \tag{4.51}$$

Expanding the second term on the right in (4.51), we have

$$I(X_1X_2M_2;Y_3) = I(M_2;Y_3) + I(X_1X_2;Y_3|M_2) \tag{4.52}$$

and one can check that $I(M_2;Y_3) > 0$ in general. Thus, a strategy with pre-assigned slots such as in Figure 4.22 is generally suboptimal.

The reason for the rate gain $I(M_2;Y_3)$ in (4.52) is that the relay sends information to the destination through its choice of operating mode. In fact, this is how we achieve the wireline network capacity in Section 4.1.3. Mode modulation further improves all the strategies considered in Section 4.2, i.e., AF, classic multi-hop, CF, and DF or MDF with network coding. At the same time, there are some challenges.

(1) $I(M_2;Y_3)$ is limited to 1 bit/clock tick because M_2 is binary. Thus, mode modulation cannot help much for channels with high capacity, e.g., wireline packet networks with large packets (see Section 4.1.3) or AWGN channels with high SNR.
(2) The rate gain requires M_2 to change rapidly. However, wireline or wireless nodes cannot always switch from "listen" to "talk" modes rapidly.

(3) A remedy for the switching problem is to constrain the M_2^n sequence to change slowly, e.g., via run-length limited codes. However, this limits the rate gain even further.

The above reasons, as well as other considerations, usually lead one to choose strategies as in Figure 4.22. We will do the same below (see also [76, 98]). First, however, we show that mode modulation can give rate gains [103].

Consider the general relay geometry of Figure 4.8(a) and suppose there is no fading with $H_{uv} = 1$ for all (u, v) and $\alpha = 4$. Suppose further that we have per-symbol power-constraints $P_1 = P_2 = 4$, and that the relay must listen at least half the time to guarantee cooperation, i.e., $P_{M_2}(L) \geq 0.5$. The DF rates (4.51), where the relay turns on and off randomly, are shown in Figure 4.23 as the curve labeled "DF, random." The DF rates (4.50), where the relay turns on and off at fixed time-intervals, are shown as the curve labeled "DF, fixed." The upper dash-dotted curve is the cut-set bound when using strategies as in Figure 4.23, and it is computed as (4.50) but with Y_2 replaced by

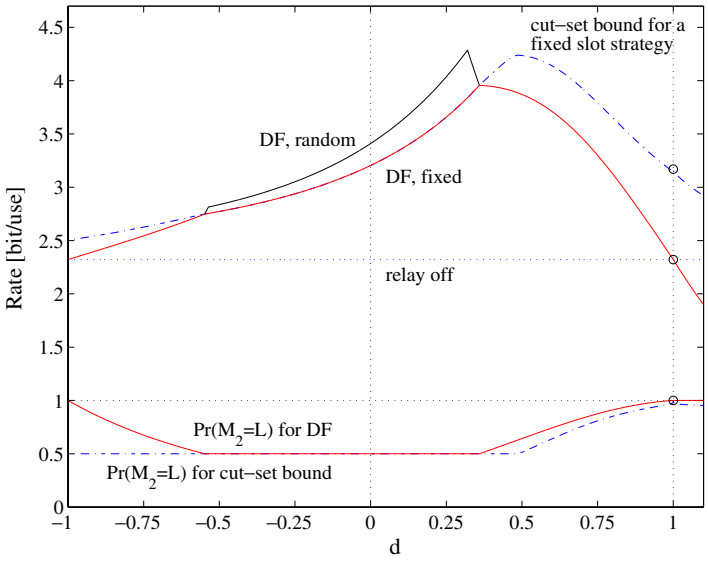


Fig. 4.23 Rates for half-duplex strategies with $P_1 = P_2 = 4$, $H_{uv} = 1$ for all (u, v), and $\alpha = 4$.

$[Y_2\, Y_3]$. Note that mode modulation outperforms any type of strategy with pre-assigned slots. The lower curves in Figure 4.23 show the optimizing $P_{M_2}(L)$ for both (4.50) and the cut-set bound.

4.4 Multi-Antenna Relaying

The information theory developed above applies to devices equipped with multiple antennas in a similar fashion as in Section 2.2.5. We again make the channel inputs and outputs vectors and write

$$\underline{Y}_2 = \frac{H_{12}}{d_{12}^{\alpha/2}}\underline{X}_1 + \underline{Z}_2, \tag{4.53a}$$

$$\underline{Y}_3 = \frac{H_{13}}{d_{13}^{\alpha/2}}\underline{X}_1 + \frac{H_{23}}{d_{23}^{\alpha/2}}\underline{X}_2 + \underline{Z}_3, \tag{4.53b}$$

where $\underline{X}_u$, $u = 1, 2$, and $\underline{Y}_u$ and $\underline{Z}_u$, $u = 2, 3$, are complex column vectors of length n_u, and H_{uv} is a complex $n_u \times n_v$ fading matrix. The $\underline{Z}_u$ have independent, Gaussian, variance N entries with the usual form. The matrix H_{uv} is independent of $\underline{X}_u$, $u = 1, 2$, $\underline{Z}_u$, $u = 2, 3$, and all other fading matrices. Rayleigh fading has H_{uv} that have independent, Gaussian, zero-mean, unit variance entries with the usual form. We consider the general geometry of Figure 4.8(a).

The block power constraints are (see (2.26))

$$\sum_{i=1}^{n} E[\|\underline{X}_{u,i}\|^2]/n \le P_u, \quad u = 1, 2. \tag{4.54}$$

However, we will here use the per-symbol constraints

$$E[\|\underline{X}_{u,i}\|^2] \le P_u, \quad u = 1, 2, \quad i = 1, 2, \ldots, n \tag{4.55}$$

to avoid power control optimization.

Consider first the point-to-point channel (4.53a) where node 1 sends a message to node 2. The best $\underline{X}_1$ are zero-mean and Gaussian by the maximum entropy theorem [34, p. 234]; recall that we write the covariance matrix of $\underline{X}_1$ as $Q_{\underline{X}_1}$. Since $Q_{\underline{X}_1}$ is Hermitian, we can write $Q_{\underline{X}_1} = U\Lambda U^\dagger$ where U is unitary, Λ is diagonal, and the eigenvalues of $Q_{\underline{X}_1}$ are on the diagonal of Λ [75, p. 171]. In particular, the trace of Λ

is the same as the trace of $Q_{\underline{X}}$. The capacity is (see (3.10) and (2.28))

$$C = \max_{p_{\underline{X}_1}(\cdot)} I(\underline{X}_1; \underline{Y}_2 H_{12}), \tag{4.56}$$

$$= \max_{p_{\underline{X}_1}(\cdot)} I(\underline{X}_1; \underline{Y}_2 \,|\, H_{12}), \tag{4.57}$$

$$= \max_{Q_{\underline{X}_1}} \int_h p_{H_{12}}(h) \log \left| I + \frac{1}{d_{12}^{\alpha} N} (hU) \Lambda (hU)^{\dagger} \right| dh, \tag{4.58}$$

where, as usual, the maximizations are interpreted as constraining $p_{\underline{X}_1}(\cdot)$ to satisfy the power constraints (4.55).

Note that $H_{12}U$ has the same distribution as H_{12} [174, Lemma 5]. We can therefore restrict attention to diagonal $Q_{\underline{X}_1}$, i.e., $\underline{X}_1$ that have independent entries. Finally, recall that $\log|A|$ is strictly concave on the convex set of positive definite Hermitian matrices A [75, p. 466]. Furthermore, we can permute the entries of Λ without changing the integral in (4.58). Averaging over all permutations and using the concavity of $\log|A|$, we find that the best $\underline{X}_1$ has independent entries that each have variance P_1/n_1. We thus have

$$C = \int_h p_{H_{12}}(h) \log \left| I + \frac{(P_1/n_1)}{d_{12}^{\alpha} N} hh^{\dagger} \right| dh. \tag{4.59}$$

We now return to our RC and model the half-duplex constraint as usual with

$$\underline{Y}_2 = \begin{cases} \dfrac{H_{12}}{d_{12}^{\alpha/2}} \underline{X}_1 + \underline{Z}_2 & \text{if } \underline{X}_2 = \underline{0}, \\ \underline{0} & \text{if } \underline{X}_2 \neq \underline{0}. \end{cases} \tag{4.60}$$

Alternatively, as in Section 4.3, we introduce a mode M_2 that takes on the values L and T. The MDF rate (4.40) is then

$$R = \max_{P_{\underline{U}\,\underline{X}_1\underline{X}_2 M_2}} \min\{ I(\underline{U}; \underline{Y}_2 | \underline{X}_2 M_2) + I(\underline{X}_1; \underline{Y}_3 | \underline{U}\,\underline{X}_2 M_2), \\ I(\underline{X}_1 \underline{X}_2; \underline{Y}_3 | M_2) + I(M_2; \underline{Y}_3) \}, \tag{4.61}$$

where $\underline{U}$ is a column vector of length n_1, and where we have implicitly augmented the $\underline{Y}_2$ and $\underline{Y}_3$ with channel gains H_{uv} as in Section 4.2.1 and Section 4.2.6. However, we will ignore the expression $I(M_2; \underline{Y}_3)$ by using pre-assigned slots, as discussed in Section 4.3.

Let $\underline{V}$ be a column vector of length n_1 and let I be an appropriately sized identity matrix. We choose $\underline{U}$, $\underline{V}$, and $\underline{X}_2$ to be independent, complex, Gaussian, zero-mean, and having covariance matrices $(1-\beta(M_2))(P_1/n_1)I$, $\beta(M_2)(P_1/n_1)I$, and $(P_2/n_2)I$, respectively, where $0 \le \beta(M_2) \le 1$ (note that (4.55) prevents using power control across modes). We further choose $\underline{X}_1 = \underline{U} + \underline{V}$. The resulting expressions in (4.61) with the model defined by (4.53b) and (4.60) are

$$I(\underline{U};\underline{Y}_2|\underline{X}_2, M_2 = L)$$
$$= \int_h p(h) \log \left| I + \frac{P_1}{d_{12}^{\alpha} n_1 N} hh^{\dagger} \right| \cdot \left| I + \frac{\beta(L) P_1}{d_{12}^{\alpha} n_1 N} hh^{\dagger} \right|^{-1} dh, \tag{4.62a}$$

$$I(\underline{X}_1;\underline{Y}_3|\underline{U}\,\underline{X}_2, M_2 = m_2) = \int_h p(h) \log \left| I + \frac{\beta(m_2) P_1}{d_{13}^{\alpha} n_1 N} hh^{\dagger} \right| dh, \tag{4.62b}$$

$$I(\underline{X}_1\underline{X}_2;\underline{Y}_3|M_2 = L) = \int_h p(h) \log \left| 1 + \frac{P_1}{d_{13}^{\alpha} n_1 N} hh^{\dagger} \right| dh, \tag{4.62c}$$

$$I(\underline{X}_1\underline{X}_2;\underline{Y}_3|M_2 = T)$$
$$= \int_{h,\tilde{h}} p(h)p(\tilde{h}) \log \left| I + \frac{P_1}{d_{13}^{\alpha} n_1 N} hh^{\dagger} + \frac{P_2}{d_{23}^{\alpha} n_2 N} \tilde{h}\tilde{h}^{\dagger} \right| dh\, d\tilde{h}, \tag{4.62d}$$

where the $p(h)$ and $p(\tilde{h})$ are matrix fading distributions. Note that for $d_{12} \le d_{13}$ it is best to choose $\beta(L) = 0$ and $\beta(T) = 1$. Moreover, this distribution is basically the same as using the MDF strategy depicted in Figure 4.24 where $\underline{X}_1$ has the same distribution irrespective of M_2. It therefore remains to optimize $P_{M_2}(\cdot)$. In fact, we shall avoid this optimization and consider only $P_{M_2}(L) = P_{M_2}(T) = 1/2$.

So suppose we use the MDF strategy of Figure 4.24.[7] We refer to the RC as $n_1 \times n_2 \times n_3$ based on the number of device antennas.

	Block 1	Block 2	Block 3	Block 4
Source	$\underline{x}_1(w_1)$	$\underline{x}_1(w_2)$	$\underline{x}_1(w_3)$	$\underline{x}_1(w_4)$
Relay	$\underline{0}$	$\underline{x}_2(w_1)$	$\underline{0}$	$\underline{x}_2(w_3)$

Fig. 4.24 An MDF strategy for a half-duplex relay channel.

[7] Observe that the relay decodes only the message blocks w_b with odd indexes b.

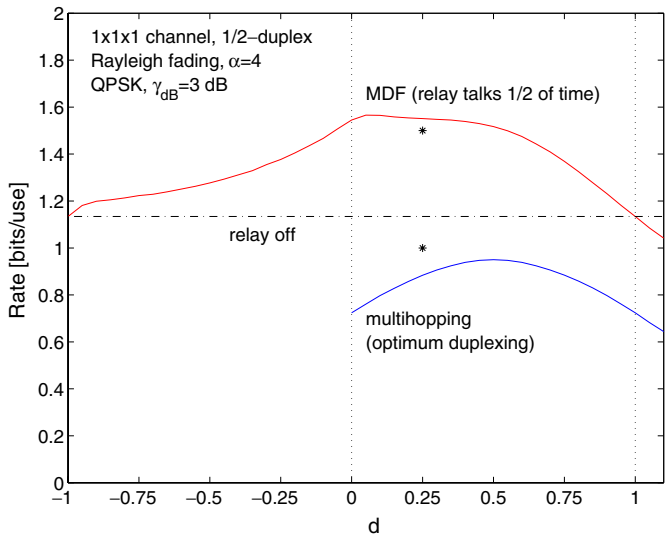


Fig. 4.25 MDF rates for a $1 \times 1 \times 1$ setup.

We consider two cases with QPSK modulation and Rayleigh fading [105, 104].

- A $1 \times 1 \times 1$ setup with $P_1/N = P_2/N = 2$ (or $\gamma_{dB} = 3$ dB). The MDF rates are shown in Figure 4.25 as a function of d. Also shown are the no-relay rate ($R \approx 1.13$ bits/use) and the traditional multi-hopping rates with optimized listen and transmit times. Observe that MDF achieves substantial rate gains over both no-relay transmission and traditional multi-hopping. For instance, the points marked with $*$ in Figure 4.25 are $(d, R) = (0.25, 1.0)$ and $(d, R) = (0.25, 1.5)$. Note that the multi-hopping curve is well below the "relay off" curve, and that the MDF curve is flat near $d = 0.25$. This happens because the source-to-relay link capacity is almost saturated at the maximum QPSK rate of 2 bits/use. One should therefore use a larger modulation signal set, e.g. 8-PSK, for the odd-numbered blocks in Figure 4.24.
- A $1 \times 1 \times 2$ setup with $P_1/N = P_2/N = 0.25$ (or $\gamma_{dB} = -6$ dB). The MDF rates are shown in Figure 4.26. The

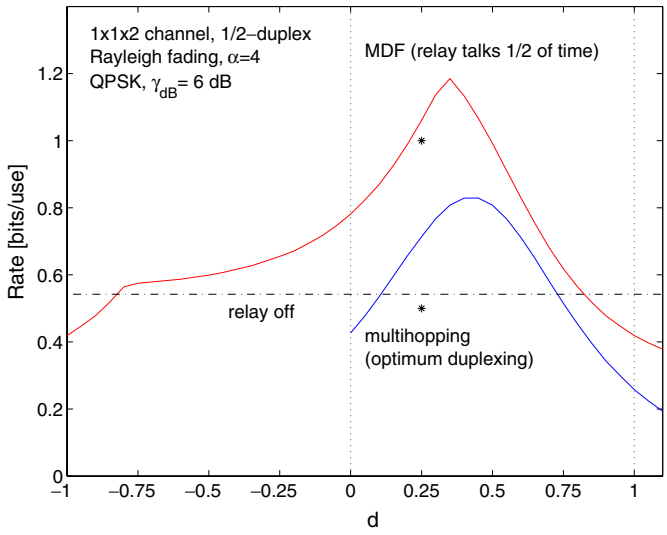


Fig. 4.26 MDF rates for a $1 \times 1 \times 2$ setup.

figure also shows the no-relay rate ($R \approx 0.54$ bits/use) and the traditional multi-hopping rates with optimized listen and transmit times. The points marked with $*$ in Figure 4.26 are $(d, R) = (0.25, 0.5)$ and $(d, R) = (0.25, 1)$.

4.5 Code Design for Relaying

We outline a code construction for half-duplex relays and for the Rayleigh fading channels of Section 4.4. The design is based on a point-to-point multi-antenna strategy known as Diagonal Bell Labs Layered Space-Time (D-BLAST) [53]. Other code constructions for relay channels are described in, e.g., [83, 200] (convolutional codes), [111] (space–time codes), [198] (turbo codes), [97, 104, 153, 50] (low-density parity-check (LDPC) codes).

We use two different codes: one for each of the two messages in the first two blocks in Figure 4.24. The relay must decode w_1 after having received only the first block of outputs from the source. We therefore require $R_c < 1/2$ for the w_1 encoder since half of the symbols

will be erased (we assume the even- and odd-numbered blocks in Figure 4.24 have the same length; we further assume the source and relay use the same signal/modulation set). We remark that MDF for half-duplex relaying does not suffer from error propagation, in contrast to MDF for full-duplex relaying and D-BLAST for point-to-point channels.

Code design is usually done by using density evolution [154] or EXIT charts [175]. We use the latter approach and design irregular low-density parity-check (LDPC) codes with the curve-fitting procedure described in [176]. The coded bits are mapped to QPSK symbols via the Gray mapping. The decoder uses the standard graph representation of an LDPC code with variable nodes on the left and check nodes on the right. The left and right nodes are connected by edges whose nodes are chosen with a random permutation that avoids 2-cycles. The decoder iterates 60 times between the left and right nodes by using an *a posteriori* probability (APP) decoder.

Consider the $1 \times 1 \times 2$ setup of Figure 4.26 and $d = 0.25$. Consider $R = 1/2$ without a relay. We design an LDPC code with rate $R_c = 1/4$ and length $n_c = 8000$ that has an (single-antenna, no fading, BPSK) AWGN decoding threshold of $(E_b/N_0)_{dB} = -0.4$ dB which is about 0.3 dB from capacity. The resulting frame error rates (FER) are shown on the right in Figure 4.27. Observe that the code operates within 1.5 dB of capacity at an FER of 10^{-3}. The extra loss (as compared to 0.3 dB for the single-antenna case) can be attributed to the short code length and the fading.

Consider next $R = 1$. We design an LDPC code with rate $R_c = 3/8$ and length $n_c = 16,000$ that has an (single-antenna, no fading, BPSK) AWGN decoding threshold of $(E_b/N_0)_{dB} = 0.1$ dB which is about 0.45 dB from capacity. The encoding and decoding procedure is as follows (see Figure 4.24).

- In the odd-numbered blocks $b = 1, 3, 5, \ldots$, the source transmits 4000 QPSK symbols (or 8000 of the 16,000 codeword bits) by using the rate $R_c = 3/8$ LDPC code.
- After every odd-numbered block b, the relay decodes the information bits of the $R_c = 3/8$ code from this block. Note

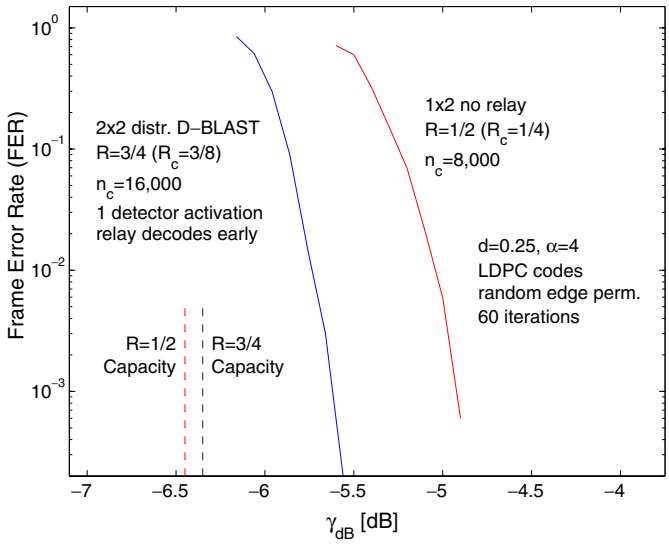


Fig. 4.27 MDF frame error rates for the $1 \times 1 \times 2$ scenario.

that the relay has received only half of this codeword's symbols.

- In the even-numbered blocks $b = 2, 4, 6, \ldots$, the source transmits using the rate $R_c = 1/4$ code described above.
- In the even-numbered blocks, the relay encodes the information bits decoded from the previous block by using the $R_c = 3/8$ encoder and transmits the last 4000 QPSK symbols of this codeword (or the last 8000 of the 16,000 codeword bits).
- After every even-numbered block, the sink decodes the information bits of the rate $R_c = 3/8$ code. The sink performs only one detector activation per codeword (we remark that multiple detector activations would improve the performance a little [176]).
- The sink cancels the interference caused by the symbols of the $R_c = 3/8$ code from the even-numbered blocks.
- After every even-numbered block, the sink decodes the information bits of the $R_c = 1/4$ code.

The overall rate is $R = 2(3/8) + 2(1/4)(1/2) = 1$ bit per use, where the leading factors 2 are due to the QPSK modulation. There are three decoding steps to consider.

- The FER of the relay decoding step is not shown in Figure 4.27 because it lies far to the left of the other two curves.
- The FER of the sink decoding the information bits from the $R_c = 3/8$ code is shown as the left curve in Figure 4.27 (labeled "2×2 distr. D-BLAST").
- The FER of the sink decoding the information bits from the $R_c = 1/4$ code is the same as the case where there is no relay, and is the curve on the right in Figure 4.27.

We see that the dominating FER is in both cases (without and with a relay) due to the direct link from the source to the sink. The reliability of the two schemes is therefore the same. However, the MDF scheme doubles the rate.

5

Cooperative Diversity

5.1 Introduction

The previous chapter examined cooperative strategies when no fading or fast fading is present. The main performance metric was rate because long codes can average out the effects of noise and fading and can make the error probability approach zero. In this chapter, we consider another extreme, namely, slow fading where the channel gains are random but are held constant for the duration of a codeword. For example, if a channel is in a deep fade then for high rates one cannot avoid making errors and we say there is an *outage* [54, 144]. We wish to characterize the tradeoff between rate and outage probability. We again assume that the nodes have CSIR (see Sections 3.3, 4.1, and 4.2.1).

Outage probability can be reduced by means of *diversity*, i.e., transmitting signals carrying the same information over different paths in time, frequency, or space. For example, if ARQ feedback is available, then one can create diversity with Hybrid-ARQ (see Section 2.3). One of our aims is to create diversity through node cooperation.

Transmit cooperation has nodes using DF to exchange each other's messages. By sharing resources, the nodes create two paths to transmit

their information (see Figure 3.1) and this is known as *cooperation diversity* [162, 163] or *cooperative diversity* [112, 114]. Receive cooperation has nodes forwarding information about their observations. For example, the nodes can use CF. A system with both transmit and receive cooperation resembles a multi-antenna or multiple-input, multiple-output (MIMO) system, and it is therefore sometimes called a *distributed* MIMO system. It is known that multiple antennas can increase capacity without sacrificing bandwidth or energy [53, 174]. For instance, if the path gains between the individual transmit and receive antennas fade independently, high data rates are achieved by sending different data streams over the independent channels, an approach known as *spatial multiplexing* [70].

Distributed MIMO systems can realize some of the usual MIMO gains. However, the absence of a high-capacity link (e.g., a cable) between the antennas limits the gains in several ways.

- The messages are known initially only to the source nodes.
- Resources such as power, bandwidth and time (delay) must be expended to enable cooperation.
- Antenna power allocation cannot be performed as in conventional MIMO systems.
- Nodes may have half-duplex constraints.
- Synchronizing signals sent from distributed antennas is more difficult than in a conventional MIMO transmitter.

The effects of such limitations will become apparent once we analyze various cooperative protocols.

5.2 Performance Metrics

Consider a point-to-point channel as in Figure 2.2 and suppose the coded modulation rate is R (see Section 2.2.3). Let γ be the average SNR at the receiver and let $I(\gamma)$ be the mutual information between the channel inputs and outputs. Observe that $I(\gamma)$ is a random variable that depends on the fading coefficients. We will usually assume that the codeword length n is sufficiently long and the codes sufficiently good so that a decoding error is negligible if $I(\gamma) > R$. The event $I(\gamma) < R$

is called an *outage* and we write

$$P_o(\gamma, R) = \Pr[I(\gamma) < R]. \tag{5.1}$$

We wish to characterize the achievable triples (γ, R, P_o). For example, for a fixed R and P_o we would like to determine the smallest possible γ. However, such an analysis is often difficult so we instead study the limiting behavior of (R, P_o) when γ is large or small. As we shall see, one obtains important insight from these extreme cases.

5.2.1 High SNR Metrics

We consider the following performance metrics for large γ: diversity gain, multiplexing gain, and the diversity-multiplexing tradeoff.

5.2.1.1 Diversity Gain

For a given rate R, we define the *diversity gain* to be

$$d = \lim_{\gamma \to \infty} \frac{-\log P_o(\gamma, R)}{\log \gamma}. \tag{5.2}$$

The diversity gain thus determines the high-SNR slope of the outage probability as a function of the average SNR when plotted on a log–log scale. For example, consider the slow Rayleigh fading channel with

$$Y_i = \frac{H}{d^{\alpha/2}} X_i + Z_i, \quad i = 1, 2, \ldots, n, \tag{5.3}$$

and the power constraint (2.4), i.e., the Z_i and H are the usual complex, Gaussian random variables. Note that H is held fixed for all i and that the expected receive SNR is $\gamma = (P/N)/d^{\alpha}$. The mutual information $I(\gamma)$ with Gaussian inputs is

$$I(\gamma) = \log_2 \left(1 + \gamma |H|^2\right) \text{ bits/use}. \tag{5.4}$$

¿From (5.4), we thus have

$$P_o(\gamma, R) = \Pr\left[|H|^2 < \frac{2^R - 1}{\gamma}\right] = 1 - \exp\left(-\frac{2^R - 1}{\gamma}\right). \tag{5.5}$$

To describe the outage behavior as a function of γ, we use the reciprocal of the *normalized SNR* (see [52, Sec. II.E])

$$g(\gamma, R) = \frac{2^{2R} - 1}{\gamma}. \tag{5.6}$$

The outage probability (5.5) is thus

$$P_o(\gamma, R) = 1 - e^{-g(\gamma, R/2)} \sim^{\gamma} g(\gamma, R/2), \tag{5.7}$$

where the notation $f_1(x, y) \sim^x f_2(x, y)$ means that

$$\lim_{x \to \infty} \frac{f_1(x, y)}{f_2(x, y)} = 1. \tag{5.8}$$

Observe from (5.6) and (5.7) that, up to a constant, the outage probability decreases as $1/\gamma$ so the diversity gain (5.2) is $d = 1$.

We remark that for fixed coded modulations such as uncoded BPSK (see Section 2.2.2) we substitute the outage probability $P_o(\gamma, R)$ in (5.2) with the error probability $P_e(\gamma)$. The BPSK error probability in Rayleigh fading behaves as

$$P_e(\gamma) = \frac{1}{4\gamma} \tag{5.9}$$

and thus BPSK has $d = 1$ and $R = 1$.

Finally, we remark that the diversity can alternatively be defined for a *family* of codes $\{\mathcal{C}(\gamma)\}$ with rates $R(\gamma)$ indexed by γ. In other words, the rate R in (5.2) changes with γ. This extension is useful when studying multiplexing gains.

5.2.1.2 Diversity Gain for MISO Channels

One can generalize the above analysis for Rayleigh-fading MIMO channels with n_1 transmit and n_2 receive antennas and show that the diversity gain is $d = n_1 n_2$ [201]. For example, suppose there are n_1 transmit antennas and one receive antenna, i.e., we have a multi-input, single-output or MISO channel. The model is (see Section 4.4)

$$Y_i = \frac{H}{d^{\alpha/2}} \underline{X}_i + Z_i, \tag{5.10}$$

where $\underline{H}$ is a $1 \times n_1$ vector with the usual independent Gaussian entries, and the $\underline{X}_i$ are $n_1 \times 1$ vectors satisfying the power constraint (4.54). Following steps outlined in Section 4.4, one can show that the best $\underline{X}_i$ are independent, zero-mean, Gaussian, and with independent entries of variance $\beta_k P$ where $\sum_{k=1}^{n_1} \beta_k = 1$. The resulting outage probability is

$$P_o(\gamma, R) = \Pr\left[\sum_{k=1}^{n_1} \beta_k |H_k|^2 < g(\gamma, R/2)\right]. \tag{5.11}$$

The choice of β_k, $k = 1, 2, \ldots, n_1$, that minimizes $P_o(\gamma, R)$ turns out to depend on γ and R [174, Section 5.1] and has not yet been characterized for $n_1 > 2$. For $n_1 = 2$, note that $\beta_1 |H_1|^2$ and $\beta_2 |H_2|^2$ are independent random variables with distribution functions

$$F_u(x) = 1 - e^{-x/\beta_u}, \quad x \geq 0,\ u = 1, 2. \tag{5.12}$$

The distribution function of their sum is

$$F(x) = \frac{\beta_1 \left(1 - e^{-x/\beta_1}\right) - \beta_2 \left(1 - e^{-x/\beta_2}\right)}{\beta_1 - \beta_2}, \quad x \geq 0. \tag{5.13}$$

Using (5.13), one can show that the optimal β_k for $n_1 = 2$ are: $\beta_1 = \beta_2 = 1/2$ if $R < \log(1 + \mu_o P)$ where $\mu_o \approx 1.2564$, and $\beta_1 = 1$ otherwise [94].

In general, however, we are mainly interested in large γ for which one can show that $\beta_k = 1/n_1$ for all k is best. To see this, note that (5.11) is lower bounded by setting $\beta_k = 1$ for all k, i.e., we have (see [148, Ch. 1])

$$P_o(\gamma, R) \geq 1 - e^{-g(\gamma, R/2)} \sum_{k=0}^{n_1 - 1} \frac{g(\gamma, R/2)^k}{k!} \tag{5.14}$$

$$\sim^{\gamma} \frac{g(\gamma, R/2)^{n_1}}{n_1!}. \tag{5.15}$$

Similarly, if we set $\beta_k = 1/n_1$ for all k then we can achieve

$$P_o(\gamma, R) = 1 - e^{-n_1 g(\gamma, R/2)} \sum_{k=0}^{n_1 - 1} \frac{n_1^k}{k!} g(\gamma, R/2)^k \tag{5.16}$$

$$\sim^{\gamma} \frac{n_1^{n_1}}{n_1!} g(\gamma, R/2)^{n_1}. \tag{5.17}$$

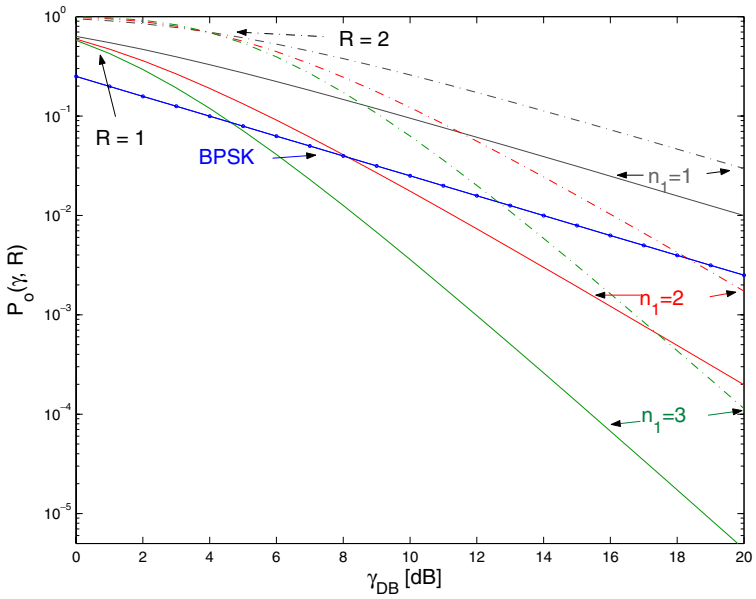


Fig. 5.1 Outage probabilities for MISO channels.

Thus, up to a constant, the outage probability decreases as γ^{-n_1} so that the diversity gain (5.2) is $d = n_1$. Figure 5.1 plots $P_o(\gamma, R)$ in (5.16) as a function of $\gamma_{dB} = 10\log_{10}\gamma$ for $n_1 = 1,2,3$ and $R = 1,2$. Also shown is the BPSK error probability (5.9). Observe that BPSK outperforms the random coding strategies with $n_1 = 1$. However, one should keep in mind that an outage for BPSK means that one bit is in error, while an outage for random codes means that an entire block of bits is in error.

5.2.1.3 Multiplexing Gain

Suppose we wish to study the limiting behavior of $I(\gamma)$ without considering the outage probability. To do this, we consider a *family* of codes $\{\mathcal{C}(\gamma)\}$ with rates $R(\gamma)$ indexed by γ, as described above. The rates $R(\gamma)$ usually increase with γ. The *multiplexing gain* (or the number of degrees-of-freedom) is defined as

$$r = \lim_{\gamma\to\infty} \frac{R(\gamma)}{\log\gamma}. \tag{5.18}$$

For example, for a Rayleigh-fading MIMO channel with CSIR the maximum multiplexing gain is $r = \min(n_1, n_2)$, which represents the rank of a randomly chosen gain matrix H. We will use $R = r \log_2 \gamma$.

5.2.1.4 Diversity-Multiplexing Tradeoff

To study the diversity and multiplexing gains together, we again consider a family of codes $\{\mathcal{C}(\gamma)\}$ with rates $R(\gamma)$, and we compute the family's diversity and multiplexing gains using (5.2) and (5.18). It is interesting to characterize the boundary of the set of achievable rate pairs (d, r) over all cooperative protocols. For example, let $d^*(r)$ be the supremum of the set of possible d for a fixed r. Similarly, let $r^*(d)$ be the supremum of the set of possible r for a fixed d. In analogy to rate-distortion theory, we call $d^*(\cdot)$ the *diversity-multiplexing function* and $r^*(\cdot)$ the *multiplexing-diversity function*.

For example, consider a MISO Rayleigh fading channel with CSIR for which we know that Gaussian inputs with independent entries are optimal. We use (5.17) and $R = r \log_2 \gamma$ to compute

$$d^*(r) = \lim_{\gamma \to \infty} \frac{-\log\left(n_1^{n_1}/(n_1!)\right) - n_1 \log\left(\gamma^{-1} \cdot \left(2^{r \log_2 \gamma} - 1\right)\right)}{\log \gamma} \tag{5.19}$$

$$= n_1(1 - r). \tag{5.20}$$

The diversity-multiplexing function is therefore linear. This result generalizes to MIMO channels with n_1 transmit and n_2 receive antennas as follows. For block lengths n with $n \geq n_1 + n_2 - 1$, we find that $d^*(r)$ is a piecewise-linear function connecting the points $(\tilde{r}, d^*(\tilde{r}))$, $\tilde{r} = 0, 1, 2, \ldots, \min(n_1, n_2)$, where $d^*(\tilde{r}) = (n_1 - \tilde{r})(n_2 - \tilde{r})$ (see [201]).

5.2.2 Low SNR Metrics

At low SNR, the system is limited by energy rather than multiplexing gain, and a meaningful metric is the rate R normalized by the SNR [182, 183, 12]. For example, consider the channel (5.3) for which (5.5) gives

$$R(\gamma, P_o) = \log_2\left(1 - \gamma \ln(1 - P_o)\right) \text{ bits/use}, \tag{5.21}$$

where $\ln(x)$ is the natural logarithm of x. We define the low-SNR and SNR-normalized rate to be

$$R'(0,P_o) = \lim_{\gamma\to 0} \frac{R(\gamma,P_o)\ln(2)}{\gamma} \text{ nats/use}, \tag{5.22}$$

where we have changed units to nats (1 bit = $\ln(2)$ nats). We compute

$$R'(0,P_o) = -\ln(1-P_o) \text{ nats/use}, \tag{5.23}$$

which gives $R'(0,P_o) \approx P_o$ nats when P_o is small.

Consider next the MISO Rayleigh fading channel (5.10) with n_1 transmit antennas and $\beta_k = 1/n_1$ for all k. We define $R' = R\ln(2)/\gamma$ nats and use (5.16) to compute

$$\lim_{\gamma\to 0} g(\gamma, R/2) = R' \tag{5.24a}$$

$$\lim_{\gamma\to 0} P_o = 1 - e^{-R'} \sum_{k=0}^{n_1-1} \frac{n_1^k}{k!} (R')^k. \tag{5.24b}$$

Equation (5.24b) can be solved for R' as in (5.23), but we will be mainly interested in small P_o and hence small R'. Furthermore, small R' is effectively the same as small $g(\gamma, R/2)$ or large γ in (5.14). Ignoring low order terms, for small R' we have (see [12])

$$R'(0,P_o) \approx \frac{(n_1!)^{1/n_1}}{n_1} P_o^{1/n_1} \text{ nats}. \tag{5.25}$$

For example, for $n_1 = 2$ and small P_o we have $R'(0,P_o) \approx \sqrt{P_0/2}$. This is much larger than $R'(0,P_o) \approx P_o$ for $n_1 = 1$ and small P_o.

We remark that we can re-use our high-SNR analysis for low-SNR if we can make the "instantaneous" received SNR large by some means. This is, in fact, possible by using "bursty" transmission where a node signals for a short period of time, or in a small frequency band, with very high power. Of course, such bursty transmission is not possible if one has per-symbol power constraints such as (4.55) or a constraint on the power per Hertz.

5.3 System Model

We study the performance of cooperative protocols for the network shown in Figure 5.2. The network outputs are

$$Y_{1,i} = \begin{cases} \frac{H_{12}}{d_{12}^{\alpha/2}} X_{2,i} + Z_{1,i} & \text{if } X_{1,i} = 0, \\ 0 & \text{if } X_{1,i} \neq 0 \end{cases} \tag{5.26a}$$

$$Y_{2,i} = \begin{cases} \frac{H_{21}}{d_{21}^{\alpha/2}} X_{1,i} + Z_{2,i} & \text{if } X_{2,i} = 0, \\ 0 & \text{if } X_{2,i} \neq 0 \end{cases} \tag{5.26b}$$

$$Y_{3,i} = \frac{H_{13}}{d_{13}^{\alpha/2}} X_{1,i} + \frac{H_{23}}{d_{23}^{\alpha/2}} X_{2,i} + Z_{3,i} \tag{5.26c}$$

for $i = 1, 2, \ldots, n$. The transmitting nodes have half-duplex constraints. The noise variables $Z_{u,i}$, $u = 1, 2, 3$, $i = 1, 2, \ldots, n$, and channel gains H_{uv} are the usual complex, Gaussian random variables (see Section 3.2 and (3.3a) and (3.3b)). Note that the H_{uv} change on a slow timescale as compared to the $X_{u,i}$ and $Z_{u,i}$. The transmitters have the average power constraints (3.5) and we set $P_1 = P_2 = P$ for simplicity. We will usually view the network of Figure 5.2 as either a RC with node 1 as the source, or a RC with node 2 as the source. Of course, this means that all the relaying strategies developed in Chapter 4 will apply here.

We add some remarks about channel and network knowledge. The DF protocols usually require CSIR only. The AF protocols additionally require the destination to know the source-to-relay channel gain

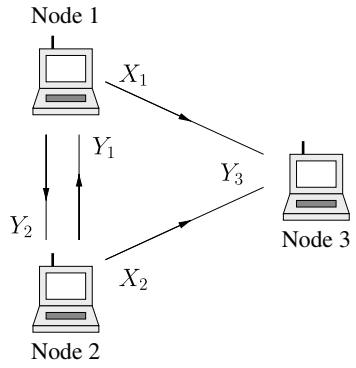


Fig. 5.2 A three-node wireless network. Nodes 1 and 2 have messages destined for node 3.

to perform maximal ratio combining [148, p. 721]. The CF protocols require both the relay and destination to know all channel gains, i.e., the relay must have CSI available before transmitting (CSIT). The CSIT usually has low rate as compared to the data, and requires feedback from the destination once this node determines the channel gains on its incoming links. Keeping this in mind, we will see that CF achieves the best possible diversity-multiplexing tradeoff in both the high and low SNR regimes. Low rate feedback in the form of an ARQ bit from the destination to the relay can also improve performance, as shown in Section 5.4.2.5.

Finally, we remark that from now on we work with

$$\tilde{\gamma} = P/N \tag{5.27}$$

rather than the received SNRs $\gamma_{uv} = (P/N)/d_{uv}^{\alpha}$. For convenience, we abuse notation and generically write mutual information expressions parameterized by $\tilde{\gamma}$ as $I(\tilde{\gamma})$ and then define $P_o(\tilde{\gamma}, R) = \Pr[I(\tilde{\gamma}) < R]$.

5.4 Cooperative Strategies for High SNR

We consider two classes of strategies by distinguishing whether the sources transmit at the same time and in the same frequency band or not.[1] Strategies for which the sources do not interfere are called *orthogonal* strategies. Such schemes are meant to achieve high diversity gains rather than high rates [112], although orthogonal strategies are good enough for low SNR (see Section 5.5). Non-orthogonal strategies use bandwidth more efficiently and thus generally exhibit better diversity-multiplexing tradeoffs [14, 139, 193]. For example, a cooperative scheme known as *dynamic DF* achieves the diversity-multiplexing function for low values of multiplexing gains [14]. For larger values, multipath DF (see Section 4.2.7) outperforms dynamic DF [11]. Interestingly, CF achieves the diversity multiplexing function for all multiplexing gains and for any number of antennas at the nodes [196], but recall that it needs extra CSI at the relay and destination.

[1] The following material is based mostly on [112].

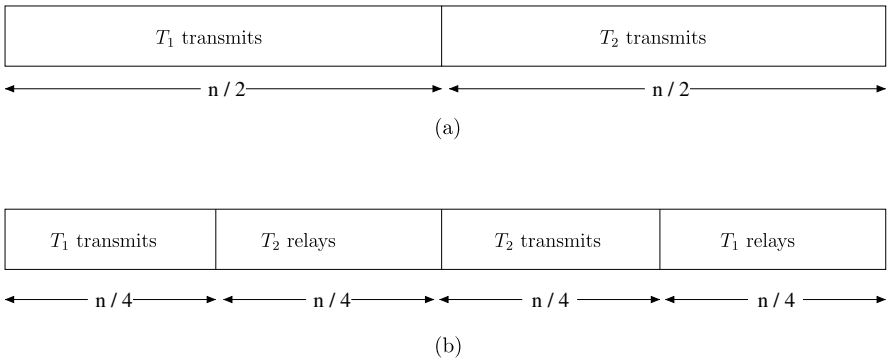


Fig. 5.3 Channel allocation for: (a) orthogonal direct transmission, (b) orthogonal relaying.

5.4.1 Direct Transmission

Consider the network in Figure 5.2 and the TDM strategy in Figure 5.3(a). This strategy is cooperative in the sense that the users do not interfere with each other, and the destination's signals are

$$Y_{3,i} = \begin{cases} \frac{H_{13}}{d_{13}^{\alpha/2}} X_{1,i} + Z_{3,i}, & i = 1, 2, \ldots, n/2 \\ \frac{H_{23}}{d_{23}^{\alpha/2}} X_{2,i} + Z_{3,i}, & i = n/2+1, \ldots, n. \end{cases} \tag{5.28}$$

The outage probability for nodes $u = 1, 2$ is (see (5.7))

$$P_o(\tilde{\gamma}, R) = 1 - e^{-g(\tilde{\gamma},R)\, d_{u3}^{\alpha}/2} \tag{5.29}$$

$$\sim^{\tilde{\gamma}} g(\tilde{\gamma}, R)\, d_{u3}^{\alpha}/2, \tag{5.30}$$

where we remark that both nodes can send with twice their average power in their time slots. Observe that, up to a constant, the outage probabilities decrease as $1/\tilde{\gamma}$ (or $1/\gamma$).

5.4.2 Orthogonal Relaying Strategies

Consider next the cooperative strategy in Figure 5.3(b). During times $i = 1, 2, \ldots, n/4$, node 1 transmits and nodes 2 and 3 receive

$$Y_{2,i} = \frac{H_{12}}{d_{12}^{\alpha/2}} X_{1,i} + Z_{2,i}, \quad i = 1, 2, \ldots, n/4 \tag{5.31a}$$

$$Y_{3,i} = \frac{H_{13}}{d_{13}^{\alpha/2}} X_{1,i} + Z_{3,i}, \quad i = 1, 2, \ldots, n/4. \tag{5.31b}$$

During times $i = n/4 + 1, \ldots, n/2$, node 2 transmits $X_{2,i}$ as a function of $Y_2^{n/4}$. Node 2 thus acts as a relay and the destination receives

$$Y_{3,i} = \frac{H_{23}}{d_{23}^{\alpha/2}} X_{2,i} + Z_{3,i}, \quad i = n/4 + 1, \ldots, n/2. \tag{5.31c}$$

During time $i = n/2 + 1, n/2 + 2, \ldots, n$, the roles of nodes 1 and 2 are reversed and the same procedure is repeated. We next consider several ways in which the relays process their received symbols.

5.4.2.1 Amplify-and-Forward

Suppose node 2 uses AF and forwards

$$X_{2,i} = aY_{2,i-n/4}, \quad i = n/4 + 1, \ldots, n/2, \tag{5.32}$$

while choosing a to satisfy its power constraint, i.e., we set (see (4.10))

$$|a|^2 = \frac{2\tilde{\gamma}}{1 + 2|H_{12}|^2 \tilde{\gamma}/d_{12}^{\alpha}}, \tag{5.33}$$

where we recall that $\tilde{\gamma} = P/N$ and the relay sends at twice its average power in its time slot. To decode, the destination combines the received signals (5.31b) and (5.31c) using maximum-ratio combining which requires knowledge of $|H_{12}|^2/d_{12}^{\alpha}$. With Gaussian codebooks, one achieves

$$I(\tilde{\gamma}) = \frac{1}{4} \log_2 \left(1 + 2\tilde{\gamma} \left(\frac{|H_{13}|^2}{d_{13}^{\alpha}} + \frac{2\tilde{\gamma} |H_{12}|^2 |H_{23}|^2 / (d_{12}^{\alpha} d_{23}^{\alpha})}{1 + 2\tilde{\gamma}(|H_{12}|^2/d_{12}^{\alpha} + |H_{23}|^2/d_{23}^{\alpha})} \right) \right), \tag{5.34}$$

where the factor $1/4$ is because nodes 1 and 2 transmit only $1/4$ of the time for each message. The outage event $I(\tilde{\gamma}) < R$ is the same as

$$\frac{|H_{13}|^2}{d_{13}^{\alpha}} + \frac{2\tilde{\gamma} |H_{12}|^2 |H_{23}|^2 / (d_{12}^{\alpha} d_{23}^{\alpha})}{1 + 2\tilde{\gamma}(|H_{12}|^2/d_{12}^{\alpha} + |H_{23}|^2/d_{23}^{\alpha})} < g(\tilde{\gamma}, 2R)/2. \tag{5.35}$$

For high SNR, one can show that the outage probability is given by (see [112])

$$P_o(\tilde{\gamma}, R) \sim^{\tilde{\gamma}} g(\tilde{\gamma}, 2R)^2 \, d_{13}^{\alpha} (d_{12}^{\alpha} + d_{23}^{\alpha})/8. \tag{5.36}$$

The diversity gain is thus $d = 2$, and this is the best possible. To see this, note that the outage probability must be larger than when the

relay knows the source message ahead of time, effectively creating a MISO channel with $n_1 = 2$ transmit antennas. We thus have $d \leq 2$ by applying (5.15).

5.4.2.2 Compress-and-Forward

Suppose the relay uses CF with the relay mode M_2 acting as a time-sharing random variable (see Sections 4.2.4 and 4.3). The rates are

$$R = \frac{1}{4} I(X_1; \hat{Y}_2 Y_3 \,|\, H, M_2 = L) \tag{5.37a}$$

subject to

$$I(\hat{Y}_2; Y_2 | Y_3 H, M_2 = L) \leq I(X_2; Y_3 | H, X_1 = 0, M_2 = T) \tag{5.37b}$$

$$p(h, m_2, x_1, x_2, y_2, y_3, \hat{y}_2) = p(h)\, p(m_2)\, p(x_1|m_2)\, p(x_2|h, m_2) \\ p(y_2, y_3|x_1, x_2, h)\, p(\hat{y}_2|x_2, y_2, h, m_2), \tag{5.37c}$$

where $H = [H_{12}\ H_{13}\ H_{23}]$. We choose $\hat{Y}_2 = Y_2 + \hat{Z}_2$ when $M_2 = L$ as in Section 4.2.4. The rate (5.37a) is a quarter of (4.28a) and $I(\hat{Y}_2; Y_2 | Y_3 H, M_2 = L)$ is the same as (4.28b) with $P_1 = P_2 = 2P$. We also have

$$I(X_2; Y_3 | H, X_1 = 0, M_2 = T) = \log_2 \left(1 + 2\tilde{\gamma} \frac{|H_{23}|^2}{d_{23}^\alpha} \right). \tag{5.38}$$

We thus choose

$$\hat{N}_2 = N \cdot \frac{1 + 2\tilde{\gamma} \left(|H_{12}|^2 / d_{12}^\alpha + |H_{13}|^2 / d_{13}^\alpha \right)}{(1 + 2\tilde{\gamma} |H_{13}|^2 / d_{13}^\alpha)\, 2\tilde{\gamma} |H_{23}|^2 / d_{23}^\alpha} \tag{5.39}$$

to satisfy (5.37b) with equality. The resulting outage event is

$$\frac{|H_{13}|^2}{d_{13}^\alpha} + \frac{2\tilde{\gamma} \frac{|H_{12}|^2}{d_{12}^\alpha} \frac{|H_{23}|^2}{d_{23}^\alpha}}{1 + 2\tilde{\gamma} \left\{ \left(1 + 2\tilde{\gamma} \frac{|H_{13}|^2}{d_{13}^\alpha} \right)^{-1} \frac{|H_{12}|^2}{d_{12}^\alpha} + \frac{|H_{23}|^2}{d_{23}^\alpha} \right\}} < \frac{g(\tilde{\gamma}, 2R)}{2}, \tag{5.40}$$

which clearly implies the AF outage event (5.35). The diversity gain is thus again $d = 2$. Note, however, that the relay and destination need full CSIT and CSIR, respectively.

5.4.2.3 Decode-and-Forward

We first consider a suboptimal DF strategy that uses a repetition code at the relay, i.e., the relay decodes the source message, re-encodes using the same codebook, and transmits. For Gaussian codebooks, we compute (see (4.32))

$$I(\tilde{\gamma}) = \min\left\{\frac{1}{4}\log_2\left(1+2\tilde{\gamma}\,\frac{|H_{12}|^2}{d_{12}^{\alpha}}\right), \right. \\ \left. \frac{1}{4}\log_2\left(1+2\tilde{\gamma}\left(\frac{|H_{13}|^2}{d_{13}^{\alpha}}+\frac{|H_{23}|^2}{d_{23}^{\alpha}}\right)\right)\right\}. \tag{5.41}$$

The outage event $I_{DF} < R$ is thus the same as

$$\min\left\{\frac{|H_{12}|^2}{d_{12}^{\alpha}}, \frac{|H_{13}|^2}{d_{13}^{\alpha}}+\frac{|H_{23}|^2}{d_{23}^{\alpha}}\right\} < g(\tilde{\gamma},2R)/2 \tag{5.42}$$

and we have

$$P_o(\tilde{\gamma},R) \sim^{\tilde{\gamma}} g(\tilde{\gamma},2R)\; d_{12}^{\alpha}/2. \tag{5.43}$$

Observe that $d=1$ and this DF strategy offers no diversity gain over non-cooperative communication. In fact, we obtain the same type of result if we use the usual DF strategy, where the only change is that the second logarithm in (5.41) increases to

$$\frac{1}{4}\log_2\left(1+2\tilde{\gamma}\,\frac{|H_{13}|^2}{d_{13}^{\alpha}}\right)+\frac{1}{4}\log_2\left(1+2\tilde{\gamma}\,\frac{|H_{23}|^2}{d_{23}^{\alpha}}\right) \tag{5.44}$$

and a similar change is made in (5.42). The deficiency in both DF strategies is that we force the relay to decode even if $|H_{12}|^2$ is small.

5.4.2.4 Selection Decode-and-Forward

To overcome the drawback of DF, we compare the measured channel gain $|H_{12}|^2/d_{12}^{\alpha}$ at the relay with a pre-specified threshold. If $|H_{12}|^2/d_{12}^{\alpha}$ is above the threshold, then the source remains silent and the relay forwards the information (either via a repetition code or another code). Otherwise, the source repeats its transmission. Observe that this method requires the relay to send one feedback bit to the source after every other transmission block. We call this a *selection DF* strategy.

We remark that the source must limit its transmit power to less than $2P$ in its time slots. However, if P_o is small, as we shall assume, then the source rarely needs to transmit twice. Hence, we have the source transmit with power $2P$ in its first time slot and with a small amount of extra power in the second time slot if need arises (alternatively, the source never transmits in the second time slot).

Suppose the relay uses a repetition code. We compute

$$I(\tilde{\gamma}) = \begin{cases} \frac{1}{4}\log_2\left(1 + 2\tilde{\gamma}\,\frac{|H_{13}|^2}{d_{13}^\alpha}\right) & \text{if } \frac{|H_{12}|^2}{d_{12}^\alpha} < g(\tilde{\gamma}, 2R)/2, \\ \frac{1}{4}\log_2\left(1 + 2\tilde{\gamma}\left(\frac{|H_{13}|^2}{d_{13}^\alpha} + \frac{|H_{23}|^2}{d_{23}^\alpha}\right)\right) & \text{if } \frac{|H_{12}|^2}{d_{12}^\alpha} > g(\tilde{\gamma}, 2R)/2. \end{cases} \tag{5.45}$$

Observe that we chose the threshold so that the direct transmission is repeated if there is an outage on the source-relay link. The overall outage event is $\mathcal{A} \cup \mathcal{B}$ where

$$\mathcal{A} = \left\{ \min\left\{ \frac{|H_{12}|^2}{d_{12}^\alpha}, \frac{|H_{13}|^2}{d_{13}^\alpha} \right\} < g(\tilde{\gamma}, 2R)/2 \right\} \tag{5.46a}$$

$$\mathcal{B} = \left\{ \frac{|H_{13}|^2}{d_{13}^\alpha} + \frac{|H_{23}|^2}{d_{23}^\alpha} < g(\tilde{\gamma}, 2R)/2 \right\} \tag{5.46b}$$

and so we have

$$P_o(\tilde{\gamma}, R) \sim^{\tilde{\gamma}} g(\tilde{\gamma}, 2R)^2\, d_{13}^\alpha (2d_{12}^\alpha + d_{23}^\alpha)/8, \tag{5.47}$$

which is identical to (5.36) except for the factor of 2 in front of d_{12}^α. This factor arises because $\mathcal{A}$ has twice the probability of $\mathcal{B}$ for large $\tilde{\gamma}$. Other DF strategies achieve similar performance [82, 84, 86, 125, 126, 172, 173]. The results are summarized in Table 5.1 where we set $d_{uv}^\alpha = 2$ for all (u, v).

Table 5.1 Diversity gains of orthogonal cooperative strategies for $d_{uv}^\alpha = 2$.

Cooperative strategy	High SNR performance	Comment	Extra CSI
Direct trans.	$g(\tilde{\gamma}, R)$	No diversity	
DF	$g(\tilde{\gamma}, 2R)$	No diversity	
Selection DF	$\frac{3}{2}g(\tilde{\gamma}, 2R)^2$	Full diversity	1 bit feedback
AF	$g(\tilde{\gamma}, 2R)^2$	Full diversity	$\lvert H_{12}\rvert^2/d_{12}^\alpha$ to dest.
CF	$g(\tilde{\gamma}, 2R)^2$	Full diversity	Relay CSIT, $\lvert H_{12}\rvert^2/d_{12}^\alpha$ to dest.

5.4.2.5 Incremental Relaying

An inspection of the selection DF strategy reveals that if $|H_{13}|^2/d_{13}^{\alpha}$ is above the threshold $g(\tilde{\gamma}, 2R)/2$ then the relay need not decode the message. Instead, the source can improve its rate by sending fresh information in the next time slot (as in hybrid ARQ or incremental redundancy ARQ). The resulting strategy operates at source power P and rate $R/2$ when the destination reception is successful, and at source power $P/2$ and rate $R/4$ otherwise. Let

$$P_{o,13}(\tilde{\gamma}, R) = \Pr\left[\frac{|H_{13}|^2}{d_{13}^{\alpha}} < \beta_1\, g(\tilde{\gamma}, R)\right], \tag{5.48}$$

where $\beta_1 = P/P_1$ and P_1 is the power the source can use when transmitting, i.e., P_1 satisfies

$$P = P_1(1 - P_{o,13}(\tilde{\gamma}, R)) + (P_1/2)P_{o,13}(\tilde{\gamma}, R). \tag{5.49}$$

We wish to compute the expected rate

$$\overline{R} = \frac{R}{2}\,(1 - P_{o,13}(\tilde{\gamma}, R)) + \frac{R}{4}\,P_{o,13}(\tilde{\gamma}, R) \tag{5.50}$$

$$= \frac{R}{2}\,e^{-\beta_1 g(\tilde{\gamma}, R/2)} + \frac{R}{4}\left(1 - e^{-\beta_1 g(\tilde{\gamma}, R/2)}\right). \tag{5.51}$$

A complication arises in that several values of R can give the same $\overline{R}$; we choose the smallest R that satisfies (5.50). One can show that incremental DF achieves full diversity, and that somewhat better results are possible with incremental AF (see [112, Sec. IV.E]). We remark, however, that this analysis implicitly requires using many fading blocks rather than just one.

5.4.3 Non-orthogonal Cooperative Strategies

The TDM constraint limits the rate more than necessary. We next consider non-orthogonal strategies that give good diversity multiplexing tradeoffs; for AF see also [139, 14, 193], for DF see [14], and for CF see [196].

5.4.3.1 Amplify-and-Forward

Consider two consecutive symbol transmissions from the source. During the first symbol transmission, the relay listens. In the next symbol period, the source transmits a new symbol and the relay uses AF. Thus, the relay effectively creates an intersymbol-interference channel (see Section 4.2.2). One can show that this AF strategy achieves the diversity-multiplexing pairs (d, r) satisfying (see [14])

$$d = (1 - r) + (1 - 2r)^+, \tag{5.52}$$

where $(x)^+ = x$ if $x > 0$ and $(x)^+ = 0$ if $x \leq 0$. This strategy achieves a better diversity-multiplexing tradeoff than orthogonal AF. However, for $r > 1/2$ this strategy is only as good as non-cooperative transmission because of the half-duplex constraint. This drawback is removed with the next strategy.

5.4.3.2 Dynamic Decode-and-Forward

Suppose the relay listens until it collects sufficient energy to decode the source message. It then re-encodes the message with its own codebook and transmits for the remaining time that the source transmits [93, 134, 136]. One can show that this *dynamic* DF scheme achieves the diversity-multiplexing pairs (d, r) satisfying

$$d = \begin{cases} 2(1 - r) & \text{if } 0 \leq r \leq 0.5, \\ (1 - r)/r & \text{if } 0.5 \leq r \leq 1. \end{cases} \tag{5.53}$$

Observe that the diversity gain is the same as for a $n_1 = 2$ MISO system for $0 \leq r \leq 1/2$. For $r > 1/2$, unfortunately, the relay cannot help enough of the time to achieve the MISO upper bound.

5.4.3.3 Compress-and-Forward

Suppose the relay listens half the time and uses CF. One can show that CF achieves the MISO upper bound for $0 \leq r \leq 1$, so we have $d^*(r) = 2(1 - r)$ [196, Sec. VI]. In fact, CF achieves the MISO upper bound for any number of antennas at the source, relay, and destination nodes. Note again, however, that the relay and destination need full

Table 5.2 Diversity-multiplexing tradeoff of non-orthogonal cooperative strategies.

Cooperative strategy	$d^*(r)$	Extra CSI
No cooperation	$1-r$	
Orthogonal AF	$2-4r$	$\lvert H_{12}\rvert^2/d_{12}^{\alpha}$ to dest.
Orthogonal DF	$2-4r$	
Non-orthogonal AF	$(1-r)+(1-2r)^+$	$\lvert H_{12}\rvert^2/d_{12}^{\alpha}$ to dest.
Dynamic DF	$\begin{cases} 2(1-r) & 0 \le r \le 0.5 \\ (1-r)/r & 0.5 \le r \le 1 \end{cases}$	Relay wait time
CF	$2(1-r)$	Relay CSIT, $\lvert H_{12}\rvert^2/d_{12}^{\alpha}$ to dest.
MISO bound	$2(1-r)$	Message to relay.

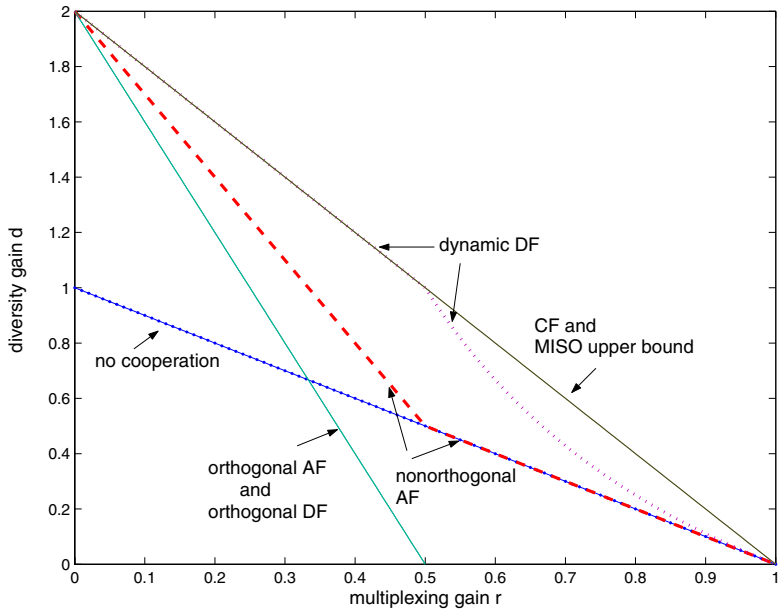


Fig. 5.4 Diversity-multiplexing tradeoff of non-orthogonal cooperative strategies.

CSIT and CSIR, respectively. The results are summarized in Table 5.2 and the diversity-multiplexing pairs are plotted in Figure 5.4.

5.5 Relaying Strategies for Low SNR

We next turn our attention to low SNR. However, for simplicity we focus on the RC rather than the network of Figure 5.2. We again consider TDM and orthogonal cooperative strategies: the source transmits

half the time, and the relay listens and talks half the time.[2] However, we will let the source transmit for a fraction β of the time with power P/β during its transmit slot (see [12, 46]).

5.5.1 Cut-Set Bound

Recall that we used the MISO diversity gain to upper bound the cooperative diversity gain. Unfortunately, for low SNR the MISO rates are too good and we need a better method. Consider the cut-set bound (3.33) for full-duplex relay channels. We compute

$$R \le \max_{\rho} \min \left\{ \log_2 \left(1 + \tilde{\gamma} \left(\frac{|H_{12}|^2}{d_{12}^\alpha} + \frac{|H_{13}|^2}{d_{13}^\alpha} \right) \left(1 - |\rho|^2 \right) \right), \right. \\ \left. \log_2 \left(1 + \tilde{\gamma} \left[\frac{H_{13}}{d_{13}^{\alpha/2}} \; \frac{H_{23}}{d_{23}^{\alpha/2}} \right] \left[\begin{array}{cc} 1 & \rho \\ \rho^* & 1 \end{array} \right] \left[\begin{array}{c} H_{13}^*/d_{13}^{\alpha/2} \\ H_{23}^*/d_{23}^{\alpha/2} \end{array} \right] \right) \right\}, \quad (5.54)$$

where $\rho = E[X_1 X_2^*]/P$. We remark that we could take ρ to be real because Rayleigh fading has uniform phase.

For example, suppose we choose $\rho = 0$. The outage probability $P_o(\tilde{\gamma}, R/2)$ is then lower-bounded by the event

$$\frac{1}{d_{13}^\alpha} |H_{13}|^2 + \min \left\{ \frac{1}{d_{12}^\alpha} |H_{12}|^2, \frac{1}{d_{23}^\alpha} |H_{23}|^2 \right\} < g(\tilde{\gamma}, R/2). \quad (5.55)$$

The distribution function of the minimum in (5.55) is

$$F(x) = 1 - e^{-x(d_{12}^\alpha + d_{23}^\alpha)}. \quad (5.56)$$

Combining (5.56) with (5.13), we obtain the distribution function of the left-hand side of (5.55) and can bound

$$P_o(\tilde{\gamma}, R/2) \ge \frac{(d_{12}^\alpha + d_{23}^\alpha)\, e^{-g(\tilde{\gamma},R/2) d_{13}^\alpha} - d_{13}^\alpha\, e^{-g(\tilde{\gamma},R/2)(d_{12}^\alpha + d_{23}^\alpha)}}{(d_{12}^\alpha + d_{23}^\alpha) - d_{13}^\alpha} \quad (5.57)$$

$$\sim^{\tilde{\gamma}} g(\tilde{\gamma}, R/2)^2\, d_{13}^\alpha (d_{12}^\alpha + d_{23}^\alpha)/2. \quad (5.58)$$

Note that (5.58) is a high SNR result, but it requires only that $g(\tilde{\gamma}, R/2)$ is small. We will soon use (5.58) for low SNR. However, we must first check that $\rho = 0$ gives the weakest cut-set bound.

[2] The following material is based mostly on [12].

Consider again (5.54) and write the second logarithm as

$$\log_2\left(1+\tilde{\gamma}\,[H_{13}\;\;H_{23}]\begin{bmatrix}\frac{1}{d_{13}^{\alpha}} & \frac{\rho}{d_{13}^{\alpha/2}d_{23}^{\alpha/2}}\\ \frac{\rho^*}{d_{13}^{\alpha/2}d_{23}^{\alpha/2}} & \frac{1}{d_{23}^{\alpha}}\end{bmatrix}\begin{bmatrix}H_{13}^*\\ H_{23}^*\end{bmatrix}\right)\Bigg\}. \quad (5.59)$$

The 2×2 matrix in (5.59) is Hermitian and decomposes as $U\Lambda U^{\dagger}$ where U is unitary and Λ is diagonal with entries λ_1 and λ_2 that satisfy

$$\lambda_1+\lambda_2 = 1/d_{13}^{\alpha}+1/d_{23}^{\alpha},\quad \lambda_1\lambda_2=\frac{1-|\rho|^2}{d_{13}^{\alpha}d_{23}^{\alpha}}. \quad (5.60)$$

But recall from Section 4.4 that $[H_{12}\;H_{13}]U$ has the same statistics as $[H_{12}\;H_{13}]$, so we can write (5.59) as

$$\log_2\left(1+\tilde{\gamma}\left(\lambda_1\,|\tilde{H}_{13}|^2+\lambda_2\,|\tilde{H}_{23}|^2\right)\right), \quad (5.61)$$

where $\tilde{H}_{13}$ and $\tilde{H}_{23}$ have the same distribution as H_{13} and H_{23} and are independent of H_{12}. The outage probability $P_o(\tilde{\gamma},R/2)$ is thus larger than the probability of the event $\mathcal{A}\cup\mathcal{B}$ where (see (5.54))

$$\mathcal{A}=\left\{\frac{1-|\rho|^2}{d_{12}^{\alpha}}\,|H_{12}|^2+\frac{1-|\rho|^2}{d_{13}^{\alpha}}\,|H_{13}|^2<g(\tilde{\gamma},R/2)\right\}, \quad (5.62a)$$

$$\mathcal{B}=\left\{\lambda_1\,|\tilde{H}_{13}|^2+\lambda_2\,|\tilde{H}_{23}|^2<g(\tilde{\gamma},R/2)\right\}. \quad (5.62b)$$

We have $\Pr[\mathcal{A}\cup\mathcal{B}]=\Pr[\mathcal{A}]+\Pr[\mathcal{B}]-\Pr[\mathcal{A}\cap\mathcal{B}]$. Consider first $\mathcal{A}$ for which it is clearly best to choose $|\rho|$ as small as possible. In fact, using (5.13) we have

$$\Pr[\mathcal{A}]\sim^{\tilde{\gamma}} g(\tilde{\gamma},R/2)^2\;\frac{d_{12}^{\alpha}\,d_{13}^{\alpha}}{2\,(1-|\rho|^2)} \quad (5.63)$$

so to improve on (5.58) for $\rho=0$ we must have $|\rho|^2<1-d_{12}^{\alpha}/(d_{12}^{\alpha}+d_{23}^{\alpha})$ (note that $\Pr[\mathcal{B}]-\Pr[\mathcal{A}\cap\mathcal{B}]\geq 0$). Consider next $\mathcal{B}$ for which we use (5.13) and (5.60) to compute

$$\Pr[\mathcal{B}]\sim^{\tilde{\gamma}} g(\tilde{\gamma},R/2)^2\;\frac{d_{13}^{\alpha}\,d_{23}^{\alpha}}{2\,(1-|\rho|^2)}. \quad (5.64)$$

Furthermore, one can check that the event $\mathcal{A}\cap\mathcal{B}$ implies the event

$$\max\left\{\frac{1-|\rho|^2}{d_{12}^{\alpha}}\,|H_{12}|^2,\lambda_1\,|\tilde{H}_{13}|^2,\lambda_2\,|\tilde{H}_{23}|^2\right\}<g(\tilde{\gamma},R/2), \quad (5.65)$$

where the three random variables are independent. We thus have

$$\begin{aligned}\Pr[\mathcal{A}\cap\mathcal{B}] &\le \left(1-e^{-g(\tilde{\gamma},R/2)d_{12}^{\alpha}/(1-|\rho|^2)}\right)\\ &\quad\times\left(1-e^{-g(\tilde{\gamma},R/2)/\lambda_1}\right)\left(1-e^{-g(\tilde{\gamma},R/2)/\lambda_2}\right) \qquad (5.66a)\\ &\sim^{\tilde{\gamma}} g(\tilde{\gamma},R/2)^3\,\frac{d_{12}^{\alpha}d_{13}^{\alpha}d_{23}^{\alpha}}{(1-|\rho|^2)^2}. \qquad (5.66b)\end{aligned}$$

Summarizing, the combination of (5.63), (5.64), and (5.66b) shows that $\rho=0$ gives the weakest high-SNR cut-set bound.

Finally, recall that we are interested in low SNR and not high SNR. We set $R'=R\ln(2)/\tilde{\gamma}$ in (5.57) and use $\lim_{\tilde{\gamma}\to 0} g(\tilde{\gamma},R/2)=R'$ to compute

$$\lim_{\tilde{\gamma}\to 0} P_o \ge \frac{(d_{12}^{\alpha}+d_{23}^{\alpha})\,e^{-R'd_{13}^{\alpha}}-d_{13}^{\alpha}\,e^{-R'(d_{12}^{\alpha}+d_{23}^{\alpha})}}{(d_{12}^{\alpha}+d_{23}^{\alpha})-d_{13}^{\alpha}}. \qquad (5.67)$$

For sufficiently small R' or P_o, we thus have (see (5.58))

$$R'(0,P_o)\le\sqrt{\frac{2P_o}{d_{13}^{\alpha}(d_{12}^{\alpha}+d_{23}^{\alpha})}}. \qquad (5.68)$$

We shall see that AF and CF can approach the rate on the right-hand side of (5.68).

5.5.1.1 Amplify-and-Forward

Consider the basic orthogonal AF strategy of Section 5.4.2.1. If $\tilde{\gamma}$ is small, then the outage event (5.35) becomes

$$\frac{|H_{13}|^2}{d_{13}^{\alpha}} < g(\tilde{\gamma},R)/2 \qquad (5.69)$$

and hence we have

$$R'(0,P_o)\approx P_o/d_{13}^{\alpha}. \qquad (5.70)$$

The rate (5.70) is less than the right-hand side (5.68) for small P_o. The reason is that the relay is amplifying a very noisy signal.

5.5.1.2 Bursty Amplify-and-Forward

To fix the basic AF strategy, let the source transmit with power P/τ for a very short fraction τ of time. The outage event is then (see (5.35))

$$\frac{|H_{13}|^2}{d_{13}^\alpha} + \frac{(\tilde{\gamma}/\tau)|H_{12}|^2|H_{23}|^2/(d_{12}^\alpha d_{23}^\alpha)}{1+(\tilde{\gamma}/\tau)(|H_{12}|^2/d_{12}^\alpha + |H_{23}|^2/d_{23}^\alpha)} < \frac{e^{R'(\tilde{\gamma}/\tau)}-1}{\tilde{\gamma}/\tau}, \quad (5.71)$$

where we have set $R' = R\ln(2)/\tilde{\gamma}$ as usual. If we now choose τ so that $\tilde{\gamma}/\tau \to \infty$ and $R'(\tilde{\gamma}/\tau) \to 0$ as $\tilde{\gamma} \to 0$, then we can mimic a high-SNR expression on the left-hand side of (5.71) and get R' on the right-hand side of (5.71). For example, we can choose (see [12])

$$\tau = \tilde{\gamma}^2, \quad R = \tilde{\gamma}^3 \text{ bits}, \quad R' = \tilde{\gamma}^2 \text{ nats}. \quad (5.72)$$

But then in (5.35) we must replace $g(\tilde{\gamma}, R)/2$ by R' and can use (5.36) to show that

$$\lim_{\tilde{\gamma}\to 0} P_o \approx (R')^2 d_{13}^\alpha (d_{12}^\alpha + d_{23}^\alpha)/2 \quad (5.73)$$

and therefore

$$R'(0, P_o) \approx \sqrt{\frac{2P_o}{d_{13}^\alpha (d_{12}^\alpha + d_{23}^\alpha)}}. \quad (5.74)$$

Bursty AF is therefore optimal at low SNR.

5.5.2 Bursty Compress-and-Forward

Our analysis of the orthogonal CF strategy showed that its outage event (5.40) implies the orthogonal AF outage event (5.35). Hence, bursty CF is also optimal at low SNR. Again, we remark that CF requires the relay and destination to have full CSIT and CSIR, respectively.

5.5.3 Decode-and-Forward

Finally, consider the selection DF strategy of Section 5.4.2.4 for which the rates are given by (5.45). Using the same steps as above, one finds

Table 5.3 Normalized outage rates at low SNR for $d_{uv}^{\alpha} = 1$.

Cooperative strategy	$R'(0, P_o)$	Comment	Extra CSI
Direct transmission	P_o	No diversity	
AF	P_o	No diversity	$\lvert H_{12}\rvert^2/d_{12}^{\alpha}$ to dest.
DF	$\sqrt{\frac{2}{3}P_o}$	Full diversity But suboptimal	
Bursty AF	$\sqrt{P_o}$	Full diversity	$\lvert H_{12}\rvert^2/d_{12}^{\alpha}$ to dest.
Bursty CF	$\sqrt{P_o}$	Full diversity	Relay CSIT $\lvert H_{12}\rvert^2/d_{12}^{\alpha}$ to dest.
Cut-set upper bound	$\sqrt{P_o}$		

that (see (5.47))

$$\lim_{\tilde{\gamma}\to 0} P_o \approx (R')^2 d_{13}^{\alpha}(2d_{12}^{\alpha} + d_{23}^{\alpha})/2, \tag{5.75a}$$

$$R'(0, P_o) \approx \sqrt{\frac{2P_o}{d_{13}^{\alpha}(2d_{12}^{\alpha} + d_{23}^{\alpha})}}. \tag{5.75b}$$

Thus, one loses rate due to the factor of 2 in front of d_{12}^{α}. The results are summarized in Table 5.3 where we chose $d_{uv}^{\alpha} = 1$ for all (u, v).

5.6 Multiplexing Gain for Wireless Networks

The results presented so far considered a single destination. Consider now a 2×2 wireless network with two source nodes and two sink nodes (see Figure 5.5). Source u, $u = 1, 2$, wishes to send a message W_u to sink node $u + 2$. We pose the question: can one achieve MIMO multiplexing gains in such cooperative systems?

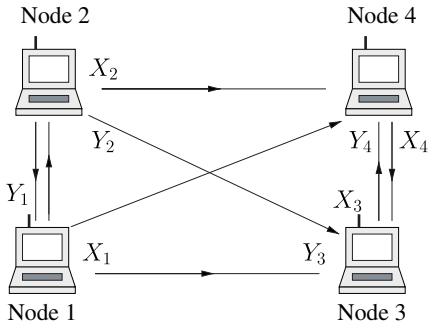


Fig. 5.5 Wireless network with two source nodes (1 and 2) and two sink nodes (3 and 4).

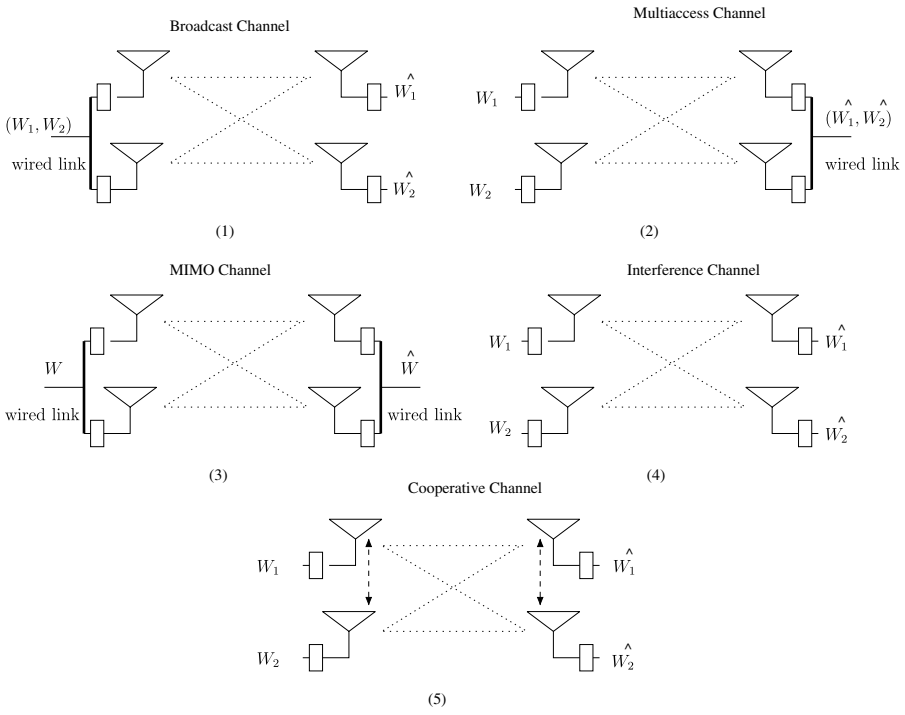


Fig. 5.6 Channel models with two transmitting and two receiving ends.

Before answering this question, we compare the model in Figure 5.5 with several other models having two transmit and two receive antennas (see Figure 5.6):

(i) MIMO broadcast channel with two transmit antennas;
(ii) MIMO multiaccess channel with two receive antennas;
(iii) MIMO channel;
(iv) interference channel.

For the MIMO channel shown in Figure 5.6(3), physical links exist between the two transmitting antennas and between the two receive antennas. At high SNR, we know that the best MIMO system rates behave as $2\log(\gamma)$. Similarly, the sum capacity of the two multiuser systems in Figures 5.6(1) and (2) also behave as $2\log(\gamma)$ [23, 34]. Thus, full cooperation at either the transmitters or the receivers suffices to

achieve the multiplexing gain of 2. On the other hand, the interference channel in Figure 5.6(4) has a multiplexing gain of 1 [77]. With respect to the amount of information shared at the transmitters and/or receivers, the cooperative system in Figure 5.6(5) falls between the MIMO and the interference channels.

Unfortunately, the sum-capacity multiplexing gain of the cooperative system in Figure 5.6(5) is only $r = 1$ even when all nodes have perfect CSIR and CSIT, perfectly synchronized transmissions, and full-duplex capabilities [78]. One can further show that this cooperative system achieves a much poorer diversity-multiplexing tradeoff than the MIMO system [196, 197].

5.7 Other Models and Methods

A variety of models has been considered to study cooperative capacity and diversity. For example, the *cognitive radio* channel shown in Figure 3.9 lets the source nodes cooperate by having one of the nodes aware of the message of the other node. The capacity region of this channel is known in some cases [89, 189]. The best cooperative schemes use both superposition coding and a sophisticated coding method known as *dirty paper coding*.

Cooperative diversity strategies can, of course, be extended to networks with many nodes. For example, each source–destination pair might use a two-phase *listen-and-transmit* scheme [113]. In the first phase, a source node transmits and several relay nodes listen; in the second phase, the relays amplify symbols or decode-and-forward. One can further identify scaling laws that characterize network performance in the limit of a large number of relays [20, 37, 58, 137, 138]. A two-hop cooperative scheme where the relays use AF achieves capacity as the number of nodes increases (see Section 4.2.2 and [58]). Cooperation also improves the energy efficiency as the number of nodes increases [37]; it allows for non-zero ergodic capacity for each source–destination pair when the number of relays scales as the square of the number of source–destination pairs; and it can "crystallize" a network, meaning that the source–destination links effectively appear as non-fading links [138].

Scaling law analyses can be done in several other ways. For instance, increasing the number of nodes in an area of fixed size increases the density of the network and one can achieve good scaling. The results are more pessimistic for *extended networks* where the node density is held constant while the area grows with the number of nodes. The throughput for n nodes decreases as $O\left((\log(n))^{-2\alpha}\right)$ in a two-dimensional extended network [127]. It is only after disconnecting some fraction of nodes that a constant growth rate can be achieved [41]. Interestingly, it suffices to disconnect an arbitrarily small fraction of nodes. An excellent review of the throughput scaling laws in *ad hoc* networks is presented in [192]. More recently, the scaling laws for both dense and extended networks have been characterized [145].

The traffic scenarios considered in this chapter have each message destined for a unique node, i.e., we considered unicast transmission. The multicast problem has a message destined for several destination nodes and is called a broadcast problem if the set of destination nodes includes all network nodes except for the source node. For both multicast and broadcast problems, cooperative strategies improve energy efficiency [132, 133] and network lifetime [134]. The asymptotic behavior of cooperative broadcast was analyzed in [170]. The simple two-phase protocol described above achieves the multicast capacity if the channels exhibit Rayleigh fading [96].

Finally, we note that *clustering* nodes at two locations lets one solve several relay problems. Examples of such results are listed below:

- CF and DF achieve capacity if the relays form a cluster with the destination and source, respectively [106]. When one set of relays is clustered with the source and another set with the destination, one can achieve MIMO-like rates. However, one cannot achieve the MIMO diversity-multiplexing rates [196, 197] unless the clustering gets tighter with SNR.
- Clustering increases the diversity gain of the systems with one source–destination pair [195].
- Cooperative schemes based on CF provide large throughput gains [92].

- If orthogonal channels are available between sources, one can exchange messages between the encoders to allow them to use dirty paper coding [87].
- For networks with n source–destination pairs, one can form hierarchies of clusters that cooperate and form virtual MIMO arrays. One can thereby achieve linear scaling of the total capacity in dense networks. In extended networks, the capacity scales as $n^{2-\alpha/2}$ for $\alpha < 3$, and $\sqrt{n}$ for $\alpha \geq 3$ [145].

6

Wireless Networking Protocols

6.1 Introduction

We have observed that cooperative networks enable a more complex set of interactions between the physical, link, and network layers. Transmissions are not point-to-point and routing is not store-and-forward. We need to distinguish between a *message* that an application wishes to communicate and the *data packets* that are transmitted in the network. At the outset, a message is a packet generated and transmitted by a source and addressed to a particular destination. Along the way, several intermediate nodes may contribute to the communication of the message to the destination, but each node may transmit its own unique data packets.

In a conventional network, a packet received in error at the PHY layer is simply discarded at the link layer. In a cooperative network, a packet received in error is not necessarily discarded; instead, such a packet may be saved by the link layer and subsequently combined with other received packets or perhaps even forwarded to the network layer as a packet with errors. Thus we distinguish between receiving a packet *unreliably* or *reliably*. Suppose a node identifies that a transmis-

sion is underway, acquires the timing of the symbol transmissions and attempts to detect what symbols are sent. If the resulting packet fails a CRC, then the packet is received unreliably. If the packet passes a CRC, then the packet is considered as having been received reliably. We say a packet is *decoded* as a synonym for received reliably.

Reliable communication of a message can involve both unreliable or reliable reception of data packets. In fact, Chapters 4 and 5 describe several ways in which a relay node generates a transmission that assists decoding at the destination even though the relay cannot decode the message itself. There are a number of ways to include these elements in the network protocol stack.

Before proceeding, we note that cooperative wireless protocols demand a tight synchronization of modems, decoders, link layer algorithms for packet combining, and MAC layer scheduling. We refer to these methods collectively as the PHY layer in the expectation that the methods will be integrated on a single physical device that appears as an interface to an operating system. By contrast, the network layer must support multiple device interfaces and will continue to be implemented as part of an operating system. A key issue is the partitioning of functions between PHY layer devices and network layer software.

6.2 Wireless Cooperation Issues

Networking traditionally defines the sequence of intermediate node transmissions as a *path* or *route*. One way to think about cooperative communication methods is to view a set of n nodes participating in delivering packets from a source to a destination as a *multi-terminal link*. We refer to this generalization as a *cooperative link*, or simply a link (the terminology *feedforward flowgraph* is used in [65]). A cooperative link $\mathcal{L}$ is a set of nodes employing coordinated actions to deliver messages reliably from a source to a set of one or more destinations. Suppose that a destination node d is the final recipient of these packets. If node d is a relay for another link, we assume that delivery of a packet for this second link to node d represents a "renewal" point in the packet delivery process. That is, node d may be the source of another cooperative link $\mathcal{L}'$ in order to deliver the packet to yet another

destination d'. However, delivery of the packet to node d' on link $\mathcal{L}'$ might not employ the transmissions associated with the link $\mathcal{L}$. In this case, a route becomes a sequence of one or more (cooperative) links.

Calling such a collection of coordinated nodes a link is consistent with the embedding of cooperative protocols in the PHY layer of the simplified IP protocol stack. This design choice would preserve the current model of mobile networks in which the PHY layer is an interface queue for IP packets and routing at the network layer is based on IP packets. On the other hand, this approach does imply an expansion in the activities of the link layer and perhaps the MAC sublayer. Now the link layer communicates with multiple network nodes in establishing a link. Moreover, a specification of the cooperative link indicates the sequence of transmissions by nodes participating in the cooperative link. That is, cooperative link establishment includes much of the functionality of routing, albeit just for nodes in a local neighborhood.

An alternate approach is to make the link layer responsible for identifying packet transmissions and whether those packets are received reliably. In this case, the network layer takes responsibility for coordinating reliable and unreliable transmissions of multiple nodes. The sequencing of such transmissions generalizes traditional store-and-forward routing.

The above approaches, cooperation via the link or network layers, differ in the network architecture to implement the following tasks:

- **Message Decoding:** using all received packets to decode,
- **Cooperative Link Establishment:** specifying a procedure to coordinate the transmissions of multiple nodes with the objective of reliable message decoding at the destination(s),
- **Cooperative Routing:** configuring a sequence of cooperative links as a route.

Message decoding is a link layer task while cooperative routing is a job for the network layer. Link establishment, however, could be executed in the link layer, the network layer, or even the MAC sublayer. A cross-layer implementation might be appropriate. What is certain is that the control signaling must have low rate relative to the gains of cooperation.

When we examine some proposed cooperative protocols in Section 6.6, we will see that overhead raises a number of implementation issues.

6.3 Networking Mechanisms

We have identified the primary tasks of cooperative networks, namely message decoding, cooperative link establishment, and cooperative routing. These tasks must be completed by peer protocols at the link layer, MAC sublayer, and network layer. In any wireless network, signaling channels are defined to facilitate these protocols. For example, a packet header that identifies the source, destination and sequence number of each packet represents an embedded control channel at the link layer. Control packets such as RTS and CTS in 802.11 systems are a type of control channel signaling at the MAC sublayer. Similarly, the *ad hoc* routing protocols DSR and AODV employ RREQ and RREP packets for network layer signaling. Alternatively, these control channels may be embedded in the PHY layer, e.g., special code sequences in 2G and 3G CDMA cellular systems or specific bits in frame headers.

In all wireless systems, the design of the signaling channels is an important engineering problem. However, from the point of view of system design, the key consideration is that the information exchanged for link establishment must be small relative to the data that is subsequently transmitted. Assuming such reasonable control mechanisms exist, the question is: How are network resources allocated?

With respect to the design of network protocols, some preliminary considerations are as follows:

- Protocols are needed to allow a source and multiple relays to decide when and how to establish a cooperative link.
- Although the payload of a packet transmission may be received unreliably, a specified set of nodes in a link must recognize and act upon this transmission. This requires a method for a node to signal a packet transmission. Conceptually, this could be done using a control message akin to RTS that is transmitted at sufficiently low rate to enable decoding by all nodes. However, a packet header must still contain an acquisition sequence, i.e., a sequence of known bits that

allows receiving nodes to acquire the symbol timing of the arriving packet. This header might as well be extended to include a low-rate data field that associates the packet with a particular message. Perhaps the header also identifies the nodes that are expected to decode the message in question.

- To handle repeated unsuccessful decoding, a mechanism is needed to allow a node to give up trying to decode and instead send a NAK. These NAKs may also be used to reconfigure a cooperative link.
- Multi-terminal link optimization will include distributed scheduling of relay transmissions, coding and power control suitable for cooperation, as well as adaptation of the forwarding strategy, e.g., AF, CF, or DF.
- Routing protocols must account for cooperative links. Suitably abstracted routing metrics must be developed.

In the following, we consider node and network architectures that begin to address the above requirements. In order to limit the discussion, we remark that although full duplex relaying and beamforming (coherent transmission) from multiple nodes benefit cooperation [33, 58], realizing these modes of operation in mobile wireless settings is at the very least challenging and arguably impossible. We thus focus on the implementation of half-duplex diversity and network coding mechanisms. In the following, we examine how cooperative communication can be implemented in the physical and link layers. We start by reviewing a conventional PHY layer architecture in Section 6.4. We then examine a cooperative PHY layer architecture in Section 6.5.

6.4 Conventional PHY Layer Architecture

A conventional architecture for the physical and link layers has a demodulator and decoder, as shown in Figure 6.1. Packet reception occurs in the DEMOD module that demodulates, samples, and puts out quantized real- or complex-valued *soft symbols*. These symbols are fed to the DECODER module to produce a data packet. If the packet has a valid CRC, then it passes to the network layer; otherwise it is discarded.

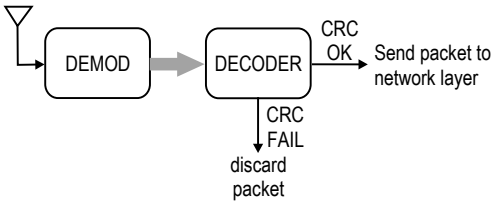


Fig. 6.1 A Conventional Physical/Link Layer Architecture. The thick arrow between DEMOD and DECODER denotes a high-rate soft symbol interface.

In the classic multihop approach, the network layer determines the future of each decoded packet. At the final destination, the network layer passes the packet to higher layers. At an intermediate node on a path, the network layer resolves the next node on the path, constructs a packet with the same data but a modified header that indicates the next node as the intended recipient, and passes this modified packet back down to the PHY layer transmitter.

The demodulator and decoder can be tightly coupled since the decoding process handles one packet at a time. This is significant because the interface that passes soft symbols is a high-rate interface. In particular, consider a coded system with M-ary modulation. A sequence of k input bits is coded at rate R into $n = k/R$ code bits. Next, $\log_2 M$ code bits at a time are passed to the modulator. This leads to the reception of a complex-valued soft symbol at the DEMOD. The real and imaginary components of this complex symbol must each be quantized to b bits. A soft symbol output is thus represented by $2b$ bits. The number of bits b must be chosen sufficiently large so that quantization noise is negligible.

The above implies that k input message bits result in

$$\frac{n}{\log_2 M} 2b = \frac{2b}{R \log_2 M} k \tag{6.1}$$

bits being passed through the receiver's soft symbol interfaces. This represents a bandwidth expansion of

$$G = \frac{2b}{R \log_2 M}, \tag{6.2}$$

which can be substantial. Generally $R < 1$ and $b = 10$ bit quantization is typical [55]. For example, under QPSK signaling with $M = 4$, a rate

$R = 1/2$ code and $b = 6$ bits quantization, the bandwidth expansion factor is $G = 12$. Values of G on the order of 10 or 20 are typical. Communication at a message bit rate of 20 Mb/s might thus require soft symbol interfaces that operate at 200–400 Mb/s. Note that it appears that transmitting with a large constellation size M reduces the expansion factor. However increasing M requires finer quantization and thus larger b. Typically, increasing M is a losing proposition in terms of the expansion factor G.

While this discussion of soft symbols may seem overly detailed, minimizing the use of soft-symbol interfaces is an important consideration for hardware designers. We will see that this consideration influences the partitioning of tasks in the network layer.

6.5 Cooperative PHY Layer Architecture

In a cooperative communication system, the DECODER is replaced by a *cooperative decoder*, often a substantially more complex device with additional storage, multiple soft-symbol interfaces, and, in some instances, a more complex interface to the network layer. The actions of the cooperative decoder depend on whether a node plays the role of an intermediate relay or a final destination of a cooperative link. A cooperative PHY layer implementation must support both roles in a single device. We next consider the role of a final destination and describe issues related to a relay role in Section 6.5.2.

6.5.1 Cooperative PHY Layer: Destination Node Receiver

Consider the orthogonal relaying strategies described in Sections 5.4 and 5.5. Such strategies perform *packet combining*, i.e., they decode a message using the possibly unreliable reception of multiple transmitted packets. Packet combining is commonly used to improve the performance of Hybrid-ARQ [120], and can be based on either hard decisions [169] or on soft (floating point) channel outputs [29].

A cooperative diversity PHY layer receiver architecture is shown in Figure 6.2. As in a conventional receiver, demodulation and sampling occurs in the DEMOD module, yielding a soft symbol output stream. What's new is that the soft symbols of previously received packets are

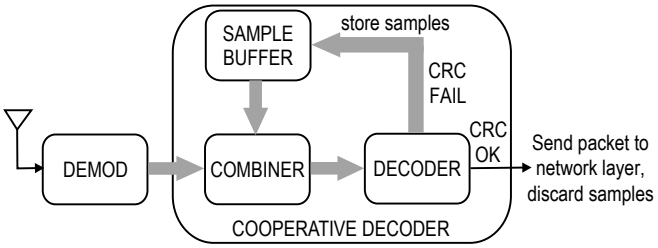


Fig. 6.2 A Cooperative Diversity Decoder Architecture. Thick arrows denote high-rate soft symbol interfaces.

stored in a SAMPLE BUFFER and a COMBINER merges the stored packets with the newly received packet. The precise action of the combiner depends on the type of cooperation.

When diversity is achieved through repetition coding, the new packet is simply a copy of a previously received packet and DEMOD performs maximal ratio combining (MRC) on the soft symbols of the packet copies. In this case, the SAMPLE BUFFER can store the soft symbols corresponding to a linear combination of past received packets, regardless of how many packet copies are received for a particular message. This applies to both the AF and DF orthogonal relaying repetition-coding schemes described in Section 5.4.2.

Observe that in Figure 6.2, there are four interfaces subject to the bandwidth expansion associated with passing soft symbols. The storage requirements of the sample buffer are subject to the same expansion factor. We further observe that a MAC protocol will result in multiplexing and asynchronous transmissions of multiple messages. Even a single cooperative link might carry traffic corresponding to multiple message streams. A cooperative node must maintain a SAMPLE BUFFER for each in-progress message communication.

Consider next the CF and incremental relaying strategies described in Section 5.4.2. These strategies are non-regenerative, as is AF, and the new packet may contain new coded symbols. Similarly, non-regenerative DF with the relay using an additional codebook would also result in new coded symbols. These approaches require the destination to store soft samples for all received packets for decoding. The storage requirements of the SAMPLE BUFFER thus increase linearly

with the number of received packets for that message. Furthermore, the DECODER component becomes considerably more complex because decoding is based on multiple transmissions from multiple transmitters employing multiple codebooks.

By comparison, the non-orthogonal strategies of Section 5.4.3 appear to be simpler since a message results in a received signal at the destination over a continuous time interval. In this case, the conventional receiver of Figure 6.1 that resolves a codeword from a vector of soft samples should be sufficient. In particular, the high-speed "loop" associated with the COMBINER, DECODER, and SAMPLE BUFFER can be omitted. On the other hand, the structure of the received signal is altered considerably when a relay begins transmitting. In particular, the received signal consisting of the initial source transmission and the subsequent combined transmission of the source and relay is based on multiple transmitters and multiple codebooks and bears resemblance to the received signal of non-regenerative orthogonal relaying. In this case, the structure of the DECODER element will be similarly complex.

In any event, for both orthogonal and non-orthogonal relaying strategies, the DECODER processes a data structure of soft samples and makes bit decisions. If the CRC check is satisfied, the decoded bits pass to the higher layers and the soft samples in the SAMPLE BUFFER are discarded. If the CRC fails, the cooperative decoder stores the soft symbols in the SAMPLE BUFFER and waits for additional received packets. In the absence of a mechanism yielding additional packets, the soft samples are discarded.

6.5.2 Cooperative PHY Layer: Relay Node

We next examine the cooperative PHY layer for a relay node. A conventional store-and-forward relay node passes reliably received packets to the network layer. The network layer reads the packet header and determines whether the packet should pass to higher layers or whether it should be forwarded. If the packet is to be forwarded, it is passed (typically with a modified header) back down to the PHY layer transmitter. This approach extends naturally to DF relaying. The key observation is that all data packet transmissions are generated by the network layer.

The situation is less clear with AF or CF. Under AF, the relay transmits a vector of received soft samples. The transmission is presumably prefixed with an appropriately descriptive header. This sort of header information is normally generated at the network layer, so it would be consistent with layering practice for this vector of unreliable soft symbols to be passed to the network layer that then constructs a "soft symbol packet" for transmission. While this may sound logical, it incurs the substantial penalty of a high-rate soft symbol interface between the PHY and network layers.

Similar issues arise for CF where the PHY layer receives a transmission block b from which it deduces a quantization codebook index s_b (see Section 4.2.4). If the relay-destination channel is very good, the quantized observation may closely approximate the received signal. In this case, forwarding even the index s_b to the network layer may result in a bandwidth expansion of the PHY-network layer interface comparable to that for AF. Avoiding this potential bandwidth expansion in the PHY-network layer interface is the chief reason to embed some network-layer functionality in the PHY layer.

6.6 Proposed Cooperative Diversity Protocols

Numerous cooperative strategies have been proposed in the literature and many of these have been examined in Chapter 5. However, only a handful of papers have considered cooperative signaling in the context of practical networks based on existing PHY layer signaling standards and supporting multiple source–destination data streams.

In this section, we examine four cooperative protocols. Each offers candidate solutions to the problems of message decoding, cooperative link establishment, and cooperative routing. Although each solution is based on a cooperative diversity mechanism, each represents different tradeoffs between complexity, backward compatibility with the 802.11 MAC, and performance.

The first three proposals (CoopMAC, HARBINGER, and VMISO) employ reliable-reception relaying in which the transmission of a message packet by a source is assisted by one or more relays that transmit additional coded packets toward the destination. In particular, a coded

packet sent by a relay is based on a single message packet from a source. Packet decoding may be unreliable and message decoding is based on the soft combining of packets as described in Section 6.5. We describe these protocols in order of estimated increasing complexity.

A fourth protocol, COPE, uses a network coding approach in which intermediate nodes transmit reliably received coded packets that are combinations of multiple message packets, often from multiple sources. This approach has a relatively simple PHY with traditional hard decision decoding of each individual received packet. The diversity gain arises from broadcasting to multiple receivers. In processing a coded packet that is a combination of multiple message packets, what each receiver decodes depends on the prior packets that receiver has decoded. Comparisons between these protocols follow in Sections 6.7 and 6.8.

6.6.1 CoopMAC

A MAC protocol known as *CoopMAC* that allows for relaying in IEEE 802.11 networks was proposed in [124] and further characterized in [122, 123]. The protocol extends the distributed coordination function (DCF) to incorporate a relay node in data exchange. The RTS/CTS protocol is extended with a "Helper-ready To Send" (HTS) message. The source decides whether to use a relay based on cached information that indicates if the relay reduces the air time needed to deliver the packet to the destination. After the source sends an RTS message, the relay can reply with an HTS message. If this HTS message is heard by the destination, then the destination replies with a CTS message that reserves time for a two-hop transmission. If the source receives both CTS and HTS messages, the data packet is sent to the relay that then forwards to the destination. If the source hears only the CTS message, then the source sends directly to the destination.

When the destination receiver has an architecture like that in Figure 6.2 and is capable of diversity combining, the relay uses a cooperative coding scheme such as that in [163]. However, CoopMAC is backward compatible with existing 802.11 radios in that the same protocol messages and cumulative air time metric can be employed even if the receiver is not capable of diversity combining. In this case, air

time calculations are based on the variable rate transmission features of 802.11. The MAC-layer forwarding is similar to rDCF in [202], but the two protocols differ in how to set up the relay connection.

6.6.2 HARBINGER

HARBINGER (Hybrid ARQ-Based INtercluster GEographic Relaying), a MAC layer protocol based on Hybrid-ARQ with incremental redundancy, has been proposed in [199]. A codeword consists of B codeword blocks and time is divided into B slots so that one codeword block requires one slot. For DF, the source transmits message block $b = 1$ in slot $b = 1$. At the start of slot b, $b > 1$, there is a set of nodes $\mathcal{D}(b)$ that have already decoded the message and thus know all past codeword blocks. Hence codeword block b, $b = 1, 2, \ldots, B$, is transmitted in slot b by nodes in the set $\mathcal{K}(b)$, $\mathcal{K}(b) \subseteq \mathcal{D}(b)$. An outage occurs if the destination cannot decode correctly after slot B.

A key routing protocol question is: What nodes in $\mathcal{D}(b)$ should transmit in slot b? HARBINGER's solution bears similarity to the Geographic Random Forwarding (GeRaF) protocol [203] where nodes use GPS receivers to identify the relays' positions relative to the destination. In HARBINGER, the node in $\mathcal{D}(b)$ geographically closest to the destination transmits in slot b. This node is identified through a contention phase for the ACKs in slot $b - 1$. Note that other metrics, such as highest instantaneous SNR at the destination, could also be used to choose a transmitter for slot b.

6.6.3 VMISO

Recently, the paper [85] introduced a framework for cooperative routing in which clusters of nodes near each transmitter form a virtual multiple-input single-output (VMISO) link to a receiver. In this VMISO protocol, cooperative link establishment and route configuration is initiated by using conventional network-layer routing protocols. The authors proposed the following stages:

- **Discovery of the Primary Path:** A traditional route or path is specified as a sequence of nodes by using an existing

ad hoc routing protocol such as AODV or DSR. For convenience, we number these nodes $1, 2, \ldots, L$ with the source s as node 1 and the destination d as node L.

- **Finding a Cooperative Path:** This step, called "Selecting the relay nodes," identifies a subsequence $u_1, \ldots, u_j$ of nodes on the primary path. These nodes create a sequence of cooperative links from a source to a destination that we call the *cooperative path*.

Once the primary path has been found, the goal of the protocol is to extract a cooperative path that bypasses hops in the primary path. The key idea is that a VMISO transmission enables the creation of links over distances that would be untenable using non-cooperative point-to-point signaling. In the following, we describe a sequential approach to the creation of VMISO cooperative links. Although the VMISO protocol does not require this sequencing, it serves to simplify our discussion.

Given that nodes $u_1, \ldots, u_j$ of the VMISO cooperative path have been identified, node u_j seeks to establish the next hop in the cooperative path through an *anycast* mechanism. That is, node u_j sends a packet addressed to a subset of nodes that are downstream on the primary path. These nodes know their hop count back to node u_j and reply in order of decreasing hop count. In particular, the most downstream node that hears the anycast and believes it is a good candidate for a cooperative link from node u_j will become node u_{j+1} on the cooperative path.

To set up a particular cooperative link $\mathcal{L}$ with a "source" node u_j, and intermediate "destination" u_{j+1}, node u_j multicasts a modified *local* RTS message to a random subset of neighbor nodes as an invitation to participate as a transmitting relay for $\mathcal{L}$. As the multicast message contains the identities (IDs) of all nodes in the chosen subset, the invited nodes can reply to node u_j in the order of their ID numbers. The replies include pilot tone transmissions that enable node u_{j+1} to estimate the channel to each relay. The participating relays then cooperatively signal an *M*-RTS message to node u_{j+1} to establish the VMISO link to the destination. Node u_{j+1} responds by establishing its own reverse MISO link. That is, node u_{j+1} signals a *local* CTS mes-

sage to a random subset of its neighbors. These neighbors reply to node u_{j+1} in ID-order with pilot tone ACKs that allow node u_j to estimate the channels from each of the reverse link relays. This is followed by the reverse transmission of the *M*-CTS message by the receiver cluster to node u_j. After receiving the *M*-CTS message, node u_j can send a data packet. This is accomplished in two steps: a data packet is sent from u_j to the transmitter cluster. The transmitter cluster then forwards to the destination over the forward-direction VMISO link. The receiver can then respond by forwarding its response to the reverse MISO cluster which then signals node u_j through the reverse MISO link.

The benefit of the VMISO approach derives from the coding gains associated with space–time coding, translating into longer transmissions and fewer hops per route. Also, the increased diversity makes the link less likely to go down. This reduces the frequency of route discovery, which is a major source of overhead in high-mobility networks.

6.6.4 COPE

COPE was introduced in [90, 91] as a way to exploit network coding in wireless environments. Although COPE is described as a coding sublayer between the MAC and network layers, we view COPE as a MAC layer extension in the following discussion. While many ideas in COPE can be applied to a variety of wireless environments, most consideration has been given to an 802.11 implementation.

At the PHY layer, COPE is perhaps the simplest cooperative strategy. The elementary COPE communication model is a broadcast channel, i.e., there is a set of receiver nodes that can decode each transmitted packet. The transmitter and its associated receivers are referred to as a hyperlink. When a hyperlink receiver node receives a packet in error, the packet is simply discarded while packets received correctly are released to the higher layers. Because of the broadcast nature of the wireless medium, a node may be a receiver in several hyperlinks in addition to overhearing packet transmissions intended for others.

In COPE, the message packet of a source is known as a *native packet* and the packets transmitted by intermediate nodes are called *coded packets*. A receiver node maintains a *packet pool* of native packets that

it has decoded and transmits *reception reports* that allow hyperlink transmitters to code opportunistically by tracking the packet pools of nearby receivers. While network coding permits a variety of ways to construct coded packets, COPE simply uses the exclusive-or (XOR) operation. That is, a coded packet is an XOR-combination of multiple native packets.

For example, consider the setup of Figure 6.3 (see also Figure 4.20). If node 3 knows that nodes 1 and 2 have decoded the respective packets p_1 and p_2, then node 3 sends the XOR-coded packet $p = p_1 \oplus p_2$. When nodes 1 and 2 decode packet p, node 1 can form $p_2 = p \oplus p_1$ while node 2 can form $p_1 = p \oplus p_2$ (see Section 4.2.8). Thus, one coded packet p has enabled each receiver to recover unique native packets, and we have a bandwidth savings by reducing the number of transmissions needed to deliver packets across the network. In this example, it may not be apparent that COPE is a cooperative protocol since the relay (node 3) is merely optimizing its own signaling strategy. However, in more complicated network scenarios, COPE becomes cooperative in that a node u may overhear a packet p and then send a coded packet that

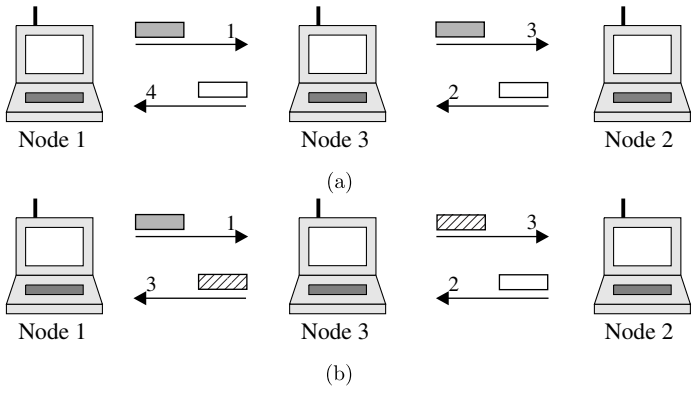


Fig. 6.3 Traditional Router vs. COPE Router: Node 3 serves as a router/relay for packets exchanged between nodes 1 and 2. (a) With node 3 as a traditional router, four transmissions are used to deliver a packet p_1 from node 1 to node 2 and a reverse packet p_2 from node 2 to node 1. (b) With node 3 as a COPE router, nodes 1 and 2 forward packets p_1 and p_2 to node 3. In the third transmission, node 3 broadcasts the XOR-coded packet $p = p_1 \oplus p_2$ to nodes 1 and 2. Node 1 forms $p_2 = p \oplus p_1$ while node 2 forms $p_1 = p \oplus p_2$, thus delivering both packets in three transmissions.

delivers p to its intended next-hop node, even though node u was not actually on the nominal route for packet p.

The basic idea behind COPE can be implemented in a variety of ways. COPE assumes, but does not specify, a routing protocol so that intermediate nodes use a method, such as DSR-type source routes in the packets or local routing tables, to identify the next-hop intended receiver for each packet. The COPE implementation employs opportunistic principles:

- **Opportunistic Listening:** Nodes operate in promiscuous mode and listen for all overheard packets. These packets are then held in the packet pool for a limited time T.
- **Opportunistic Coding:** Nodes aim to maximize the number of native packets delivered in each transmission. Suppose a hyperlink transmitter has native packets $p_1, p_2, \ldots, p_n$ intended for nodes $u_1, u_2, \ldots, u_n$. The transmitter sends

$$p = p_1 \oplus p_2 \oplus \cdots \oplus p_n \tag{6.3}$$

  only if each intended receiver u_j can decode its native packet p_j from the coded packet p.

Ideally, reception reports identify each neighbor receiver's packet pool. However, in the event of network congestion, reception reports may lag behind packet deliveries. In this case, a COPE sender may use routing topology information such as link delivery probabilities to estimate the neighbor state.

For 802.11 networks, there are a number of implementation details. In particular, COPE adds a packet header that includes an identifier (hashed source IP address and IP sequence number) for each native packet in an XOR-coded packet. Other header components include reception reports and neighbor ACK mechanisms. While COPE is based on broadcast transmission, the 802.11 broadcast mode does not provide a link layer ARQ that ensures reliable packet delivery. Hence COPE uses a *pseudo-broadcast* in which packets are sent in unicast mode addressed to a particular next-hop receiver and the COPE header identifies the other next-hop receivers. The unicast-addressed receiver

can use the 802.11 fast ACK to return an ACK while the other next-hop receivers must use the reception report mechanism.

6.7 Experimental Performance Comparisons

A reader may wish to compare the above proposals and ask: Is one of them truly superior? We note that encouraging results have been reported for all three protocols:

- **CoopMAC:** In [122], it was reported that CoopMAC provides network capacity gains ranging from 40% to 60% relative to 802.11g, with the higher figure requiring receiver combining. Improvements in channel access delay and user energy efficiency on the order of 20%–40% were also observed.
- **HARBINGER:** In [199], HARBINGER was found to provide a power savings ranging from 7 to 20 dB over direct transmission and from 1 to 3 dB over multihop routes. The savings increase as the number of codeword blocks increases, as one would expect [25].
- **VMISO:** In static networks of 200 nodes and CBR traffic, it was observed in [85] that VMISO increases throughput from 30% to 100%. VMISO also decreases average delay by roughly 50%. The largest throughput improvements occur at light loads with few sessions. Similar improvements are reported under scenarios with mobility. The authors argue that the diversity afforded by a VMISO link improves the link lifetime. The resulting savings in the overhead of route repair and discovery more than offset the overhead in setting up the VMISO links.
- **COPE:** Experimentation with a 20 node testbed showed that COPE performance depends on the link topology and traffic flow characteristics [91]. For example, in a congested wireless network serving UDP flows, COPE provides a throughput increase of a factor 3 to 4. In other circumstances, the improvements are less dramatic. A mesh network connected to an Internet gateway yields throughput increases of 5%–70%, with the gains increasing with the fraction of down-

> link traffic. In general, two kinds of gains are noted: *coding gains* in which COPE delivers packets with fewer transmissions and MAC *gains* in which COPE compensates for packet dropping that would otherwise be induced by the fairness properties of the 802.11 MAC.

Head-to-head comparisons are difficult because each protocol requires evaluation in a custom environment with a mix of analytic modeling, simulation, and testbed evaluation.

Although CoopMAC could be implemented in an *ad hoc* network, published results have employed MAC channels with a group of stations transmitting to an access point. A key barrier is that a cooperative diversity receiver cannot be implemented without low-level access to the wireless interface. Thus, the evaluation of CoopMAC employs a custom simulator "to faithfully model all the critical MAC and PHY layer features of IEEE 802.11" [122].

Since HARBINGER makes no particular assumption on the routing algorithm, none was tested. Instead, instructive predetermined configurations of nodes and routes were used, with an emphasis on nodes or clusters of nodes along a line. The MAC protocol was not explicitly modeled and interference from other network links was approximated by noise. Finally, signaling overhead that would have been incurred by a practical routing protocol was neglected.

Using an OPNET simulation, VMISO received an extensive performance evaluation. However, symbol-level effects were modeled coarsely: for each cooperative link, each transmitter u at distance d_u from the receiver was characterized by distance-dependent attenuation and a Rayleigh fading variable α_u. The received SNR was $\sum_u \alpha_u / d_u^4$, where the sum is over the transmitters participating in the VMISO link. A packet is decoded correctly if the received SNR is above a threshold $(\mathrm{SNR}_{\mathrm{TH}} - D)$ dB where $\mathrm{SNR}_{\mathrm{TH}}$ is a SISO threshold and D is a diversity gain. Interference was neglected under the assumption that a node can use RTS/CTS to gain access to the channel while silencing potential interferers. This abstraction allowed experimentation in a network with 200 nodes and various mobility models.

COPE does not employ PHY layer cooperative signaling, a simplification that enabled a testbed implementation. Although limited to a 20 node network, the COPE testbed evaluation was perhaps the most complete. The experimental results highlighted how system performance depends on the network configuration and on the traffic induced by higher layer protocols.

Unfortunately, a unified framework for comparing cooperative protocols is unlikely to emerge. For protocols that implement PHY layer cooperation, symbol-level simulations are computationally too tedious to permit generating sufficient numbers of packets for even a single network session, much less dozens or hundreds of sessions. On the other hand, it is non-trivial to accurately abstract PHY layer effects. The modeling task becomes especially difficult if interference from simultaneous transmissions cannot be neglected.

6.8 Design Comparisons

The CoopMAC, HARBINGER, VMISO, and COPE protocols offer very different mechanisms for relay selection and packet delivery. However, they do share some common features:

- Receiver cooperation among multiple nodes for packet decoding is not used.
- Cooperation is handled *below* the network layer (see Figure 2.1).
- Establishing a multihop route from a source to a destination is handled at the network layer.

Handling cooperation below the network layer implies that the additional storage and computation of cooperative relaying falls to the MAC layer. For example, the CoopMAC and VMISO protocols both require the MAC layer to track which nodes in its neighborhood are useful relays. In effect, a form of local routing table is stored at the MAC layer. In addition to caching information regarding potential relays, relay connections in CoopMAC, HARBINGER, and VMISO require the following data structures at the MAC layer:

- At the source: a table listing relays for each destination.

- At a relay: a table listing those links or routes that are being helped.
- At the destination: a table of relays for each incoming link on a cooperative route. This information prevents the MAC layer from prematurely sending a NAK.

In comparison, COPE avoids these data structures at the expense of storing a packet pool and lists to track the packet pools of neighbor nodes.

Additional differences emerge in how the interactions with routing are handled. Each protocol, either implicitly or explicitly, starts with a primary path from a source to a destination established by a network-layer routing protocol. CoopMAC simply accepts the primary route and seeks to improve the diversity on each link by adding a relay. Given a primary path, VMISO then extracts a path with fewer hops but longer distance per hop. The designers of HARBINGER endorse cross-layer routing methods that employ side information such as geographic proximity or channel quality to the destination. This side information is used to construct a route (in fact, a sequence of transmissions) dynamically as codeword blocks are transmitted and intermediate nodes decode a message. HARBINGER selects the next transmitter among those nodes that have decoded the message by a contention resolution process that favors nodes according to the side information. For example, with geographic side information, nodes closer to the destination are given higher priority for transmitting the next codeword block. This approach bears similarity to the anycast employed by VMISO for route configuration, especially if HARBINGER were to employ as side information the hop count to the destination, as derived from the primary path found by a conventional routing protocol.

For all four protocols, the creation of a cooperative route as a sequence of cooperative links is based on first using a conventional method to find a primary path followed by MAC-layer methods to construct cooperative links over or within the primary path. This two-step procedure is not guaranteed to find the best path. In particular, CoopMAC forfeits the option to create longer distance hops. While VMISO and HARBINGER can construct long hops, these hops still depend

on the heuristic initial choice of the conventional primary path. In the same way, "routes" for COPE, i.e., sequences of specified next-hops for delivery of a packet, are constructed using an ordinary routing protocol, although COPE uses those routes more efficiently through opportunistic listening and coding.

The local two-step route optimization describes the difference between relaying and routing. Both relays and routers are forwarding source messages but they help at different timescales. A router keeps its forwarding table on the network layer, while the relay has its table on the MAC layer. Since route discovery and route repair are slow processes that generate a lot of overhead, the routing algorithm cannot adapt to rapid changes in the network due to fading or high mobility. However, given a route that was formed based on the past network state, variations in network topology can be accommodated by using intermediate nodes that overhear the transmissions and act as relays. In that sense, relaying and routing offer complementary ways to improve network performance.

6.9 A Contrarian View

In this text, we have observed that cooperation enables non-trivial network performance benefits. But one may wonder if cooperative communications will be implemented in mainstream networking technologies or if cooperation will remain just a research curiosity. From the perspective of practice, a contrarian can identify many potential barriers to implementation. We examine two such barriers.

6.9.1 Channel Variation Speed

> *Coherent combining will never work because distributed phase synchronization is too difficult.*

This statement is likely valid in mobile environments, since phase synchronization requires tracking movements to within a fraction of a wavelength. However, for fixed wireless systems with a dominant line-of-sight path, this issue needs further investigation. Another potentially feasible scenario would be mixed wireline/wireless networks in which

multiple base stations are connected by a fiber-optic backbone that enables sufficient information exchange for coherent transmission.

> *Cooperative diversity mechanisms are useful only if a node is unable to exploit the temporal variation of the channel. In typical mobile environments, the time variation is more than sufficient for a node to achieve diversity gains using coding and/or ARQ without employing a cooperative relay.*

The main issue is still the speed of channel variations. If the channel varies sufficiently quickly, the value of cooperation seems to diminish. However, for applications requiring small delay, the value of cooperative diversity is easily recognized.

6.9.2 Processing Energy

> *The energy consumption in decoding overheard packets is such that no relay strategy beyond basic forwarding makes sense for energy-constrained nodes.*

Before directly addressing this issue, we first observe that the benefits of cooperation go beyond reducing transmit power. For example, COPE increases throughput by using channel bandwidth more efficiently. In high load scenarios, these benefits should not be ignored, even if energy consumption for listening to overheard packets increases.

Returning to the issue at hand, the benefit of reduced transmit powers must be balanced against the receive energy costs at the relays. To examine this issue, we compare the data-handling capabilities of some familiar devices and the energy costs that are incurred. These comparisons are based on current practices and technology for transceiver design, as opposed to any fundamental limits on the energy cost of computation. Nevertheless, the following discussion highlights that receive energy is an issue meriting further study.

We begin with the ubiquitous mobile phone that advertises a talk time on the order of 4 h. As the optimized sleep mode of the phone enables a battery lifetime of several days in the absence of talk, we can

conclude that the energy used for non-communication tasks is negligible. Thus, in the simplest analysis, a phone can consume its battery in cooperation if it spends 4 h relaying the packets of other nodes.

Continuing, consider a Motorola RAZR with a 3.7 V battery rated at 740 mAh that stores 9.8 kJ. During a call, the phone is either transmitting and/or receiving data bits at roughly 10^4 b/s. In fact, the phone may be transmitting and receiving simultaneously but we ignore this factor of two since it is roughly cancelled by a 40% voice activity factor [63]. Thus, in 4 h (14,400 s), 1.44×10^8 bits (18 MB) are communicated. In using 9.8 kJ, the energy cost is 544 J/MB.

Now consider the 802.11 WLAN interface. An Atheros whitepaper [10] found that typical WLAN interfaces consumed 2–8 W while actively communicating. As the actual transmit power of a 802.11 device is in the range of 20–100 mW, signal transmission uses only a small fraction of the power budget of even a 2 W WLAN interface. That is, *the power required for signal processing dominates the transmit power* and thus the power consumption is roughly the same for transmission and reception. In fact, this should not be surprising as the energy consumption of receiver processing is closely tied to the bandwidth expansion associated with the receiver's soft-symbol interfaces as discussed in Section. 6.4. In any event, it follows that a dedicated WLAN interface communicating at 3 MB/s consuming 3 W power will operate at 1 J/MB.

Mobile phones and 802.11 laptops represent two extreme points in mobile communication. The low energy/bit efficiency of the mobile phone derives from communication distances on the order of 1 km, roughly 10–100 times typical WLAN communication distances. With an $\alpha = 3$ path-loss exponent, the resulting transmit energy per bit for the mobile phone is higher by a factor ranging from 10^3 to 10^6. In the case of a cellular phone, a mitigating factor is that a cellular base station can use a large directional antenna to reduce the relative loss by perhaps a factor of 10 or more. Thus, our initial estimate that a cellphone may require 500 times the Joules/bit of a WLAN interface is in the right ballpark.

The cellphone's dramatically higher per-bit energy costs arise because cellphones and WLAN devices operate in different regimes.

The required transmit power dominates a cellphone's energy budget, while signal processing dominates for WLAN devices. These power consumption regimes are well recognized. In the context of Berkeley motes, [72] models the energy cost per bit for a reliable 1 Mb/s link over a distance d with path-loss exponent $\alpha = 2$ by a transmitter cost of

$$\mathcal{E}_{tx} = \mathcal{E}_t + \mathcal{E}_{pa} d^{\alpha} \quad \text{J/MB}, \tag{6.4}$$

where $\mathcal{E}_t = 0.36$ J/MB is the energy dissipated in the transmitter electronics and $\mathcal{E}_{pa} = 8 \times 10^{-5}$ J/m^2/MB scales the required transmit energy per bit. In addition, as signal detection is more complex than signal synthesis, the receiver was found to expend $\mathcal{E}_{rx} = 1.08$ J/MB in signal processing.

We observe that bit processing costs are comparable for the mote and the 802.11 WLAN. This is not surprising since processing costs are dictated by current circuit technology. Thus, with respect to energy consumption per bit, wireless links can be thought of as occurring in two distinct flavors:

- **Long Distance Links:** The transmit power requirements dominate.
- **Local Links:** Signal processing dominates and the power consumption is proportional to the communication rate (sum of bit rate inflow and outflow).

For the mote model in [72], transmission and processing energy costs are equal at the transition distance of $d = \sqrt{\mathcal{E}_t/\mathcal{E}_{pa}} = 67$ m. For other systems and environments, the transition distance depends on system factors such as the receiver design and antenna size and environmental factors such as the path-loss exponent. Nevertheless, the transition distance is typically on the order of a hundred meters for common systems. For example, [36] studied the energy efficiency of MIMO and cooperative MIMO systems. For each signaling strategy, there is a long distance regime, typically starting around 50 m, in which the energy cost of additional receiver complexity is more than compensated by the reduction in transmitter power.

Advances in circuit technology are likely to reduce the transmitter and receiver bit processing costs $\mathcal{E}_t$ and $\mathcal{E}_{rx}$. Nevertheless, per-bit

transmit energy costs will increase as d^{α} and thus there is always a long-distance regime in which transmit power is the dominant cost. These energy consumption figures demonstrate that irrespective of the technology and communications system (cellular, WLAN, motes, etc.), there is always a long-distance regime in which conserving transmit power is the primary consideration. In these scenarios, one cannot ignore the savings in transmit power afforded by cooperative relaying.

6.10 Final Remarks

We have observed that cooperative communication holds considerable promise. Some of this promise is evidenced in the proposals examined in this chapter. However, much remains to be done. For example, the proposals do not incorporate many of the ideas in Chapter 4.

It is tempting to argue that cooperation demands a complete redesign of the networking protocol stack. An abstract study might suggest that substantial benefits will be obtained by a tighter integration of the network layer with the MAC, link, and PHY layers. In particular, such an approach is likely to yield the greatest flexibility in optimizing the global use of wireless network resources.

However, this approach requires a significant expansion in the function of the network layer and more complex high-rate interfaces between PHY layer devices and network layer software. Demanding such a reconfiguration might well preclude the practical implementation of cooperative techniques. That is, requiring a more complex network layer represents a barrier to the adoption of cooperative wireless protocols. The practical value of cooperative communication is likely to depend on whether we can achieve the gains associated with cooperation without demanding more from the network layer.

Acknowledgments

We thank Tony Ephremides for persuading us to start this project and encouraging us to finish it. Though the task was a challenge, it was overall a pleasant experience.

We thank the reviewers for their many suggestions both big and small, and, in particular, for proposing substantial revisions in the organization of the material.

We thank James Finlay of NOW publishers for his cheerful patience despite our considerable delays.

G. Kramer gratefully acknowledges the support of the Board of Trustees of the University of Illinois Subaward no. 04-217 under National Science Foundation Grant CCR-0325673 and the Army Research Office under ARO Grant W911NF-06-1-0182.

I. Maric gratefully acknowledges the support of the DARPA ITMANET program under grant 1105741-1-TFIND and the Army Research Office under the MURI award W911NF-05-1-0246.

R. Yates gratefully acknowledges the support of the National Science Foundation under NSF Grants ANI-338805 and CNS-0434854.

References

[1] I. C. Abou-Faycal, M. D. Trott, and S. Shamai, "The capacity of discrete-time memoryless Rayleigh-fading channels," *IEEE Transactions on Information Theory*, vol. 47, no. 4, pp. 1290–1301, May 2001.

[2] R. Ahlswede, "Multi-way communication channels," in *Proc. 2nd Int. Symp. Inform. Theory (1971)*, pp. 23–52, Tsahkadsor, Armenian S.S.R., Publishing House of the Hungarian Academy of Sciences, 1973.

[3] R. Ahlswede, "The capacity region of a channel with two senders and two receivers," *The Annals of Probability*, vol. 2, no. 5, pp. 805–814, October 1974.

[4] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, "Network information flow," *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1204–1216, July 2000.

[5] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications*. Upper Saddle River, New Jersey: Prentice Hall, 1993.

[6] M. Aleksic, P. Razaghi, and W. Yu, "Capacity of a class of modulo-sum relay channels," http://arxiv.org/abs/0704.3591v1, April 2007.

[7] V. Anantharam and S. Verdu, "Bits through Queues," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 4–18, January 1996.

[8] AODV ns-2 Implementation, http://core.it.uu.se/adhoc/AodvUUImpl.

[9] M. R. Aref, *Information Flow in Relay Networks*. PhD thesis, Stanford University, Stanford, CA, October 1980.

[10] Atheros Communications, "Power consumption and energy efficiency comparisons of WLAN products," www.atheros.com/pt/whitepapers/atheros_power_whitepaper.pdf.

[11] A. S. Avestimehr and D. N. C. Tse, "Optimal cooperative communication in the low SNR regime," in *MSRI Workshop: Mathematics of Relaying and Cooperation in Communication Networks*, April 2006.

[12] A. S. Avestimehr and D. N. C. Tse, "Outage capacity of the fading relay channel in the low-SNR regime," *IEEE Transactions on Information Theory*, vol. 53, no. 4, pp. 1401–1415, February 2006.

[13] B. Awerbuch, D. Holmer, and H. Rubens, "The medium time metric: High throughput route selection in multi-rate ad hoc wireless networks," *Mobile Networks and Applications*, vol. 11, no. 2, pp. 253–266, April 2006.

[14] K. Azarian, H. El Gamal, and P. Schniter, "On the achievable diversity-multiplexing tradeoff in half-duplex cooperative channels," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4152–4172, December 2005.

[15] A. Bakre and B. R. Badrinath, "I-TCP: Indirect TCP for Mobile Hosts," *15th International Conference Distributed Computing Systems*, 1995.

[16] H. Balakrishnan, V. N. Padmanabhan, S. Seshan, and R. H. Katz, "A comparison of mechanisms for improving TCP performance over wireless links," *IEEE/ACM Transaction Networking*, vol. 5, no. 6, pp. 756–769, 1997.

[17] R. J. Benice and A. H. Frey, "An analysis of retransmission systems," *IEEE Transactions on Communication Technology*, vol. 12, no. 4, pp. 135–145, December 1964.

[18] D. P. Bertsekas and R. G. Gallager, *Data Networks*. Englewood Cliffs, NJ: Prentice-Hall, second edition, 1992.

[19] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: information-theoretic and communications aspects," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2619–2692, October 1998.

[20] H. Bölcskei, R. U. Nabar, Ö. Oyman, and A. J. Paulraj, "Capacity scaling laws in MIMO relay networks," *IEEE Transactions on Wireless Communication*, vol. 5, no. 6, pp. 1433–1444, June 2006.

[21] J. Broch, D. A. Maltz, D. B. Johnson, Y.-C. Hu, and J. Jetcheva, "A performance comparison of multi-hop wireless ad hoc network routing protocols," in *Mobile Computing and Networking*, pp. 85–97, October 1998.

[22] G. Caire and S. Shamai, "On the capacity of some channels with channel state information," *IEEE Transactions on Information Theory*, vol. 45, no. 6, pp. 2007–2019, September 1999.

[23] G. Caire and S. Shamai (Shitz), "On the achievable throughput of the multi-antenna gaussian broadcast channel," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1691–1706, July 2003.

[24] G. Caire, G. Taricco, and E. Biglieri, "Optimum power control over fading channels," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1468–1489, July 1999.

[25] G. Caire and D. Tuninetti, "The throughput of hybrid-ARQ protocols for the gaussian collision channels," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1971–1988, July 2001.

[26] J. Cannons, R. Dougherty, C. Freiling, and K. Zeger, "Networy routing capacity," *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 777–788, March 2006.

[27] A. B. Carleial, *Capacity of Multiple-Terminal Communication Networks*. PhD thesis, Stanford University, Stanford, CA, August 1975.
[28] A. B. Carleial, "Multiple-access channels with different generalized feedback signals," *IEEE Transactions on Information Theory*, vol. 28, no. 6, pp. 841–850, November 1982.
[29] D. Chase, "Code combining-a maximum-likelihood decoding approach for combining an arbitraty number of noisy packets," *IEEE Transactions on Communication*, vol. 33, no. 5, pp. 385–393, May 1985.
[30] K.-W. Chin, J. Judge, A. Williams, and R. Kermode, "Implementation experience with Manet routing protocols," *SIGCOMM Computer Communication Review*, vol. 32, no. 5, pp. 49–59, 2002.
[31] D. Costello Jr., J. Hagenauer, H. Imai, and S. Wicker, "Applications of error control coding," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2531–2560, October 1998.
[32] T. M. Cover, "Broadcast channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 2–14, January 1972.
[33] T. M. Cover and A. El Gamal, "Capacity theorems for the relay channel," *IEEE Transactions on Information Theory*, vol. 25, no. 5, pp. 572–584, September 1979.
[34] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley & Sons, 1991.
[35] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Channels*. Budapest: Akadémiai Kiadó, 1981.
[36] S. Cui, A. Goldsmith, and A. Bahai, "Energy-efficiency of MIMO and cooperative MIMO techniques in sensor networks," *IEEE Journal on Selected Areas Communication*, vol. 22, no. 6, pp. 1089–1098, August 2004.
[37] A. Dana and B. Hassibi, "On the power efficiency of sensory and ad-hoc wireless networks," *IEEE Transactions on Information Theory*, vol. 52, no. 7, pp. 2890–2914, July 2006.
[38] S. R. Das, C. E. Perkins, and E. M. Royer, "Performance Comparison of Two On-Demand Routing Protocols for Ad Hoc Networks," in *Proceedings of INFOCOM*, March 2000.
[39] D. S. J. De Couto, D. Aguayo, J. Bicket, and R. Morris, "A high-throughput path metric for multi-hop wireless routing," in *MobiCom '03: Proceedings of the 9th Annual International Conference on Mobile Computing and Networking*, pp. 134–146, New York, NY, USA: ACM Press, 2003.
[40] N. Devroye, P. Mitran, and V. Tarokh, "Achievable rates in cognitive radio channels," *IEEE Transactions on Information Theory*, vol. 52, no. 5, pp. 1813–1827, May 2006.
[41] O. Dousse, M. Franceschetti, and P. Thiran, "On the throughput scaling of wireless relay networks," *IEEE Transaction on Information Theory*, vol. 52, no. 6, pp. 2756–2761, June 2006.
[42] DSR Implementation in Linux, http://core.it.uu.se/adhoc/DsrUUImpl.
[43] A. El Gamal, "On information flow in relay networks," in *Proceedings of IEEE National Telecommunicaiton Conference*, vol. 2, pp. D4.1.1–D4.1.4, November 1981.

[44] A. El Gamal and M. Aref, "The capacity of the semideterministic relay channel," *IEEE Transactions on Information Theory*, vol. 28, no. 3, p. 536, May 1982.

[45] A. El Gamal, M. Mohseni, and S. Zahedi, "Bounds on capacity and minimum energy-per-bit for AWGN relay channels," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1545–1561, April 2006.

[46] A. El Gamal and S. Zahedi, "Minimum energy communication over a relay channel," in *IEEE Interational Symposium Information Theory*, p. 344, July 2003.

[47] A. El Gamal and S. Zahedi, "Capacity of a class of relay channels with orthogonal components," *IEEE Transactions on Information Theory*, vol. 51, no. 5, pp. 1815–1817, May 2005.

[48] U. Erez, S. Shamai (Shitz), and R. Zamir, "Capacity and lattice strategies for canceling known interference," *IEEE Transactions on Information Theory*, vol. 51, no. 11, pp. 3820–3833, November 2005.

[49] U. Erez and R. Zamir, "Achieving 1/2 log(1+SNR) on the AWGN channel with lattice encoding and decoding," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2293–2314, October 2004.

[50] J. Ezri and M. Gastpar, "On the performance of independently designed Ldpc codes for the relay channel," in *IEEE Interational Symposium on Information Theory*, pp. 977–981, July 2006.

[51] N. Feamster, J. Winick, and J. Rexford, "A model of BGP routing for network engineering," in *SIGMETRICS '04/Perf. '04: Proceedings of Joint International Conference Measurement and Modeling Computer Systems*, pp. 331–342, New York, NY, USA: ACM Press, 2004.

[52] G. D. Forney Jr. and G. Ungerboeck, "Modulation and coding for linear gaussian channels," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2384–2415, October 1998.

[53] G. J. Foschini, "Layered space–time architecture for wireless communication in a fading environment when using multi–element antennas," *Bell Laboratories Technical Journal*, vol. 1, no. 2, pp. 41–59, 1996.

[54] G. J. Foschini and J. Salz, "Digital communications over fading radio channels," *Bell System Technical Journal*, vol. 62, no. 2, pp. 429–456, February 1983.

[55] J.-F. Frigon and B. Daneshrad, "Field measurements of an indoor high-speed QAM wireless system using decision feedback equalization and smart antenna array," *IEEE Transactions on Wireless Communication*, vol. 1, no. 1, pp. 134–144, January 2002.

[56] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[57] M. Gastpar, G. Kramer, and P. Gupta, "The multiple-relay channel: coding and antenna-clustering capacity," in *Proceedings of 2002 IEEE International Symposium on Information Theory*, (Lausanne, Switzerland), p. 136, June 30–July 5 2002.

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

[59] M. Gastpar and M. Vetterli, "On the capacity of large Gaussian relay networks," *IEEE Transactions on Information Theory*, vol. 51, no. 3, pp. 765–779, March 2005.
[60] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Problemy Peredachi Informatsii*, vol. 9, no. 1, pp. 19–31, 1980.
[61] A. Goldsmith, *Wireless Communications*. Cambridge, UK: Cambridge University Press, 2005.
[62] A. J. Goldsmith and P. P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Transactions on Information Theory*, vol. 43, no. 6, pp. 1986–1992, November 1997.
[63] D. J. Goodman, *Wireless Personal Communication Systems*. Reading, MA: Addison-Wesley, 1997.
[64] S. Gopal and D. Raychaudhuri, "Experimental evaluation of the TCP simultaneous-send problem in 802.11 wireless local area networks," in *E-WIND '05: Proceedings of 2005 ACM SIGCOMM Workshop on Experimental Approaches to Wireless Network Design and Analysis*, pp. 23–28, New York, NY, USA: ACM Press, 2005.
[65] P. Gupta and P. R. Kumar, "Toward an information theory of large networks: An achievable rate region," *IEEE Transactions on Information Theory*, vol. 49, no. 8, pp. 1877–1894, August 2003.
[66] M. C. Gursoy, H. V. Poor, and S. Verdu, "The noncoherent Rician fading channel-part I: structure of the capacity-achieving input," *IEEE Transactions Wireless Communication*, vol. 4, no. 5, pp. 2193–2206, September 2005.
[67] J. Hagenauer, "Rate compatible punctured convolutional codes (RCPC) and their applications," *IEEE Transactions on Communication*, vol. 36, no. 4, pp. 389–400, April 1988.
[68] T. S. Han and K. Kobayashi, "A new achievable rate region for the interference channel," *IEEE Transactions on Information Theory*, vol. 27, no. 1, pp. 49–60, January 1981.
[69] S. Hanly and D. Tse, "Multi-access fading channels: Part II: Delay-limited capacities," *IEEE Transactions on Information Theory*, vol. 44, no. 7, pp. 2816–2831, November 1998.
[70] R. Heath Jr. and A. Paulraj, "Switching between multiplexing and diversity based on constellation distance," in *Proceedings of Allerton Conference on Communication, Control and Computing*, October 2000.
[71] C. Heegard and A. El Gamal, "On the capacity of computer memory with defects," *IEEE Transactions on Information Theory*, vol. 29, no. 5, pp. 731–739, September 1983.
[72] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient routing protocols for wireless microsensor networks," in *Proceedings of 33rd Hawaii International Conference on Systems and Sciences*, pp. 1–10, 2000.
[73] T. Ho, M. Medard, R. Koetter, D. R. Karger, M. Effros, J. Shi, and B. Long, "A random linear network coding approach to multicast," *IEEE Transaction on Information Theory*, vol. 52, no. 10, pp. 4413–4430, October 2006.
[74] G. Holland and N. Vaidya, "Analysis of TCP performance over mobile ad hoc networks," *Wireless Networks*, vol. 8, no. 2/3, pp. 275–288, 2002.

[75] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, UK: Cambridge University Press, 1985.

[76] A. Høst-Madsen, "On the capacity of wireless relaying," in *Proceedings of IEEE Vehicular Technology Conference (VTC)*, pp. 1333–1337, September 2002.

[77] A. Høst-Madsen, "Capacity bounds for cooperative diversity," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1522–1544, April 2006.

[78] A. Høst-Madsen and A. Nosratinia, "The multiplexing gain of wireless networks," in *Proceedings of IEEE International Symposium Information Theory*, pp. 2065–2069, September 2005.

[79] A. Høst-Madsen and J. Zhang, "Capacity bounds and power allocation for the wireless relay channel," *IEEE Transactions on Information Theory*, vol. 51, no. 6, pp. 2020–2040, June 2005.

[80] T. C. Hu, "Multi-commodity network flows," *Operations Research*, vol. 11, no. 3, no. 3, pp. 344–360, 1963.

[81] J. Huang and S. P. Meyn, "Characterization and computation of optimal distributions for channel coding," *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2336–2351, July 2005.

[82] T. Hunter, S. Sanayei, and A. Nosratinia, "Outage analysis of coded cooperation," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 375–391, February 2006.

[83] T. E. Hunter and A. Nosratinia, "Cooperation diversity through coding," in *Proceedings IEEE International Symposium Information on Theory*, p. 220, Lausanne, Switzerland, June 30–July 5 2002.

[84] T. E. Hunter and A. Nosratinia, "Diversity through coded cooperation," *IEEE Transactions on Wireless Communication*, vol. 5, no. 2, pp. 283–289, February 2006.

[85] G. Jakllari, S. V. Krishnamurthy, M. Faloutsos, P. V. Krishnamurthy, and O. Ercetin, "A Framework for distributed spatio-temporal communications in mobile ad hoc networks," in *Proceedings of INFOCOM*, pp. 1–13, April 2006.

[86] M. Janani, A. Hedajat, T. E. Hunter, and A. Nosratinia, "Coded cooperation in wireless communications: Space time transmission and iterative decoding," *IEEE Transactions on Signal Proceedings*, vol. 52, no. 2, pp. 362–371, February 2004.

[87] N. Jindal, U. Mitra, and A. Goldsmith, "Capacity of ad-hoc networks with node cooperation," in *Proceedings IEEE International Symposium Information on Theory*, p. 271, June 2004.

[88] D. B. Johnson and D. A. Maltz, "Dynamic source routing in ad hoc wireless networks," in *Mobile Computing*, (Imielinski and Korth, eds.), pp. 153–181, Kluwer Academic Publishers, 1996.

[89] A. Jovičić and P. Viswanath, "Cognitive radio: An information-theoretic perspective," in *Proceedings of IEEE International Symposium on Information Theory*, pp. 2413–2417, July 2006.

[90] S. Katti, D. Katabi, W. Hu, H. Rahul, and M. Medard, "The importance of being opportunistic: Practical network coding for wireless environments," in

Proceedings of Allerton Conference on Communication Control and Computing, 2005.

[91] S. Katti, H. Rahul, W. Hu, D. Katabi, M. Medard, and J. Crowcroft, "XORs in the air: Practical wireless network coding," in *Proceedings of SIGCOMM*, pp. 243–254, September 2006.

[92] M. Katz and S. Shamai (Shitz), "Relaying protocols for two co-located users," in *Proceedings IEEE International Symposium Information on Theory*, pp. 936–940, September 2005.

[93] M. Katz and S. Shamai (Shitz), "Transmitting to colocated users in wireless ad hoc and sensor networks," *IEEE Transaction on Information Theory*, vol. 51, no. 10, pp. 3540–3563, October 2005.

[94] M. Katz and S. Shamai (Shitz), "On the outage probability of a multiple-input single-output communication link," *IEEE Transactions on Information Theory*, submitted, 2006.

[95] V. Kawadia, Y. Zhang, and B. Gupta, "System Services for ad-hoc routing: Architecture, implementation and experiences," in *Proceedings of MobiSys'03*, pp. 99–112, May 2003.

[96] A. Khisti, U. Erez, and G. W. Wornell, "Fundamental limits and scaling behavior of cooperative multicasting in wireless networks," *IEEE Transaction on Information Theory*, vol. 52, no. 6, pp. 2762–2770, June 2006.

[97] M. A. Khojastepour, N. Ahmed, and B. Aazhang, "Code design for the relay channel and factor graph decoding," in *Proceedings of 38th Annual Conference on Signals, Systems, Computer*, pp. 7–10, Pacific Grove, CA, November 2004.

[98] M. A. Khojastepour, A. Sabharwal, and B. Aazhang, *On the capacity of 'cheap' relay networks*. March 2003.

[99] R. C. King, *Multiple Access Channels with Generalized Feedback*. PhD thesis, Stanford University, Stanford, CA, March 1978.

[100] R. Knopp, "Two-way radio networks with a star topology," in *Proceedings of 2006 International Zurich Seminar*, pp. 154–157, February 2006.

[101] R. Koetter and M. Medard, "An algebraic approach to network coding," *IEEE/ACM Transactions on Networking*, vol. 11, no. 5, pp. 782–795, October 2003.

[102] G. Kramer, "Capacity results for the discrete memoryless network," *IEEE Transactions Information Theory*, vol. 49, no. 1, pp. 4–21, January 2003.

[103] G. Kramer, "Models and theory for relay channels with receive constraints," in *Proceedings of 42nd Annual Allerton Conference on Communication, Control, and Computing*, (Monticello, IL), pp. 1312–1321, September 29–October 1 2004.

[104] G. Kramer, "Distributed and layered codes for relaying," in *Asilomar Conference on Signals, Systems, and Computers*, pp. 1752–1756, Pacific Grove, CA, October 30–November 2 2005.

[105] G. Kramer, "Communication strategies and coding for relaying," in *Wireless Communications, Vol. 143 of the IMA Volumes in Mathematics and its Applications*, pp. 163–175, New York, NY: Springer, 2007.

[106] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3037–3063, September 2005.

[107] G. Kramer and S. A. Savari, "Capacity bounds for relay networks," in *Proceedings of 2006 Workshop on Information Theory and its Application*, January 2005.

[108] G. Kramer and S. A. Savari, "On networks of two-way channels," *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, vol. 68, pp. 133–143, American Mathematical Society, 2005.

[109] G. Kramer and S. A. Savari, "Edge-cut bounds on network coding rates," *Journal of Network and Systems Management*, vol. 14, no. 1, pp. 49–67, March 2006.

[110] G. Kramer and A. J. van Wijngaarden, "On the white Gaussian multiple-access relay channel," in *Proceedings of IEEE International Symposium on Information Theory*, p. 40, Sorrento, Italy, June 25–30 2000.

[111] J. N. Laneman, *Cooperative Diversity in Wireless Networks: Algorithms and Architectures*. PhD thesis, M.I.T., Cambridge, MA, August 2002.

[112] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 3062–3080, December 2004.

[113] J. N. Laneman and G. W. Wornell, "Distributed space-time coded protocols for exploiting cooperative diversity in wireless networks," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2415–2525, October 2003.

[114] N. Laneman, G. Wornell, and D. N. C. Tse, "An efficient protocol for realizing cooperative diversity in wireless networks," in *IEEE International Symposium on Information Theory*, p. 294, June 2001.

[115] A. Lapidoth and S. M. Moser, "Capacity bounds via duality with applications to multiple-antenna systems on flat fading channels," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2426–2467, October 2003.

[116] S.-Y. R. Li, R. W. Yeung, and N. Cai, "Linear network coding," *IEEE Transactions on Information Theory*, vol. 49, no. 2, pp. 371–381, February 2003.

[117] Y. Liang and V. V. Veeravalli, "Capacity of noncoherent time-selective rayleigh-fading channels," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3095–3110, December 2004.

[118] Y. Liang and V. V. Veeravalli, "The impact of relaying on the capacity of broadcast channels," in *Proceedings of IEEE International Symposium Information Theory*, p. 403, Chicago, IL, June 2004.

[119] H. Liao, "A coding theorem for multiple access communications," in *Proceedings of IEEE International Symposium on Information Theory*, Asilomar, CA, 1972.

[120] S. Lin and D. Costello, *Error Control Coding: Fundamental and Applications*. Upper Saddle River, NJ.: Prentice Hall, 1983.

[121] S. Lin, D. Costello, and M. J. Miller, "Automatic-repeat-request error control schemes," *IEEE Communications Magazing*, vol. 22, no. 12, pp. 5–17, December 1984.

[122] P. Liu, Z. Tao, Z. Lin, E. Erkip, and S. Panwar, "Cooperative wireless communications: A cross-layer approach," *IEEE Communications Magazine*, vol. 13, no. 4, pp. 84–92, August 2006.

[123] P. Liu, Z. Tao, S. Narayanan, T. Korakis, and S. Panwar, "CoopMAC: A cooperative MAC for wireless LAN," *IEEE Journal of Selected Areas Communication*, vol. 25, no. 2, pp. 340–354, February 2007.

[124] P. Liu, Z. Tao, and S. Panwar, "A cooperative MAC protocol for wireless local area networks," in *Proceedings of IEEE International Conference on Communication (ICC)*, June 2005.

[125] R. Liu, P. Spasojević, and E. Soljanin, "User cooperation with punctured turbo codes," in *Proceedings of Allerton Conference on Communication, Control and Computing*, October 2003.

[126] R. Liu, P. Spasojević, and E. Soljanin, "Cooperative diversity with incremental redundancy turbo coding for quasi-static wireless networks," in *IEEE International Workshop Signal Proceedings of Advances for Wireless Communication*, pp. 791–795, June 2005.

[127] X. Liu and R. Srikant, "An information-theoretic view of connectivity in wireless sensor networks," in *Proceedings of IEEE SECON*, October 2004.

[128] M. Luby, "LT-codes," in *Proceedings of 43rd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 271–280, November 2002.

[129] H. Lundgren, *Implementation and Real-world Evaluation of Routing Protocols for Wireless Ad hoc Networks*. PhD thesis, Uppsala University, 2002.

[130] D. Maltz, J. Broch, J. Jetcheva, and D. Johnson, "The effects of on-demand behavior in routing protocols for multi-hop wireless ad hoc networks," *IEEE Journal of Selected Areas Communication*, vol. 17, no. 8, pp. 1439–1453, August 1999.

[131] D. Maltz, J. Broch, and D. Johnson, "Experiences designing and building a multi-hop wireless ad hoc network testbed," CMU Technical Report CS-99-116, School of Computer Science, Carnegie Mellon University, Pittsburgh PA, March 1999.

[132] I. Marić and R. D. Yates, "Efficient multihop broadcast for wideband systems," in *Multiantenna Channels: Capacity, Coding and Signal Processing, DIMACS Workshop on Signal Processing for Wireless Transmission*, (G. J. Foschini and S. Verdu, eds.), pp. 285–299, October 2002. DIMACS Series in Discrete Mathematics and Theoretical Computer Science.

[133] I. Marić and R. D. Yates, "Cooperative multihop broadcast for wireless networks," *IEEE Journal of Selected Areas Communication*, vol. 22, no. 6, pp. 1080–1088, August 2004.

[134] I. Marić and R. D. Yates, "Cooperative multicast for maximum network lifetime," *IEEE Journal of Selected Areas Communication*, vol. 23, no. 1, pp. 127–135, January 2005.

[135] T. L. Marzetta and B. M. Hochwald, "Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading," *IEEE Transactions on Information Theory*, vol. 45, no. 1, pp. 139–157, January 1999.

[136] P. Mitran, H. Ochiai, and V. Tarokh, "Space-time diversity enhancements using collaborative communications," *IEEE Transaction on Information Theory*, vol. 51, no. 6, pp. 2041–2057, June 2005.

[137] V. Morgenshtern and H. Bölcskei, "On the value of cooperation in interference relay networks," in *Proceedings of Allerton Conference on Communication, Control and Computing*, September 2005.

[138] V. Morgenshtern and H. Bölcskei, "Crystallization in large wireless networks," *IEEE Transactions on Information Theory*, to appear, 2007.

[139] R. U. Nabar, H. Bölcskei, and W. Kneubühler, "Fading relay channels: Performance limits and space-time signal design," *IEEE Journal of Selected Areas Communication*, vol. 22, no. 6, pp. 1099–1109, August 2004.

[140] K. Nahm, A. Helmy, and C.-C. J. Kuo, "TCP over multihop 802.11 networks: issues and performance enhancement," in *MobiHoc '05: Proceedings of 6th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pp. 277–287, New York, NY, USA: ACM Press, 2005.

[141] K. R. Narayanan and G. L. Stuber, "A Novel ARQ technique using the turbo coding principle," *IEEE Communication Letters*, vol. 1, no. 2, pp. 49–51, March 1997.

[142] S. Narayanaswamy, V. Kawadia, R. S. Sreenivas, and P. R. Kumar, "Power Control in ad-hoc networks: Theory, architecture, algorithm and implementation of the COMPOW protocol," in *Proceedings of European Wireless 2002*, pp. 179–186, February 2002.

[143] T. Oechtering, I. Bjelakovic, C. Schnurr, and H. Boche, "Broadcast capacity region of two-phase bidirectional relaying," http://arxiv.org/abs/cs/0703078v1, March 2007.

[144] L. Ozarow, S. Shamai (Shitz), and A. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Transaction on Vehicular Technology*, vol. 43, no. 2, pp. 359–378, May 1994.

[145] A. Özgur, O. Leveque, and D. Tse, "Hierarchical cooperation achieves optimal capacity scaling in ad hoc networks," *IEEE Transactions on Information Theory*, Submitted, 2006.

[146] C. E. Perkins and P. Bhagwat, "Highly dynamic destination-sequenced distance-vector routing (DSDV) for mobile computers," in *SIGCOMM '94: Proceedings of Conference Communication Architectures, Protocols, Application*, pp. 234–244, New York, NY, USA: ACM Press, 1994.

[147] C. E. Perkins and E. M. Royer, "Ad Hoc on-demand distance vector routing," in *Proceedings of 2nd IEEE Workshop Mobile Computing and Applications*, pp. 90–100, February 1999.

[148] J. G. Proakis, *Digital Communications*. New York: McGraw-Hill, second edition, 1983.

[149] B. Rankov and A. Wittneben, "Achievable rate regions for the two-way relay channel," in *Proceedings IEEE International Symposium on Information Theory*, pp. 1668–1672, July 2006.

[150] B. Rankov and A. Wittneben, "Spectral efficient protocols for half-duplex fading relay channels," *IEEE Journal of Selected Areas Communication*, vol. 25, no. 2, pp. 379–389, February 2007.

[151] T. S. Rappaport, *Wireless Communications Principles and Practice*. Englewood Cliffs, NJ: Prentice Hall, 1997.

[152] N. Ratnakar and G. Kramer, "The multicast capacity of deterministic relay networks with no interference," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2425–2432, June 2006.

[153] P. Razaghi and W. Yu, "Parity forwarding for multiple-relay networks," in *Proceedings IEEE International Symposium on Information Theory*, pp. 1678–1682, Seattle, WA, July 2006.

[154] T. J. Richardson, A. Shokrollahi, and R. L. Urbanke, "Design of capacity-approaching low-density parity-check codes," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 619–637, February 2001.

[155] E. M. Royer and C. E. Perkins, "An implementation study of the AODV routing protocol," in *Proceedings of IEEE Wireless Communications and Networking Conference*, pp. 1003–1008, September 2000.

[156] L. Sankaranarayanan, G. Kramer, and N. B. Mandayam, "Capacity theorems for the multiple-access relay channel," in *Proceedings of Allerton Conference on Communication, Control and Computing*, pp. 1782–1791, Monticello, IL, September 29–October 1 2004.

[157] L. Sankaranarayanan, G. Kramer, and N. B. Mandayam, "Hierarchical sensor networks: Capacity bounds and cooperative strategies using the multiple-access relay channel model," in *Proceedings of 2004 IEEE Conference on Sensor and Ad Hoc Communication Networks*, pp. 191–199, October 2004.

[158] L. Sankaranarayanan, G. Kramer, and N. B. Mandayam, "Hierarchical sensor networks: capacity bounds using the constrained multiple-access relay channel model," in *Asilomar Conference on Signals, Systems, and Computers*, pp. 1912–1916, November 2004.

[159] L. Sankaranarayanan, G. Kramer, and N. B. Mandayam, "Cooperation vs. hierarchy: An information-theoretic comparison," in *Proceedings IEEE International Symposium on Information Theory*, pp. 411–415, September 2005.

[160] L. Sankaranarayanan, G. Kramer, and N. B. Mandayam, "Cooperative diversity in wireless networks: A geometry-inclusive analysis," in *Proceedings Allerton Conference on Communication, Control and Computing*, pp. 1598–1607, October 2005.

[161] B. E. Schein, *Distributed Coordination in Network Information Theory*. PhD thesis, M.I.T., Cambridge, MA, October 2001.

[162] A. Sendonaris, E. Erkip, and B. Aazhang, "Increasing uplink capacity via user cooperation diversity," in *IEEE International Symposium on Information Theory*, p. 156, August 1998.

[163] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity – Part I: System description," *IEEE Transactions on Information Theory*, vol. 51, no. 11, pp. 1927–1938, November 2003.

[164] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423 and 623–656, Reprinted in *Claude Elwood Shannon: Collected Papers*, pp. 5–83, (N. J. A. Sloane and A. D. Wyner, eds.) Piscataway: IEEE Press, 1993, July and October 1948.

[165] C. E. Shannon, "Communication in the presence of noise," *Proceedings of IRE*, vol. 37, pp. 10–21, Reprinted in *Claude Elwood Shannon: Collected Papers*, pp. 160–172, (N. J. A. Sloane and A. D. Wyner, eds.) Piscataway: IEEE Press, 1993., January 1949.

[166] C. E. Shannon, "Channels with side information at the transmitter," *IBM Journal of Research Development*, vol. 2, pp. 289–293, September 1958.

[167] C. E. Shannon, "Two-way communication channels," in *Proceedings of 4th Berkeley Symposium on Mathematical Statistics and Probability*, (J. Neyman, ed.), vol. 1, pp. 611–644, Berkeley, CA: University of California Press, 1961. Reprinted in *Claude Elwood Shannon: Collected Papers*, pp. 351–384, (N. J. A. Sloane and A. D. Wyner, eds.) Piscataway: IEEE Press, 1993.

[168] A. Shokrollahi, "Raptor codes," *IEEE Transaction on Information Theory*, vol. 52, no. 6, pp. 2551–2567, June 2006.

[169] P. S. Sindhu, "Retransmission error control with memory," *IEEE Transaction on Communication*, vol. 25, no. 5, pp. 473–479, May 1977.

[170] B. Sirkeci, A. Scaglione, and G. Mergen, "Asymptotic analysis of multistage cooperative broadcast in wireless networks," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2531–2550, June 2006.

[171] Special Issue on Networking and Information Theory, *IEEE Transactions on Information Theory*, vol. 52, no. 6, June 2006.

[172] A. Stefanov and E. Erkip, "Cooperative space-time coding for wireless networks," in *IEEE Information on Theory Workshop*, April 2003.

[173] A. Stefanov and E. Erkip, "Cooperative coding for wireless networks," *IEEE Transactions on Communications*, vol. 52, no. 9, pp. 1470–1476, September 2004.

[174] E. Telatar, "Capacity of multi-antenna gaussian channels," *European Transactions on Telecommunications*, pp. 585–595, November 1999.

[175] S. ten Brink, "Convergence of iterative decoding," *Electronic Letters*, vol. 35, no. 10, pp. 806–808, May 1999.

[176] S. ten Brink, G. Kramer, and A. Ashikhmin, "Design of low-density parity-check codes for modulation and detection," *IEEE Transactions on Communication*, vol. 52, no. 4, pp. 670–678, April 2004.

[177] The Network Simulator – Ns-2, http://www.isi.edu/nsnam/ns/.

[178] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, UK: Cambridge University Press, 2005.

[179] E. Tuncel, "Slepian-wolf coding over broadcast channels," *IEEE Transaction on Information Theory*, vol. 52, no. 4, pp. 1469–1482, April 2006.

[180] E. C. van der Meulen, *Transmission of Information in a T-Terminal Discrete Memoryless Channel*. PhD thesis, University of California at Berkeley, Berkeley, CA, June 1968.

[181] E. C. van der Meulen, "Three-terminal communication channels," *Advances in Applied Probability*, vol. 3, no. 1, no. 1, pp. 120–154, 1971.

[182] S. Verdú, "On channel capacity per unit cost," *IEEE Transaction on Information Theory*, vol. 36, no. 5, pp. 1019–1030, September 1990.

[183] S. Verdú, "Spectral efficiency in the wideband regime," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1319–1343, June 2002.

[184] A. J. Viterbi, *Principles of Coherent Communication*. New York: McGraw-Hill, 1966.

[185] A. J. Viterbi, *CDMA Principles of Spread Spectrum Communication*. Reading, MA: Addison-Wesley, 1995.

[186] S. B. Wicker, *Error Control Systems for Digital Communication and Storage*. Upper Saddle River, NJ.: Prentice Hall, 1987.

[187] F. M. J. Willems, *Informationtheoretical Results for the Discrete Memoryless Multiple Access Channel. Doctor in de Wetenschappen Proefschrift*, Leuven, Belgium: Katholieke Universiteit Leuven, October 1982.

[188] J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*. New York: Wiley, 1965.

[189] W. Wu, S. Vishwanath, and A. Arapostathis, "On the capacity of Gaussian weak interference channels with degraded message sets," in *Proceedings of Conference on Information Sciences and Systems*, pp. 1703–1708, March 2006.

[190] Y. Wu, P. A. Chou, and S.-Y. Kung, "Minimum-energy multicast in mobile ad hoc networks using network coding," *IEEE Transactions on Communication*, vol. 53, no. 11, pp. 1906–1918, November 2005.

[191] L.-L. Xie and P. R. Kumar, "A network information theory for wireless communication: Scaling laws and optimal operation," *IEEE Transactions on Information Theory*, vol. 50, no. 5, pp. 748–767, May 2004.

[192] F. Xue and P. R. Kumar, *Scaling Laws for Ad Hoc Wireless Networks: An Information Theoretic Approach*. Delft, The Netherlands: NOW Publishers, 2006.

[193] S. Yang and J.-C. Belfiore, "Towards the optimal amplify-and-forward cooperative diversity scheme," *IEEE Transactions on Information Theory, submitted*, 2006.

[194] X. Yu, "Improving TCP performance over mobile ad hoc networks by exploiting cross-layer information awareness," in *MobiCom '04: Proceedings of 10th Annual International Conference Mobile Computing and Networking*, pp. 231–244, New York, NY, USA: ACM Press, 2004.

[195] M. Yuksel and E. Erkip, "Diversity gains and clustering in wireless relaying," in *Proceedings of IEEE International Symposium on Information Theory*, p. 400, June 2004.

[196] M. Yuksel and E. Erkip, "Cooperative wireless systems: A diversity-multiplexing tradeoff perspective," *IEEE Transactions on Information Theory, submitted*, 2006.

[197] M. Yuksel and E. Erkip, "Diversity-multiplexing tradeoff in cooperative wireless systems," in *Proceedings of Conference on Information Sciences and Systems*, pp. 1062–1067, March 2006.

[198] Z. Zhang and T. M. Duman, "Capacity-approaching turbo coding and iterative decoding for relay channels," *IEEE Transactions on Information Theory*, vol. 53, no. 11, pp. 1895–1905, November 2005.

[199] B. Zhao and M. Valenti, "Practical relay networks: A generalization of Hybrid-ARQ," *IEEE Journal of Selected Areas Communication*, vol. 23, no. 1, pp. 7–18, January 2005.

[200] B. Zhao and M. C. Valenti, "Distributed turbo coded diversity for relay channel," *Electronic Letters*, vol. 39, no. 10, pp. 786–787, May 2003.

[201] L. Zheng and D. N. C. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple antenna channels," *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1073–1096, May 2003.

[202] H. Zhu and G. Cao, "rDCF: A relay-enabled medium access control protocol for wireless ad hoc networks," *IEEE Transaction Mobile Computing*, vol. 5, no. 9, pp. 1201–1214, September 2006.

[203] M. Zorzi and R. Rao, "Geographic random forwarding (GeRaF) for ad hoc and sensor networks: Multihop performance," *IEEE Transactions Mobile Computing*, vol. 2, no. 4, pp. 337–348, October 2003.