1

Introduction

Wireless has become as much part of our lives as have houses, cars and computers. Mobile phones, an example application of wireless technologies, are indispensable today as they allow us to be connected anywhere at any time. We have, in fact, taken so much liking to wireless technologies that system capacity is reaching saturation levels. This is aggravated by recently emerged bandwidth hungry applications ranging from web browsing to multimedia transmissions. Network designers are struggling to meet this ever increasing demand in capacity and any means to increase capacity is hence welcome.

Interestingly, Martin Cooper of Arraycomm has observed that ‘the wireless capacity has doubled every 30 months over the last 104 years.’ This translates into an approximately million-fold capacity increase since the 1960s, which has been broken down [7] to yield a 25-times improvement from wider spectrum, a fivefold improvement by dividing the spectrum into smaller slices, a fivefold improvement by designing better modulation schemes, and an impressive 1600-fold gain through reduced cell sizes and transmit distance.

Among the many possible approaches to capitalize on these enticing gains, this book focuses on cooperation and we will show how cooperative techniques at the physical layer can potentially be of great benefit to capacity as well as coverage.

1.1 Book Structure

To facilitate a coherent understanding of cooperation for people working in the field as well as people acquainting themselves with the topic, the book is structured as follows:

- **Chapter 1: Introduction.** A basic introduction to cooperative systems is given first. We will then discuss typical application scenarios of cooperative techniques. Furthermore, we will discuss pros and cons of relaying and briefly quantify the capacity gains. We will also revise definitions and terminologies typically used in the context of cooperative systems. Finally, some historical background as well as key milestones are discussed.
- **Chapter 2: Wireless Channel.** In contrast to traditional wireless systems, the wireless channel becomes part of the cooperative system, which has a profound impact on channel statistics, including temporal and correlation behaviors. We will hence discuss in detail the channel behavior of regenerative, transparent and distributed multiple-input, multiple-output (MIMO) systems. This is pivotal in understanding and quantifying the performance of cooperative protocols at the physical (PHY) layer.

Cooperative Communications Mischa Dohler and Yonghui Li

- **Chapter 3: Transparent PHY.** We will discuss and quantify the performance of different transparent architectures here. We will introduce some tools that facilitate the characterization of general architectures, which range from transparent relaying to transparent distributed space–time block coding, distributed space–time trellis coding, distributed multiplexing and distributed beamforming. We will also dwell on issues pertaining to distributed system optimization, such as distributed power allocation and distributed relay selection.
- **Chapter 4: Regenerative PHY.** We will discuss and quantify the performance of a plethora of regenerative architectures here. We will mainly deal with different estimate-and-forward, decode-and-forward, compress-and-forward and soft-information relaying protocols, among others, with the aim characterizing their performance. We will also dwell on recently emerged advanced topics related to distributed coding, such as distributed space–time block coding, distributed space–time trellis coding, distributed turbo coding, distributed network-channel coding for single as well as multiple source–destination pairs, etc.
- **Chapter 5: Hardware.** We will discuss how hardware facilitates as well as limits any implementation of cooperative relaying schemes. These limitations in hardware render, for instance, the implementation of some of the recently proposed cooperative protocols infeasible. We will discuss the implementation of transparent and regenerative schemes. We will also compare their costs and implementation complexities. We then derive complexities and power consumption of various relaying architectures based on 3G and 4G standards. Finally, we discuss some available hardware platforms and testbeds implementing relaying or distributed space–time processing techniques.
- **Chapter 6: Road Ahead.** Finally, the book is concluded by summarizing its contributions and also highlighting important open research directions and still-to-be-explored impact areas. We also describe how to include real-world impairments into the analysis, such as interference, etc. Finally, some business issues are briefly dealt with.

1.2 Quick Introduction

Wireless communication systems are traditionally conceptualized such that users individually communicate with the associated base station and vice versa. Cooperation is referred to as any architecture that deviates from this traditional approach, that is where a user's communication link is enhanced in a supportive way by relays or in a cooperative way by other users. Due to the many available degrees of freedom of such systems, that is the many ways in which supportive relays and cooperative users can be deployed, an enormous amount of different architectures exist. Some canonical architectures are briefly introduced below, after having discussed the wireless channel they experience and the gains one can expect from such deployments.

1.2.1 Channel

The wireless channel is pivotal to the understanding of the gains of cooperative systems. Whilst discussed in more detail in Chapter 2, we shall expose here some of its fundamental properties. As such, the transmitted signal is impaired by three effects:

- **Pathloss.** Averaging the received power at a particular distance over a sufficiently large area, yields the loss in power or the pathloss versus distance. The pathloss law is (almost) deterministic and traditionally behaves linearly in decibels, that is as an inverse law in linear scale. Pathloss limits interference but also rapidly diminishes the useful signal power. Any technique improving on the pathloss is hence highly appreciated by network planners.
- **Shadowing.** Averaging the received power at a particular distance over an area of radius of approximately shadowing coherence distance yields a variation in the received power around the pathloss. This variation is referred to as shadowing. The shadowing law is random and is traditionally modeled as Gaussian in decibels, that is lognormal in linear scale. Shadowing is one of the most

detrimental performance factors in modern communications systems since it cannot be absorbed by suitable channel codes, thus causing non-availabilities of links referred to as outages. Any technique improving on the shadowing outage is hence highly appreciated.

- **Fading.** Not averaging the signal at all allows one to observe fading as a signal fluctuation around pathloss and shadowing. It is caused by the constructive and destructive addition of the signal traveling via multiple propagation paths.

  If channel fading realizations change from symbol to symbol due to high mobility in the channel, then the channel is referred to as fast-fading; otherwise, it is referred to as slow-fading. A fast-fading channel offers temporal diversity, which can be picked up by a suitable channel coder; a slow-fading channel does not offer temporal diversity, which yields large outages.

  If the multiple symbol copies can be resolved because the mutual propagation delays are larger than the symbol duration, then the channel is referred to as a frequency-selective fading channel; otherwise, it is referred to as a frequency-flat or simply flat fading channel. A frequency-selective fading channel offers frequency diversity, which can be picked up by suitable signal processing coupled with a channel coder; a flat fading channel does not offer frequency diversity, which again yields large outages.

  In terms of outages, the worst case scenario is hence a slow and flat fading channel and the best case is a fast and frequency-selective fading channel. Modern communications systems can be fast or slow fading, depending on the mobility they are subjected to; however, due to the increasingly short symbol durations, they are typically frequency selective and hence inherently offer a healthy amount of diversity. Techniques improving on diversity therefore seem unnecessary. Yet, as will be shown in Section 1.5.3, diversity (reliability) can be traded against communication rate (capacity), thereby encouraging the development of techniques that can improve on either.

In decibels (dB), these effects add up, whereas in linear scale they are multiplicative. Denoting pathloss, shadowing and fading by L, S and F respectively, the received signal power is hence given in dB as $P_\mathrm{r} = P_\mathrm{t} + L + S + F$. Cooperative communications generally evolves around diminishing these losses associated with the wireless channel as well as using the detrimental effects to its benefit.

1.2.2 Typical Gains

As will be quantified throughout this book, cooperative communications yields several gains. Three of the most important gains are given below:

- **Pathloss Gain.** Due to the non-linear pathloss behavior, splitting the propagation path yields transmit power gains because the aggregate pathloss of the split path is less than the pathloss of the full path. The pathloss, and hence the signal-to-noise ratio (SNR), is known to be inversely proportional to the propagation distance. More precisely,

  $$\mathrm{SNR} \propto \frac{1}{x^n}, \tag{1.1}$$

  where x is the distance between transmitter and receiver, and n is the pathloss exponent. The latter typically differs for different propagation environments between 2 (line-of-sight) and 6 (highly cluttered environments). Splitting, for example, the communication path between source and destination into two equal distances and allocating to each part half of the power, the gain w.r.t. direct communication assuming a pathloss coefficient of $n = 4$ is quantified as

  $$\frac{\frac{1/2}{(x/2)^4} + \frac{1/2}{(x/2)^4}}{\frac{1}{x^4}} = 16, \tag{1.2}$$

  that is yielding a $10\log_{10} 16 = 12\,\mathrm{dB}$ power saving. This is a significant gain and one of the main incentives for using cooperative techniques.

- **Diversity Gain.** Providing additional independent copies of the *same information* via independent shadowing and fading channels yields diversity gains. These gains stem from the fact that with the amount of additional copies increasing, the probability of all of them being illegible decreases. The provision of such copies can be achieved, for example, by having a relay provide a copy in addition to the information received already via the direct link; or by having several relays provide copies in parallel. Diversity gains improve the performance of the system, such as the probability of error P_e or outage P_{out}. Either probability depends on the rate R in a similar fashion and is known to be related to the diversity gain at asymptotically high SNRs as

$$P_{\mathrm{out}}(R) = \frac{\mathrm{const}(R)}{\mathrm{SNR}^d}, \tag{1.3}$$

  where d is the diversity gain or diversity order, and const(R) is some constant depending on R. It effectively equals the number of copies of the same information; for instance, having one relay and a direct link providing two independent copies of the same information, yields $d = 2$. At the same level of performance metric, for example, requiring an outage probability of 1%, increasing the diversity order allows a saving in transmit power; for instance, with $d = 2$ about 3 dB transmission power can be saved in a flat Rayleigh fading channel.
- **Multiplexing Gain.** In an asymptotic SNR regime, the achievable data rate R is known to be proportional to the logarithm of the SNR, that is [8]

$$R = r \log_2 \mathrm{SNR} + \mathrm{const}. \tag{1.4}$$

  Here, r is referred to as the rate or multiplexing gain and is equal to the degrees of freedom of the channel, that is the number of independent channels over which *different information* can be sent. In a system with n_t transmit and n_r receive antennas, the maximum number of independent channels under favorable propagation conditions is $r = \min(n_t, n_r)$. For instance, it is clear that, if a source node possesses one transmit and a destination node two receive antennas or vice versa, they can communicate at a multiplexing gain of one only since only one independent channel can be provided. Using a relay in addition to the direct link, allows under certain conditions the creation of a second channel, hence doubling the multiplexing gain and thus the communication rate, assuming the SNR is kept constant.

These gains are not independent and hence can be traded one against another:

- **Power versus Data Rate.** As per Equation (3.77), the power gains due to pathloss and diversity gains can be used to increase the data rate R at a constant multiplexing gain r. This is typically done by increasing the cardinality of the signal constellation. For instance, gaining about 6 dB in power, one could switch from QPSK to 16QAM.
- **Diversity versus Multiplexing Gain.** This fundamental trade-off, also known as the diversity-multiplexing-tradeoff (DMT), will be discussed in Section 1.5.3. The DMT can be quantified by inserting Equation (3.77) into Equation (3.76), solving w.r.t. d and letting the SNR grow asymptotically large to arrive at [8]

$$d = -\lim_{\mathrm{SNR}\to\infty} \frac{\log P_{\mathrm{out}}(r \log_2 \mathrm{SNR})}{\log \mathrm{SNR}}. \tag{1.5}$$

  Due to the minus sign, the logarithm being a monotonically increasing function and the probability P_{out} increasing with r, increasing the multiplexing gain r necessarily requires the diversity gain d to be diminished. Each system configuration will hence have a maximum diversity gain d_{max} at minimum multiplexing gain r_{min}, and conversely a maximum multiplexing gain r_{max} at minimum diversity gain d_{min}. The above relationship quantifies these maxima and minima and also how they trade in-between, that is how reliability (diversity gain) trades against capacity (multiplexing gain) for a given channel and system configuration.

The above gains and trade-offs justify the research into different cooperative architectures, some canonical forms of which are described below.

1.2.3 Canonical Architectures

Whilst the degree of freedom of cooperative systems is very large, we list below some choices/parameters impacting most the realization of a particular architecture and thus used protocols:

- **Transparent versus Regenerative Relaying.** One of the foremost design drivers in cooperative systems is due to the choice between transparent and regenerative relaying approaches. Transparent relaying generally implies that the relay only amplifies the signal before retransmitting it. It is also possible, however, that the relay performs some other linear and non-linear operations in the analog domain, such as phase shifting, etc. Regenerative relaying, on the other hand, requires the relay to change the waveform and/or the information contents by performing some processing in the digital domain. An example is the relay receiving the information from the source, decoding, re-encoding and finally retransmitting it.
- **Traditional versus Distributed Space–Time Relaying.** Another important factor is the choice between traditional relaying and spatially distributed space–time processing relaying architectures. Traditional relaying has been around for some decades already and is realized by means of an arbitrary number of serial and/or parallel relays delivering the information from source towards destination. Space–time processing relaying, however, is realized by means of a distributed deployment of arbitrary number of (typically but not necessarily) synchronized nodes performing one of the many possible forms of distributed space–time processing. Known space–time techniques are thus applicable either directly or in modified form to these architectures, such as space–time coding, BLAST type algorithms or beamforming.
- **Dual-Hop versus Multi-Hop Networks.** The choice of the number of relaying stages is very important to system designers. As such, relays can be connected in series or operated in parallel. Increasing the number of serial relaying nodes increases the pathloss gain. Increasing the number of parallel relaying nodes increases the maximum diversity gain. Note that the relay channels ought to but do not necessarily have to be orthogonal so as to minimize interference; this can be achieved by using different frequencies, time slots, codes, etc.
- **Availability of Direct Link.** Depending on the propagation conditions, there may or may not be a direct link between source and destination or various relaying stages that is sufficiently strong to facilitate data transmission. Without the direct link, only pathloss gains can be achieved; with the direct link, the maximum diversity gain can also be increased. The direct link is usually available in situations where the system is capacity limited and not available where coverage limited.
- **Degree of Cooperation.** One generally distinguishes between the cases of supportive relaying and cooperative relaying. Typically, placing a relay node in-between a source and destination node is referred to as supportive relaying or simply relaying. Supportive relaying can be extended to cooperative communications, where at least two cooperative nodes are each other's respective relays at the same time to boost the other's communication links. Cooperative deployments clearly boost the maximum diversity and maximum multiplexing gains and, albeit not for every node, also the pathloss gain.

These choices and parameters have been visualized in Figure 1.1, where we have not shown the choice between transparent and regenerative architectures as any could be either of these two. The application of combinations of these canonical cooperative architectures to some practical scenarios is briefly discussed in the subsequent section.

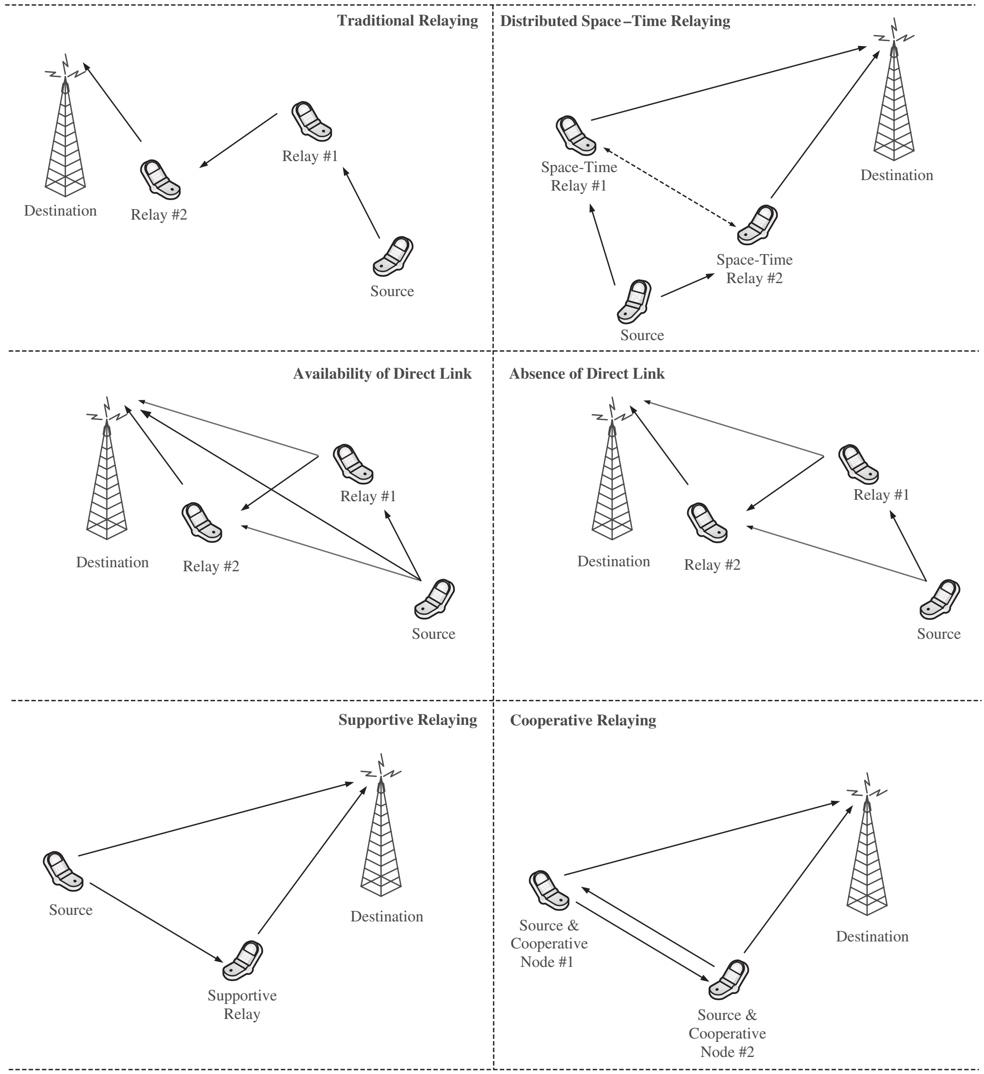

Figure 1.1 Exemplification of canonical relay architectures with the choice between traditional and distributed space–time processing relaying (top); between the availability of a direct link versus its absence (middle); and between supportive and cooperative relaying (bottom). All of these example architectures could realize transparent or regenerative relaying as well as any combination thereof

1.3 Application Scenarios

Before discussing the theoretical nature of cooperative communications, we shall give a few practical applications of cooperation by means of a set of chosen scenarios, ranging from cellular to wireless sensor networks.

1.3.1 Cellular Capacity and Coverage Extension

A network composed of adjacent cells, each served by a base station, is known as cellular network. Typical examples of today's cellular systems are GSM and UMTS; future roll-outs will include technologies from 3GPP LTE and WiMAX. Cellular networks generally suffer from three fundamental problems:

1. **Capacity:** Each cell is assigned a limited amount of resources in terms of bandwidth and allowed transmission power. These resources suffice to serve a given number of users; however, due to the ever-increasing number of users connecting to the same base station, the offered cell capacity is starting to become insufficient w.r.t. the required accumulated capacity of all associated users. The system is said to be capacity limited.
2. **Coverage:** The limit in transmission power has a profound impact on the coverage of the cell. Users at the cell edge experience insufficient power levels to support communication due to the weak signal of interest from the associated base station. The system is said to be coverage limited.
3. **Interference:** The necessity to roll out and support more than one cell leads to interference between these cells. Users at the cell edge hence not only experience insufficient power levels from their associated base station but also interference power from adjacent cells using the same frequency starts to become a dominant detrimental performance factor. The system is said to be interference limited.

These three effects are not independent, that is interference impacts coverage and capacity, etc. To alleviate these problems, however, the use of relays has been proposed where communication between a base station (BS) and a mobile station (MS) not only happens directly but also or exclusively via a relay station (RS). As explained in previous sections, such a deployment yields significant gains which, as per Figure 1.2, translates into the following [9]:

1. **Capacity Gains:** MSs not suffering from coverage problems can gain extra capacity since the BS can switch to higher modulation orders and thus boost system capacity.

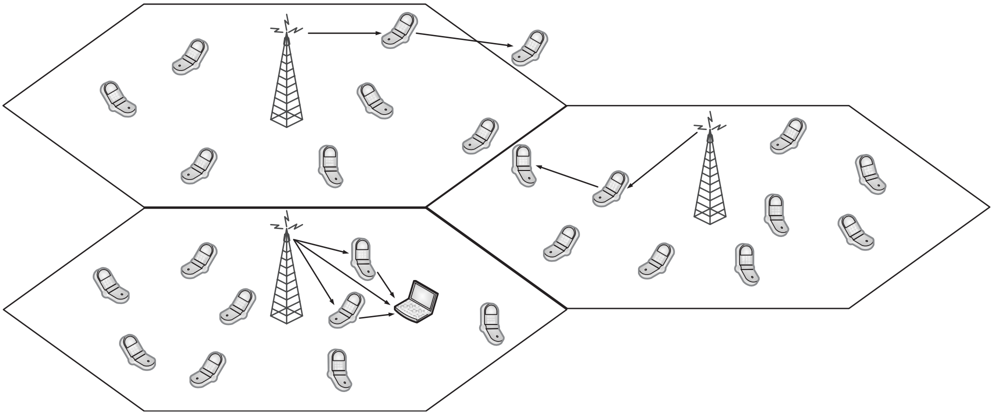

Figure 1.2 Cellular scenario: relaying boost performance of users that are capacity limited (bottom-left cell); coverage limited due to range (top-left cell) or interference (middle-right cell)

2. **Coverage Gains:** MSs at or beyond the cell edge as well as in coverage holes receive a signal of sufficient quality to communicate and hence the cell coverage is increased.
3. **Interference Gains:** Part of the capacity and/or coverage gains can be used to decrease the transmission power and hence the general interference temperature in the system. This reduction in interference, however, has to be gauged against the increased interference caused by the additional relay traffic.

Due to these obvious and potential gains, Vodafone pushed to include such a relaying approach into the 3GPP standard via ETSI's UTRA air interface technology proposals group Epsilon. The proposed access method was termed Opportunity Driven Multiple Access (ODMA) relaying protocol [10–15]. For various reasons, however, it was not selected as a candidate for 3GPP but included as an optional protocol in several early standard releases of 3GPP; it was eventually discontinued in the R99 standard release.

Many lessons have been learned ever since and current standards developing bodies are picking up on the relaying concept. In particular, IEEE 802.16j has developed a relay enabled mode to WiMAX in the hope of giving it a competitive edge in 3GPP LTE developments. The latter currently only standardizes repeaters to leverage coverage problems; however, it is envisaged that LTE Advanced will include cooperative relay features. This and related topics are discussed in [4, 5] and will also be dealt with in Chapter 5 of this book.

1.3.2 WLAN Capacity and Coverage Extension

Built on IEEE 802.11 technology, WLANs are a unique communication phenomenon of recent times. This is corroborated by the millions of WLAN stations in use today in homes, offices, cafes, train stations, airports, etc. [16–27]. In fact, usage in urban environments is already so dense that almost complete coverage could be provided and interference commences to become a serious performance-limiting factor.

Similarly to the cellular case, cooperative techniques are promising to boost capacity and increase coverage even further. This is exemplified by means of Figure 1.3, where a home WLAN access point (AP) communicates with a user in the street via another user acting as a relay. The question on the optimum number of allowed hops arises which, according to some preliminary studies using realistic system parameters, is apparently only two on average.

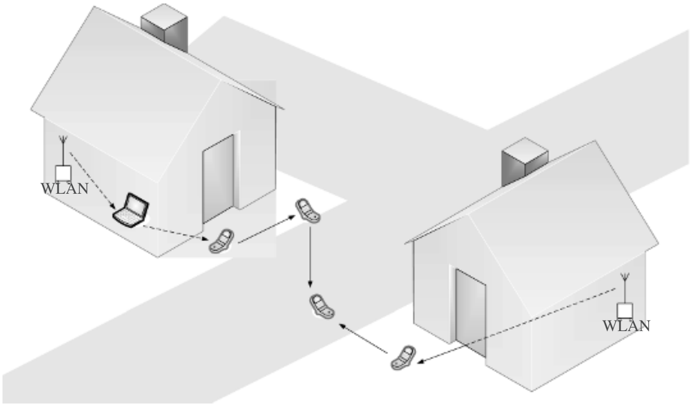

Figure 1.3 A WLAN station installed inside a house provides access to users in the street via relays

1.3.3 Vehicle-to-Vehicle Communication

Future vehicles will allow for platooning (automated steering within a group of cars), in-vehicle internet access, inter-vehicle communication, etc. [28–45]. The increasing density of vehicles allows for the deployment of cooperative vehicle systems. The advantage of cooperative techniques, as per Figure 1.4, becomes apparent here since the redundant links established in cooperation offer high link stability in such volatile and dynamic propagation conditions.

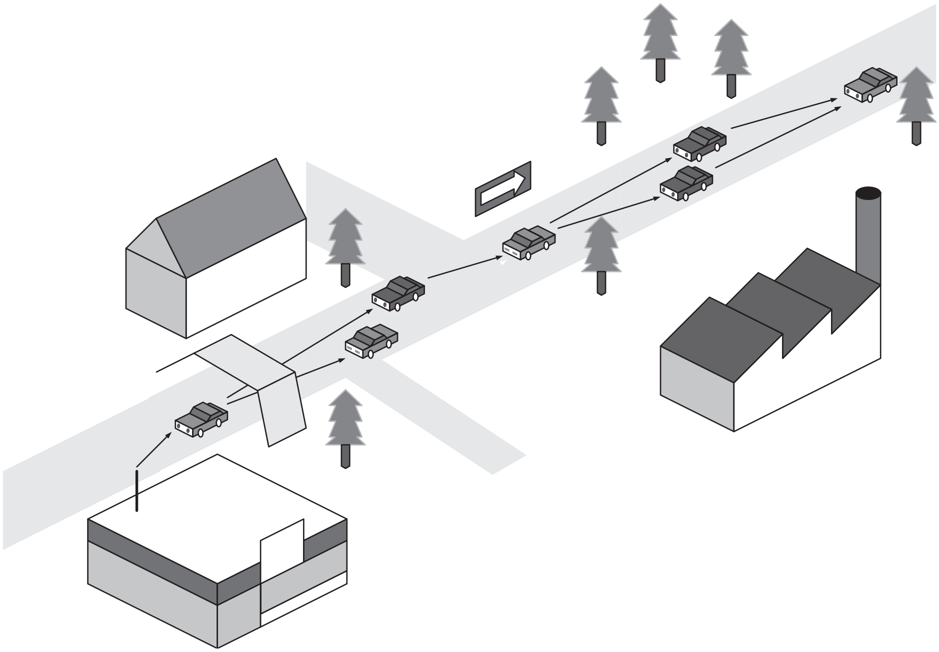

Figure 1.4 Distributed vehicle-to-vehicle communication scenario, where cars cooperate to facilitate the relay and delivery of information

1.3.4 Wireless Sensor Networks

Sensor networks have been researched and deployed for decades; their wireless extension, however, has witnessed a tremendous upsurge in recent years [25, 46–62]. This is mainly attributed to the unprecedented operating conditions of WSNs, that is an enormous number of nodes of extremely low complexity and cost reliably operating under stringent energy constraints.

As of today, a major problem in deploying WSNs is their dependence on limited battery power. A main design criterion is to optimize the lifetime of the network without jeopardizing reliable and efficient communications from sensor nodes to other nodes as well as data sinks. Limiting the transmission power is hence a must and coverage is therefore a predominant problem. As exemplified in Figure 1.5, cooperative techniques are beneficial if not crucial in closing coverage gaps and therefore maintaining network integrity.

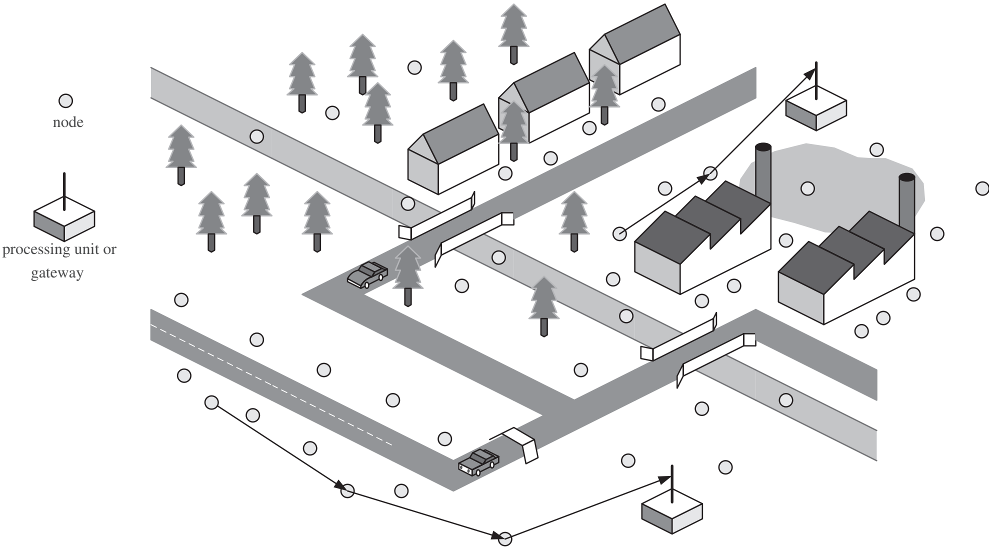

Figure 1.5 WSN application where sensors are heavily coverage limited and hence rely on relaying to deliver sensed information

Hybrid solutions are also foreseen, such as unmanned aerial vehicles (UAVs) and wireless sensor networks. As in Figure 1.6, a set of sensors in areas difficult to reach communicates with a set of UAVs, which in turn cooperate between each other prior to sending the data to a processing unit. In [63–65], it has been shown that such cooperative UAVs considerably increase the reliability of the transmission of sensor readings under real-world impairments.

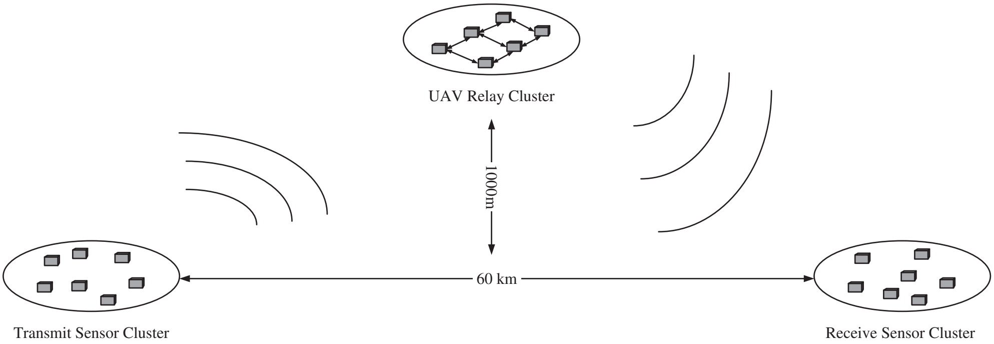

Figure 1.6 Distributed and cooperative UAVs acting as relays, which can utilize beamforming, STCs, multiplexing, etc., to relay sensor readings

1.4 Pros and Cons of Cooperation

Although partially discussed before, the advantages and disadvantages of the canonical cooperative architectures of Section 1.2.3 are summarized below. We will not only discuss the PHY layer but also dwell on some system deployment aspects.

1.4.1 Advantages of Cooperation

The key advantages of using supportive, cooperative or space–time relays in the system can be summarized as follows:

- **Performance Gains.** Large system-wide performance gains can be achieved due to pathloss gains as well as diversity and multiplexing gains. These translate into decreased transmission powers, higher capacity or better cell coverage.
- **Balanced Quality of Service.** Whilst in traditional systems users at the cell edge or in shadowed areas suffered from capacity and/or coverage problems, relaying allows to balance this discrepancy and hence give (almost) equal quality of service (QoS) to all users.
- **Infrastructure-Less Deployment.** The use of relays allows the roll-out of a system that has minimal or no infrastructure available prior to deployment. For instance, in disaster-struck areas, relaying can be used to facilitate communications even though the cellular system is nonfunctioning. For hybrid deployments, that is a cellular system coupled with relays, the deployment and maintenance costs can be lowered as has been shown in [66].
- **Reduced Costs.** Compared to a purely cellular approach to provide a given level of QoS to all users in the cell, relaying is a more cost effective solution. In [66], however, it has also been shown that whilst savings are not as dramatic as hoped for, the capital and operational expenditures are generally lower when relays are used.

1.4.2 Disadvantages of Cooperation

Some major disadvantages of using supportive, cooperative or space–time relays in the system are given below:

- **Complex Schedulers.** Whilst maintaining a single cooperative relaying link is a fairly trivial task, at system level with many users and relays this quickly becomes an arduous task. As such, relaying requires more sophisticated schedulers since not only traffic of different users and applications needs to be scheduled but also the relayed data flows. Any gains due to cooperation at the physical layer dissipate rapidly if not handled properly at medium access and network layers.
- **Increased Overhead.** A full system functioning requires handovers, synchronization, extra security, etc. This clearly induces an increased overhead w.r.t to a system that does not use relaying.
- **Partner Choice.** To determine the optimum relaying and cooperative partner(s) is a fairly intricate task. Also, the complexity of maintaining such cooperative partnership is higher w.r.t. noncooperative relaying.
- **Increased Interference.** If the offered power savings are not used to decrease the transmission power of the relay nodes but rather to boost capacity or coverage, then relaying will certainly generate extra intra- and inter-cell interference, which potentially causes the system performance to deteriorate. An optimum trade-off needs, therefore, to be found at system level.
- **Extra Relay Traffic.** The relayed traffic is, from a system throughput point of view, redundant traffic and hence decreases the effective system throughput since in most cases resources in the form of extra frequency channels or time slots need to be provided.
- **Increased End-to-End Latency.** Relaying typically involves the reception and decoding of the entire data packet before it can be re-transmitted. If delay-sensitive services are being supported, such as voice or the increasingly popular multimedia web services, then the latency induced by the decoding may become detrimental. Latency increases with the number of relays and also with the use of interleavers, such as utilized in GSM voice traffic. To circumvent this latency, either simple transparent relaying or novel decoding methods need to be used.
- **Tight Synchronization.** A tight synchronization needs generally be maintained to facilitate cooperation. This in turn requires expensive hardware and potentially large protocol overheads since nodes need to synchronize regularly by using some form of beaconing or other viable techniques.

- **More Channel Estimates.** The use of relays effectively increases the number of wireless channels. This requires the estimation of more channel coefficients and hence more pilot symbols need to be provided if coherent modulation was to be used.

The list of disadvantages is clearly longer than the list of advantages, some of which have been summarized for the most important canonical cooperative architectures in Table 1.1. A careful system design is therefore needed to realize the full gains of cooperative relaying systems and to ensure that cooperation does not cause deterioration of the system performance. This is the prime reason why many of the latest research efforts concentrate on above-listed issues.

Table 1.1 Principal pros and cons of some canonical cooperative architectures

	Main advantages	Main disadvantages
Supportive relaying	Pathloss gains	Increased interference
	Balanced user QoS	Complex schedulers
Cooperative relaying	Diversity gains	Optimum partner choice
	Balanced user QoS	Complex schedulers
Space–time relaying	Diversity gains	Increased overhead
	Multiplexing gains	Tight synchronization
	Available space–time codes	More channel estimates
	Balanced user QoS	Complex schedulers

1.4.3 *System Tradeoffs*

From this list of advantages and disadvantages, it is obvious that many system design parameters can be traded against one another. Some important system trade-offs are discussed below:

- **Coverage versus Capacity.** As already discussed in some detail, cooperative systems allow coverage to be traded against capacity or, equivalently, outage versus rate or diversity versus multiplexing gains. Therefore, the system designer has the choice to let a relay help boost capacity or increase the coverage range. Increasing one inherently diminishes the other.
- **Algorithmic versus Hardware Complexity.** Solving the coverage and capacity problem by means of more cellular base stations requires more complex and hence costly hardware, not to mention the expensive real estate to physically place the base stations. Relays, on the other hand, are of relatively low hardware complexity. However, the decrease in hardware complexity by using relays yields an increase in algorithmic complexity since scheduling, synchronization, handover and other algorithms become significantly more complex. An optimum solution trading algorithmic with hardware complexity hence needs to be determined, likely leading to a hybrid deployment.
- **Interference versus Performance.** As discussed in Section 1.2.2, cooperative communications yields gains which can either be used to decrease the transmission power and hence generated interference or increase capacity/coverage. Furthermore, relaying generates extra traffic, which is an additional source of interference.
- **Ease-of-Deployment versus Performance.** Relays can be deployed in a planned and unplanned manner. In the former, the network designer optimizes the placement and parameterization of the static relay node; this is a complex task but leads to superior performance. In the latter, relays are deployed in an unplanned manner and hence can be stationary or mobile; deployment is therefore significantly simplified at the cost of decreased performance w.r.t. the planned roll-out.

- **Cost versus Performance.** Being a traditional trade-off, the cost of the chosen cooperative solution has a profound impact on its performance. Clearly, deploying highly complex relaying nodes that allow, say, for cooperative space–time relaying induces high costs but also improved performance.

These trade-offs are also visualized in Figure 1.7.

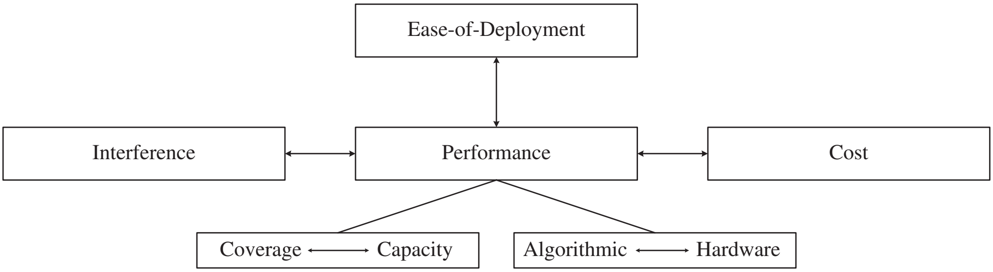

Figure 1.7 At a given performance level, coverage can be traded capacity and algorithmic with hardware complexity. Performance can also be traded against amount of interference, ease-of-deployment and cost

1.5 Cooperative Performance Bounds

So far, we have qualitatively described the potentials of cooperative relaying systems. In this section, we will only briefly quantify achievable performance bounds for cooperative relaying systems; for a more detailed treatment on this matter, the interested reader is referred to [67].

Although not the focus of this book, capacity bounds are very useful as they indicate whether a given cooperative relaying scheme performs well, that is close to the bound, or not. These bounds therefore impact the design of hardware as well as protocols of the entire OSI stack. Hence they serve as a justification and inspiration for any subsequent exposures in this book.

Staying at an introductory level and thus without going into great mathematical detail, we will discuss the achievable rate gains, outage probability gains, and the DMT in the context of cooperative systems.

1.5.1 Capacity Gains

Shannon proved that one can design codes facilitating a communication rate of R bits/symbol with arbitrarily small error [68]. His proof holds for codes that are of infinite duration (or, in practice, very long), so as to average out the effect of noise. His theory, however, is not concerned with the code construction or code complexity, nor with decoding delay.

The maximum data rate at which reliable communication is possible is referred to as capacity C of the channel. Given a signal power S and noise power N at the receiver input, the capacity offered by an additive white Gaussian noise (AWGN) channel is:

$$C = \log_2\left(1 + \frac{S}{N}\right), \tag{1.6}$$

which has been normalized by the bandwidth W; the units are hence $[C] =$ bits/s/Hz. Signals, however, typically traverse a wireless channel that impacts the received signal power S and hence the capacity offered by such a channel. Whilst the deterministic effect of pathloss only scales the useful signal power

S in Equation (1.6), the randomness introduced by the wireless channel changes the capacity since the useful signal power effectively varies over the duration of the transmitted codeword. Depending on the type of channel variations, one typically distinguishes ergodic and nonergodic fading channels.

1.5.1.1 Ergodic Channel

A channel characteristic typically assumed in the context of Shannon capacity is that of an ergodic channel. Rigorously speaking, a process is ergodic if the time averages may be used to replace ensemble averages. In more practical terms, this means that the channel varies sufficiently rapidly over the duration of the transmission and hence traverses all fading states.

Ergodicity allows one to apply the concept of averages since the channel's average mutual information over all (infinitely long codewords) is the same. This means that an ergodic channel can support the following maximum error-free transmission rate with 100% reliability:

$$C = \mathbb{E}_g \left\{ \log_2 \left(1 + g \frac{S}{N} \right) \right\}, \tag{1.7}$$

where g is the instantaneous channel gain/power and $\mathbb{E}_g \{\cdot\}$ denotes the expectation operator w.r.t. λ. The random channel fluctuation g may comprise the effects of fading and shadowing but typically only includes fading due to shadowing as being a fairly slowly varying effect. The notion of an ergodic channel w.r.t. two transmitted codewords of infinite duration is illustrated in Figure 1.8.

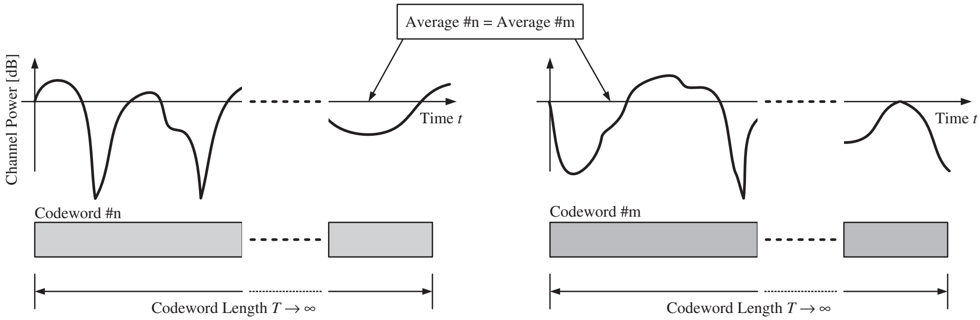

Figure 1.8 In the case of an ergodic channel all fading states are traversed over the duration of a Shannon codeword, thereby facilitating the provision of an average channel capacity

For example, if the channel obeys flat Rayleigh fading with its channel gain hence having a probability density function (PDF) $p_g(g) = 1/\overline{g} \cdot \exp(-g/\overline{g})$, the average capacity can be calculated as [69, 70]:

$$\begin{aligned} C &= \mathbb{E}_g \left\{ \log_2 \left(1 + g \frac{S}{N} \right) \right\} && \text{(1.8a)} \\ &= \int_0^\infty \log_2 \left(1 + g \frac{S}{N} \right) \cdot \frac{1}{\overline{g}} \, \mathrm{e}^{-g/\overline{g}} \, \mathrm{d}g && \text{(1.8b)} \\ &= -\mathrm{e}^{1/\overline{\gamma}} \mathrm{Ei}\left(-1/\overline{\gamma}\right) / \log(2), && \text{(1.8c)} \end{aligned}$$

where $\mathrm{Ei}(\cdot)$ is the exponential integral, $\overline{g}$ is the average fading power usually normalized to unity, $\gamma = gS/N$ is the instantaneously experienced SNR at the decoder, and $\overline{\gamma} = \overline{g}S/N$ its average. The step from Equation (1.8b) to Equation (1.8c) follows from [71, Section 4.337.2].

1.5.1.2 Capacity Gains

To illustrate the capacity gains, we will follow the analysis exposed in [72–74]. To this end, the transceiver architecture as shown in Figure 1.1 (bottom-right) is assumed. The capacity gains of this simple cooperative relaying schemes assuming Rayleigh fading channels are exemplified by means of Figures 1.9 and 1.10, where the former assumes a symmetric communication scenario and the latter assumes the average fading power between the first node and destination to be significantly better than between the second node and destination.

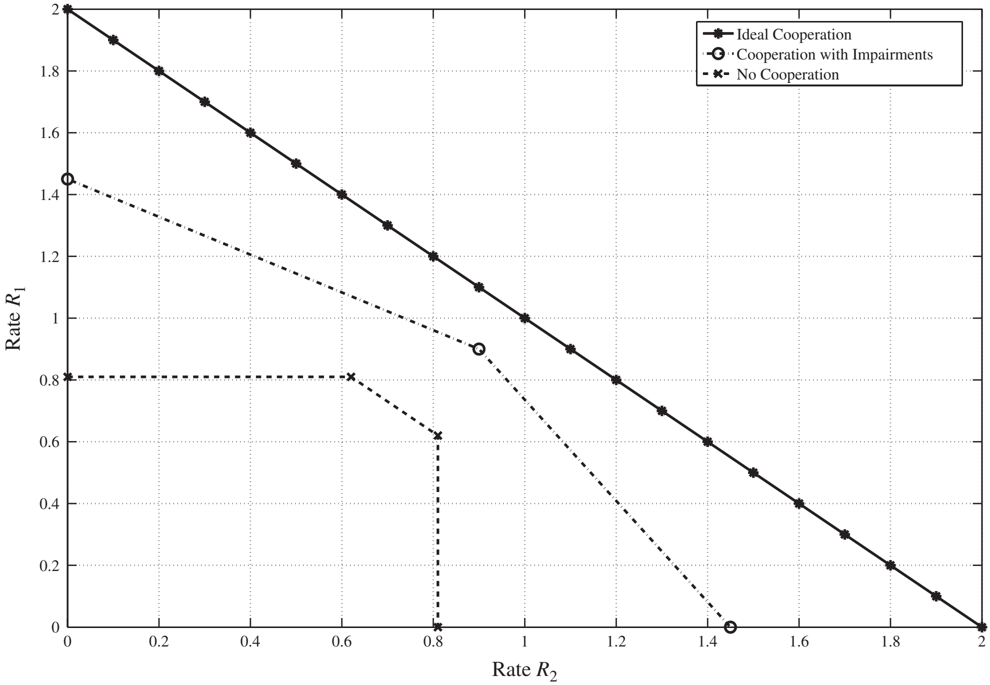

Figure 1.9 Symmetric rate region, where both users have equal channel conditions to the destination, assuming no cooperation, ideal cooperation with error-free inter-user channel, and cooperation with good inter-user channel [72]

With reference to these figures, the case of ideal cooperation is for a noiseless inter-user channel and serves as an upper bound of cooperation. No cooperation between the nodes yields the typical multiple access channel. In the cooperative case, as the inter-user channel degrades, performance approaches that of no cooperation. Typical points of interest are listed below:

- **Equal Rate Point $R_1 = R_2$.** These rates are achieved along the 45° line from the origin. It is a useful indicator if the system treats both users with equal priority. With reference to the symmetric and asymmetric communication scenarios of Figures 1.9 and 1.10, cooperation yields an equal rate gain of about 20% and 200%, respectively.
- **Maximum Rate Sum Point $\max(R_1 + R_2)$.** These rates are achieved along the points where the sum of both nodes is maximal. It is a useful indicator if the total system capacity is of importance and not how much rate each user is guaranteed. With reference to the symmetric and asymmetric communication cases, cooperation yields a maximum rate sum gain of about 20% and 10%, respectively.

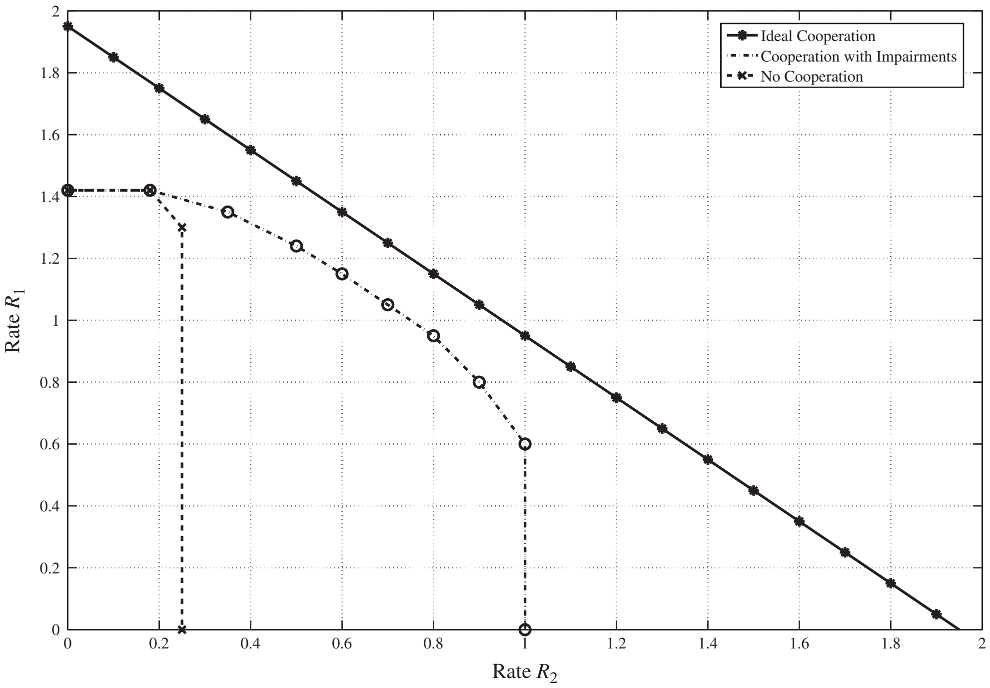

Figure 1.10 Asymmetric rate region, where the first user has much better channel conditions to the destination than the second user, assuming no cooperation, ideal cooperation with error-free interuser channel, and realistic cooperation with bad interuser channel [72]

- **Degraded Relay Rate Points $R_1 = 0$, $R_2 \neq 0$ and $R_1 \neq 0$, $R_2 = 0$.** These rates are achieved when setting one of the rates to zero. It is a useful indicator of how much maximum rate can be delivered to a user whilst the other user is not transmitting own data but only supportively relaying. With reference to the symmetric communication scenario, cooperation almost doubles the degraded relaying rate points; in the asymmetric case, however, only the user with bad channel conditions really profits.

It becomes apparent that using a cooperative relaying approach significantly improves the capacity of each user as well as the network. The gains are of particular importance in the asymmetric case where one user suffers from bad channel conditions. These results, in fact, inspired research into practical communication schemes that are capable of achieving the promised rate gains. To this end, it has already been shown [74] that in the design region of interest, an increase in sum capacity approximately equals the increase in coverage area and that a simple repetition-based coding scheme using CDMA spreading sequences performs well within the rate regions. The remainder of this book is thus dedicated to various techniques able either to perform well within or even approach the bounds of this rate region.

1.5.2 Rate Outage Gains

Shannon's information theory is not designed to cater for communication scenarios where average channel conditions change from codeword to codeword. To this end, the concept of rate outage probability has been successfully introduced in the information theory community. Its application to the context of cooperative relaying systems is discussed below.

1.5.2.1 Nonergodic Channel

In contrast to an ergodic channel, in the case of a nonergodic channel the channel does not vary sufficiently fast to traverse all fading states over the duration of the communication. In other words, a process is nonergodic if samples help meaningfully to predict values that are very far away in time from that sample, that is the stochastic process is sensitive to initial conditions. Practical situations where this occurs are when the channel is very slow fading and/or experiences severe and long-lasting shadowing.

The concept of averages is not very meaningful in the context of nonergodic channels since it will be different for each transmitted codeword. A nonergodic channel can hence not support a maximum error-free transmission rate with 100% reliability; however, it can support any given rate R with a certain probability $P_{\text{out}}(R)$, which is referred to as the rate outage probability. The notion of a nonergodic channel w.r.t. two transmitted codewords of infinite duration is illustrated in Figure 1.11.

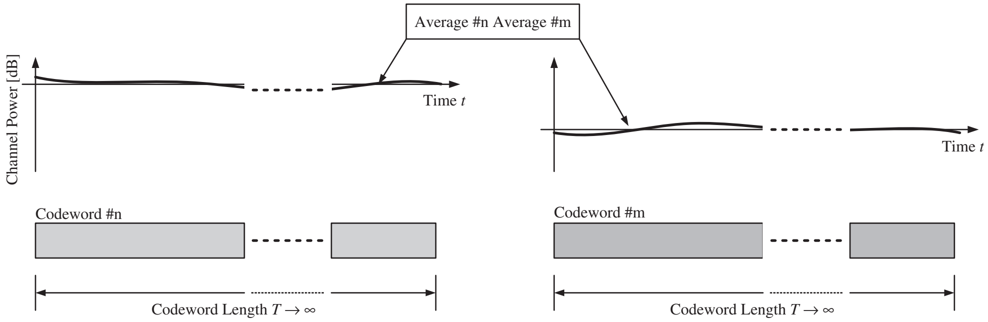

Figure 1.11 In the case of a nonergodic channel not all fading states are traversed over the duration of a Shannon codeword, thereby preventing the provision of an average channel capacity and hence requiring the concept of outage to be invoked

To quantify the outage probability, we assume an instantaneous power $\gamma = gS/N$ at the decoder input due to which an information rate of $C(\gamma) = \log_2(1+\gamma)$ bits/s/Hz can be supported. The channel is in outage if this rate falls below a threshold information rate R; the corresponding outage event is $C(\gamma) < R$ or $\gamma < \left(2^R - 1\right)$. The outage probability is hence

$$P_{\text{out}} = \Pr\left(\gamma < \left(2^R - 1\right)\right) = \int_0^{2^R-1} p_\gamma(\gamma)\mathrm{d}\gamma, \tag{1.9}$$

where $\Pr(\cdot)$ denotes the probability and $p_\gamma(\gamma)$ the PDF of the SNR. For instance, for a Rayleigh fading process with mean power $\overline{\gamma}$ and PDF given in the previous section, the outage probability is

$$P_{\text{out}} = 1 - \exp\left(-(2^R - 1)/\overline{\gamma}\right). \tag{1.10}$$

From Equation (1.10) it is clear that the outage probability decreases quickly with increasing SNR; since the system designer is interested in low outage probabilities, it is therefore customary to plot the outage or its compliment in logarithmic scale.

1.5.2.2 Rate Outage Gains

To illustrate the rate outage probability gains, or simply outage gains, we will utilize the same communication model as depicted in Figure 1.1 (bottom-right) and follow the analysis first exposed in [75–77]. The protocol is fairly simple: Both users send their data to the destination and hence to each other. If a user manages correctly to decode the other user's information, it relays it to the destination; if not, it continues sending its own information.

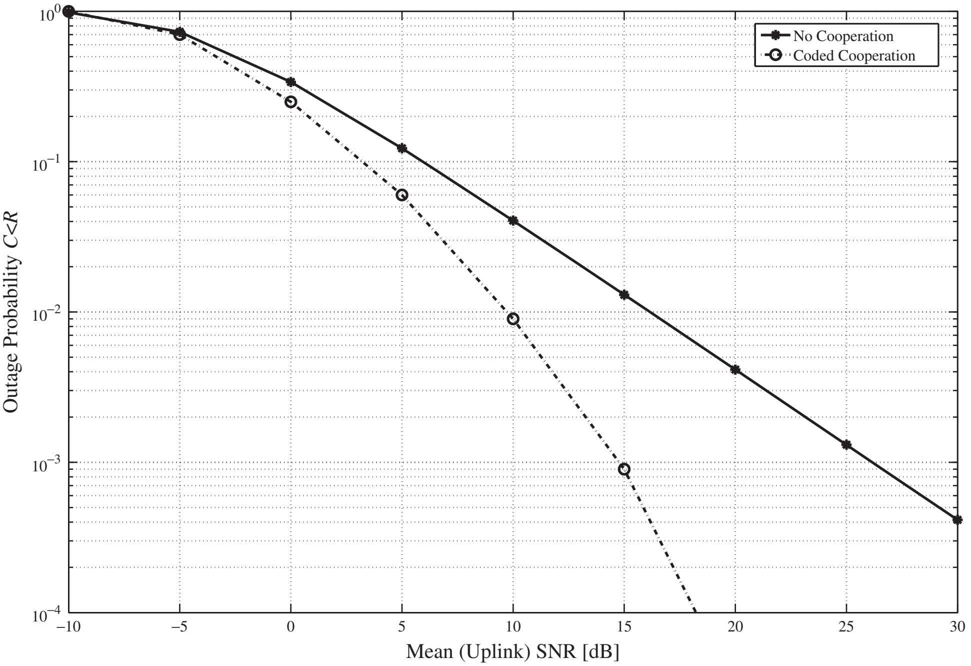

Figure 1.12 Outage versus average uplink SNR, where inter-user channel is significantly weaker; $R = 0.5$ bits/s/Hz [77]

As per Figure 1.12, the outage gains due to cooperation are significant. In the region of interest, that is typically between 1 and 10% outage probability, up to 6 dB transmission power can be saved on average. These gains are attributed to the fact that the probability of both direct and cooperative relaying link being in outage is much lower than just the direct link being in outage. More complex topologies and protocols follow the same trend and cooperation is generally able to provide significant outage gains, whether the random channel fluctuations be due to fading or shadowing.

1.5.3 Diversity-Multiplexing Tradeoff

The diversity-multiplexing trade-off (DMT) has been intuitively presumed by system designers for decades already but was first quantified in [8]. It essentially describes how quickly the probability of outage decreases and the communication rate increases with an increase in SNR. Since the concept of outage is not applicable over ergodic channels, the DMT in the Shannon sense is only applicable to nonergodic channels. However, as will be discussed below, the DMT is also applicable to any real-world system operating over slow- or fast-fading channels, which allows one to trade reliability against rate.

With reference to Equation (3.76), the diversity gain relationship can be reformulated as

$$d = -\lim_{\mathrm{SNR}\to\infty} \frac{\log P_{\mathrm{out}}(R, \mathrm{SNR})}{\log \mathrm{SNR}}, \tag{1.11}$$

where P_{out} (R, SNR) is the outage probability in the Shannon sense at a given average SNR and required rate R, and d is the diversity gain. At an asymptotically high SNR, it is equivalent to the gradient of the outage curves. If $d = 0$ then with an increasing SNR no decrease in outage is achieved, that is the gains one can potentially obtain from increasing the SNR are used somewhere else (most likely to increase the rate). Taking the example of Figure 1.12, the noncooperative case has a diversity order of $d = 1$ whereas the cooperative protocol can achieve the double diversity order of $d = 2$. Clearly, a steeper gradient yields increasing gains with an increasing SNR.

With reference to Equation (3.77), the multiplexing gain relationship can be reformulated as

$$r = \lim_{\mathrm{SNR}\to\infty} \frac{R(\mathrm{SNR})}{\log_2 \mathrm{SNR}}, \tag{1.12}$$

where r is the rate or multiplexing gain. At an asymptotically high SNR, it is equivalent to the gradient of the capacity or rate curves. If $r = 0$ then with an increasing SNR no increase in rate R is achieved, that is the gains one can potentially obtain from increasing the SNR are used somewhere else (most likely to decrease the outage).

By inserting Equation (1.12) into Equation (1.11) and solving w.r.t. d, we arrive at the general DMT expression

$$d = -\lim_{\mathrm{SNR}\to\infty} \frac{\log P_{\mathrm{out}}(r \log_2 \mathrm{SNR})}{\log \mathrm{SNR}}, \tag{1.13}$$

which implies that increasing the rate multiplexing capabilities inherently requires the reliability of these rates to be decreased, and vice versa.

Similar arguments can be used to derive the DMT for real-world systems operating over slow- or fast-fading channels, where for slow-fading channels P_{out} in the Shannon sense needs to be replaced by the P_{out} for the given real-world scheme and for fast-fading channels by the average error rate P_{e} [8]. We will use the example of a system deploying quadrature amplitude modulation (QAM) with variable modulation index over fast fading channels to illustrate the DMT in such a context [8]. The average symbol error probability (SEP) of QAM over a Rayleigh fading channel at asymptotically high SNRs is given by $P_{\mathrm{e}}(R, \mathrm{SNR}) = \lim_{\mathrm{SNR}\to\infty} 2^R/\mathrm{SNR}$ which, having inserted (1.12), yields

$$P_{\mathrm{e}}(r \log \mathrm{SNR}) = \lim_{\mathrm{SNR}\to\infty} \frac{2^{r \log_2 \mathrm{SNR}}}{\mathrm{SNR}}. \tag{1.14}$$

Finally, inserting Equation (1.14) into Equation (1.13), gives us the DMT for the real-world communication scheme using QAM:

$$d = 1 - r, \tag{1.15}$$

which is depicted in Figure 1.13. For the degraded point at $d_{\mathrm{max}} = 1$ and $r_{\mathrm{min}} = 0$, that is at no increase in rate, this means that with increasing SNR reliability can exhibit a maximum gradient of one; or, in other words, doubling the SNR (in dB) doubles the reliability (in log scale). For the other degraded point at $r_{\mathrm{max}} = 1$ and $d_{\mathrm{min}} = 0$, that is at no increase in reliability, this means that with increasing SNR the data multiplexing capability can exhibit a maximum gradient of one; or, in other words, doubling the SNR (in dB) doubles the rate capabilities. Any other point along the line and in-between these degraded points can, for example, be achieved by simple time multiplexing; for instance, if one wishes to have a diversity gain of $d = 0.5$ and a multiplexing gain of $r = 0.5$, then – for increasing SNR – one should communicate half of the time using a constant modulation, which gives the reliability benefits, and the other half of the time using an adaptive modulation that gives the rate benefits.

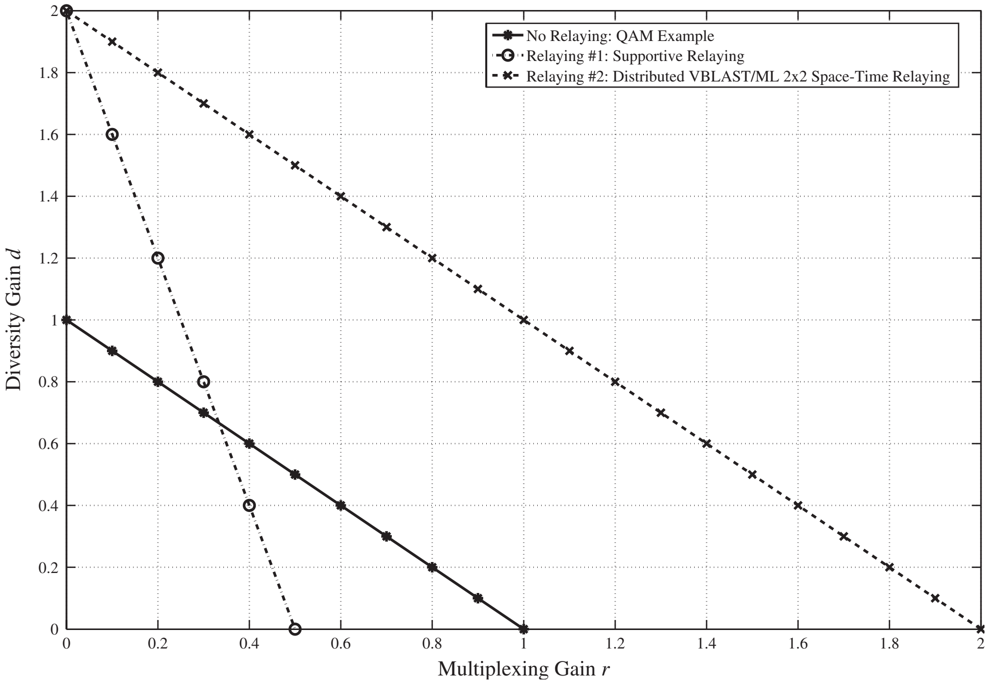

Figure 1.13 The diversity gain shows how fast the outage or error probability decreases with SNR and the multiplexing gain shows how fast the rate can be increased with SNR. A simple repetition-based relaying protocol can increase the diversity order but not the multiplexing gain. A more sophisticated 2 × 2 space-time relaying protocol using V-BLAST and ML can increase both diversity and multiplexing gains

Figure 1.13 also shows the impact of two example cooperative relaying protocols. The first protocol is a simple repetition-based relaying protocol that clearly yields double diversity gain but half multiplexing gain since the same data is repeated twice. The second protocol aims at increasing the multiplexing capabilities by means of relaying, which is achieved, for example, by using distributed 2 × 2 space–time relaying where one cooperative relay acts as a transmitting antenna array close to the source and another cooperative relay acts as a receive antenna close to the sink. If the channel between source and transmit relay as well as destination and receive relay is ideal, a V-BLAST scheme with maximum likelihood (ML) receiver achieves the depicted DMT [8].

Apart from the power and shadowing macro-diversity gains of such cooperative relaying schemes, the ability to trade reliability against rate is of great help to system designers. For instance, slow narrowband fading channels offer little diversity, which can be improved by a relaying system at the expense of a rate improvement. Wideband channels, on the other hand, already offer a healthy amount of diversity that can be used either to increase the reliability (CDMA-based transceivers) or rate (OFDM-based transceivers). Depending on this choice, cooperative relaying can then further boost the rate (CDMA-based transceivers) or reliability (OFDM-based transceivers), or both. These trade-offs have motivated research into practical cooperative relaying protocols as outlined in this book.

1.6 Definitions and Terminology

We now move on to some definitions and terminologies for cooperative communication systems. Indeed, different terms are currently being used for the same communication principle, and vice versa.

It is hence vital to harmonize an understanding of such systems, where some first excellent steps have been taken by [78]. We proceed to discuss terminologies in the context of three areas: (i) the relaying node itself; (ii) its one and two-hop neighborhood, which needs to be controlled by suitable multiple access methods; and (iii) the thus formed cooperative relaying network at system level.

1.6.1 Relaying Node

We will now characterize the different types of behavior of relaying nodes, available relaying families and how information is being handled in the relaying node. For the sake of clarity, this is also summarized in Figure 1.15.

1.6.1.1 Node Behaviors

Nodes that relay or cooperate play a central role in cooperative networks. Their behavior has a profound impact onto the system performance and, as exemplified in Figure 1.14, can generally be classified as follows:

- **Egoistic Behavior (no help).** This is the most typical node behavior found in today's wireless communication systems. Here, each node communicates with the base station separately if it has data to transmit or it remains idle if it has no own traffic to transmit, even though it could help another node that has. Other nodes are generally seen as competitors, that is increasing the resources for one node requires it to be decreased for other nodes. Nodes of networks following such design usually strongly experience the impact of the wireless channel in that nodes with good channel conditions enjoy large rates, whereas nodes with a bad channel suffer from low rates.
- **Supportive Behavior (unidirectional help).** Such a behavior is well known in the *ad hoc* community, where data is delivered from a source towards a destination via relay(s) that do not have data of their own to transmit. Supportive relaying-based designs, however, slowly find their way into cellular systems; for instance, LTE and WiMAX are likely to integrate supportive relaying

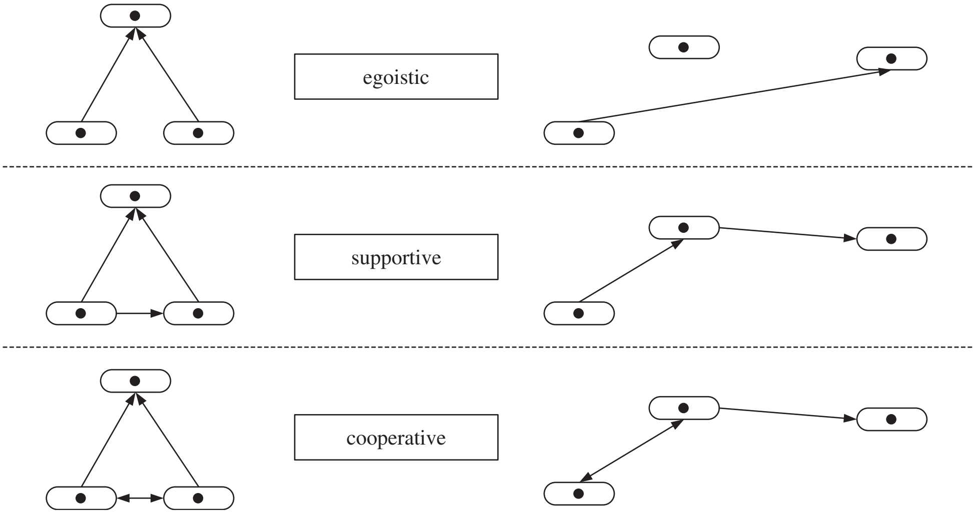

Figure 1.14 Typical forms of node behavior. In the symmetric case (left), either node could be supportive or cooperative; in the asymmetric case (right), the better option is typically chosen

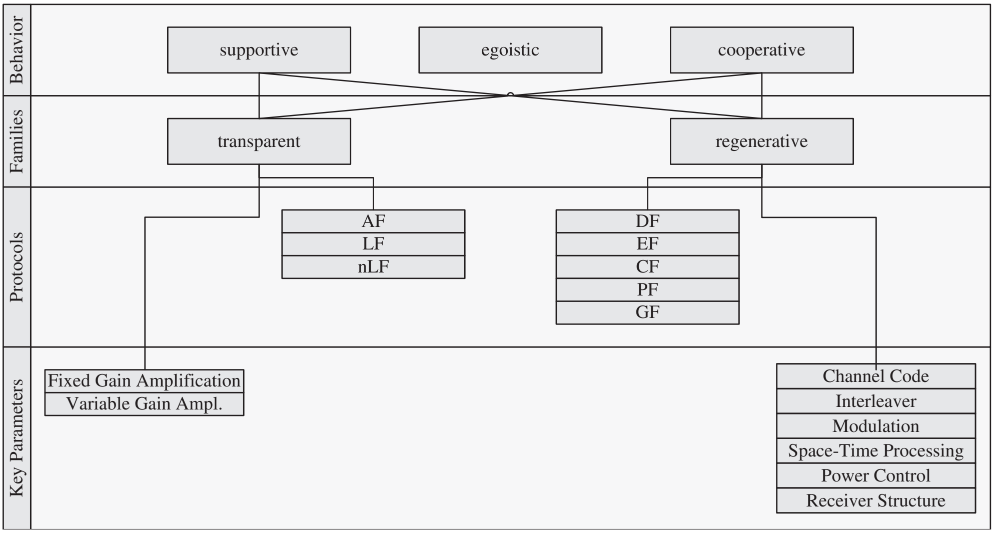

Figure 1.15 Taxonomies and definitions related to the relaying node

nodes. Clearly, the relay is not gaining in performance at that particular instant since it is only helping the source node; however, in the long run, the relay (if not part of a planned or unplanned infrastructure) may also be in the situation of requiring help and hence enjoying the supportive help of other relaying nodes.

- **Cooperative Behavior (mutual help).** The truly cooperative behavior is exhibited by nodes mutually helping each other, that is all involved nodes have data to transmit and mutually try to get it successfully delivered. System designs following a cooperative node paradigm are still a long way off. Nonetheless, cooperation generally smoothes the impact of the wireless channel in that even nodes in bad channel conditions experience an acceptable channel quality and hence sufficient communication rates.

Future networks are likely to be composed of nodes of all three behaviors. One can generally state that the higher the level of cooperation, the better the performance but also the more complicated the set-up and maintenance of the scheme.

1.6.1.2 Transparent Relaying Protocols

As is well documented throughout available literature on this subject, a whole gamut of different relaying methods exists today. They can roughly be classified into two groups, that is transparent and regenerative relaying protocols.

Using the family of transparent relaying, the relay does not modify the information represented by a chosen waveform. Very simple operations are usually performed, such as simple amplification, phase rotation, etc. Since no digital operations are performed on the signal, the analog signal is received in one frequency band, amplified and momentarily retransmitted on another frequency band. Example protocols belonging to the transparent relaying family are:

- **Amplify and Forward (AF).** Constituting one of the simplest and most popular relaying methods, the signal received by the relay is amplified, frequency translated and retransmitted. Different amplification factors can be used as will be discussed in later chapters.

- **Linear-Process and Forward (LF).** This relaying method includes some other simple linear operations, which are performed on the signal in the analog domain after amplification. An example of such a linear operation is phase shifting, which facilitates the implementation of distributed beamforming.
- **Nonlinear-Process and Forward (nLF).** Not yet fully explored, this method performs some nonlinear operations on the received analog method prior to retransmission. An example application is the nonlinear amplification of the received signal which minimizes the end-to-end error rate [79, 80].

An important design issue related to transparent relaying protocols is the choice of amplification factor in the relay, where these options are generally available:

- **Constant Output Power.** In this case, the transparent relaying node transmits at a constant output power that has been set during node manufacturing. Whilst this is the simplest way of realizing a transparent relay, it is also suboptimum compared to the fixed and variable gain amplification.
- **Fixed Gain Amplification.** In this case, the node fixes the amplification factor over a given time window to a value that depends on long-term statistics in the channel or network. For instance, the amplification factor is typically an inverse function of the average channel gain between source and relay. In poor channel conditions, this may lead to very large amplification factors and hence high output powers; in this case, the retransmitted signal is delimited to the maximum transmission power, leading to clipping effects.
- **Variable Gain Amplification.** This case differs from the case of fixed gain amplification in that the amplification gain is adapted to instantaneous changes in the channel and network. For instance, the amplification factor is typically an inverse function of the instantaneous channel gain between source and relay. Again, clipping effects may occur due to large amplification factors. As will be shown in Chapter 3 of this book, a variable gain amplification is needed for a transparent architecture to realize its full diversity gain.

1.6.1.3 Regenerative Relaying Protocols

In the case of regenerative relaying protocols, information (bits) or waveform (samples) is modified. This requires digital baseband operations and thus more powerful hardware. Hence, regenerative relays usually outperform transparent ones. The most prominent examples of regenerative relaying are:

- **Estimate and Forward (EF).** The analog signal is amplified and down-converted to baseband, after which some detection algorithms aim at recovering the original representation of the signal. This estimate is then retransmitted. For instance, the EF relay estimates the modulated symbol and retransmits its estimate using the same or a different modulation order.
- **Compress and Forward (CF).** This protocol is similar to the above EF protocol in that it relays a compressed version of the detected information stream to the destination. This involves some form of source coding on the sampled signal samples and was shown to be capacity/performance optimum for the compressing node being close to the destination.
- **Decode and Forward (DF).** Being the prominent counter protocol to the transparent AF protocol, DF detects the signal, decodes it and re-encodes it prior to retransmission. A vast amount of different DF protocols exists today and we will dedicate large parts of Chapter 4 to this protocol. Over a wide gamut of application scenarios, DF is known to be performance optimum w.r.t. typical metrics such as error rate.
- **Purge and Forward (PF).** Modern communication systems are usually designed to be interference rather than noise limited. This design principle also applies to cooperative systems where PF allows for interference between the different relaying streams and deals with it by eliminating as much of it as possible at each relay node.

Table 1.2 Example protocols belonging to the two relaying families

Transparent relaying protocols	Regenerative relaying protocols
Amplify and forward (AF)	Decode and forward (DF)
Linear-process and forward (LF)	Estimate and forward (EF)
Nonlinear-process and forward (nLF)	Compress and forward (CF)
	Purge and forward (PF)
	Gather and forward (GF)

- **Gather and Forward (GF).** Also sometimes referred to as aggregate and forward protocol, this protocol is an extension to CF in that a relay node not only performs source coding over the sampled information but also on the information itself, which is aggregated over a few communication slots.

The two relaying protocol families are summarized in Table 1.2. The list is far from exhaustive but includes the most prominent examples, the majority of which will be dealt with later in this book.

Regenerative relaying protocols also obey a large set of parameters that can be optimized, the more important of which are listed below:

- **Choice of Channel Code.** Channel coding has a profound impact on the performance of a communication system. It basically trades encoding/decoding complexity and power with coding gains in form of transmit power reduction. Using no channel code in a relay is the least complex solution but also the worst performing one. If a channel code is used, then the network designer can choose from a variety of block codes, trellis codes and concatenations thereof. Block codes can correct a fixed amount of error but not more (for example, three errors in a block of 255 bits) and trellis codes can correct with a given probability a density of errors (for example, an error every 10 bits).
- **Choice of Interleaver.** The interleaver rearranges the output bit stream w.r.t. the input bit stream. The role of the interleaver is to break long sequences of errors so that they can be corrected more easily. Since it breaks long error bursts into several short ones, the application of interleavers is useful in block fading environments where the channel remains constant over a few symbols. Interleavers trade performance gains against memory requirements, that is larger interleavers randomize errors better at the expense of needing more memory to store the bits. One therefore has the choice of using no interleaver (preferably in static channel conditions or embedded systems with low memory), block interleaver (writing the input into the columns of the interleaver matrix and reading it from the rows) or random interleaver (randomly rearranging the order of the bits).
- **Choice of Waveform and Modulation.** An important design factor in relays is the choice of modulation, which comprises three important issues. First, one has to decide between single carrier or multicarrier modulation schemes. The former includes the traditional approach to modulation and the latter for example, orthogonal frequency division modulation (OFDM). Multicarrier modulation schemes are usually susceptible to nonlinearities in the transmitter and hence require fairly expensive amplifiers. Second, there is the choice between coherent and differential modulation. The former encodes information in amplitude and phase, whereas the latter in phase only. This has an impact on the receiver design and performance. Coherent modulation outperforms differential modulation at the expense of requiring a reliable channel estimate at the receiver. Differential modulation, conversely, does not require any channel estimates and is hence suitable in environments where the channel varies rapidly. Third, the modulation order is yet another design parameter where higher modulation orders exhibit a higher spectral efficiency at the expense of being more susceptible to noise and interference.
- **Choice of Space–Time Processing.** If the relay has multiple antennas and/or relays to realize distributed space–time relaying, a vast gamut of available space–time processing techniques is available. As such, one could use distributed space–time block codes (STBCs), space–time trellis

codes (STTCs), layered space–time codes (LSTCs) where data is multiplexed over transmit antennas, and beamforming techniques. Beamforming is the only one really requiring a feedback channel; the other schemes, however, profit from feedback since waterfilling and precoding can be applied to boost performance. STBCs yield diversity gains but no coding gains; STTCs yield both diversity and multiplexing gains; and LSTC yield multiplexing gains. More general classes of codes exist that trade diversity and multiplexing gains. These codes require a rich scattered channel in order to use the full potential of the MIMO channel. Conversely, beamforming sends the information in a given spatial direction and hence benefits from sparsely scattered channels.
- **Power Control.** In addition, the regenerative relay may use adaptive amplification factors, mainly to facilitate power control and hence manage interference. Clearly, larger transmit powers facilitate larger communication ranges but also create interference at more nodes.
- **Choice of Receiver.** A large variety of receivers and detectors is available to date, which are optimized for a given choice of transmitter and system configuration. Available techniques include simple threshold detectors, zero forcing (ZF) and minimum mean square error (MMSE) receivers, sophisticated interference cancelation receivers, etc.

The regenerative relay clearly exhibits more design degrees than the transparent relay. This is an opportunity because the regenerative generally outperforms the transparent relay; however, it is also a challenge in that optimizing regenerative relaying systems is a cumbersome task.

1.6.2 Multiple Access Resolution

The wireless medium is inherently broadcast in nature, that is if a node transmits, everybody in its vicinity overhears the signal. To facilitate communication between nodes and minimize interference, suitable multiple access methods thus need to be put in place. The prime roles of these multiple access methods are to take care of the following:

- **Duplexing Method.** Duplexing traditionally refers to the way a transmitter and a receiver communicate, that is whether they are able to communicate simultaneously or not. Since a cooperative systems utilizes relays, the traditional definition of duplexing breaks down and novel concepts need to be applied.
- **Coordination of Access.** Such coordination is vital since it guarantees that potential transmitters can access the channel and reach the receiver without causing interference or being interfered with.
- **Resource Allocation.** Another important part of the multiple access is to decide how many and which resources are allocated to each transmitter.

We will now discuss these issues in the context of cooperative relaying systems in greater detail.

1.6.2.1 Duplexing Methods

Having originated in wireline communications, a duplex communication system is a system composed of two connected devices that can communicate with one another in both directions. These said devices can either communicate simultaneously or not. The former is also referred to as full-duplex and the latter as half-duplex. A prominent example of a full-duplex system is the old plain telephone system (POTS) and of a half-duplex system an apartment's intercom system.

The extension of this concept to the wireless domain has caused a lot of confusion and with it differing duplexing definitions emerged. It is important to note, however, that the concept of simultaneity, as required by a full-duplex link, is very subjective and dependent on the traffic. For instance, POTS is a truly full-duplex system from an engineering point of view but signal delays incurred by long communication distances do not give the feeling of full-duplexity from a user perception point of view. Therefore, as long as these traffic-dependent delays stay within given bounds, two connected users feel simultaneity, which allows full-duplex systems to be emulated by different techniques in practice:

- **Frequency Division Duplex (FDD).** Reception and transmission of the information stream between the connected devices happens at different frequency bands. In theory, this allows both devices to communicate simultaneously and hence the system operates in true full-duplex. In practical systems, however, the receiver often needs to decode and process the information received on one band before responding on the other band–since this delay is often negligible, practical FDD systems efficiently emulate a full-duplex system. Furthermore, in practical contexts, these frequency bands are often referred to as frequency carriers separated by a frequency offset. Generally, a system is designed in FDD if both ends have approximately the same traffic to transmit or the signal travel duration is large, leading to long guard times if TDD is used.
- **Time Division Duplex (TDD).** Here, reception and transmission of the information stream between the connected devices happens at the same frequency band. This allows one device to communicate at a time only and hence the system theoretically to operate in half-duplex. In practical digital systems, however, the delays between both half-duplex streams is often so small that a full-duplex system is efficiently emulated. For instance, the cordless telephone system DECT relies on TDD and yet successfully manages to convey the perception of a full-duplex system. Generally, if traffic flows are strongly imbalanced, TDD approaches are preferred since one device can be allocated more time to communicate.
- **Division Free Duplex (DFD).** To communicate at the same time in the same frequency band in both directions is widely applied in the wired but so far not in the wireless domain. The main problem is related to the hardware implementation, which needs to prevent the high-power transmitted signal overshadowing the low-power received signal [81]. A possible approach is to place the transmitting and receiving antennas spatially sufficiently far apart, which is clearly not a very attractive solution. Other methods, however, are possible and will be discussed in Chapter 5. DFD, if implemented, facilitates both true and emulated full-duplex systems and also exhibits a superior spectral efficiency when compared with FDD and TDD.

The introduction of cooperative relays complicates the duplexing procedure, mainly because the term originated from the notion of two and only two devices being connected. The concept of multiplexing can also not be applied since relaying does not necessarily involve the simultaneous transmission of multiple signals over one medium. It has, however, become customary to refer to a full-duplex relay if it can receive and retransmit in the same frequency band at the same time, and half-duplex if it can only do this in the same band or the same time [82]. This is conceptually different from the duplex concept applied to a link since it bears no notion of simultaneity. In analogy to the practical implementations of the duplex principles at link level, we use the following duplex principles at the relay:

- **Time Division Relaying (TDR).** The incoming and outgoing information streams at a cooperative relay are separated in time, that is they are allocated to different slots and hence realize a half-duplex relay. The incoming and outgoing streams are typically but not necessarily in the same frequency band. This leaves enough time for processing the information stream on a slot or packet basis. TDR is hence popular with regenerative relaying protocols. Note that the incoming and relayed time slots do not necessarily need to be of equal length; for instance, if the channel from the source to the relay is worse than from the relay to the sink then a DF protocol could use a higher modulation index in the second hop and thus relay a packet of shorter duration. Furthermore, there will generally be a small time gap between received and relayed frame that accounts for the packet processing delay and the time needed for radio circuitry to switch from reception to transmission.
- **Frequency Division Relaying (FDR).** Here, reception and transmission of the relay stream happen at different frequency bands, that is realizing a half-duplex relay. The incoming and outgoing streams

are typically but not necessarily scheduled at the same time. If they are scheduled at the same time, then this reduces to the well-known out-of-band or out-of-channel relaying [83]. FDR is applicable to both regenerative and transparent relaying families. Again, incoming and relayed frequency bands and packet duration do not necessarily have to be of the same size.

- **Division Free Relaying (DFR).** Here, reception and transmission of the relay stream happen at the same time using the same band, which realizes a full-duplex relay and is referred to as in-band relaying. This technique is currently being used in the context of simple repeaters in cellular systems to provide sufficient coverage, where it is referred to as on-frequency repeater (OFR) [84–86]; it is also being used in the context of more sophisticated repeaters capable of canceling interference, where it is referred to as interference canceling system (ICS) repeaters [87]; finally, broadcasting systems have also used such an approach, which is referred to as on-channel repeater (OCR) [88]. Since available technologies spatially separate receive and transmit antennas, this technique ought to be referred to as spatial division relaying. However, since such an approach is clearly inappropriate for envisaged relays of small size, we hope that viable single antenna solutions will be found soon and will hence stick to the notion of DFR. In any case, today's DFR systems usually use a simple AF protocol, which only puts tough requirements on the hardware as will be discussed in more detail in Chapter 5. Future systems may also choose to decode and re-encode the slightly delayed retransmitted signal stream, which requires special coding and decoding techniques to be used as will be discussed in Chapter 4.

DFR is thus spectrally efficient but very demanding from a technology point of view. On the other hand, TDR and FDR require the introduction of an extra time slot and extra frequency band, respectively, which clearly diminishes the spectral efficiency of a cooperative relaying system. The spectral efficiency decreases even further with an increasing number of relays. It will however be demonstrated later in this book that the loss in spectral efficiency is more than compensated for by the gains due to cooperation.

With the aid of FDD/TDD/DFD and FDR/TDR/DFR we are able to support truly full and half duplex as well as emulated full duplex systems, which is exemplified in Figure 1.17 and to be discussed in subsequent sections.

1.6.2.2 Multiple Access Protocols

The existence of more than one user or at least one relay in the system requires suitable multiple access protocols, because otherwise interference between their transmission occurs. As per Figure 1.16, this leads to the following set of canonical channel configurations:

- **Point-to-Point (unicast).** Not actually requiring multiple access methods, the point-to-point channel is formed by means of a direct channel between source and destination.

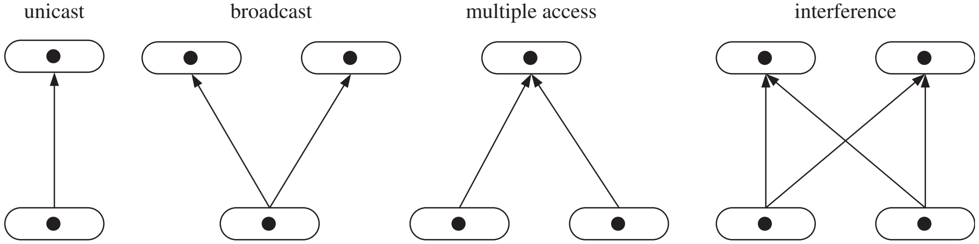

Figure 1.16 Canonical access channels: unicast, broadcast, multiple access, interference channel

- **Point-to-Multipoint (broadcast).** In this case, a single source communicates with multiple destinations in a broadcast fashion. An example is a base station communicating with several mobile stations, either transmitting the same information to all of them or different information to each of them.
- **Multipoint-to-Point (multiple access).** Here, multiple sources communicate with a single destination. A typical example is a set of users communicating with a base station.
- **Multipoint-to-Multipoint (interference channel).** This is the most general case, where multiple sources communicate with multiple destinations. This case is also sometimes referred to as interference channel because the two information streams stemming from the two sources interfere at both destinations.

The actual access for the broadcast, multiple access and interference channels can be facilitated by traditional multiple access protocols. These are generally divided into reservation-based and contention-based protocols, where the former is useful in the context of centralized systems where resources can be reserved *a priori* and where traffic is regular; the latter is applied in decentralized systems or where traffic is bursty and resources need to be contended for prior to communication. Examples of reservation-based protocols are:

- **Time Division Multiple Access (TDMA).** Different user and relay information streams are scheduled in a time slotted fashion. For instance, the direct link from source to destination is assigned the first time slot and the relay link the second time slot.
- **Frequency Division Multiple Access (FDMA).** Different user and relay information streams are scheduled at different frequency bands. For instance, the direct link from source to destination is assigned one band whereas the relay link uses another one. In practical systems, these frequency bands are often referred as frequency channels.
- **Code Division Multiple Access (CDMA).** Different user and relay information streams are assigned different (ideally orthogonal) spreading sequences. For instance, the direct link from source to destination is assigned one spreading code whereas the relay link uses another one. Whilst this method allows communication of all information streams in the same bands at the same time, power control is often needed, which is not easy to implement in a distributed cooperative network.
- **Orthogonal Frequency Division Multiple Access (OFDMA).** Different user and relay information streams are assigned different subcarriers. For instance, the relayed and own information of a cooperative node are assigned to different subcarriers. Also, the cooperative relay may decide to reshuffle the incoming subcarriers into a differently arranged set of relayed subcarriers so as to maximize the end-to-end performance.
- **Multicarrier CDMA (MC-CDMA).** This access method combines both OFDM and CDMA and could also be used in conjunction with OFDMA. The key idea is to apply a user or relay unique spreading sequence in frequency across several subcarriers or, in time, over several data symbols at a given subcarrier. Such an access method is very flexible but difficult to optimize.

In the case of contention-based protocols, it is generally assumed that users use the same frequency (and code) and contend for resources in time. This means that contention-based protocols typically use TDD and TDR. Prominent examples belonging to this access family are:

- **Aloha.** Probably the first contention-based protocol, the prime idea is that the sender transmits data when there is data in the buffer and if the message does not reach the receiver (due to unfavorable wireless channel or collision), just resend it later. Whilst very simple, the efficiency of Aloha is fairly low.
- **Carrier Sense Multiple Access (CSMA).** An improvement to Aloha is CSMA, where the transmitter first senses the channel before transmitting. If the channel is detected as occupied, it will refrain from transmission and – depending on the backoff algorithm – try again at a later moment. If the channel is detected to be free, transmission is initiated. Efficiencies of the various variants of CSMA are well above that of Aloha, but generally inferior to reservation-based protocols.

1.6.2.3 Resource Allocation Strategies

Once the duplexing methods and multiple access protocols have been determined for the system, each source and relay node can be allocated different resources in terms of time, frequency, number of codes, power, etc. The most typical approaches are discussed below:

- **Time Slots.** If one device has more traffic to transmit then it can be allocated more time slots. An example is when a device needs to transmit its own traffic in addition to relayed traffic.
- **Frequency Bands.** Devices can also be allocated more frequency bands or a larger frequency band if needed. In today's practical systems, this means that an optimum choice of frequency carriers and channels needs to be found.
- **Transmission Power.** Another way of increasing a device's ability to reach the destination or a relay is to allow it to transmit at higher transmission powers. Whilst detrimental to the overall network, since this generates excess interference, this method allows prioritizing of devices and information flows.
- **Number of Codes.** Yet another method of allocating more resources to a device in need is to assign it multiple codes in the case of CDMA. This is currently being used in the multicode transmission of HSDPA in 3GPP.
- **Choice of Subcarrier.** If some form of subcarrier scheduling has originally been performed at the source, that is allocating higher modulation indexes to subcarriers of better quality, then it is beneficial that the relay also performs such an allocation according to the subcarrier quality of the relay destination channel.
- **Probability of Persistency.** In the context of contention-based protocols, an important design parameter influencing the effective resources allocated to a device is the probability of persistency used in the backoff algorithm. This probability depends on many factors, such as the amount of contending traffic and nodes.

The discussed factors are summarized and visualized in Figure 1.17.

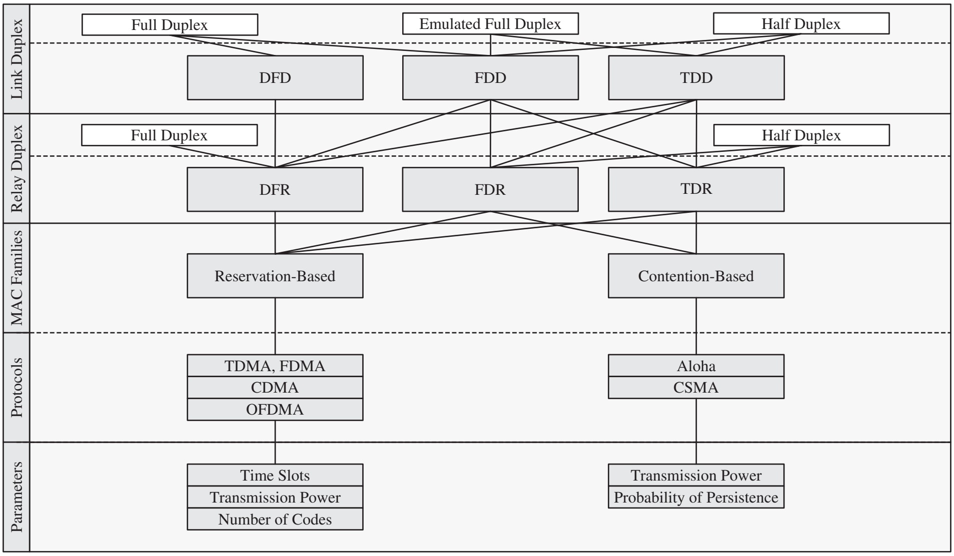

Figure 1.17 Taxonomies and definitions related to multiple access resolution

1.6.2.4 Typical Access Configurations

The degrees of freedom for applying various duplexing methods, access protocols and resource allocation algorithms to cooperative relaying networks is clearly infinite. To illustrate the above exposure, we shall demonstrate a few typical access configurations in the context of two system designs, that is a system where relays are designed to be part of the system *a priori* and a system where the relays are inserted into the system later:

- **Systems With Prior Relay Insertion.** Performing the system design with relays in mind has an impact on access mechanisms. First, to limit interference, the system ought to be power controlled in such a manner that direct links between source and destination do not occur. Whilst suboptimal at link level, this yields significant gains at system level as shown, for example, for contention-based MACs in [89] and from a capacity point of view in [90]. Second, the relay can be seen as a full system entity to which duplexing rules apply, that is the notion of full or half-duplex does not apply to source and destination but rather to any pair of communicating nodes, be it source and relay or relay and destination. This yields a typical scenario as shown in Figure 1.18 with two relaying hops, two active up and downlinks and no direct communication between BS and MSs.

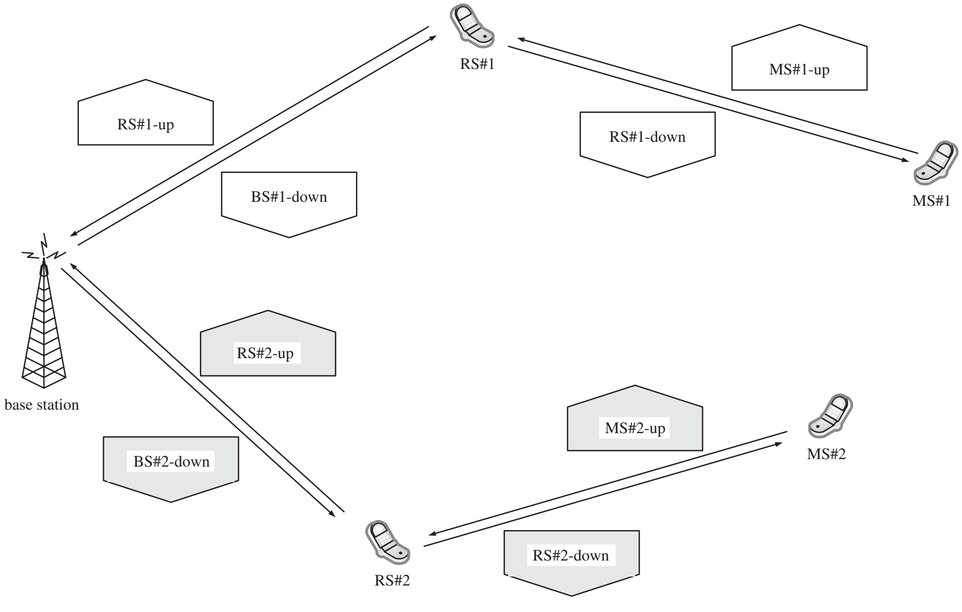

Figure 1.18 Example access scenario with one BS, two RSs and two MSs. The packet transmitted from the BS to the RS#1 has been marked 'BS#1-down'; the one from the RS#1 towards the MS#1 has been marked 'RS#1-down', etc.

The cases of TDD/TDR and particular realizations of the TDMA, FDMA and CDMA multiple access protocols are shown in Figure 1.19. Clearly, many more combinations are possible; for instance, FDMA could be implemented with more frequency channels; the uplink of user #1 and downlink of user #2 could be scheduled simultaneously in the FDMA and CDMA cases to minimize interference, etc. The cases of TDD/FDR, FDD/TDR and FDD/FDR with TDMA, FDMA and CDMA as multiple access protocols are shown in Figures 1.20–1.22, respectively. Clearly, not all combinations are optimal nor is this list exhaustive (even for this fairly simple case). The combinations with DFD and DFR have been omitted here as they are easily derived from above frame structures.

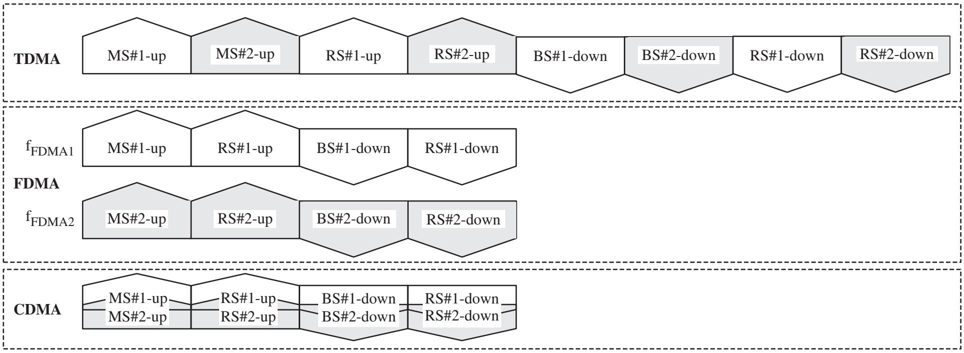

Figure 1.19 An example TDD/TDR configuration with TDMA, FDMA and CDMA as access protocols. TDMA requires only one frequency channel but takes twice the duration compared with other solutions; FDMA implementation requires two frequency channels; CDMA requires half power per user to be used to facilitate a fair comparison

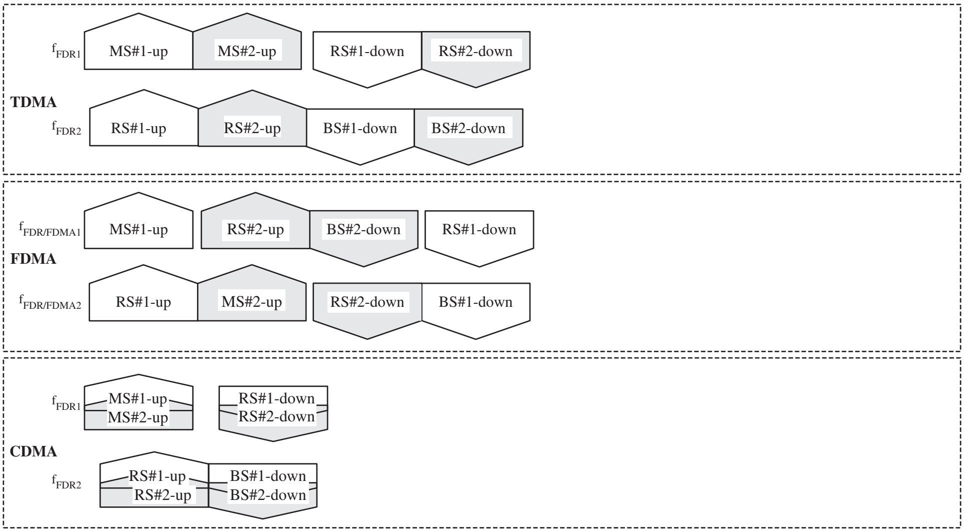

Figure 1.20 An example TDD/FDR configuration with TDMA, FDMA and CDMA as access protocols. For delay-constrained applications, CDMA is clearly the best choice

- **Systems With Subsequent Relay Insertion.** Inserting relays into an already existing and designed system with a fixed carrier and channel structure restricts the use and applicability of certain access mechanisms. This is mainly because a direct link between source and destination is needed and hence a relay has to obey the duplexing and access rules of this existing link. Assuming the same scenario as that in Figure 1.18 but with additional direct links between MSs and BS and the notion of duplexing applied to the MSs and BS, only the access configurations shown in Figures 1.19 and 1.21 are possible. Again, this is far from exhaustive since more combinations are feasible.

It shall be noted here that it is generally easier to build and maintain a mesh network with its respective schedule between cooperating nodes when only the temporal domain is used, that is TDD combined with TDR and TDMA.

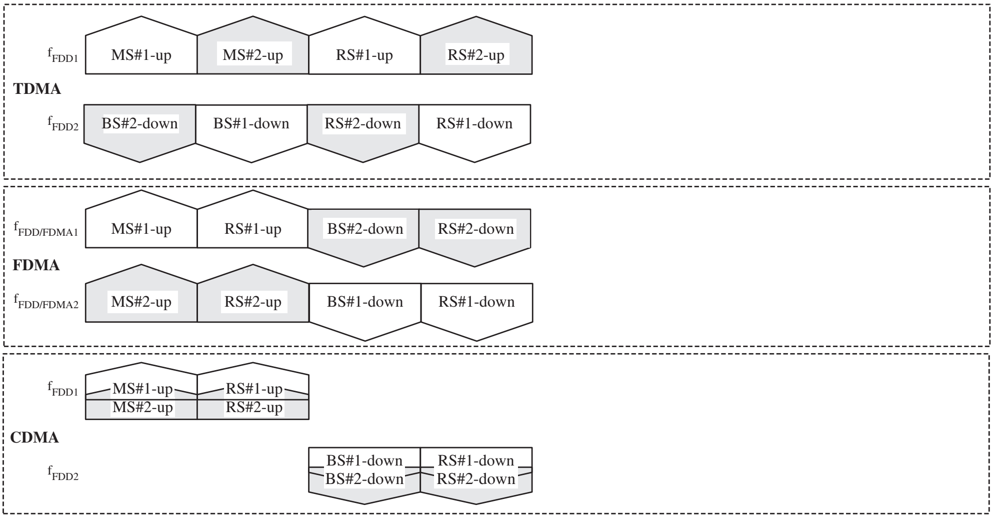

Figure 1.21 An example FDD/TDR configuration with TDMA, FDMA and CDMA as access protocols

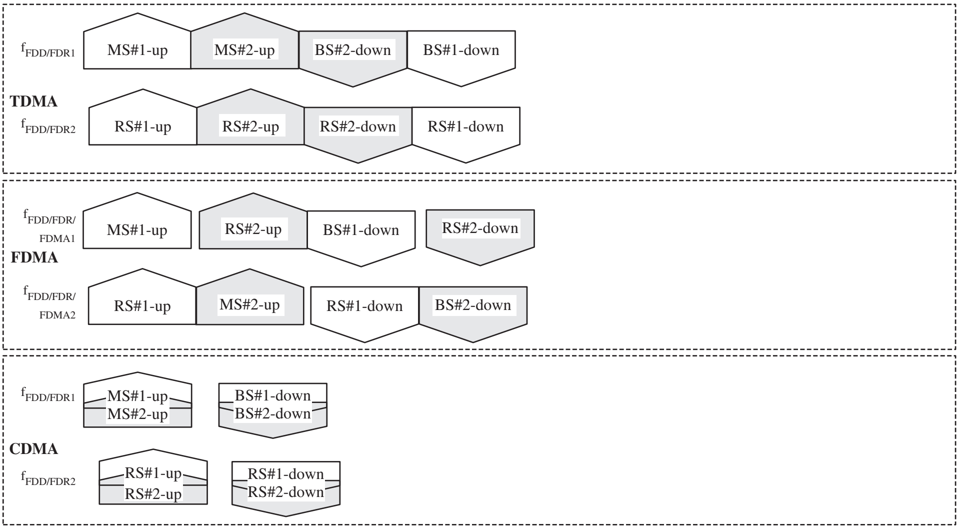

Figure 1.22 An example FDD/FDR configuration with TDMA, FDMA and CDMA as access protocols. Again, CDMA is shortening the delivery delay

1.6.3 Cooperative Networking Aspects

With the node behavior determined and the multiple access rules established, it is important to discuss how information is handled at system and network level. To this end, we will now discuss canonical information flows resulting from cooperative relaying systems. We will also discuss important design parameters that have a serious impact on the performance of such networks. Again, for the sake of completeness, we have summarized the taxonomy related to networking aspects in Figure 1.24. Note that this discussion only pertains to elements most relevant to the PHY layer and does not include a discussion on routing protocols, etc.

1.6.3.1 Canonical Information Flows

From above-discussed node behaviors, relaying protocols, duplexing and access methods, we can construct different information flows and architectures, some of which have already been discussed at the beginning of this chapter. Subsequent discussions relate to the case of a noncooperative single source, single destination systems only; the extension to the case of cooperation, multiple sources and destination follows the same recipe.

- **Direct Link.** Information from a source can of course reach the destination by means of a single direct link.
- **Serial Relaying.** As per Figure 1.23 (top left), serial relaying connects the source and the destination by means of a chain of relays that are assumed to use orthogonal channels to relay the information. Note that in this and subsequent cases a direct link may or may not be available.
- **Parallel Relaying.** As per Figure 1.23 (top right), parallel relaying connects the source and the destination by means of a parallel set of relays that are assumed to use orthogonal channels to relay the information at the same time.
- **Space–Time Relaying.** As per Figure 1.23 (bottom left), space–time relaying connects the source and the destination by means of a parallel set of relays that are assumed to use space–time encoded channels to relay the information.
- **Composites Thereof.** As per Figure 1.23 (bottom right), an information flow can be realized by building hybrids from the above-discussed relaying methods.

Such networks as just described can be infrastructure based (that is an infrastructure is available prior to deployment) or infrastructureless (that is it emerges after deployment or remains unavailable). If an infrastructure is available, it can be managed in a centralized or noncentralized fashion. Note that

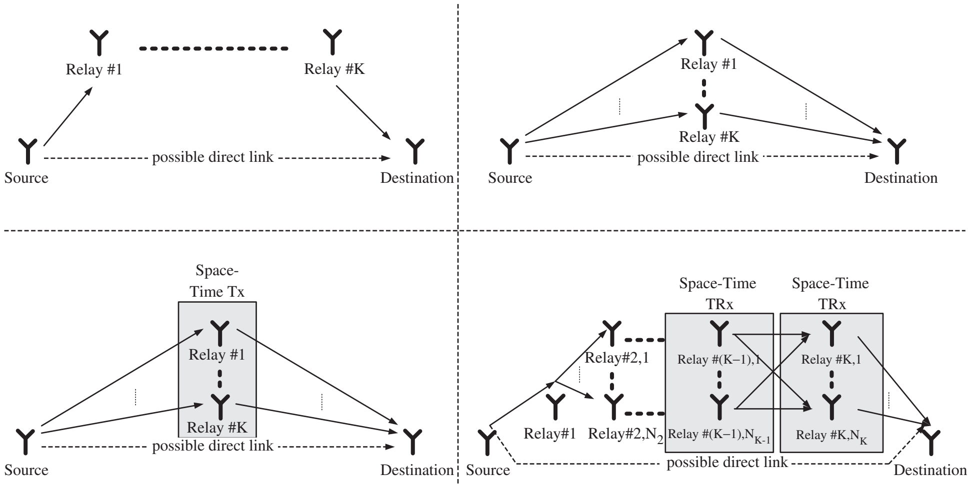

Figure 1.23 Canonical network information flows by means of serial relaying (top left), parallel relaying (top right), space–time relaying (bottom left) and hybrids thereof (bottom right)

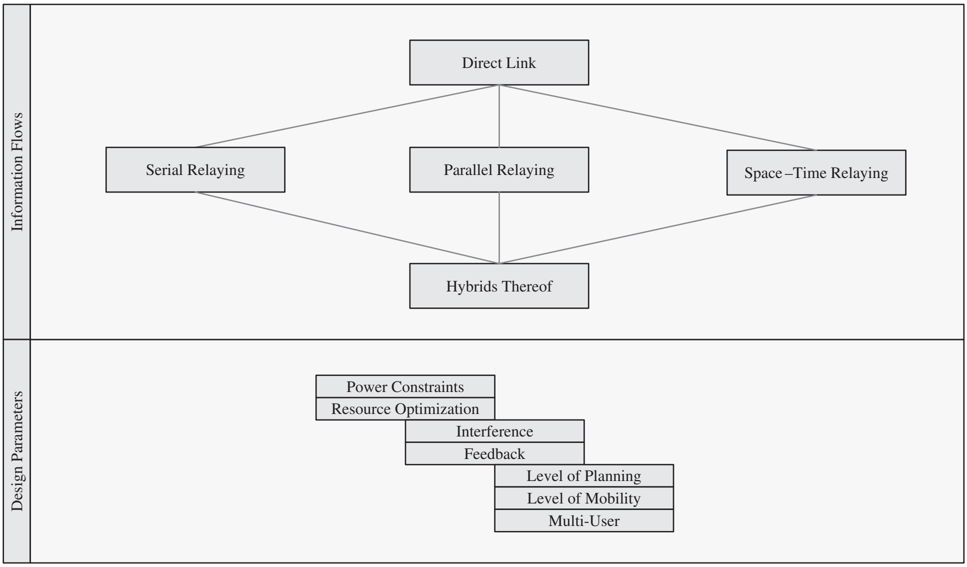

Figure 1.24 Taxonomies and definitions related to networking issues

you may have a decentralized infrastructure-based system (for example, a cooperative system using decentralized radio resource management algorithms) as well as a centralized infrastructureless system (for example, a WSN applying clustering).

1.6.3.2 Important Design Parameters

The following nonexhaustive list of issues has an impact on the optimization process as well as the performance of cooperative relaying systems:

- **Power Constraints.** It is important to know whether a power constraint applies per node or per information flow. For instance, if the power constraint applies to the flow, then introducing more relays requires the given power to be shared between these relays. Most practical applications obey a power constrained per node, that is adding a new relay node injects extra power into the system.
- **Resource Optimization.** It is also of importance to determine *a priori* whether resources for nodes in the cooperative relaying system are optimized depending on of various network parameters. For instance, one could decide to allocate more power to the relay node compared with the source if the distance between relay and destination is larger than between source and relay. Most practical applications, however, still refrain from performing such an optimization as the involved overhead is still too large.
- **Interference between Flows.** The system can be designed such that there is no interference between the information flows generated by sources and relays. From a Shannon point of view, however, designing a system to be interference limited is a better choice.
- **Feedback.** The performance will be strongly impacted by the availability of feedback channels, that is the ability of a receiver to report back to its transmitter. Useful information to be fed back is full channel state information (for example, instantaneous channel realizations) or partial channel state information (for example, averaged channel realizations), interference temperature, etc. It is generally well accepted that a feedback channel can drastically improve the performance of the system because it facilitates spatial waterfilling (that is optimal allocation of power per transmit

antenna) and precoding (that is optimal transmit code design). Practical communication systems generally have a feedback channel, even if it is only a simple acknowledgement (ACK) packet.

- **Level of Planning.** A planned roll-out of cooperative relays can significantly boost performance. However, this comes at the price of an increased planning cost and time, as well as the necessity to acquire the planned relay sites. An unplanned roll-out is clearly performance sub-optimal but saves a lot of time and money since relays can simply be placed on, say, lampposts.
- **Level of Relay Mobility.** Another important parameter is the level of mobility of the relays. For instance, if relays are fixed, they typically belong to an extended infrastructure. Alternatively, if users are used as relays, their mobility will strongly impact the performance of the cooperative system.
- **Level of Synchronization.** Synchronization generally stretches from the hardware to the network level. At hardware level, nodes may or may not be synchronized down to the carrier frequency; synchronization here is vital for distributed space–time codes to work properly. At access and network level, nodes can be synchronized by means of slots, which is known to boost system capacity but also requires suitable synchronization algorithms.
- **Multiuser Scenarios.** A recently emerging concept is the application of relaying nodes to multiuser scenarios. This has a profound impact on the design process as multiple transmitting users are not considered as competitors but as facilitators.

As they are not the focus of this book, many other design factors at system and network level have been omitted here. Nonetheless, above list serves to highlight that link and system level issues are closely interrelated in cooperative systems.

1.6.4 System Analysis and Synthesis

We also briefly summarize important quantities which are derived when analyzing cooperative systems. These quantities are often used to optimize the system performance, that is to derive optimum system configurations, protocols, etc., which we also summarize below.

1.6.4.1 Performance Analysis

The following issues and quantities play a central role when analyzing cooperative relaying systems:

- **Average Error Rates.** The ‘instantaneous’ bit (BER), symbol (SER) or packet error rate (PER) is calculated assuming a given channel realization and average noise power. When the channel varies sufficiently fast or, otherwise put, all channel realizations are traversed over the duration of the transmission, the average BER, SER and PER can be calculated and used for system characterization. Whilst SERs are usually easily calculated in closed form using the moment generating function (MGF) approach [91], exact BER and PER expressions are generally not easy to derive. Approximations and asymptotic expressions are hence typically involved, in particular for the PER, which is the most important quantity in real-world systems.
- **Outage Probabilities.** If the channel does not vary sufficiently fast or, in other words, not all statistical states are traversed over the duration of the transmission, then invoking the average does not make sense since it would yield a different value for each transmission. The concept of outage is thus typically involved, which quantifies the probability that a certain performance cannot be met. Typically, outage probabilities can be calculated for channel realizations, and information rates as well as bit, symbol and packet errors. Since fading is generally assumed to vary sufficiently quickly whereas shadowing not, the averages are derived for the former and its outages for the latter. For example, in a cellular system obeying fast fading and shadowing, one would first calculate the average PER over the fading statistics and then calculate the packet error outage (PEO) for the

calculated average PER over the shadow statistics [92, 93]. Cell dimensioning is then typically done with the latter, which is why the quantification of outages are of prime importance in real-world systems.

- **Throughput.** Once the average rates and/or outages are calculated, the average end-to-end and system throughput can be calculated. This quantity gives an insight to the system designer of how much capacity can be offered to the user when using a particular communication scheme, be it simple direct communication or cooperative relaying.
- **Delay/Latency.** Another quantity of importance is delay or latency, which quantifies the time needed to deliver a packet to the destination from the moment it has been generated. Delay typically comprises (i) the time the packet spends in the buffer queue of the device; (ii) the time it is required to get it successfully delivered via the wireless channel to the destination; and (iii) the time needed to process it. The queuing time typically depends on the amount of traffic entering and the amount of traffic leaving the queue. The successful delivery time includes the contention or transmission time, as well as retransmissions in the case of delivery failure. If relays are present in the system, then this delay needs to be multiplied by the number of relays to obtain an average end-to-end latency.
- **Real-World Impairments.** Of augmenting importance is the quantification of the impact of real-world impairments on the performance of cooperative relaying systems. Prime concerns are the impact of channel estimation errors, synchronization errors, phase errors, erroneous feedback information, etc.

The discussed factors are summarized and visualized in Figure 1.25.

1.6.4.2 System Synthesis

The above-described performance analysis clearly gives useful insights into cooperative relaying systems; however, changing a few underlying assumptions often requires analysis and simulations to be redone. It is hence desirable to take the inverse approach, where design guidelines are synthesized

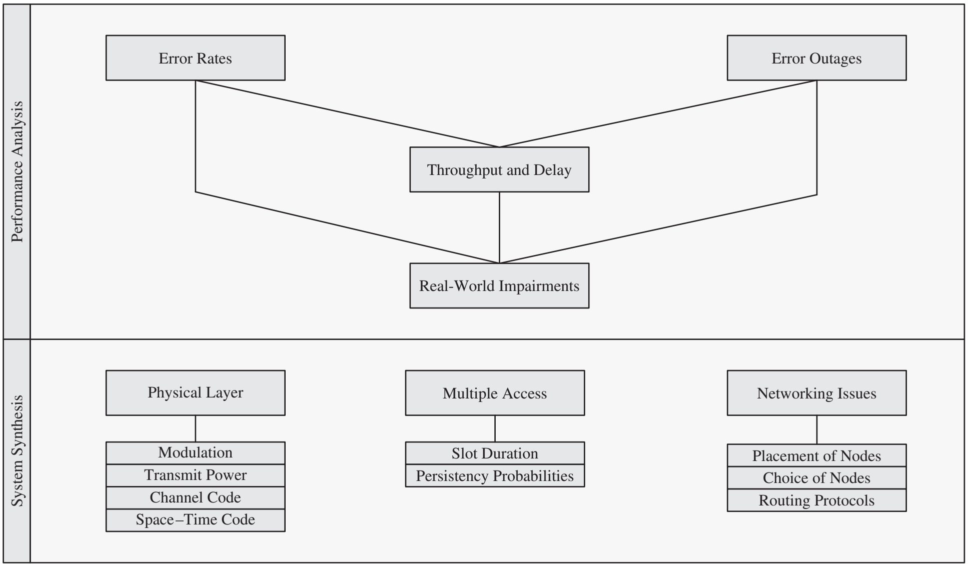

Figure 1.25 Taxonomies and definitions related to analysis and synthesis of cooperative relaying networks

for a broad range of underlying assumptions. For instance, instead of offering the spectral efficiency versus the SNR for different modulations, a synthetic approach would yield the modulation that is optimum for a given SNR and number of users. System synthesis is a very challenging area that is mainly complicated by the often very complex expressions resulting from the performance analysis. The following nonexhaustive list of issues are typically considered when performing a system synthesis:

- **Optimum Set of PHY Parameters.** Important PHY parameters, which a system synthesis should yield in explicit form, are the optimum choice of modulation order, the optimum choice of transmission power, the optimum choice of channel and space–time code and its generator matrix, etc.
- **Optimum Multiple Access Rules.** Of importance at access level are the optimum distribution of resources (for example, duration or number of time slots), the persistency probability for a given traffic load and user density, the packet duration to minimize collisions in the case of contention-based protocols, etc.
- **Optimum Networking Protocols.** Closely related to the physical and access level are the optimum choice of relaying nodes as well as their placement. Of further importance is the optimal choice of routing protocols and associated parameters.

As is clearly evident from this entire section, the number of degrees of freedom in constructing a cooperative system is virtually unbounded, somehow explaining the large amount of literature available on this subject.

1.7 Background and Milestones

This introductory chapter is concluded by summarizing key contributions to cooperative communications. Note that this does not constitute a complete state-of-the-art review but rather the exposure of early milestones that helped in shaping today's research landscape in cooperative systems. Important state-of-the-art contributions will be mentioned in the respective technical chapters.

1.7.1 First Key Milestones

Early developments concerning supportive, cooperative and space–time relaying were related but have largely emerged independently:

- **Supportive Relaying.** This simplest form of cooperation is not exactly new. Information theoretical developments stem back to the seminal contribution by van der Meulen in 1968 [94, 95] and by Cover and Gamal in 1979 [96]. Whilst some information theoretical contributions emerged here and there, the communication and protocol developments received a revival in the early 1990s with the 3GPP Concept Group Epsilon, driven by Vodafone [10]. Back then, communication engineers argued that no user would agree to relay data for another user since the short-term gains of the relaying user are nil. Harrold and Nix [97, 98] were the first to prove by means of simulations that – whilst short-term gains were indeed sometimes unfavorable – every user gained in the long run by cooperating; they also showed that by using simple relaying, coverage holes could largely be closed in a cellular deployment. Similar insights were provided a little later in [9, 99].
- **Cooperative Relaying.** Cooperative relaying, that is the case where at least two users help each other to boost each other's performance, has been pioneered by Sendonaris *et al.* in 1998 [73]. Later, around 2000, Laneman and coworkers rigorously formalized various types of supportive and cooperative relaying protocols and proved that significant performance and outage gains can be achieved [100–102]. It is largely due to Laneman's seminal work that the area of cooperative

communication systems commenced to flourish. A little later, Hunter and coworkers [75, 103, 104] and Stefanov and Erkip [105] were the first to propose a viable cooperative scheme based on channel coding and special code designs.

- **Space–Time Relaying.** Space–time relaying had been pioneered by Dohler and coworkers in 1999 and made public to the audience of the Mobile Virtual Centre of Excellence (M-VCE), a UK national research initiative, from 2000 onwards [106, 107]. Their work was based on then just emerged works on space–time codes by Foschini [119], Alamouti [120] and Tarokh [121, 122]. Subsequently and sometimes in parallel, pioneering key contributions related to distributed space–time codes and their design emerged from Laneman and Wornell [108] and Stefanov and Erkip [109, 110].

These early key contributions are summarized in a time chart in Figure 1.26. We will now describe these and other early contributions in some greater detail.

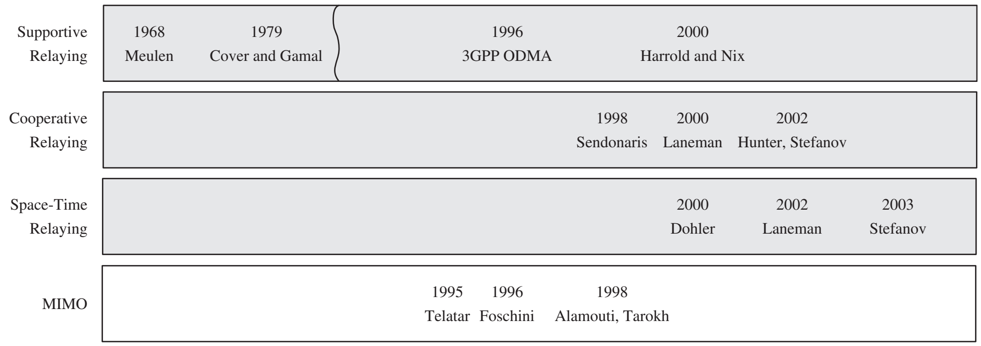

Figure 1.26 Early key contributions in the field and their timeline of publication

1.7.2 Supportive Relaying

The method of relaying has been introduced in 1968 by van der Meulen [94, 95] and has also been studied by Sato [111]. A first rigorous information theoretical analysis of the relay channel, however, has been published by Cover and Gamal [96], a more detailed description of which can be found in his excellent book [112].

In these contributions, a source MS communicates with a destination MS directly *and* via a RS. In [96] the maximum achievable communication rate was derived for various communication scenarios, including the cases with and without feedback to either source MS or RS, or both. The capacity of such a relaying configuration was shown to exceed the capacity of a simple direct link.

It should be noted that the analysis was performed for Gaussian communication channels only; therefore, neither the wireless fading channel has been considered, nor were the power gains due to shorter relaying communication distances explicitly incorporated into the analysis.

Only in the middle of the 1990s, research in and around the Concept Group Epsilon revived the idea of utilizing relaying to boost the capacity of wireless networks, thereby leading to the concept of ODMA [10]. The power gains due to the shorter relaying links were the main incentive in investigating such systems to reach MSs out of BS coverage. The emphasis of the study was its applicability to cellular systems, as well as a suitable protocol design; no theoretical investigations into capacity bounds, etc., have been performed.

1.7.3 Cooperative Relaying

Seminal work in the area of cooperative relaying has been the contribution by Sendonaris, Erkip and Aazhang, dating back to 1998 [73]. In their study, a very simple but effective user cooperation protocol was suggested in order to boost the uplink capacity and reduce the uplink outage probability for a given rate. The designed protocol stimulates a MS to broadcast its data frame to the BS and to a spatially adjacent MS, which then retransmits the frame to the BS. Such a protocol certainly yields a higher degree of diversity because the channels from both MSs to the BS can be considered as uncorrelated.

The simple cooperative protocol has been extended by the same authors to more sophisticated schemes, which can be found in the excellent contributions [72] and [74]. Note that in its original formulation [73], no distributed space–time coding has been considered.

The contributions by Laneman and Wornell in 2000 [100] are a conceptual and mathematical extension to [73], where energy-efficient multiple access protocols are suggested based on decode-and-forward and amplify-and-forward relaying technologies. It has been shown that significant diversity and outage gains are achieved by deploying the relaying protocols when compared to the direct link. Note again, that no distributed space–time coding was considered at that time.

Gupta and Kumar were the first to analyze statistically the information throughput theoretically offered for large scale relaying networks [113]. They showed that under somewhat ideal situations of no interference, hop-by-hop transmission and predefined terminal locations, capacity per MS decreases by $1/\sqrt{M}$ with an increasing number of MSs M in a fixed geographic area. They also showed that if the terminal and traffic distributions are random, then the capacity per terminal decreases even in the order of $1/\sqrt{M \log M}$. The analysis in [113] has been extended by the same authors to more general communication topologies, where the interested reader is referred to the landmark paper [114].

Furthermore, Grossglauser and Tse have shown that mobility counteracts the decrease in throughput for an increasing number of users in a fixed area [115]. The protocols suggested therein benefit from the decreased power for a hop-per-hop transmission for decreasing transmission distances. It also benefits from the location variability due to mobility, that is a packet is picked up from the source MS by any passing by RS and only re-transmitted (and hence delivered) when passing by the target MS.

1.7.4 Space–Time Relaying

To understand the contributions and timings of space–time relaying schemes, seminal works on traditional MIMO systems will be revisited first.

Contributions on MIMO systems have flourished ever since the publication of the landmark papers by Telatar [116, 117] and Foschini and Gans [118] on capacity, and by Foschini [119], Alamouti [120] and Tarokh [121, 122] on the construction of suitable space–time transceivers. In the BLAST system, introduced by Foschini in 1996 [119], a transmitter spatially multiplexes signal streams onto different transmit antennas, which are then iteratively extracted at the receiving side using the fact that the fades from any transmit to any receive antenna are uncorrelated and of different strength. The BLAST concept has ever since been extended to more sophisticated systems, a good summary of which can be found in [123]. Alamouti introduced a very appealing transmit diversity scheme by orthogonally encoding two complex signal streams from two transmit antennas, thereby achieving a rate one space–time block code [120]. His work was then mathematically enhanced by the landmark paper of Tarokh [121], who essentially exposed various important properties of space–time block codes. He also showed how to construct suitable space–time trellis codes, which were shown to yield diversity *and* coding gain [122].

A system utilizing the advantages of both MIMO and relaying was suggested by Dohler in 1999 and hence became one of the main research topics within the M-VCE from 2000 onwards. It has been suggested that spatially adjacent mobile terminals be used to form distributed antenna arrays in hot-spot areas. The thus formed distributed antenna array had been termed an artificial antenna array but – since the abbreviation AAA was already in use – was renamed as virtual antenna array or VAA. Numerous studies [106] have led to a set of patents [107], which are backed by about 20 industrial members, such as Vodafone, Nokia, Philips, Nortel Networks, Samsung, etc. The studies encompassed the following (in chrondogical order):

- downlink distributed receive diversity in cellular systems;
- downlink distributed MIMO in cellular systems;
- uplink distributed MIMO in cellular systems;
- introduction of distributed relaying to cellular systems;
- extension of the above to WLAN and hot-spot systems;
- generalization to arbitrary distributed relaying topologies.

The case of distributed space–time coding has also been analyzed by Laneman in [108] and in his PhD dissertation [102]. In his thesis, information theoretical results for distributed single-input single-output (SISO) channels with possible feedback were utilized to design simple communication protocols, taking into account systems with and without temporal diversity, as well as various forms of cooperation. He has demonstrated that cooperation yields full spatial diversity, which allows drastic transmit power savings at the same level of outage probability for a given communication rate.

Specific distributed space–time coding schemes have also been suggested, for example, by Stefanov and Erkip [109, 110]. In this publication, two spatially adjacent MSs cooperate to achieve a lower frame error rate to one or more destination(s), where a quasi-static fading channel has been assumed. Distributed space–time trellis codes have been designed that maximize the performance for the direct link from either of the MSs to the destination *and* the relaying link. Furthermore, distributed beamforming has also been introduced [124].

Finally, it should be mentioned that recently, Ozgur *et al*. [90, 125] have shown for the first time that linear transport capacity scaling is possible in a large network by means of cooperative hierarchies and space–time relaying.

1.8 Concluding Remarks

This chapter has served to introduce the topic of cooperative relaying. This area has recently received a lot of attention, which has often been coupled with a lot of hype and confusion. We therefore paid a lot of attention to introduce key notions properly and apply viable taxonomies. This, hopefully, has shed some light on an otherwise complicated topic and paved a solid basis for subsequent chapters and hardware, wireless channel and PHY algorithms.

Despite its recent revival, relaying is – strictly speaking – old hat in the information theory community where first developments emerged about four decades ago [94]. However, since the information theory community is historically not concerned with the wireless channel and usually assumes AWGN channels only, exclusively regenerative relaying families over Gaussian channels have been looked at. Other relaying protocols became popular only recently within the information theory community, mainly due to the cooperative relaying protocols' success in the communications community.

From an implementation and engineering point of view, cooperative relaying protocols are also not exactly new. In fact, the now reborn AF and DF relaying protocols have been known in parts by the satellite community for nearly five decades [126–130] and by the radio community for almost a century already [131, 132]. Here, the main incentive to use relays was to overcome coverage and range problems rather than capacity problems.

What has been truly new and responsible for the renaissance of relaying protocols in recent years is the capacity and performance benefits of cooperative relaying protocols under realistic channel and system conditions. These gains are in fact the justification and inspiration for the subsequent chapters of this book.