

A tutorial survey of topics in wireless networking: Part I

ANURAG KUMAR¹ and D MANJUNATH²

¹Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore 560 012

²Department of Electrical Engineering, Indian Institute of Technology, Bombay, Mumbai 400 076
e-mail: anurag@ece.iisc.ernet.in; dmanju@ee.iitb.ac.in

MS received 24 October 2006; revised 13 April 2007; accepted 12 June 2007

Abstract. In this two part paper, we provide a survey of recent and emerging topics in wireless networking. We view the area of wireless networking as dealing with problems of resource allocation so that the various connections that utilise the network achieve their desired performance objectives. In the first part of the paper, we first survey the area by providing a taxonomy of wireless networks as they have been deployed. Then, we provide a quick tutorial on the main issues in the wireless ‘physical’ layer, which is concerned with transporting bits over the radio frequency spectrum. Then, we proceed to discuss some resource allocation formulations in CDMA (code division multiple access) cellular networks and OFDMA (orthogonal frequency division multiple access) networks.

In the second part of the paper, we first analyse random access wireless networks and pay special attention to 802.11 (Wi-Fi) networks. We then survey some topics in ad hoc multihop wireless networks, where we discuss arbitrary networks, as well as some theory of dense random networks. Finally, we provide an overview of the technical issues in the emerging area of wireless sensor networks.

Keywords. Wireless networks; resource allocation; CDMA networks; OFDMA networks.

1. Introduction

The idea of sending information over radio waves (i.e. *wireless communication*) is over a hundred years old. When several devices with radio transceivers share a portion of the radio spectrum to send information to each other, we say that we have a *wireless communication network*, or simply a wireless network. Hence, *wireless networking* is concerned with all the mechanisms, procedures or algorithms for efficient sharing of a portion of the radio spectrum so that all the instances of communication between the various devices obtain their desired *quality of service (QoS)*. In this paper we will survey the issues, concepts and techniques for wireless networking. Our approach will be tutorial and the emphasis will be on recent trends in the field.

The following is an outline of the paper. We begin this part with an overview of the area of wireless networking as it is currently practised. We do this in § 2. Section 4 is a descriptive, tutorial discussion of wireless physical layer communications, a technology that has been evolving rapidly. In the remainder of the paper we provide a survey of recent topics in wireless networking. In § 5 we discuss some resource allocation problems in CDMA access networks. The second part of the paper begins with a discussion of access networks based on random multiple access in § 6. In § 7 we provide a survey of wireless ad hoc internets. The emerging area of ad hoc wireless sensor networks (WSNs) is covered in § 8.

2. A taxonomy of current practice

We begin our survey with a look at a taxonomy of wireless networks as they exist today. Figure 1 provides such a taxonomy. Several commonly used terms of the technology will arise, as we discuss this taxonomy. These will be highlighted by the *italic* font, and their meanings will be clear from the context.

Fixed wireless networks include line-of-sight microwave links, which until recently were popular for long distance transmission. Such networks basically comprise point-to-point line-of-sight digital radio links. When such links are set up, with properly aligned high gain antennas on tall masts, the links can essentially be viewed as point-to-point bit pipes, albeit with a higher bit error rate than wired links. Thus, in such fixed wireless networks no essentially new issues arise than in a network of wired links.

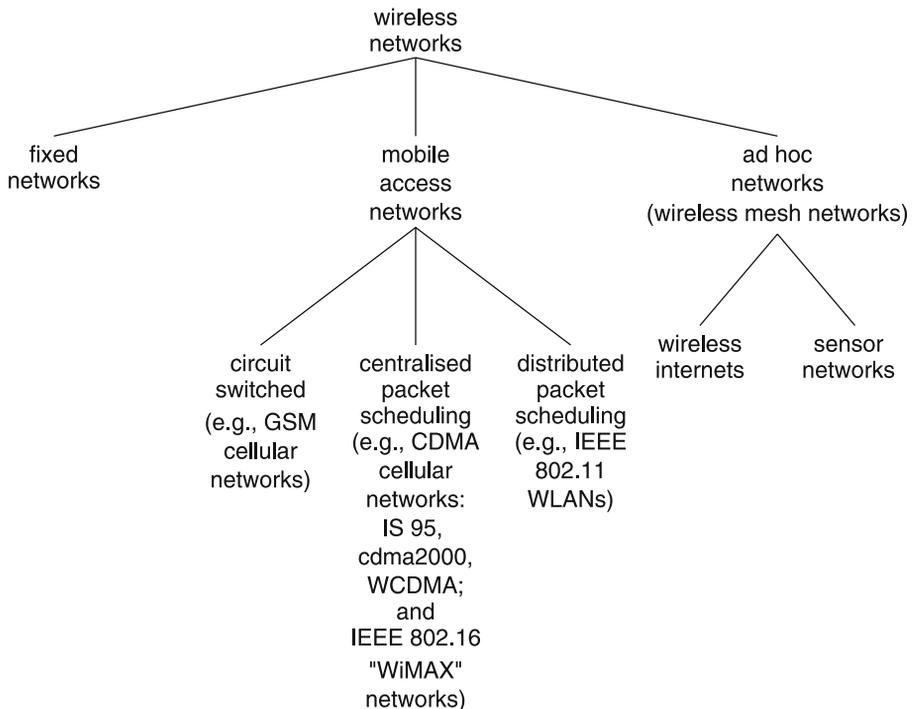


Figure 1. A taxonomy of wireless networks.

On the other hand, the second and third categories shown in the first level of the taxonomy (i.e. access networks, and ad hoc networks) involve *multiple access* where, in the same geographical region, several devices share a radio spectrum to communicate among themselves. Currently, the most important role of wireless communications technology is in mobile access to wired networks. We can further classify such access networks into two categories: circuit switched and packet switched, and the latter further into systems where the packet scheduling is centrally controlled and those in which the scheduling is distributed.

Cellular wireless networks were introduced in the early 1980s as a technology for providing access to the wired phone network to mobile users. The network coverage area is partitioned into regions (with diameters ranging from 100s of meters to a few kilometers) called *cells*, hence the term ‘cellular’. In each cell there is a *base station* (BS), which is connected to the wired network, and through which the mobile devices in the cell communicate over a one hop wireless link. The cellular systems that have the most widespread deployment are the ones that share the available spectrum using *frequency division multiple access and time division multiple access* (FDMA/TDMA) technology (Goldsmith 2005). Among such systems by far the most commercially successful has been the GSM system, developed by an European consortium. The available spectrum is first partitioned into a contiguous *uplink* band and another contiguous *downlink* band. Each of these bands is statically or dynamically partitioned into re-use sub-bands, with each cell being allocated such a sub-band (this is the FDM aspect). The partitioning of the uplink and downlink bands is done in a paired manner so that each cell is actually assigned a pair of sub-bands. Each sub-band is further partitioned into channels or carriers (also an FDM aspect), each of which is digitally modulated and then slotted in such a way that a channel can carry a certain number of calls (e.g. 8 calls) in a TDM fashion. Each arriving call request in a cell is then assigned a slot in one of the carriers in that cell; of course, a pair of slots is assigned in paired uplink and downlink channels in that cell. Thus, since each call is assigned dedicated resources, the system is said to be *circuit switched*, just like the wireline phone network. These are *narrowband systems* (i.e. users bit streams occupy frequency bands just sufficient to carry them), and the radio links operate at a high signal-to-interference-plus-noise-ratio (SINR), and hence careful *frequency planning* (i.e. partitioning of the spectrum into re-use sub-bands, and allocation of the sub-bands to the cells) is needed to avoid *co-channel interference*. The need for allocation of frequency bands over the network coverage area (perhaps even dynamic allocation over a slow time scale), and the grant and release of individual channels as individual calls arrive and complete, requires the control of such systems to be highly centralised.

Another cellular technology that has developed over the past decade is the one based on *code division multiple access* (CDMA). The entire available spectrum is re-used in every cell. These are *broadband systems*, which means that each user’s bit stream (a few kilobits per second) occupies the entire available radio spectrum (a few megahertz). This is done by *spreading* each user’s signal over the entire spectrum by multiplying it by a *pseudorandom sequence*, that is allocated to the user. This makes each user’s signal appear like noise to other users. The knowledge of the spreading sequences permits the receivers to separate the users’ signals, by means of *correlation receivers*. Although no frequency planning is required for CDMA systems, the performance is *interference limited* as every transmitted signal is potentially an interferer of every other signal. Thus at any point of time there is an allocation of powers to all the transmitters sharing the spectrum, such that their desired receivers can decode their transmissions, in the presence of all the cross interferences. These desired power levels need to be set depending on the locations of the users, and the consequent channel conditions between the users and the base stations, and need to be adjusted as users move about

and channel conditions change. Hence, tight control of transmitter power levels is necessary. Further, of course, the allocation of spreading codes, and management of movement between cells needs to be done. We note that, unlike the FDMA/CDMA system described earlier, there is no dedicated allocation of resources (frequency and time-slot) to each call. Indeed, during periods when a call is inactive no radio resources are utilised, and the interference to other calls is reduced. However, if there are several calls in the system, each needing certain *quality of service* (bit rate, maximum bit error rate) then the number of calls in the system needs to be controlled. This requires *call admission control* which is an essential mechanism in CDMA systems. Evidently, these are all centrally coordinated activities and hence even CDMA cellular systems depend on central intelligence that resides in the *base station controllers* (BSCs).

Until recently cellular networks were driven primarily by the needs of circuit switched voice telephony; on demand, a mobile phone user is provided a wireless digital communication channel on which is carried compressed telephone quality (though not 'toll' quality) speech. Above, we have described two technologies for *second generation (2G)* cellular wireless telephony. Recently, with the growing need for mobile Internet access, there have been efforts to provide packetised data access on these networks as well. In the FDMA/TDMA systems, low bit rate data can be carried on the digital channel assigned to a user; as is always the case in circuit switched networks, flexibility in the allocation of bandwidth is limited to assigning multiple channels to each user. This is, of course, inefficient, since data is intrinsically bursty (Kumar *et al* 2004). On the other hand, there is considerable flexibility in CDMA systems where there is no dedicated allocation of resources (spectrum or power). In fact, both voice and data can be carried in the packet mode, with the user bit rate, the amount of spreading, and the allocated power changing on a packet-by-packet basis. This is what is envisaged for the *third generation (3G)* cellular systems (Goldsmith 2005), which will be entirely based on CDMA technology, and will carry *multimedia traffic* (i.e. store and forward data, packetised telephony, interactive video, and streaming video).

Cellular networks were developed with the primary objective of providing wireless access for mobile users. With the growth of the Internet as the de facto network for information dissemination, access to the Internet has become an increasingly important requirement in most countries. In large congested cities, and in developing countries without a good wireline infrastructure, *fixed wireless access* to the Internet is seen as a significant market. It is with such an application in mind that the IEEE 802.16 standards have been developed, and are known in the industry as WiMAX. The major technical advance in WiMAX is in the adoption of several high performance physical layer (PHY) technologies to provide several 10s of Mbps between a base station (BS) and fixed subscriber stations (SS) over distances of several kilometres. The PHY technologies that have been utilised are *orthogonal frequency division multiple access (OFDMA)* and multiple antennas at the transmitters and the receivers. The latter are commonly referred to as *MIMO (multiple input multiple output)* systems. The spectrum is shared between uplink and downlink transmissions, by dividing time into frames, and each frame is divided into an up-link and a down-link part (this is called *time division duplexing (TDD)*). The SSs request for allocation of time on the up-link. The BS allocates time to various down-link flows in the down-link part of the frame and, based on SS requests, in the up-link part of the frame. This kind of MAC structure has been used in several earlier systems, e.g. satellite networks involving very small aperture satellite terminals (VSATs), and even in wireline systems such as those used for the transmission of digital data over cable television networks.

We now discuss the third class of networks in the mobile access category in the first level of the taxonomy shown in figure 1, i.e. distributed packet scheduling. While cellular networks have emerged from centrally managed point-to-point radio links, another class of wireless networks has emerged from the idea of *random access* whose prototypical example is the Aloha network (Bertsekas & Gallager 1992). Spurred by advances in digital communication over radio channels, random access networks can now support bit rates close to desk-top wired Ethernet access. Hence random access wireless networks are now rapidly proliferating as the technology of choice for wireless Internet access with limited mobility. The most important standards for such applications are the ones in the IEEE 802.11 series. Networks based on this standard now support physical transmission speeds from a few Mbps (over 100s of meters) up to 100 Mbps (over a few meters). The spectrum is shared in a *statistical TDMA* fashion (as opposed to slotted TDMA, as discussed in the context of first generation FDMA/TDMA systems, above), with nodes contending for the channel, colliding, then backing off for random amounts of time, and then re-attempting. When a node is able to acquire the channel it can send at the highest of the standard bit rates that can be decoded given the channel condition between it and its receiver. This technology is predominantly deployed for creating *wireless local area networks* (WLANs) in campuses and enterprise buildings, thus basically providing a one hop untethered access to a building's Ethernet network. In the latest enhancements to the IEEE 802.11 standards, MIMO-OFDM physical layer technologies are being employed in order to obtain up to 100 Mbps transmission speeds in indoor environments.

With the wide spread deployment of IEEE 802.11 WLANs in buildings, and even public spaces (such as shopping malls and airports), an emerging possibility is that of carrying interactive voice and streaming video traffic over these networks. The emerging concept of 4th generation wireless access networks envisions mobile devices that can support multiple technologies for physical digital radio communication, along with the resource management algorithms that would permit a device to seamlessly move between 3G cellular networks, IEEE 802.16 access networks and IEEE 802.11 WLANs, while supporting a variety of packet mode services, each with its own QoS requirements.

With reference to the taxonomy in figure 1, we now turn to the category labelled 'ad hoc networks' or 'wireless mesh networks.' Wireless access networks provide mobile devices with a one hop wireless access to a wired network. Thus, typically, in the path between two such mobile devices there is only one or at most two wireless links. On the other hand, a wireless ad hoc network comprises several devices arbitrarily located in a space (e.g. a line segment, or a two-dimensional field). Each device is equipped with a radio transceiver, all of which typically share the same radio frequency band. In this situation, the problem is to communicate between the various devices. Nodes need to discover neighbours in order to form a topology, good paths need to be found, and then some form of time scheduling of transmissions needs to be employed in order to send packets between the devices. Packets going from one node to another may need to be forwarded by other nodes. Thus, these are *multihop* wireless packet radio networks, and they have been studied as such over several years. Interest in such networks has again been revived in the context of *multihop wireless internets* (Perkins 2001) and *wireless sensor networks* (Akyildiz *et al* 2002). We discuss these briefly in the following two paragraphs.

In some situations it becomes necessary for several mobile devices (such as portable computers) to organise themselves into a multihop wireless packet network. Such a situation could arise in the aftermath of a major natural disaster such as an earthquake, when emergency management teams need to coordinate their activities and all the wired infrastructure has been damaged. Notice that the kind of communication that such a network would be required to

support would be similar to what is carried by regular public networks, i.e. point-to-point store and forward traffic such as electronic mails and file transfers, and even low bit rate voice and video communication. Thus, we can call such a network a *multihop wireless internet*. In general, such a network could attach at some point to the wired Internet.

While multihop wireless internets have the service objective of supporting instances of point-to-point communication, an ad hoc wireless sensor network has a global objective. The nodes in such a network are miniature devices, each of which carries a microprocessor (with an energy efficient operating system), one or more sensors (e.g. light, acoustic, or chemical sensors), a low power, low bit rate digital radio transceiver, and a small battery. Each sensor monitors its environment and the objective of the network is to deliver some global information or an inference about the environment to an operator who could be located at the periphery of the network, or could be remotely connected to the sensor network. An example is the deployment of such a network in the border areas of a country to monitor intrusions. Another example is to equip a large building with a sensor network comprising devices with strain sensors in order to monitor the building's structural integrity after an earthquake. Yet another example is the use of such sensor networks in monitoring and control systems such as those for the environment of an office building or hotel, or a large chemical factory.

3. Technical elements

In the previous section we have provided an overview of the current practice of wireless networks. We organised our presentation around a taxonomy of wireless networks shown in figure 1. Although the technologies that we have discussed may appear to be disparate, there are certain common technical elements that comprise these wireless networks. The efficient realisation of these elements constitutes the area of wireless networking.

The following is an enumeration and preliminary discussion of the technical elements. In the remaining sections of this tutorial survey, we will focus on these three elements, as they are essential to all types of wireless networks.

3.1 *Transport of the users' bits over the shared radio spectrum*

There is, of course, no communication network unless bits can be transported between users. Digital communication over mobile wireless links has rapidly evolved over the past 2 decades (Tse & Viswanath 2005) and (Goldsmith 2005). Several approaches are now available, with various trade-offs and areas of applicability. An important technique is that even in a given system the digital communication mechanism can be adaptive. Firstly, for a given digital modulation scheme the parameters can be adapted (e.g. the transmit power, or the amount of error protection), and, secondly, sophisticated physical layers actually permit the modulation itself to be changed even at the packet or burst timescale (e.g. if the channel quality improves during a call then a higher order modulation can be used, thus helping in store and forward applications that can utilise such time varying capacity). This adaptivity is useful in the mobile access situation where the channels and interference levels are rapidly changing.

3.2 *Neighbour discovery, association and topology formation, routing*

Except in the case of fixed wireless networks, we typically do not 'force' the formation of specific links in a wireless network. For example, in an access network each mobile device could be in the vicinity of more than one BS or *access point* (AP). To simplify our writing, we will refer to a BS or an AP as an access device. It is a non-trivial issue as to which

access device a mobile device connects through. First, each mobile needs to determine which access devices are in its vicinity, and through which it can potentially communicate. Then each mobile should associate with an access device such that certain overall communication objectives are satisfied. For example, if a mobile is in the vicinity of two BSs and needs certain quality of service, then its assignment to only a particular one of the two BSs may result in satisfaction of the new requirement, and all the existing ones.

In the case of an access network the problem of routing is trivial; a mobile associates with a BS and all its packets need to be routed through that BS. On the other hand, in an ad hoc network, after the associations are made and a topology is determined, good routes need to be determined. A mobile would have several neighbours in the discovered topology. In order to send a packet to a destination, an appropriate neighbour would need to be chosen, and this neighbour would further need to forward the packet towards the destination. The choice of the route would depend on factors such as: the number of hops on the route, the congestion along the route, and the residual battery energies in devices along the route.

We note that association and topology formation is a procedure whose timescale will depend on how rapidly the relative locations of the network nodes is changing. However, one would typically not expect to associate and re-associate a mobile device, form a new topology, or recalculate routing at the packet timescale.

If mobility is low, for example in wireless LANs and static sensor networks, one could consider each fixed association, topology and routing and compute the performance measures at the user level. Note that this step requires a scheduling mechanism, discussed as the next element. Then that association, topology and routing would be chosen that optimises, in some sense, the performance measures. In the formulation of such a problem, first we need to identify one or more performance objectives (e.g. the sum of the user utilities for the transfer rates they get), then we need to specify whether we seek a co-operative optimum (e.g. the network operator might seek the global objective of maximising revenue; for one approach, (Kumar & Kumar 2005) or a non-cooperative equilibrium (the more practical situation, since users would tend to act selfishly, attempting to maximise their performance while reducing their costs; e.g. (Shakkottai *et al* 2006) and finally, whatever the solution of the problem, we need an algorithm (centralised or distributed) to compute it on-line.

If the mobility is high, however, the association problem would need to be dynamically solved as the devices move around (Hanly 1995). Such a problem may be relatively simple in a wireless access network, and, indeed, necessary since cellular networks are supposed to handle high mobility users. On the other hand, such a problem would be hard for a general mesh network; highly mobile wireless mesh networks, however, are not expected to be 'high performance' networks.

3.3 Transmission scheduling

Given an association, a topology and the routes, and the various possibilities of adaptation at the physical layer, the problem is to schedule transmissions between the various devices so that the users' QoS objectives are met. In its most general form, the schedule needs to dynamically determine which transceivers should transmit, how much they should transmit, and which physical layer (including its parameters, e.g. transmit power) should be used between each transceiver pair. Such a scheduler would be said to be *cross-layer* if it took into account state information at multiple layers, e.g. channel state information, as well as higher layer state information, such as link buffer queue lengths. Note that a scheduling mechanism will determine the schedulable region for the network, i.e. the set of user flow rates of each type that can be carried so that each flow's QoS is met.

In general, the above three technical elements are interdependent and the most general approach would be to jointly optimise them.

3.4 Specialised elements

In addition to the above elements that provide the basic communication functionality, some wireless networks require other functional elements, that could be key to the networks' overall utility. The following are two important ones, that are of special relevance to ad hoc wireless sensor networks.

3.4a Location determination: In an ad hoc wireless sensor network the nodes make measurements on their environment, and then these measurements are used to carry out some global computation. Often, in this process it becomes necessary to determine which location a measurement came from. Sensor network nodes may be too small to carry a GPS receiver. Hence GPS-free techniques for location determination become important (Doherty *et al* 2001). Even in cellular networks, there is a requirement in some countries, that, if needed, a mobile device should be geographically locatable. Such a feature can be used to locate someone who is stranded in an emergency situation and is unaware of the exact location.

3.4b Distributed computation: This issue is specific to wireless sensor networks. It may be necessary to compute some function of the values measured by sensors (e.g. the maximum or the average) (Giridhar & Kumar 2006) and (Khude *et al* 2005). Such a computation may involve some statistical signal processing functions such as data compression, detection or estimation. Since these networks operate with simple digital radios and processors, and have only small amounts of battery energy, the design of efficient self-organising wireless ad hoc networks and distributed computation schemes on them is an important emerging area. Since in such networks there is communication delay and also data loss, existing algorithms may need to be re-designed to be robust to information delay and loss.

4. The wireless PHY: Technologies and concepts

As discussed earlier, the most basic element of a wireless network is the ability to send bit streams over the available portion of the radio spectrum, i.e. wireless digital communication. In the OSI model these functionalities reside in the physical layer, hence the often used abbreviation 'PHY.' An excellent up-to-date coverage of this topic is provided in the books (Tse & Viswanath 2005) and (Goldsmith 2005). In order to understand the basic approaches to wireless digital communication, and the trade-offs involved, let us consider the following standard linear model for digital communication over a bandlimited channel between a pair of devices

$$Y_k = \sum_{l=0}^{L-1} A_k(l)X_{k-l} + I_k + Z_k \quad (1)$$

The following is an explanation of this expression. Time is divided into a sequence of *symbol times*, indexed by k . The symbol duration, or equivalently the symbol rate, depends on the width of the radio spectrum (usually called the *bandwidth*¹) that the signal occupies, and is given by $\frac{1}{W}$ if W is the bandwidth. The various terms are understood as follows.

¹The term 'bandwidth' has varied and confusing usage in the wireless networking literature. The RF spectrum in which a system operates has a bandwidth. When a digital modulation scheme is

- (1) The sequence X_k is a sequence of complex numbers into which the user's bit stream is encoded. For example, $X_k \in \{1, j, -1, -j\}$ with two user bits being mapped into each symbol, e.g. $00 \rightarrow 1, 01 \rightarrow j, 11 \rightarrow -1, 10 \rightarrow -j$. The set $\{1, j, -1, -j\}$ is called a *constellation*; this is the QPSK (quadrature phase shift keying) constellation. The combination of the choice of the constellation and the way we map bits into the symbols is called *modulation*. Evidently, the modulation scheme has to be agreed on between the receiver and the transmitter before data transfer can take place.
- (2) For every k , $A_k(l)$, $0 \leq l \leq L$, are complex random variables that model the way the channel attenuates and phase shifts the transmitted symbols (they are called channel 'gains', and model the phenomenon of multiplicative *fading*). $A_k(l)$ models the influence that the input l symbols in the past has on the channel output at k . Thus, in general, a channel has memory; in the model, the memory extends over L symbols. This is also called the *delay spread*, T_d , since memory arises as a consequence of there existing several paths from the transmitter to the receiver, with the different paths having different delays. For example, if the path lengths differ by no more than 100s of meters then the delay spread would be in 100s of nanoseconds. The notation shows that the channel gain at the k th symbol could be a function of the symbol index k ; this models the fact that fading is a time varying phenomenon. As the devices involved in the communication move around the radio channel between them also keeps changing.
- (3) I_k and Z_k are also complex random variables that model the interference (from other transmissions in the same or nearby spectrum²) and the random additive noise, respectively. The additive noise term models, for example, the thermal noise in the electronic circuitry of the receiver, and is taken to be a white Gaussian random process. A commonly used simplification is to use the same model even for the interference process, with the noise and interference processes being modelled as being independent.

It follows that the received sequence Y_k is also a complex valued random process. The problem for the receiver, on receiving the sequence of complex numbers Y_k , is to carry out a *detection* of which symbols X_k were sent and hence which user bits were sent. This problem is particularly challenging in mobile wireless systems since the channel is randomly changing with time. Thus, in designing and understanding modulation schemes it is important to have an effective but simple model of the channel attenuation process.

The delay spread is a time domain concept. We can also view this in the frequency domain as follows. The symbols X_k are carried over the RF spectrum by first multiplying them with a (baseband) pulse of bandwidth approximately W (e.g. 200 KHz), and then upconverting the resulting signal to the carrier frequency (e.g. 900 MHz). Due to the delay spread in the channel some of the frequency components in the modulated signal can get selectively attenuated, resulting in the corruption of the symbols they carry. On the other hand, if the delay spread

used over this spectrum then a certain bit rate is provided; often this aggregate bit rate may also be referred to as bandwidth, and we may speak of users sharing the bandwidth. This latter usage is unambiguous in the wireline context. In multi-access wireless networks, however, users would be sharing the same RF spectrum bandwidth, but would be using different modulation schemes and thus obtaining different (and time varying) bit rates, rendering the use of a phrase such as 'bandwidth assigned to a user' very inappropriate.

²Note that we are taking the simplified approach of treating other users' signals as interference. More generally, it is technically feasible to extract multiple users' symbols even though they are superimposed. This is called *multiuser detection*.

is very small compared to $\frac{1}{W}$, then the pulse would be passed through with only an overall attenuation; this is called *flat fading*. The reciprocal of the delay spread is called the *coherence bandwidth*, W_c . Thus, if the coherence bandwidth is larger than the system bandwidth W then all the frequencies fade together and we have flat fading. On the other hand, if the coherence bandwidth is small compared to the system bandwidth (delay spread is large compared to $\frac{1}{W}$) then we say that the channel is *frequency selective*, as within the band different frequencies will fade differently. In relation to the model in Equation 1, and recalling that the symbol duration is $\frac{1}{W}$, we observe that frequency selectivity corresponds to the channel memory extending over more than 1 symbol time, and hence to the existence of *intersymbol interference (ISI)*. Thus when high bit rates are carried over wideband channels (i.e. large W) then techniques have to be used to combat ISI, or to avoid it altogether. We will encounter CDMA and OFDMA later, as two wideband schemes that actually exploit delay spread or frequency selectivity to achieve *diversity*.

The random process $A_k(l)$ also has correlations in time k . The fading can be taken to be weakly correlated over times separated by more than the *coherence time*, T_c . The coherence time is related to the *Doppler frequency*, f_d , which is related to the carrier frequency, f_c , the speed of movement, v , and the speed of light, c , by $f_d = f_c \frac{v}{c}$. Roughly, the coherence time is the inverse of the Doppler frequency. For example, if the carrier frequency is 900 MHz, and $v = 20$ meters/sec, then $f_d = 60$ Hz, leading to a coherence time of 10s of milliseconds. In the indoor office or home environment, the Doppler frequency could be just a few Hz (e.g. 3 Hz), with coherence times of 100s of milliseconds.

4.1 Narrowband systems

4.1a The basic linear model: In a system where the signal bandwidth W is small, since the symbol times are large, it is possible that the delay spread is small compared to the symbol times, and then the delay spread can be ignored. This could be a reasonable simplification for a narrowband system, where the available radio spectrum is channelised and each bit stream occupies one channel. Then the model of Equation 1 simplifies to

$$Y_k = A_k X_k + I_k + Z_k \quad (2)$$

where we write A_k for $A_k(0)$. Let us now think of this system as a bit carrier or a digital link. Several issues are brought up by this model of the bit carrier.

4.1b Link performance: The bit stream being encoded into this channel will carry some user application, for example, packet voice, or TCP controlled file transfers. In order to perform well, each such application will expect some performance levels from the bit carrier. Because of the random disturbances (fading, interference and noise) a certain amount of data loss is inevitable. Some systems attempt to reduce data loss by using various forms of *Automatic Repeat Request (ARQ)*; (Bertsekas & Gallager 1992) and these introduce delay. Thus from the point of view of the user, the bit carrier will have:

- A bit rate, which is possibly time varying (see below)
- A data loss rate (measured in terms of packet loss rate to the application)
- A random delay.

For packet voice, the link will need to provide an average bit rate that exceeds the rate of the voice coder-and-packetiser output, and, depending on the amount of voice carried in each

packet, a certain maximum packet loss rate. If packets carry 20 ms of voice, for example, then a 1 % packet loss rate is acceptable. For an application such as streaming video, the average bit rate over the radio link must be larger than the average coded rate of the video source. If the average bit rate over the link is too close to the source rate, then there could be large buffering delays or buffer overflow loss. Buffering delays can be compensated by playout buffering at the receiver, while packet loss will need to be concealed in the video decoder. Clearly, there are limits to how much delay and loss can be tolerated, and this will put a requirement on the minimum average rate at which the buffer must be served. TCP is sensitive to packet loss as well; for example, for local area transfers a better than 1 % packet loss rate is required (Kumar *et al* 2004) for a TCP throughput better than 90 % of the link's bit rate.

4.1c *Channel coding*: We see that the transmitted symbol can be corrupted by additive and multiplicative random disturbances. In order to combat this, the transmitter could send blocks of symbols, each block carrying fewer than the maximum possible information bits. For example, if a block of 16 QPSK symbols (see above) is used, then a rate half code would send only 16 information bits over these 16 symbols. Thus 2^{16} blocks of 16 QPSK symbols (out of a possible 4^{16} such blocks) need to be chosen. These would constitute the code, which would be known to the receiver as well. The code would be designed so as to provide a high probability of recovering the transmitted string, in spite of the fading and additive noise plus interference. Notice that in frame based or packet based systems block coding does not really introduce any additional delay since a frame or packet worth of bits has to be accumulated anyway. However, the user level bit rate gets reduced.

4.1d *Exploiting diversity*: It turns out that commonly used codes provide a better chance of successful decoding if the channel error process is uncorrelated over the code symbols. We saw earlier that the channel fade process, A_k , is correlated over periods called the channel coherence time, which depends on the speed of movement of the mobile device. *Interleaving* is a way to obtain an uncorrelated fade process from a correlated one. Basically, the transmitter does not send successive symbols of a code word over contiguous channel symbols, but successive symbols are separated out so that they see uncorrelated fading. In between, other code words are interleaved. We say that interleaving exploits *time diversity*, i.e. the fact that channel times separated by more than the coherence time fade independently. Observe that interleaving introduces *interleaving delay* which adds to the link delay, and hence to the end-to-end delay over the wireless network. Also, interleaving fails if the fading is very slow, for example if the relative motion stops, and the transmitter–receiver pair are caught in a bad fade.

4.1e *Adaptive modulation and coding*: Given that the channel is time varying and the power is limited, it is possible for the system to use one of several different modulation and coding schemes. When the channel is good then a higher order constellation can be used and the coding can be weak (more information bits per symbol). On the other hand, when the channel is bad then a lower order constellation can be used with strong coding. Note that a channel could become 'bad' even if the channel coherence time becomes large, thus making interleaving ineffective, and requiring either stronger coding or a lower constellation. Of course, if adaptive modulation and coding is done then the receiver has to be informed as to which modulation and coding scheme is being used. One way of doing this is for the transmitter to always start by sending a small amount of information using a fixed modulation and coding; this is used

to inform the receiver of the modulation and coding to be used for the actual data to follow. In centrally coordinated cellular systems, the base station can periodically broadcast such information.

Having chosen a constellation, the transmitter could boost its power and thus spread out the symbols of the constellation. Note, however, that a high transmitter power not only drains the battery of a mobile device, but also causes interference to other concurrent cochannel transmissions.

Note that adaptive coding and modulation implies that the bit rate of the link is time varying. The way in which the rate varies depends on the way in which the channel and the interference varies over time. It is certainly preferable to have links that provide an essentially constant bit rate over time. In some systems, such as CDMA the interference becomes fairly constant because of the law of large numbers averaging effect. However, in systems where this is not the case (e.g. high speed downlink data in 3rd generation CDMA systems) and where channel variations are exploited to achieve higher bit rates, the effect of a time varying bit rate on the application becomes important to analyse (see Section 5.2).

4.1f *Signal-to-interference-plus-noise ratio (SINR)*: Having chosen a particular modulation and coding scheme, interleaving depth and transmission power level, the ratio of the received symbol power to interference plus noise power, the SINR, will determine the probability of successful detection of the transmitted data. The way in which the average probability of error on SINR depends on whether we account for fading or not. If the channel fade level is taken to be constant then the error probability decreases exponentially with increasing SINR. On the other hand, if fading is taken into account then the average error probability can decrease only reciprocally with SINR (for the Rayleigh model for the amplitude of the fading process). The application being carried on the wireless link will set the requirement for the link error rate, which, along with the channel model, will set a lower limit on the SINR. Thus the design of the link, for a given application, can be reduced to achieving a target SINR.

4.2 Channel capacity without fading

Consider the following simpler version of the basic linear model that was shown in Equation 2

$$Y_k = X_k + Z_k \quad (3)$$

where we notice that we have removed the multiplicative fading term, and the additive interference term, leaving just a model in which the output random variable at symbol k , is the input symbol X_k with an additive noise term Z_k . Thus, there is no attenuation of the transmitted symbol, but there is perturbation by additive noise. A common model for the random process $Z_k, k \geq 1$, is that these are i.i.d. Gaussian random variables with mean 0 and variance σ^2 .

Suppose that the input has the following *power constraint*:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n |x_k|^2 \leq \bar{P} \quad (4)$$

i.e. the average energy per symbol is bounded by \bar{P} . This is a practical constraint as power amplifiers can only generate limited power. Also it is desirable to limit the rate of battery drain. Hence, some form of power constraint is usually placed in wireless communication problems.

If the input symbols are allowed to be only real numbers, then it has been shown that the maximum rate at which information can be transmitted over this *AWGN channel*, in bits/symbol, is given by

$$C = \frac{1}{2} \log_2 \left(1 + \frac{P_{rcv}}{\sigma^2} \right) \text{ bits/symbol} \quad (5)$$

where, P_{rcv} is the *received* signal power, and $\frac{P_{rcv}}{\sigma^2}$ is the received signal to noise power ratio. Evidently, here in the no fading case, we have $P_{rcv} = P$. If the input symbols are allowed to be complex numbers (as is practically, always the case, by using two 90° phase shifted carriers), and the additive noise sequence comprises i.i.d. circularly symmetric complex Gaussian random variables, then the capacity formula takes the simple form

$$C = \log_2 \left(1 + \frac{P_{rcv}}{\sigma^2} \right) \text{ bits/symbol} \quad (6)$$

We note that the above capacity expressions gave the answer in bits per symbol. Often, in analysis it is better to work with natural logarithms. With this in mind we can rewrite Equation 6 as

$$C = \ln \left(1 + \frac{P_{rcv}}{\sigma^2} \right) \text{ nats/symbol}$$

Since $\ln x = \log_2 x \times \ln 2$, the capacity in nats per symbol is obtained by multiplying the capacity in bits per symbol by $\ln 2 \approx 0.693$.

4.3 Channel capacity with fading

Consider again the model shown in Equation 2, but now let us retain the multiplicative fading term, while removing the interference term, thus yielding

$$Y_k = A_k X_k + Z_k \quad (7)$$

Given the power constraint in Equation 4, the question naturally arises as to how to choose the energy to use in the k th symbol, for each k . Suppose that the transmitter knows the channel attenuation at each symbol time. This is practically possible (by using *pilots*) if the channel is varying slowly. If the channel happens to be highly attenuating in a symbol, we can try to combat the attenuation by boosting the symbol power, or we can wait the fade out and use the power only when the channel is good again. If the random process A_k , $k \geq 1$, is ergodic then the fraction of time it will be in each state will be the same as the probability of that state. Hence, if the distribution of A_k , $k \geq 1$, gives positive mass to good states, then eventually we will get a good period to transmit in. Of course, this argument ignores the fact that there could be urgent information waiting to be sent out, but we will come back to this issue later, and, for the moment, seek a *power control* that will maximise the rate at which the channel can send bits.

Since the performance of the channel depends on the signal power to noise power ratio, let us write $H_k = |A_k|^2$. Thus, H_k , $k \geq 1$, is the random sequence of power ‘gains’ (or attenuations) of the channel. Now consider those times at which $H_k = h$, for some non-negative number h . Suppose, at these times we use the transmit power $P(h)$ (i.e. over these

symbols, $\mathbb{E}(|X|^2) = P(h)$, or, more precisely, $\lim_{n \rightarrow \infty} \frac{\sum_{k=1}^n I_{\{H_k=h\}} |X_k|^2}{\sum_{k=1}^n I_{\{H_k=h\}}} = P(h)$. Then, over such times, the channel capacity achievable is

$$C = \ln \left(1 + \frac{hP(h)}{\sigma^2} \right)$$

where σ^2 continues to be the noise power³.

Let the channel power gain process $H_k, k \geq 1$, take values in the finite set $\mathcal{H} = \{h_1, h_2, \dots, h_J\}$, where $J := |\mathcal{H}|$. Let us denote by $P_j = P(h_j)$, the power used when the channel power gain is h_j , and write the power control as the J vector \mathbf{P} . Then, it can be shown that, for this power control, the average channel capacity over n symbols is given by

$$\begin{aligned} C_n(\mathbf{P}) &= \frac{1}{n} \sum_{h_j \in \mathcal{H}} \sum_{k=1}^n I_{\{H_k=h_j\}} \ln \left(1 + \frac{h_j P(h_j)}{\sigma^2} \right) \\ &= \frac{1}{n} \sum_{k=1}^n \ln \left(1 + \frac{H_k P(H_k)}{\sigma^2} \right) \end{aligned}$$

where the first equality is obtained by summing over those symbols at which the channel attenuation is $h_j \in \mathcal{H}$, for each such h . The second equality is obviously the same calculation. Now, taking n to ∞ on both sides, using the Birkhoff Strong Ergodic Theorem on the right hand side, and defining the resulting limit as $C(\mathbf{P})$, we find that $C_n(\mathbf{P})$ converges with probability 1 to

$$C(\mathbf{P}) := \mathbb{E} \left(\ln \left(1 + \frac{HP(H)}{\sigma^2} \right) \right) \quad (8)$$

It can be shown that this capacity is achievable by a coding scheme, albeit with large coding delays.

We wish to ask the question: ‘What is the best power control?’ First, as explained earlier we impose a power constraint on the input symbols, i.e.

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n |X_k|^2 \leq \bar{P}$$

Consider the left hand side of this expression

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n |X_k|^2 &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{h_j \in \mathcal{H}} \sum_{k=1}^n I_{\{H_k=h_j\}} |X_k|^2 \\ &= \lim_{n \rightarrow \infty} \sum_{h_j \in \mathcal{H}} \left(\frac{1}{n} \sum_{k=1}^n I_{\{H_k=h_j\}} \right) \left(\frac{\sum_{k=1}^n I_{\{H_k=h_j\}} |X_k|^2}{\sum_{k=1}^n I_{\{H_k=h_j\}}} \right) \\ &= \sum_{h_j \in \mathcal{H}} P_j g_j \end{aligned}$$

³We note that, to achieve this rate, we will need to *code across* symbols with the same power gains h , thus adding even more delay.

where $g_j, h_j \in \mathcal{H}$, is the fraction of symbols that find the channel in the power attenuation h_j , i.e. $\Pr(H = h_j) = g_j$, and H denotes the marginal random variable of the process $H_k, k \geq 1$. We have also used the fact, discussed above, that in those symbols in which the power gain is h_j , the average transmitter power used is P_j . Notice that, in terms of $g_j, h_j \in \mathcal{H}$, we can write

$$C(\mathbf{P}) = \sum_{h_j \in \mathcal{H}} \ln \left(1 + \frac{h_j P_j}{\sigma^2} \right) g_j$$

We are thus led to the following optimisation problem

$$\max \sum_{h_j \in \mathcal{H}} \ln \left(1 + \frac{h_j P_j}{\sigma^2} \right) g_j$$

subject to

$$\begin{aligned} \sum_{h_j \in \mathcal{H}} P_j g_j &\leq P \\ P_j &\geq 0 \quad \text{for every } h_j \in \mathcal{H} \end{aligned} \tag{9}$$

This is a nonlinear optimisation problem with linear constraints. We will solve it from first principles. For each power control \mathbf{P} , and a number $\lambda \geq 0$, consider the function defined as follows:

$$L(\mathbf{P}, \lambda) := \sum_{h_j \in \mathcal{H}} \ln \left(1 + \frac{h_j P_j}{\sigma^2} \right) g_j - \lambda \sum_{h_j \in \mathcal{H}} P_j g_j$$

It is as if we are penalising ourselves for the use of power, and λ is the *price* per unit power⁴. It can now be shown that, for fixed λ , $L(\mathbf{P}, \lambda)$ is a strictly concave function of the vector argument \mathbf{P} . Let us maximise the function $L(\mathbf{P}, \lambda)$, for a given λ , over the power controls \mathbf{P} , while requiring only the non-negativity of these powers⁵. The strict concavity of $L(\mathbf{P}, \lambda)$ in \mathbf{P} , implies that there is a unique maximising power vector in \mathbb{R}^+ . Re-writing,

$$L(\mathbf{P}, \lambda) := \sum_{h_j \in \mathcal{H}} \left(\ln \left(1 + \frac{h_j P_j}{\sigma^2} \right) - \lambda P_j \right) g_j$$

we make the simple observation, that, since the power constraint is no longer imposed, we can maximise the above expression term by term for each channel gain h_j . Then it easily follows that the power vector \mathbf{P}_λ^* that maximises $L(\mathbf{P}, \lambda)$ has the form

$$P_{\lambda,j}^* = \left(\frac{1}{\lambda} - \frac{\sigma^2}{h_j} \right)^+$$

⁴The dimension of the function $L(\cdot)$ could be monetary, in which case we should multiply the capacity term by a monetary value per unit capacity. But on dividing across by this value we again get the same form as displayed.

⁵We say that the power constraint has been *relaxed*, leaving only the non-negativity constraint.

Define

$$P_\lambda := \sum_{h_j \in \mathcal{H}} P_{\lambda,j}^* g_j$$

and consider another vector \mathbf{P} , with

$$\sum_{h_j \in \mathcal{H}} P_j g_j \leq P_\lambda$$

i.e. we are asking for power controls whose average power is no more than that of the power control that maximises $L(\mathbf{P}, \lambda)$, for a given power price λ . Since \mathbf{P}_λ^* maximises $L(\mathbf{P}, \lambda)$ for the given λ , for all power controls \mathbf{P} , we have

$$C(\mathbf{P}_\lambda^*) - \lambda \sum_{h_j \in \mathcal{H}} P_{\lambda,j}^* g_j \geq C(\mathbf{P}) - \lambda \sum_{h_j \in \mathcal{H}} P_j g_j$$

i.e.

$$\begin{aligned} C(\mathbf{P}_\lambda^*) &\geq C(\mathbf{P}) + \lambda \left(P_\lambda - \sum_{h_j \in \mathcal{H}} P_j g_j \right) \\ &\geq C(\mathbf{P}) \end{aligned}$$

since \mathbf{P} was chosen to make the second term on the right non-negative. We conclude that \mathbf{P}_λ^* is optimal for the Problem 9 among all power controls that satisfy the constraint \bar{P}_λ . It follows that we just need to choose a λ such that $P_\lambda = P$, i.e.

$$\sum_{h_j \in \mathcal{H}} \left(\frac{1}{\lambda} - \frac{\sigma^2}{h_j} \right)^+ g_j = P$$

Let us denote the resulting power price by $\lambda(P)$. Then the optimal power control becomes $\mathbf{P}_{\lambda(\bar{P})}^*$, i.e.

$$P_{\lambda(\bar{P}),j}^* = \left(\frac{1}{\lambda(\bar{P})} - \frac{\sigma^2}{h_j} \right)^+.$$

4.4 Wideband systems

Unlike the narrow-band digital modulation discussed above, in CDMA and OFDMA the available spectrum is not partitioned, but all of it is dynamically shared among all the users. The simplest viewpoint is to think of CDMA in the time domain and OFDMA in the frequency domain. In a wideband system, a user's symbol rate is much smaller than the symbol rate that the channel can carry (i.e. $\frac{1}{W}$).

4.4a CDMA: In CDMA a user's symbol (say, ± 1 , for simplicity), which could be of duration G channel symbols (also called *chips*), is multiplied by a *spreading code* of length G channel symbols. This is called *direct sequence spread spectrum (DSSS)*, since this multiplication by the high rate spreading code results in the signal spectrum being spread out to cover the system bandwidth. If the user's bit rate is R and the channel symbol rate is W , then $G = \frac{W}{R}$

and is called the *spreading factor*. The spreading codes are chosen so that each code is approximately orthogonal to all time shifts of the other code, and also to its own time shifts. Then these code words are transmitted. If the delay spread is L channel symbols, and G is much larger than L , then the neighbouring user symbols do not overlap by much. We can still write the output of each channel symbol using Equation 1. Now the received signal Y_k can be viewed as a superposition of shifts of the transmitted code. Because of the orthogonality property mentioned above, at the receiver, multiplication of the received signal by various shifts of the spreading code yields a detection statistic that is the sum of L faded copies of the user symbol (i.e. ± 1). Since the L shifts correspond to as many paths from the transmitter to the receiver, this is called *multipath resolution*. If the paths fade independently then multipath diversity gain is obtained.

There is no channelisation, hence all user transmissions occupy the system bandwidth at the same time. Thus from the point of view of one transmission the other overlapping transmissions appear as co-channel interference. If there are many users, since the signal from each looks like a random sequence, the superposition of all the interference can be approximately modelled as white Gaussian noise. It can then be shown that the error performance for a user's flow is governed by the received signal power to interference + plus + noise power ratio (SINR) multiplied by $G = \frac{W}{R}$, which is, hence, also called the *processing gain*. The following is the intuition behind this: The correlation receiver for a particular user will collect all the power in the symbol of the user, that is spread over G chips, on the other hand, the interference will not correlate with the receiver and will remain spread, thus the effective detection signal to noise ratio for the user is the received SINR multiplied by G .

Scheduling transmissions in a CDMA systems involves a decision as to the constellation, coding and power level to use. These determine the rate at which a particular bit flow can be transmitted. Of course, this decision will have to be jointly made for all users, since the decision for one user impacts every other user.

4.4b OFDMA: OFDMA is based on OFDM (orthogonal frequency division multiplexing) (Goldsmith 2005), which can be viewed as statically partitioning the available spectrum into several, e.g. 128 or 512, subcarriers. If there are n subcarriers, then this is also called the OFDM *block length*. If the bandwidth of each subcarrier, say B , is chosen so that $B \ll W_c$, the coherence bandwidth, then each subcarrier has flat fading. First consider a single user occupying the entire system bandwidth. The user's symbols are modulated using some constellation. Then a block of n user symbols is transmitted over each of the carriers in such a way that the following model is achieved

$$Y_k = A_k X_k + I_k + Z_k \quad (10)$$

where $k, 1 \leq k \leq n$, indexes the sub-carrier over which the symbol X_k is sent rather than the time. Then A_k denotes the fading on that sub-carrier. Observe that this means that the user's symbols are transmitted in parallel over the sub-carriers. The term *orthogonal* in OFDM refers to the fact that the sub-carriers are chosen to be overlapping, but with center frequencies separated by the reciprocal of the OFDM block time; this makes the carriers approximately orthogonal over the block time. The fading is uncorrelated between sub-carriers that are spaced more than the coherence bandwidth apart. Hence, just as time diversity is exploited in TDMA systems, frequency diversity can be exploited in OFDMA systems, i.e. successive symbols of a user's code word can occupy independently fading sub-carriers. It is easy to see how the above concept can be used to share the flows from multiple users over a single OFDM link. Depending on the rate requirement of each user, a certain

number of sub-carriers can be allotted to each of the users. Scheduling transmissions over an OFDMA link involves a decision as to how many sub-carriers to assign to a user, and what constellations, coding and power levels to use from time to time, depending on the channel conditions and user rate requirements. Again, the decisions for various users are interrelated.

4.5 Cross layer control

In the above discussion we saw that in each system design there is some flexibility that can be used to adapt the modulation and coding to the time varying channel. Now, instead of blindly adapting the modulation and coding to optimise link capacity, the adaptation can be coupled to the higher layer requirements. For example, if the application is delay sensitive and if the queue has become large, then it may be beneficial to use some power and serve the queue even if the channel is bad. TCP controlled transfers over a wide area network are sensitive to variations in the bottleneck link service rate. When the Internet is accessed over a wireless access link, this is often the bottleneck. Hence it would be of interest to design a cross layer scheduling scheme that somehow resolves the poor interaction between the TCP control loop and a time varying bottleneck service rate (Karnik & Kumar 2005). Thus a cross layer control would be one in which state information from several layers is pooled in order to make decisions at these layers.

5. Access networks: Centralised packet scheduling

Recalling the taxonomy in figure 1, we now proceed with a survey of some recent topics in wireless access networks. The circuit switched technologies used in the first generation systems are well established, and resource allocation in OFDMA systems is still an emerging topic. Hence, we will focus on multimedia CDMA systems as these have recently seen a lot of research activity and are being employed for the third generation access networks. By a 'multimedia' access network we mean one that carries not only voice, but also other services such as interactive video, streaming video and also high speed Internet access.

5.1 Association and power control for guaranteed QoS calls in CDMA cellular systems

As explained in § 4.5, in CDMA access networks the link performance obtained by each mobile station (MS) is governed by the strength of its signal and the interference experienced by the MS's signal at the intended receiver. Hence it is important to associate MSs with base stations (BSs) in such a way that signal strengths of intended signals are high and interference from unintended signals is low. It is evident that increasing the transmit power to help one MS may not solve the overall problem, as this increase may cause intolerable interference at the intended receiver of another MS. It is also clear that in some situations there may be no *feasible power allocation* to each of the MSs such that all SINR requirements are met. Here we will discuss some important seminal results on the problem of optimal association and power allocation in CDMA systems.

There are m MSs and n BSs, with $\mathcal{B} = \{1, 2, 3, \dots, n\}$ denoting the set of BSs; see figure 2. Let $h_{i,j}$, $1 \leq i \leq m$, $1 \leq j \leq n$, denote the power 'gains' (attenuations) from MS i to BS j . Let $A = (a_1, a_2, \dots, a_m)$, $a_i \in \mathcal{B}$, denote an *association* of MSs with the BSs; thus, in the association A , MS i is associated with BS a_i . Let p_i be the average transmit signal power used by MS i , $1 \leq i \leq m$. For the most part of the following discussion, we will assume that

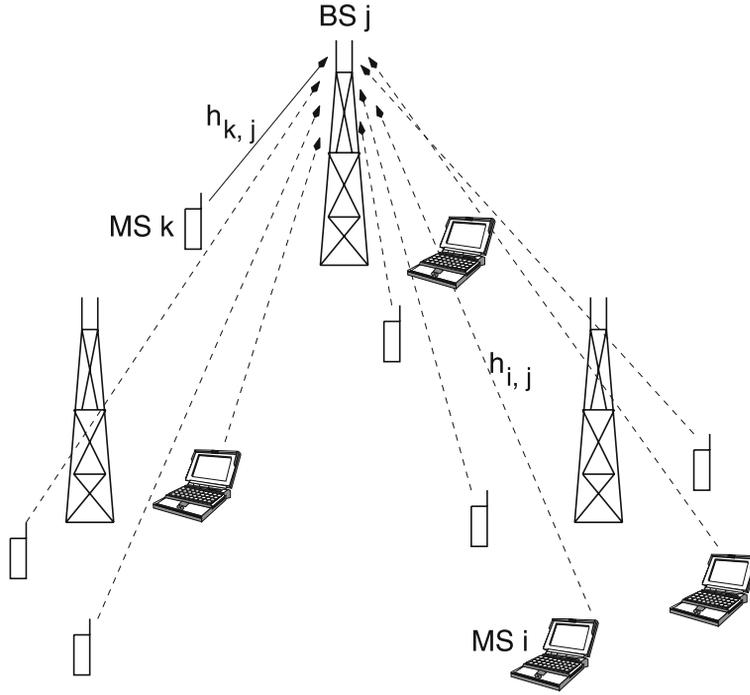


Figure 2. A depiction of the power control and association problem of several MSs in the vicinity of several BSs. The traffic model is that MS k has a connection (say, voice or video) requiring an uplink rate R_k .

the power gains and the association are fixed. With these definitions we can write the uplink received signal power to interference plus noise ratio for MS k as

$$(\text{SINR})_k = \frac{h_{k,a_k} p_k}{\sum_{\{i:1 \leq i \leq m, i \neq k\}} h_{i,a_k} p_i + N_0 W}$$

where N_0 is the one-sided power spectral density of the additive noise, and W is the radio spectrum bandwidth. Assume that the interference plus noise is well modelled by a white Gaussian noise process. In order to ensure a target bit-error-rate (BER) (which, as discussed above in § 4, is governed by the required QoS for the application being carried), we need to lower bound SINR_k multiplied by the processing gain $\frac{W}{R_k}$ (see Section 4), where the bit rate of MS k is R_k . We recall from Section 4 that, for traffic such as voice or video, the value of R_k will be determined by the coding rate of the source, and the desired quality of service, such as mean delay or some stochastic delay bound. If the desired lower bound is γ_k , then we obtain, for MS k ,

$$\frac{h_{k,a_k} p_k}{\sum_{\{i:1 \leq i \leq m, i \neq k\}} h_{i,a_k} p_i + N_0 W} \geq \Gamma_k$$

where $\Gamma_k := \gamma_k \frac{R_k}{W}$. For a given association and given channel gains, we thus obtain m linear inequalities in the m uplink powers of the m users. Before we proceed, let us write these

inequalities in matrix form as follows

$$\mathbf{p} \geq \mathbf{F}\mathbf{p} + \mathbf{g}$$

where \mathbf{p} is the (column) vector of powers, \mathbf{F} is an $m \times m$ matrix with a zero diagonal, $F_{k,i} = \frac{\Gamma_k h_{i,a_k}}{h_{k,a_k}}$, and $g_k = \frac{N_0 W \Gamma_k}{h_{k,a_k}}$, $1 \leq k \leq m$. Notice that g_k is the uplink power required at MS k if there was no interference from other users. For fixed \mathbf{F} and \mathbf{g} , we need to know if the power allocation problem is *feasible*, i.e. if the set

$$\{\mathbf{p} : \mathbf{p} \geq \mathbf{F}\mathbf{p} + \mathbf{g}\}$$

is nonempty. The answer to this question has been shown to depend on the matrix \mathbf{F} and hence on the association and the channel gains (Hanly 1995; Yates 1995; Bambos 1998). The matrix \mathbf{F} is non-negative, and it can also be shown that it is primitive for $m > 2$ (i.e. there exists k such that $\mathbf{F}^k > \mathbf{0}$, where $\mathbf{0}$ is the all-zero matrix). Then \mathbf{F} has an eigenvalue ρ that is real, simple, positive and greater than the magnitude of any other eigenvalue. If this, so-called Perron–Frobenius eigenvalue, ρ , is strictly less than 1 ($\rho < 1$), then it can be shown that $(\mathbf{I} - \mathbf{F})$ is a non-singular matrix, and hence there exists \mathbf{p}^* such that

$$\mathbf{p}^* = (\mathbf{I} - \mathbf{F})^{-1} \mathbf{g}$$

Thus $\mathbf{p}^* \in \{\mathbf{p} : \mathbf{p} \geq \mathbf{F}\mathbf{p} + \mathbf{g}\}$, and the power allocation problem for the given association and channel gains is feasible. Further, the following results can be proved about \mathbf{p}^* (Yates 1995):

- (1) \mathbf{p}^* is a Pareto power allocation, i.e. there does not exist any feasible power allocation \mathbf{p} that is strictly smaller than \mathbf{p}^* for every MS. In other words \mathbf{p}^* is on the ‘lower left boundary’ of the set of feasible power vectors, and hence cannot be strictly reduced for all MSs.
- (2) \mathbf{p}^* is the unique Pareto power allocation. Thus, \mathbf{p}^* is called the Pareto optimum power allocation.

It can, in fact, be shown that $\rho < 1$ is a necessary condition as well, and hence this can be used as an admission control criterion. Suppose that the associations are fixed, there is a large number of MSs associated with each BS, and there is a homogeneous distribution of MSs over the BSs. Then it is reasonable to model the uplink other cell interference in a cell as an η times the intracell interference. It then turns out that the admission control rule *into a cell* becomes: admit calls with requirements $\Gamma_1, \Gamma_2, \dots, \Gamma_m$ into a cell, provided that (Evans & Everitt 1999a) and (Evans & Everitt 1999b)

$$\sum_{k=1}^m \frac{\Gamma_k}{1 + \Gamma_k} < \frac{1}{1 + \eta} \quad (11)$$

Notice that the channel gains do not appear in this admission control rule. If the Inequality 11 is strictly met at all times, we have ‘hard’ admission control. In practice, calls are not active at all times even after being admitted. For example, a voice source is quiet during listening periods. For the SINR requirements to be met, the inequality needs to hold only for calls that are active. Thus, ‘soft’ admission control admits so many calls such that, given their active-inactive statistics, the probability of Inequality 11 not being met (i.e. the probability of *outage*) is small.

In order to minimise battery drain, the MSs should operate with the power vector \mathbf{p}^* . The following is an iterative algorithm that converges to \mathbf{p}^* starting from an initial feasible power allocation $\mathbf{p}^{(0)}$. Since $\mathbf{p}^{(0)}$ is feasible, we have

$$\mathbf{p}^{(0)} \geq \mathbf{F}\mathbf{p}^{(0)} + \mathbf{g}$$

Define, for $i \geq 1$,

$$\mathbf{p}^{(i)} = \mathbf{F}\mathbf{p}^{(i-1)} + \mathbf{g} \quad (12)$$

Hence, if $\mathbf{p}^{(i-1)}$ is feasible, it follows that

$$\mathbf{p}^{(i)} \leq \mathbf{p}^{(i-1)}$$

i.e. $\mathbf{p}^{(i)}$ is a non-negative and non-increasing sequence, and hence converges. Since (by non-negativity of terms) $\mathbf{F}\mathbf{p}^{(i-1)} + \mathbf{g} \geq \mathbf{F}\mathbf{p}^{(i)} + \mathbf{g}$ it follows that $\mathbf{p}^{(i)}$ is also feasible. Further, taking limits as $i \rightarrow \infty$ in both sides of Equation 12 it follows that the iterations converge to \mathbf{p}^* . There remains the question of how the computation on the right side of Equation 12 is carried out without knowledge of \mathbf{F} and \mathbf{g} . It is easily seen from the derivation that led to this matrix expression that this computation can be done if, with the power vector set at $\mathbf{p}^{(i-1)}$, the total interference plus noise experienced by the signal from each MS can be measured, as also the channel attenuation h_{k,a_k} for each k . Note that this yields a distributed synchronous algorithm. An asynchronous version has also been studied (Yates 1995).

The above discussion showed how the optimal power allocation can be achieved for a given association. However, there are several alternative associations and the complete problem is to find the association that yields the smallest power allocation vector. This problem is addressed in (Hanly 1995). Let \mathcal{A} denote the set of all feasible associations, i.e. those for which the set of feasible power allocations is non-empty. Let $\mathbf{p}^*(A)$ denote the optimal power allocation for the association A . It can be shown that among all the feasible associations $A \in \mathcal{A}$, there is an association A^* such that $\mathbf{p}^*(A^*) \leq \mathbf{p}^*(A)$. In (Hanly 1995), provides an iterative distributed algorithm for achieving this optimal association and the corresponding optimal power allocation.

The above discussion was with regard to uplink power allocation. Let us now turn to a discussion of some issues in downlink power allocation. The approach is to allocate to each BS a certain amount of average power budget which the BS then allocates to the MSs that associate with it. Since transmissions from a BS to all of its MSs are chip synchronous, intracell interference is less of a problem. However, multipath propagation does result in some intracell interference. This is because, even though the BS transmits the user symbols synchronously, multipath propagation causes multiple phase shifted copies of the transmitted signal to arrive at the receiver. Thus, since cross-correlations of the spreading sequences are not perfectly zero, some residual intracell interference is obtained at the correlation receiver. Turning to intercell interference, since there is universal spectrum re-use, the power radiated by all other BSs potentially interferes with the transmission from a BS to one of its MSs. However, now there are a few large interferers rather than several small ones, and hence the interference levels can be more variable than in the uplink, and also the white Gaussian interference assumption is less valid. Yet, the analysis of downlink power allocation is usually done using the same modelling approach as discussed above for the uplink.

5.2 High speed internet access in WCDMA

Wireless Internet access typically involves *elastic transfers* from servers on the Internet to mobile clients. The term elastic transfers refers to applications that simply involve the transfer

of a file from one computer to another, and includes email, web browsing, and file transfers. The reason why such transfers are called ‘elastic’ is that they do not need any fixed intrinsic transfer rate (as would a real time voice call); the source of the transfer can adjust its sending rate to adapt to the available rate in the network. Evidently, this requires some sort of explicit or implicit feedback control between the sources and the network (Kumar *et al* 2004).

In the previous section we reviewed the optimal association and power control problem for calls that require a guaranteed bit rate and bit error rate, such as a voice call. Cellular networks were originally designed for mobile telephony, and have been primarily used for this since their inception over two decades ago. With the rapid developments in digital communication over fading wireless channels, the most eagerly awaited service is ubiquitous wireless access to the Internet. Hence considerable attention is being paid to high speed wireless Internet access in the next generation cellular systems. In this section we will survey the problem of power control and scheduling in high speed down-link elastic data transfers in a WCDMA system (Bonald *et al* 2004) and (Bonald & Proutiere 2003).

The down-link power allocation problem was discussed briefly in § 5.1. Each BS is assigned a certain total average power, P_d , which it allocates among the ongoing downlink transmissions. We use the notation defined in § 5.1. Further define \mathcal{S} to be the set of all MSs, and, for $1 \leq j \leq n$, let \mathcal{S}_j denote the set of MSs associated with BS j . The sets \mathcal{S}_j , $1 \leq j \leq n$, constitute a partition of \mathcal{S} , and such a partition is equivalent to an association A . For $i \in \mathcal{S}_j$, let p_i be the power assigned by BS j to MS i . Thus

$$\sum_{i \in \mathcal{S}_j} p_i \leq P_d$$

Now, ignoring the intracell interference (see the discussion in the previous section), the down-link received signal power to interference plus noise power ratio is given by

$$(SINR)_i = \frac{h_{i,a_i} p_i}{\sum_{j:1 \leq j \leq n, j \neq a_i} h_{i,j} P_d + N_0 W}$$

where the first term in the denominator is the total interference power received at MS i from the other BSs, assuming that they are all transmitting at their maximum down-link power P_d . For a given association, define, for $1 \leq i \leq m$,

$$\eta_i := \frac{\sum_{j:1 \leq j \leq n, j \neq a_i} h_{i,j} P_d + N_0 W}{h_{i,a_i}}$$

Defining, as in the previous section, the SINR target $\Gamma_i = \gamma_i \frac{R_i}{W}$, for user i , the power allocation needs to satisfy the following inequality

$$\Gamma_i \eta_i \leq p_i$$

Γ_i can be viewed as summarising the performance that is being provided to user i ; note that $\frac{E_b}{N_0}$ relates to bit error rate, which in turn relates to packet error probability, which in turn affects the performance of TCP controlled transfers (Kumar *et al* 2004).

For a given association, we now need to obtain the power allocation that is optimal in some sense. Allocating all the down-link power from a BS to the best user in that cell will maximise the overall throughput carried by the network but will make a lot of users unhappy. One approach is to evaluate the *utility* obtained by an MS when a certain rate is allocated to

it, and then optimise the total network utility. The utility function can be chosen to capture the desired trade-off between network throughput and fairness between users for example, (Kumar *et al* 2004).

Let $U(\cdot)$ be the utility function, so that the utility to user i is evaluated as $U(\Gamma_i)$. Let us fix an association and ask for a power allocation in each cell so that the constraints $\Gamma_i \eta_i \leq p_i$ are met for the users, and the network utility is maximised. This leads to the following optimisation problem.

Problem P1

$$\max \sum_{j=1}^n \sum_{i \in S_j} U(\Gamma_i)$$

subject to

$$\Gamma_i \eta_i \leq p_i$$

$$\sum_{i \in S_j} p_i \leq P_d \quad \text{for } 1 \leq j \leq n$$

$$\Gamma_i \geq 0, p_i \geq 0, \quad \text{for } 1 \leq j \leq n$$

Note that in Problem P1, in each cell, some power may be allocated to every user. In order to avoid the problem of intracell interference (which we have ignored in the above formulation) an alternative is to allocate the entire power in each cell (i.e. P_d) to a user at a time, and obtain a power allocation over the users by *time sharing*. Let ϕ_i be the fraction of time power is allocated to MS i , by the BS a_j ; then, of course, $\sum_{i \in S_j} \phi_i = 1$, and we obtain the following optimisation problem.

Problem P2

$$\max \sum_{j=1}^n \sum_{i \in S_j} U(\phi_i \Gamma_i)$$

subject to

$$\sum_{i \in S_j} \phi_i = 1 \quad \text{for } 1 \leq j \leq n$$

$$\Gamma_i \eta_i \leq P_d \quad \text{for } 1 \leq i \leq m$$

$$\Gamma_i \geq 0, p_i \geq 0, \quad \text{for } 1 \leq j \leq n$$

For the utility function $U(\cdot) = \log(\cdot)$, it can be shown that the two problems above yield the same solution $\Gamma_i = \frac{P_d}{m_j \eta_j}$, where m_j is the number of MSs associated with BS j , i.e. the number of elements in the set S_j . This solution can be implemented in the framework of Problem P1 by always allocating to an STA $i \in S_j$ the power $p_i = \frac{P_d}{m_j}$, or, in the framework of Problem P2, by allocating power P_d to each station in cell j a fraction $\frac{1}{m_j}$ of the time.

The latter solution also avoids intracell interference and is the one that is preferred in practice. We note that the solution obtained is such that each user obtains a time average performance that is proportional to the performance it can get if it was allocated all the resources. Thus, this solution is also called *proportionally fair*.

Let us now examine this time sharing solution and obtain the mean file transfer delay under a certain traffic model (Bonald & Proutiere 2003). Assuming that the same value of γ_i is required for all users, when a user is being served (and is therefore allocated the full down-link power P_d in its cell), the user receives a down-link physical bit rate of $R_i = \frac{WP_d}{\gamma_i \eta_i}$. If the γ_i is appropriately chosen then the TCP packet loss probability will be small and the TCP throughput will be close to R_i (Kumar *et al* 2004); let us assume this to be the case. Now since each user served by BS j is assigned the full power an equal fraction of the time, it is clear that the average down-link transfer throughputs obtained by the users in cell j are *proportional* to their peak throughputs R_i , $i \in S_j$. Thus we say that the allocation achieves *proportional fairness*.

Let us now assume that the k th request has transfer volume V_k , and is from a user whose peak transfer rate (because of its location) is R_k . Define $T_k = \frac{V_k}{R_k}$, the time taken to complete the k th transfer if the full power was allocated to it without any interruption. However, the BS time-shares equally among all the ongoing transfers. In practice there is a positive time-slice, but in the limit as the time-slice goes to 0, we have a standard processor sharing (PS) queuing model (Wolff 1989). If the transfer requests arrive in a Poisson process of rate λ , and the random sequence T_k , $k \geq 1$, is i.i.d. with some general distribution with finite mean, then we have an M/G/1 PS model for the number of ongoing transfers. This process is stable when $\lambda E(T) < 1$, i.e. the average amount of transfer time brought in per second by the users is less than 1 second. Defining $\rho = \lambda E(T)$, the mean transfer delay is then easily seen to be

$$\frac{E(T)}{1 - \rho}$$

where we have used the the fact that the distribution of the number of customers in an M/G/1 PS queue depends on the distribution of the service time only through its mean, and hence is the same as that of the M/M/1 PS queue (Wolff 1989). From the point of view of traffic engineering, note that we will need to limit how close ρ is to 1 in order to provide some mean transfer delay guarantee to the users.

6. Summary

In the first part of the paper we have first provided a taxonomy of the various wireless network systems. We then reviewed wireless physical layer techniques, and surveyed a variety of resource allocation and network design problems in wireless networks with a centralised resource allocation mechanism like those in CDMA and OFDMA networks. In the next part we will consider random access networks, ad hoc multi-hop networks and also provide a brief overview of wireless sensor networks.

References

- Akyildiz I F, Su W, Sankarasubramaniam Y, Cayirci E 2002 Wireless sensor networks: A survey. *Computer Networks*, 38: 393–422

- Bambos Nicholas 1998 Toward power-sensitive network architectures in wireless communications: Concepts, issues, and design approaches. *IEEE Per. Comm.* June, 50–59
- Bertsekas D, Gallager R 1992 *Data Networks*. Second edn. PHI
- Bonald Thomas, Proutiere A 2003 Wireless downlink data channels: User performance and cell dimensioning. In *Proc. ACM Mobicom*
- Bonald Thomas, Borst Sem, Proutiere A 2004 Intercell scheduling in wireless data networks. Tech. rept. ISSN 1386–3711, CWI, Amsterdam
- Doherty L, Pister K, Ghaoui L El 2001 Convex position estimation in wireless sensor networks. In *Proc. IEEE Infocom*
- Evans Jamie S, Everitt, David 1999a Effective bandwidth-based admission control for multiservice CDMA cellular networks. *IEEE Trans. on Vehicular Technol.* 48(1): 36–46
- Evans Jamie S, Everitt, David 1999b On the teletraffic capacity of CDMA cellular networks. *IEEE Trans. on Vehicular Technol.* 48: 153–165
- Giridhar Arvind, Kumar P R 2006 Towards a theory of in-network computation in wireless sensor networks. *IEEE Commun. Mag.* 44(4): 98–107
- Goldsmith Andrea 2005 *Wireless Communications* (New York: Cambridge University Press)
- Hanly Stephen V 1995 An algorithm for combined cell-site selection and power control to maximise cellular spread spectrum capacity. *IEEE J. on Selected Areas in Commun.* 13(7): 1332–1340
- Karnik Aditya, Kumar Anurag 2005 Performance of TCP congestion control with explicit rate feedback. *IEEE/ACM Trans. on Networking* 13(1): 108–120
- Khude Nilesh, Kumar Anurag, Karnik Aditya 2005 Time and energy complexity of distributed computation in wireless sensor networks. In *Proc. IEEE Infocom. IEEE.*
- Kumar Anurag, Manjunath D, Kuri Joy 2004 *Communication networking: An analytical approach*. San Francisco: Morgan-Kaufmann (an imprint of Elsevier)
- Kumar A, Kumar V 2005 Optimal association of stations and APs in an IEEE 802.11 WLAN. In *Proc. of the Nat. Conf. on Comm. (NCC)*
- Perkins Charles E 2001 *Ad Hoc Networking* (NY: Addison Wesley)
- Shakkottai Srinivas, Altman Eitan, Kumar Anurag 2006 The case for non-cooperative multihoming of users to access point in IEEE 802.11 WLANs. In *Proc. IEEE Infocom.*
- Tse David, Viswanath Pramod 2005 *Fundamentals of wireless communication* (New York: Cambridge University Press)
- Wolff Ronald W 1989 *Stochastic modelling and the theory of Queues* (New Jersey: Prentice Hall)
- Yates Roy D 1995 A framework for uplink power control in cellular radio systems. *IEEE J. Selected Areas in Comm.* 13(7): 1341–1347