Escola d'Enginyeria de Telecomunicació i Aeroespacial de Castelldefels
UNIVERSITAT POLITÈCNICA DE CATALUNYA

TREBALL DE FI DE CARRERA

TÍTOL DEL TFC: Wireless Sensor Networks

TITULACIÓ: Enginyeria Tècnica de Telecomunicació, especialitat Telemàtica

AUTOR: Marta Jurado Vilar

DIRECTOR: Ari Rantala

DATA: 3 de juny de 2011

Títol: Wireless Sensor Networks

Autor: Marta Jurado Vilar

Director: Ari Rantala

Data: 3 de juny de 2011

Resum

Les xarxes de sensors sense fils prometen una interfície segura entre el món virtual i el físic. Són una de les més ràpides evolucions en noves tecnologies de la informació, amb aplicacions en una àmplia gamma de camps, incloent el control de processos industrials, de seguretat i vigilància, sensors ambientals, i la seguiment de la salut estructural. El tema d'aquest projecte està motivat per la necessitat urgent de proporcionar una visió integral i organitzada del camp. Es mostra com els principals reptes de l'eficiència energètica, la robustesa i l'autonomia s'aborden en aquests sistemes mitjançant tècniques de xarxa a través de múltiples capes. Els temes coberts inclouen el desplegament de xarxa, característiques del wireless, sincronització de temps, la congestió i control d'errors, d'accés al medi, normes, control de topologia, seguretat, enrutament, la transferència de dades, protocols de transport i les noves tecnologies i materials en la fabricació de sensors.

Title: Wireless Sensor Networks

Author: Marta Jurado Vilar

Director: Ari Rantala

Date: June, 3rd 2011

Overview

Wireless sensor networks promise an unprecedented fine-grained interface between the virtual and the physical world. They are one of the most rapidly developing new information technologies, with applications in a wide range of fields including industrial process control, security and surveillance, environmental sensing, and structural health monitoring.
The subject of this project is motivated by the urgent need to provide a comprehensive and organized survey of the field. It shows how the core challenges of energy efficiency, robustness, and autonomy are addressed in these systems by networking techniques across multiple layers.
The topics covered include network deployment, wireless characteristics, time synchronization, congestion and error control, medium access, standards, topology control, routing, security, data transfer, transport protocols and new technologies and materials in fabricating sensors.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1. INTRODUCTION

1.1. WIRELESS SENSOR NETWORKS

Wireless Sensor Networks is a fast growing and exciting research area that has attracted considerable attention in the recent past. This area has been fueled by the recent tremendous technological advances in the development of low-cost sensor devices equipped with wireless network interfaces. The creation of large-scale sensor networks interconnecting several hundred to a few thousand sensor nodes opens up many technical challenges and immense application possibilities. Wireless sensor networks have moved from the research domain into the real world with the commercial availability of sensors with networking capabilities.

Recent technological advances allow us to envision a future where large number of low-power, inexpensive sensor devices are densely embedded in the physical environment, operating together in a wireless network. The envisioned applications of these wireless sensor networks range widely: ecological habitat monitoring, structure health monitoring, environmental contaminant detection, industrial process control, and military target tracking, among others. So we can say sensor networks find applications spanning several domains including military, medical, industrial, and home networks.

Wireless sensor networks provide bridges between the virtual world of information technology and the real physical world. They represent a fundamental paradigm shift from traditional inter-human personal communications to autonomous inter-device communications. They promise unprecedented new abilities to observe and understand large-scale, real-world phenomena at a fine spatio-temporal resolution. As a result, wireless sensor networks also have the potential to engender new breakthrough scientific advances.

While the notion of networking distributed sensors and their use in military and industrial applications dates back at least to the 1970s, the early systems were primarily wired and small in scale. It was only in the 1990s that researchers began envisioning and investigating large-scale embedded wireless sensor networks for dense sensing applications.

Maybe one of the earliest research efforts in this direction was the **low-power wireless integrated microsensors** (LWIM) project at UCLA funded by DARPA (1). The LWIM project focused on developing devices with low-power electronics in order to enable large, dense wireless sensor networks. This project was succeeded by the Wireless Integrated Networked Sensors (WINS) project a few years later, in which researchers at UCLA collaborated with Rockwell Science Center to develop some of the first wireless sensor devices. Other early projects in this area, starting around 1999-2000, were also primarily in academia, at several places including MIT, Berkeley, and USC.

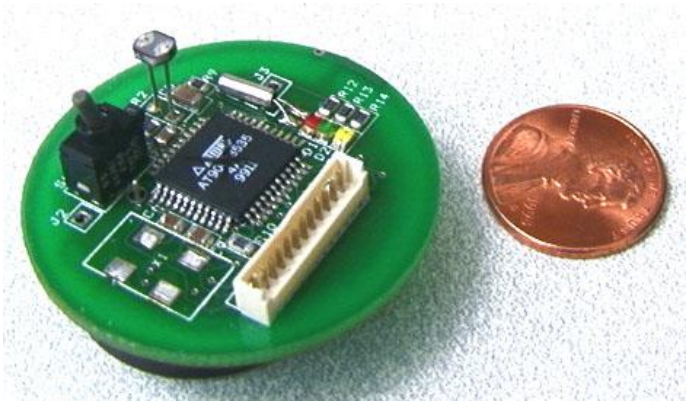

Figure 1-1 A Berkeley mote

Researchers at Berkeley developed embedded wireless sensor networking devices called **motes**, which were made publicly available commercially, along with TinyOS, and associated embedded operating system that facilitates the use of these devices (2). Figure 1-1 shows an image of a Berkeley mote device. The availability of these devices as an easily programmable, fully functional, relatively inexpensive platform for experimentation, and real deployment has played a significant role in the ongoing wireless sensor networks revolution.

1.2. HOW MOTES WORK

1.2.1. THE BASIC IDEA

The **mote concept** creates a new way of thinking about computers, but the basic idea is pretty simple:

- The core of a mote is a small, low-cost, low-power computer.
- The computer monitors one or more sensors. It is easy to imagine all sorts of sensors, including sensors for temperature, light, sound, position, acceleration, vibration, stress, weight, pressure, humidity, etc. Not all mote applications require sensors, but sensing applications are very common.
- The computer connects to the outside world with a radio link. The most common radio links allow a mote to transmit at a distance from 3 to 61 meters. Power consumption, size and cost are the barriers to longer distances. Since a fundamental concept with motes is tiny size (and associated tiny cost), small and low-power radios are normal.

Motes can either run off of batteries, or they can tap into the power grid in certain applications. As motes shrink in size and power consumption, it is possible to imagine solar power or even vibration power to keep them running.

All of these parts are packaged together in the smallest container possible. In the future, people imagine shrinking motes to fit into something just a few millimeters on a side. It is more common for motes today, including batteries and antenna, to be the size of a stack of five or six quarters, or the size of a pack of cigarettes. The battery is usually the biggest part of the

package right now. Figure 1-2 shows the **MICA2DOT mote**, typically powered by a circular “button” battery, which is not much bigger than a quarter.

It is hard to imagine something as small and innocuous as a mote sparking a revolution, but that's exactly what they have done.

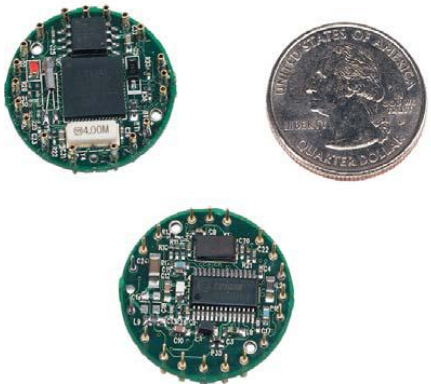

Figure 1-2 The MICA2DOT mote

1.2.2. TYPICAL APPLICATIONS

If we survey the literature for different ways that people have thought of to use motes, we find a huge assortment of ideas. It is possible to think of motes as lone sensors. For example:

- We could embed motes in bridges when we pour the concrete. The mote could have a sensor on it that can detect the salt concentration within the concrete. Then once a month we could drive a truck over the bridge that sends a powerful magnetic field into the bridge. The magnetic field would allow the motes, which are buried within the concrete of the bridge, to power on and transmit the salt concentration. Salt (perhaps from deicing or ocean spray) weakens concrete and corrodes the steel rebar that strengthens the concrete. Salt sensors would let bridge maintenance personnel gauge how much damage salt is doing. Other possible sensors embedded into the concrete of a bridge might detect vibration, stress, temperature swings, cracking, etc., all of which would help maintenance personnel spot problems long before they become critical.
- We could connect sensors to a mote that can monitor the condition of machinery – temperature, number of revolutions, oil level, etc. and log it in the mote's memory. Then, when a truck drives by, the mote could transmit all the logged data. This would allow detailed maintenance records to be kept on machinery (for example, in an oil field), without maintenance personnel having to go measure all of those parameters themselves.
- We could attach motes to the water meters or power meters in a neighborhood. The motes would log power and water consumption for a customer. When a truck drives by, the motes get a signal from the truck and they send their data. This would allow a person to read all the meters in a neighborhood very easily, simply by driving down the street.

All of these ideas are good; some allow sensors to move into places where they have not been before (such as embedded in concrete) and others reduce the time needed to read sensors individually.

However, much of the greatest excitement about motes comes from the idea of using large numbers of motes that communicate with each other and form ad hoc networks.

1.2.3. AD HOC NETWORKS

The **Defense Advanced Research Projects Agency** (DARPA) was among the original patrons of the mote idea. One of the initial mote ideas implemented for DARPA allows motes to sense battlefield conditions.

For example, imagine that a commander wants to be able to detect truck movement in a remote area. An airplane flies over the area and scatters thousands of motes, each one equipped with a magnetometer, a vibration sensor and a GPS receiver. The battery-operated motes are dropped at a density of one every 30 meters or so. Each mote wakes up, senses its position and then sends out a radio signal to find its neighbors.

All of the motes in the area create a giant, amorphous network that can collect data. Data funnels through the network and arrives at a collection node, which has a powerful radio able to transmit a signal many miles. When an enemy truck drives through the area, the motes that detect it transmit their location and their sensor readings. Neighboring motes pick up the transmissions and forward them to their neighbors and so on, until the signals arrive at the collection node and are transmitted to the commander. The commander can now display the data on a screen and see, in real time, the path that the truck is following through the field of motes. Then a remotely-piloted vehicle can fly over the truck, make sure it belongs to the enemy and drop a bomb to destroy it.

This concept of **ad hoc networks** – formed by hundreds or thousands of motes that communicate with each other and pass data along from one to another – is extremely powerful.

1.2.4. A TYPICAL MOTE

MICA mote is a commercially available product that has been used widely by researchers and developers. It has all of the typical features of a mote and therefore can help us understand what this technology makes possible today. MICA motes are available to the general public through a company called Crossbow. These motes come in two form factors:

- Rectangular, measuring 5.7 x 3.18 x 0.64 cm, it is sized to fit on top of two AA batteries that provide it with power.
- Circular, measuring 2.5 x 0.64 cm, it is sized to fit on top of a 3 volt button cell battery.

Figure 1-3 A rectangular MICA mote

The MICA mote uses an **Atmel ATmega 128L** processor running at 4 MHz. The 128L is an 8-bit microcontroller that has 128 kB of onboard flash memory to store the mote's program. This CPU is about as powerful as the 8088 CPU found in the original IBM PC (circa 1982). The big difference is that the ATmega consumes only 8 mA when it is running, and only 15 µA in sleep mode.

This low power consumption allows a MICA mote to run for more than a year with two AA batteries. A typical AA battery can produce about 1,000 mA-hours. At 8 mA, the ATmega would operate for about 120 hours if it operated constantly. However, the programmer will typically write his/her code so that the CPU is asleep much of the time, allowing it to extend battery life considerably. For example, the mote might sleep for 10 seconds, wake up and check status for a few µA, and then go back to sleep.

MICA motes come with 512 kB of flash memory to hold data. They also have a 10-bit A/D converter so that sensor data can be digitized. Separate sensors on a daughter card can connect to the mote. Sensors available include temperature, acceleration, light, sound and magnetic. Advanced sensors for things like GPS signals are under development.

The final component of a MICA mote is the radio. It has a range of several hundred feet and can transmit approximately 40,000 bps. When it is off, the radio consumes less than one µA. When receiving data, it consumes 10 mA. When transmitting, it consumes 25 mA. Conserving radio power is key to long battery life.

1.2.5. THE FUTURE

A few years ago, researchers managed to cram all of the parts needed for a mote onto a single chip less than 3 millimeters on each side. The total size is about 5 square mm, meaning that you could fit more than a dozen of these chips onto a penny.

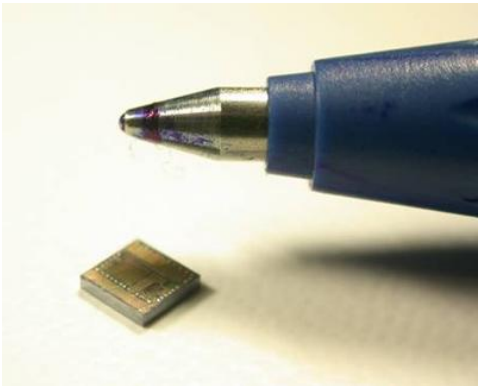

Figure 1-4 Mote pictured beside the tip of a ballpoint pen

The chip contains all of the components found in a mote: a CPU, memory, and A/D converter for reading sensor data and a radio transmitter. To complete the package we attach the sensor(s), a battery and an antenna. The cost of the chip will be less than a dollar when it is mass produced.

1.3. NETWORKED WIRELESS SENSOR DEVICES

As shown in Figure 1-5, there are several key components that make up a typical wireless sensor network (WSN) device:

1. **Low-power embedded processor**: the computational tasks on a WSN device include the processing of both locally sensed information as well as information communicated by other sensors. At present, primarily due to economic constraints, the embedded processors are often significantly constrained in terms of computational power (for example, many of the devices used currently in research and development have only an eight-bit 16-MHz processor). Due to the constraints of such processors, devices typically run specialized component-based embedded operating systems, such as TinyOS. However, it should be kept in mind that a sensor network may be heterogeneous and include at least some nodes with significantly greater computational power. Moreover, future WSN devices may possess extremely powerful design techniques, such as efficient sleep modes and dynamic voltage scaling to provide significant energy savings.

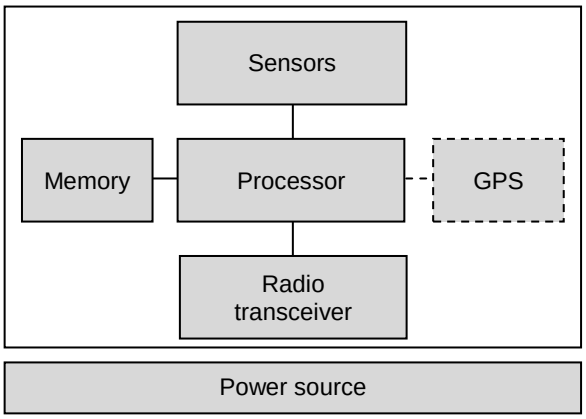

Figure 1-5 Schematic of a basic wireless sensor network device

2. **Memory/storage**: storage in the form of random access and read-only memory includes both program memory (from which instructions are executed by the processor), and data memory (for storing raw and processed sensor measurements and other local information). The quantities of memory and storage on board a WSN device are often limited primarily by economic considerations, and are also likely to improve over time.

3. **Radio transceiver**: WSN devices include a low-rate, short-range wireless radio (10-100 kbps, <100 m). While currently quite limited in capability too, these radios are likely to improve in sophistication over time – including improvements in cost, spectral efficiency, tunability, and immunity to noise, fading, and interference. Radio communication is often the most power-intensive operation in a WSN device, and hence the radio must incorporate energy-efficient sleep and wake-up modes.

4. **Sensors**: due to bandwidth and power constraints, WSN devices primarily support only low-data-rate sensing. Many applications call for multi-modal sensing, so each device may have several sensors on board. The specific sensors used are highly dependent on the application: for example, they may include temperature sensors, light sensors, humidity sensors, pressure sensors, accelerometers, magnetometers, chemical sensors, acoustic sensors, or even low-resolution imagers.

5. **Geopositioning system**: in many WSN applications, it is important for all sensor measurements to be location stamped. The simplest way to obtain positioning is to pre-configure sensor locations at deployment, but this may only be feasible in limited deployments. Particularly for outdoor operations, when the network is deployed in an ad hoc manner, such information is most easily obtained via satellite-based GPS. However, even in such applications, only a fraction of the nodes may be equipped with GPS capability, due to environmental and economic constraints. In this case, other nodes must obtain their locations indirectly through network localization algorithms.

6. **Power source**: for flexible deployment the WSN device is likely to be battery powered (for example, using LiMH AA batteries). While some of the nodes may be wired to a continuous power source in some applications, and energy harvesting techniques may provide a degree of energy renewal in some cases, the finite battery energy is likely to be the most critical resource bottleneck in most WSN applications.

Depending on the application, WSN devices can be networked together in a number of ways. In basic data-gathering applications, for instance, there is a node referred to as the sink to which all data from source sensor nodes are directed. The simplest logical topology for communication of gathered data is a single-hop star topology, where all nodes send their data directly to the sink. In networks with lower transmit power settings or where nodes are deployed over a large area, a multi-hop tree structure may be used for data-gathering. In this

case, some nodes may act both as sources themselves, as well as routers for other sources.

One interesting characteristic of wireless sensor networks is that they often allow the possibility of intelligent in-network processing. Intermediate nodes along the path do not act merely as packet forwarders, but may also examine and process the content of the packets going through them. This is often done for the purpose of data compression or for signal processing to improve the quality of the collected information.

1.4. APPLICATIONS OF WIRELESS SENSOR NETWORKS

The several envisioned applications of WSN are still very much under active research and development, in both academia and industry. We can describe a few applications from different domains briefly to give a sense of the wide-ranging scope of this field.

1.4.1. ECOLOGICAL HABITAT MONITORING

Scientific studies of ecological habitats (animals, plants, micro-organisms) are traditionally conducted through hands-on field activities by the investigators. One serious concern in these studies is what is sometimes referred to as the "**observer effect**" – the very presence and potentially intrusive activities of the field investigators may affect the behavior of the organisms in the monitored habitat and thus affect observed results. Unattended wireless sensor networks promise a cleaner remote-observer approach to habitat monitoring. Further, sensor networks, due to their potentially large scale and high spatio-temporal density, can provide experimental data of an unprecedented richness.

One of the earliest experimental deployments of wireless sensor networks was for habitat monitoring, on Great Duck Island, Maine (3). A team of researchers from the Intel Research Lab at Berkeley, University of California at Berkeley, and the College of the Atlantic in Bar Harbor deployed wireless sensor nodes in and around burrows of Leach's storm petrel, a bird which forms a large colony on that island during the breeding season. The sensor-network-transmitted data were made available over the web, via a base station on the island connected to a satellite communication link.

1.4.2. MILITARY SURVEILLANCE AND TARGET TRACKING

As with many other information technologies, wireless sensor networks originated primarily in military-related research. Unattended sensor networks are envisioned as the key ingredient in moving towards network-centric warfare systems. They can be rapidly deployed for surveillance and used to provide battlefield intelligence regarding the location, numbers, movement, and identity of troops and vehicles, and for detection of chemical, biological, and nuclear weapons.

Much of the impetus for the fast-growing research and development of wireless sensor networks has been provided though several programs funded

by DARPA, most notably through a program known as Sensor Information Technology (SensIT) (4) from 1999 to 2002. Indeed, many of the leading US researchers and entrepreneurs in the area of wireless sensor networks today have been and are being funded by these DARPA programs.

1.4.3. STRUCTURAL AND SEISMIC MONITORING

Another class of applications for sensor networks pertains to monitoring the condition of civil structures (5). The structures could be buildings, bridges, and roads; even aircraft. At present the health of such structures is monitored primarily through manual and visual inspections or occasionally through expensive and time-consuming technologies, such as X-rays and ultrasound. Unattended networked sensing techniques can automate the process, providing rich and timely information about incipient cracks or about other structural damage. Researchers envision deploying these sensors densely on the structure – either literally embedded into the building material such as concrete, or on the surface. Such sensor networks have potential for monitoring the long-term wear of structures as well as their condition after destructive events, such as earthquakes or explosions. A particularly compelling futuristic vision for the use of sensor networks involves the development of controllable structures, which contain actuators that react to real-time sensor information to perform "echo-cancellation" on seismic waves so that the structure is unaffected by any external disturbance.

1.4.4. INDUSTRAL AND COMMERCIAL NETWORKED SENSING

In industrial manufacturing facilities, sensors and actuators are used for process monitoring and control. For example, in a multi-stage chemical processing plant there may be sensors placed at different points in the process in order to monitor the temperature, chemical concentration, pressure, etc. The information from such real-time monitoring may be used to vary process control, such as adjusting the amount of a particular ingredient or changing the heat settings. The key advantage of creating wireless networks of sensors in these environments is that they can significantly improve both the cost and the flexibility associated with installing, maintaining, and upgrading wired systems (6). As an indication of the commercial promise or wireless embedded networks, it should be noted that there are already several companies developing and marketing these products, and there is a clear ongoing drive to develop related technology standards, such as the IEEE 802.15.4 standard (7), and collaborative industry efforts such as the Zigbee Alliance[1].

[1] The Zigbee Alliance, http://www.zigbee.org

[2] The terminology of upstream/downstream can sometimes be confusing. Consider the one-way data-gathering, where the sink is the root to which all flows are directed. A node i is the parent of node j if it is the next hop from j towards the sink. We then refer to node j as being *below* i on

1.5. DESIGN CHALLENGES

Wireless sensor networks are interesting from an engineering perspective, because the present a number of serious challenges that cannot be adequately addressed by existing technologies:

1. **Extended lifetime**: as mentioned above, WSN nodes will generally be severely energy constrained due to the limitations of batteries. A typical alkaline battery, for example, provides about 50 watt-hours of energy; this may translate to less than a month of continuous operation for each node in full active mode. Given the expense and potential infeasibility of monitoring and replacing batteries for a large network, much longer lifetimes are desired. In practice, it will be necessary in many applications to provide guarantees that a network of unattended wireless sensors can remain operational without any replacements for several years. Hardware improvements in battery design and energy harvesting techniques will offer only partial solutions. This is the reason that most protocol designs in wireless sensor networks are designed explicitly with energy efficiency as the primary goal. Naturally, this goal must be balanced against a number of other concerns.

2. **Responsiveness**: a simple solution to extending network lifetime is to operate the nodes in a duty-cycled manner with periodic switching between sleep and wake-up modes. While synchronization of such sleep schedules is challenging in itself, a larger concern is that arbitrarily long sleep periods can reduce the responsiveness and effectiveness of the sensors. In applications where it is critical that certain events in the environmental be detected and reported rapidly, the latency induced by sleep schedules must be kept within strict bounds, even in the presence of network congestion.

3. **Robustness**: the vision of wireless sensor networks is to provide large-scale, yet fine-grained coverage. This motivates the use of large numbers of inexpensive devices. However, inexpensive devices failure will also be high whenever the sensor devices are deployed in harsh or hostile environments. Protocol designs must therefore have built-n mechanisms to provide robustness. It is important to ensure that the global performance of the system is not sensitive to individual device failures. Further, it is often desirable that the performance of the system degrade as gracefully as possible with respect to component failures.

4. **Synergy**: Moore’s law-type advances in technology have ensured that device capabilities in terms of processing power, memory, storage, radio transceiver performance, and even accuracy of sensing improve rapidly (given a fixed cost). However, if economic considerations dictate that the cost per node is reduced drastically from hundreds of dollars to less than a few cents, it is possible that the capabilities of individual nodes will remain constrained to some extent. The challenge is therefore to design synergistic protocols, which ensure that the system as a whole is more capable than the sum of the capabilities of its individual components. The protocols must

provide an efficient collaborative use of storage, computation, and communication resources.

5. **Scalability**: for many envisioned applications, the combination of fine-granularity sensing and large coverage area implies that wireless sensor networks have the potential to be extremely large scale (tens of thousands, perhaps even millions of nodes in the long term). Protocols will have to be inherently distributed, involving localized communication, and sensor networks must utilize hierarchical architectures in order to provide such scalability. However, visions of large numbers of nodes will remain unrealized in practice until some fundamental problems, such as failure handling and *in-situ* reprogramming, are addressed even in small settings involving tens to hundreds of nodes. There are also some fundamental limits on the throughput and capacity that impact the scalability of network performance.

6. **Heterogeneity**: there will be heterogeneity of device capabilities (with respect to computation, communication, and sensing) in realistic settings. This heterogeneity can have a number of important design consequences. For instance, the presence of a small number of devices of higher computational capability along with a large number of low-capability devices can dictate a two-tier, cluster-based network architecture, and the presence of multiple sensing modalities requires pertinent sensor fusion techniques. A key challenge is often to determine the right combination of heterogeneous device capabilities for a given application.

7. **Self-configuration**: because of their scale and the nature of their applications, wireless sensor networks are inherently *unattended* distributed systems. Autonomous operation of the network is therefore a key design challenge. From the very start, nodes in a wireless sensor network have to be able to configure their own network topology; localize, synchronize, and calibrate themselves; coordinate inter-node communication; and determine other important operating parameters.

8. **Self-optimization and adaption**: traditionally, most engineering systems are optimized *a priori* to operate efficiently in the face of expected or well-modeled operating conditions. In wireless sensor networks, there may often be significant uncertainty about operating conditions prior to deployment. Under such conditions, it is important that there be in-built mechanisms to autonomously learn from sensor and network measurements collected over time and to use this learning to continually improve performance. Also, besides being uncertain *a priori*, the environment in which the sensor network operates can change drastically over time. WSN protocols should also be able to adapt to such environmental dynamics in an online manner.

9. **Systematic design**: as we shall see, wireless sensor networks can often be highly application specific. There is a challenging tradeoff between *ad hoc*, narrowly applicable approaches that exploit application-specific characteristics to offer performance gains and more flexible, easy-to-generalize design methodologies that sacrifice some performance. While

performance optimization is very important, given the severe resource constraints in wireless sensor networks, systematic design methodologies, allowing for reuse, modularity, and run-time adaptation, are necessitated by practical considerations.

10. **Privacy and security**: the large scale, prevalence, and sensitivity of the information collected by wireless sensor networks (as well as their potential deployment in hostile locations) give rise to the final key challenge of ensuring both privacy and security.

CHAPTER 2. DATA TRANSFER AND TOPOLOGIES

Networking of hardware and software resources is essential to bring multiple devices together. Networking introduces efficiency by enabling the exchange of information, creating collaborative operations, and sharing the functions of equipment and devices. Networks are a collection of interoperational devices linked together by a communication medium and supported by suitable software. The software may be responsible for the functionality of part of the system or the entire system. A **system** is a group of interrelated parts with the focus on establishing an interrelationship between them to ensure efficiency, to facilitate integration of the application, and to share the resources.

Connecting devices together to form networks is a concept that has been used for many decades in a wide range of applications. In earlier networks, almost all the communicating devices were connected by wires, thus they were largely fixed in space. The devices in modern networks, as discussed in this thesis, can be interconnected with the use of wireless communication technology. Applications of wireless technology create mobility in space while still maintaining the network. Therefore modern networks can be viewed as wired networks in which the communication devices are connected by wires and are largely fixed in space; wireless networks in which devices communicate wirelessly and can move in space; and hybrid networks in which both wired and wireless techniques provide primarily voice-based services, but these are increasingly handling data and other forms of information. Wireless networks can perform functions similar to those of fixed network, plus they offer many advantages such as reduced cost for initial setup and maintenance.

A great deal of commonality exists between wired and wireless networks. Wireless networks are built on top of the existing network technology, thus making use of the vast accumulated knowledge gained over many years.

2.1. DATA TRANSFER

The term **data** refers to alphabetical, numerical, or special purpose characters that are appropriately grouped in binary form to constitute words, messages, or information. Data communication is primarily concerned with the transfer of data from a device in one location to a device in another one. Two or more devices communicating with each other form a system and the devices are said to be networked. Networks can be wired or wireless, or a combination of the two.

The **transfer of data** from one device to another is measured as the baud rate or bit rate. The baud rate indicates the number of symbols transmitted in a unit of time, usually per second. The bit rate indicates the number of bits transmitted per second. The baud rate and bit rate are the same only when one bit is allocated per symbol. But symbols are usually expressed as a series of bits forming words, streams, and codes. For example, Murray codes, which are used for numbers and alphanumeric characters, contain five bits per symbol.

Digital information is often transmitted in **data frames**. A data frame is a collection of characters conveying a complete message that can be understood by the transmitting and receiving devices. A typical data frame is shown in

Figure 2-1. When data frames are used, the information rate is not the same as the bit rate or baud rate because it contains overhead data addresses, error checks, and starts and stops information. The type of information in the frames is governed by the protocols and standards used in that particular application. The protocols are configured within some reference model. Understandably, when protocols are used, the information rate may be much less than the quoted transmission rate.

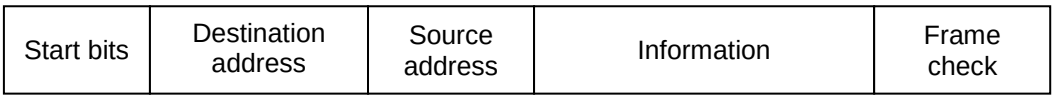

Figure 2-1 Digital data transmission in frames

The theory, the protocols, and the implementation of digital communication systems and associated networks that rely on physical connections such as wires or optical cables are well established and have been in use for many years. However, when compared to wired techniques, wireless data transmission and networking of instruments and sensors in relatively new and the wireless components of most networks behave like their wired counter-parts. Thus the operational principles of wired and wireless networks have many common points, but wireless networks are developing as a separate entity in technological developments and applications.

In both wired and wireless communication systems, data can be transmitted either in **parallel or in serial forms**; with synchronous or asynchronous information flow; or with simplex, half-duplex, or full-duplex data transmission modes. Therefore the discussions presented on these concepts are applicable to wired as well as wireless data communication.

2.1.1. SERIAL AND PARALLEL DATA TRANSMISSION

Data can be transmitted from one device to another in serial or parallel forms. In **serial** data transmissions, each bit of a code is sent sequentially, as shown in Figure 2-2. Consequently serial transmission can be achieved only by one pair of conductors connecting a receiver and a transmitter.

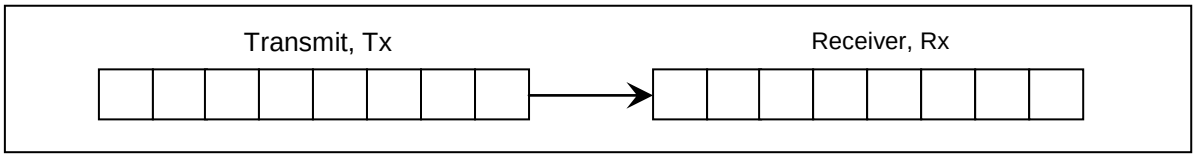

Figure 2-2 Serial data transmission

In **parallel** data transmission, all bits or a number of bits of a code are transmitted simultaneously. Therefore the number of wires required equals the number of bits sent plus the return wire. For example, for an eight-bit code at least eight parallel wires must connect between the transmitter and the receiver, as illustrated in Figure 2-3.

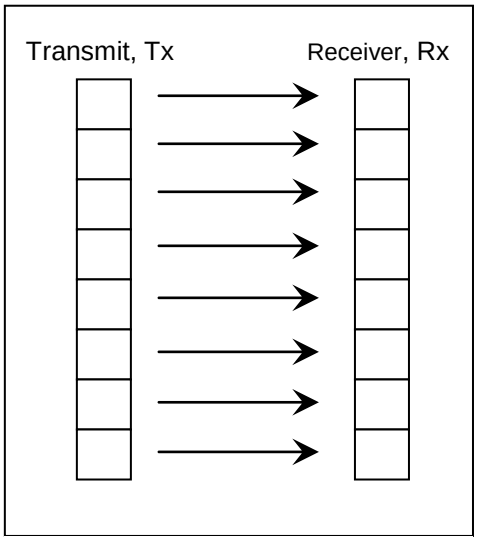

Figure 2-3 Parallel data transmission

2.1.2. SYNCHRONOUS AND ASYNCHRONOUS TRANSMISSION

Serial data can be transmitted in two forms via asynchronous or synchronous transmission. **Asynchronous** transmission sends messages in blocks. This form of transmission may contain significant idle periods between blocks and is often used where high-speed data transmission is not required. Asynchronous data transmission uses data characters that contain information on the synchronization process, the nature and length of the data, and the locations of the first and last bit of the data block, so that the receiver knows the characteristics of the information coming from the transmitter. Since the receiver knows the start and stop bits of the block, the block can be sent at any time and at any rate. Each block between the transmitter and receiver is synchronized in its own right by the use of start and stop elements. The length of the data stream and the time gap between blocks are not usually fixed, but are decided on a per synchronization basis. Naturally, asynchronous transmission is slower than synchronous transmission because of the added synchronization needs.

Figure 2-4 shows a typical binary character transmitted in an asynchronous transmission form. When the character is transmitted, it is preceded by a start bit (binary 0) and followed by an optional parity bit and one or more stop bits. The stop bit is usually a mark or a binary 1.

In asynchronous transmission, the receiver detects the start bit by nothing the transition form a mark to a space, and then decodes the next seven bits as a character. If more characters are to be transmitted, this process is repeated. The receiver and transmitter provide their own internal clocks, both being at about the dame rate, but these clocks are not necessarily synchronized. Also, asynchronous transmission allows variable intervals between the transmitted characters.

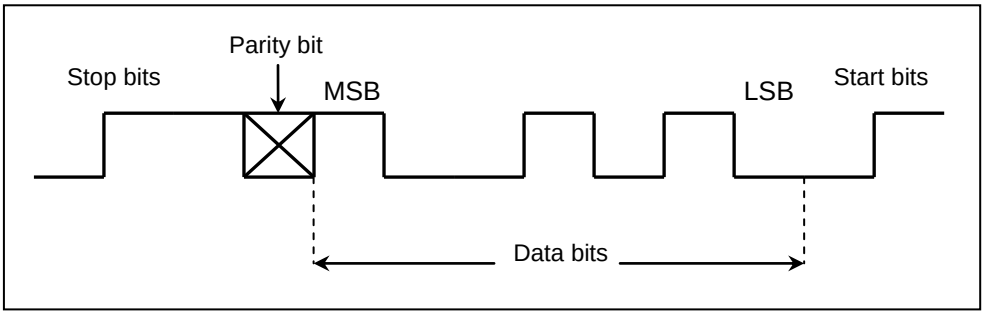

Figure 2-4 Asynchronous transmission of serial data

Synchronous transmission is a message-based transmission technique; it does not employ start and stop bits as in asynchronous transmission. It requires common clock pulse at the transmitting and receiving ends to achieve synchronization. The use of a common clock also helps character pulse identification. The receiver is able to recognize a unique code in the receive bits of the incoming data stream. The receiver must be clocked at exactly the same clock rate as the transmitter. The synchronization of the clock is known as bit synchronization. Synchronous operation can be characterized as follows:

- There are no start and stop bits to synchronize each character.
- Every bit in the transmitter and receiver must be synchronized to a common clock.
- Data are sent in blocks that consist of many elements without separation in between.
- The entire block is framed by codes that indicate the beginning and end.
- The receiver must know the codes, the length of the block, and other relevant and control information.
- It is not sensitive to possible distortion of transmitted signals, since timing is done in a synchronized manner.

With synchronous transmission, synchronization is dealt with on a message basis rather than on a character basis. Once synchronized, it does not allow for a break or an interval between characters. This may limit the effective communication in devices that do not have continuous information flow or devices that do not have buffers to hold messages in case continuous transmission cannot be maintained.

Synchronous and asynchronous transmissions are handled by dedicated devices such as **universal synchronous-asynchronous receiver/transmitters** (USARTs) and **universal asynchronous receiver/transmitters** (UARTs). USARTs and UARTs are an important part of serial data transmission. A USART is a device that converts parallel bits info into a continuous serial data stream, or vice versa. A USART can operate in synchronous or asynchronous form. A UART is a device that handles asynchronous serial communication. A typical UART is a 40-pin programmable device that transmits and receives asynchronous data in either half-duplex of

full-duplex mode. A UART accepts parallel data and converts it into asynchronous mode to make it ready for a serial transmission.

2.1.3. SIMPLEX, HALF-DUPLEX, AND FULL-DUPLEX DATA TRANSMISSION

The transmission of data between two devices can be characterized as simplex, half-duplex, and full-duplex, as shown in Figure 2-5. **Simplex** operation indicates that transmission can take place in only one direction from one device to the next. In this mode, one of the devices can transmit but cannot receive, or it receives but does not transmit. **Half-duplex** operation indicates that transmission can take place in either direction, but in only one direction at a time. **Full-duplex** operation indicates that transmission can take place in both directions simultaneously.

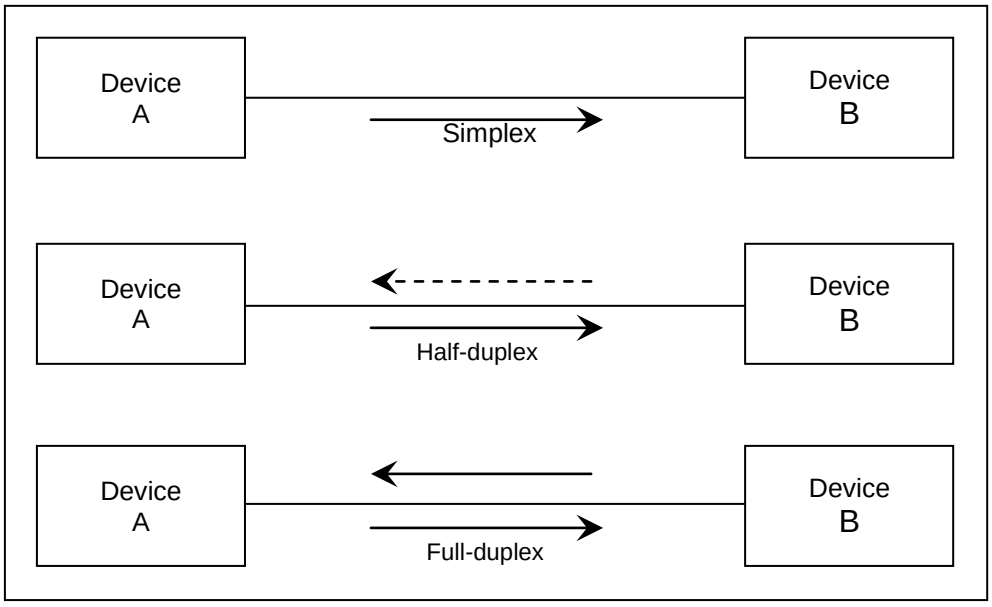

Figure 2-5 Three modes of channel operation

In networks where many devices are involved in communication, transmission uses multiple channels. A channel is defined as a single path on a line through which signals flow. Lines are defined as the components and parts that extend between the terminals of the communicating devices.

An example of simplex operation is a paging system. In paging, messages are received but not necessarily acknowledged. An example of half-duplex operation is a walkie-talkie. In walkie-talkies, the operator pushes a button to talk and releases the button to listen. Both operators cannot communicate at the same time. A full-duplex system provides simultaneous, but separate channels by techniques such as frequency division duplexing (FDD) or time division duplexing (TDD). FDD uses different frequency channels and TDD uses adjacent time slots on a single channel.

2.1.4. WIRELESS DATA TRANSMISSION

Devices can communicate by wired or wireless connections forming local area networks (LANs). The wireless components of most LANs behave like their wired counterparts, but they use space as the transmission media. The operational principles of wired and wireless networks are much the same: it is necessary to attach a network interface for the transmitting and receiving devices. In the case of wireless networks, the interface takes place mainly by radio frequency (RF) transceivers rather than cables. In many cases, wired and wireless systems are used in a mix-and-match form. When interfacing wireless systems to wired networks, devices called access points are used to connect both sides. This allows the shuttling of data traffic back and forth between the wired and wireless components.

In wireless communication systems, the **frequency** used for transmission affects the amount of data and the speed at which the data can be transmitted. The strength or power level of the transmission signal determines the distance over which the data can be sent and received without errors and corruption. In general, the principle that governs wireless transmissions dictates that a lower channel frequency can carry less data, more slowly, but over long distances. Although higher frequencies can carry more data at faster rates, the distance of effective transmission becomes shorter.

Modern wireless communication systems largely use the middle part of the electromagnetic spectrum. The middle part of the electromagnetic spectrum is divided into several frequency ranges, or bands, for communication purposes. These frequency bands are radio frequencies (10 kHz to 1 GHz), microwave frequencies (1 GHz to 500 GHz), and visible and infrared frequencies (500 GHz to 1 THz).

2.1.5. RADIO FREQUENCY DATA TRANSMISSION

There is an inverse relationship between the frequency and the distance that an electromagnetic wave can carry data. There is also a direct relationship between the frequency and data transfer rate and the bandwidth. In data transfer, wireless networks make use of three primary frequency bands: radio frequency (RF), infrared and laser, and microwave.

RF communication systems are designed to operate as narrowband, spread spectrum, or broadband systems both in long-distance and short-distance operations. However, the use of this frequency band and the powers that can be transmitted at these frequencies are strictly regulated. In the United States, government agencies such as the **Federal Communications Commission** (FCC) regulate nearly all radio frequencies. Any commercial or government organization that wishes to use a particular frequency band must apply for permission. This may be granted to use that frequency for broadcasting in a specific location, usually with maximum transmission power limits. However, the FCC set aside certain frequencies for unregulated frequency ranges are 902-928 MHz, 2.4 GHz, and 5.72-5.85 GHz, with some maximum broadcasting distances, typically about 70 meters.

Narrowband radio or single-frequency radio networks use low-power and two-way radio communication systems in a half-duplex format. Radios in taxi

cabs and base stations are good examples of such systems. In these systems, both the receiver and the transmitter must be tuned to a specific frequency to handle incoming and outgoing calls. Some single-frequency systems are made to operate at higher power ratings. Systems of this type can usually transmit over long distances and use repeaters and signal bouncing techniques to increase their coverage distance.

Spread spectrum radio systems address several weaknesses of single-frequency communication systems, both in high and low-power operations. Spread spectrum uses multiple frequencies simultaneously, thereby improving the reliability and reducing the susceptibility to interference. Multiple frequencies make eavesdropping in data transfer much more difficult, if not impossible.

Two main types of spread spectrum communication are frequency hopping and direct sequence modulation. Frequency hopping is based on switching the data among multiple frequencies at regular intervals. The transmitter and receiver must be carefully synchronized to maintain communication. In many systems, hardware handles the timing of hops and chooses the next frequency for transmission.

Direct sequence modulation breaks data into fixed-size segments, called chips, and transmits them on several different frequencies at the same time. The receiving equipment can identify the frequencies that are carrying data. Once the data is received from the identified frequencies, the receiver reassembles the arriving chips into properly arranged sequences of information as sent by the transmitter. For security purposes, some systems transmit dummy data on one or more channels along with the real data on another channel in order to make life even more difficult for eavesdroppers.

2.1.6. INFRARED DATA TRANSMISSION

Infrared wireless transmitters use light beams at infrared frequencies to send communication signals from the transmitter to the receiver. Infrared transmitters generate strong signals to prevent interference from other light sources. The communication systems work well mainly because of their high bandwidth. These systems can deliver data at speeds of 10 Mbps to 100 Mbps. There are four primary types of infrared systems:

- **Line-of-sight** systems require a clear line of sight, or an unobstructed view, between the transmitter and receiver.
- **Reflective infrared** systems signals are generated by an intermittent device called the central hub. The central hub then forwards the messages to the intended recipients.
- **Scatter infrared** systems operate by bouncing the transmitted signals off of walls or other solid objects. The bounced signal is then picked up by the receiver. This approach limits the distance of transmission to typically 30 m or less, depending on the strength of the transmitted signal, the sensitivity of the receiver, and the presence of interference from other sources. Bounce technologies introduce delays in signal transmission;

therefore scatter infrared systems operate on smaller bandwidths than line-of-sight systems.

Laser-based transmission also requires a clear line of sight between the sender and the receiver. In many applications, laser technology-based communication systems are subject to use limitations because excessive radiation can affect human vision and health.

2.1.7. MICROWAVE DATA TRANSMISSION

Microwave data transmission is an established technology that is used extensively worldwide. However, the infrastructure for microwave systems is expensive. Microwave data transmission may not be practical for small systems with show-distance data transfer, but the technology is available. It is used in aerospace communication, television broadcasting, and in military and some civilian applications for long-distance and high-rate data transmission.

2.2. SECURITY IN DATA FLOW

Security between the transmitter and receiver is very important in all types for communication systems. Data needs to be transferred without corruption (also called jamming) that can be caused intentionally by other parties and without being intercepted and listened to by third parties. Information delivered between the transmitter and receiver should be reliable, without any losses, erasures, additive noise, or fading, and it should not be intercepted by unauthorized parties.

There are many effective methods for the reliable delivery of information, such as channel coding, spread spectrum, multiplexing, and encryption.

2.2.1. CHANNEL CODING

For successful information flow between two devices, the receiver must be able to recover the original signal from a received signal that might have gone through a number of changes during transmission. Channel coding, also known as error control coding, is a method of protecting message signals from signal impairment by adding redundancies in the message signal, as illustrated in Figure 2-6. Channel coding can substantially reduce the probability of errors that can occur during transmission.

In **channel coding**, the use of redundancy helps to distinguish the intended message even though there might have been significant corruption during transmission. The introduction of controlled redundancy creates subsets that contain portions of the original message, thus, in a sense, hiding the message. The subset that contains the redundant portion is called the code and the valid messages are called code words or code vectors. A good code contains protected code words, so the likelihood of errors due to corruption during transmission is minimal.

The use of channel coding reduces the error probability as the redundancy in the code increases. However, adding redundancy increases the number of transmission bits from n bits to $n + k$ bits. This causes inefficiency in the system by reducing the transmission rate of the useful information. In applications, the codes are denoted by n and k, the code rate is given by $\frac{n}{k}$, and the increase in the data rate is $\frac{k}{n+k}$. Nevertheless, although errors are reduced, the use of channel coding does not guarantee the elimination of all errors.

There are two basic types of channel coding: block codes and convolution codes. Block coding partitions the source data into blocks of n bits. The encoder adds redundancy and converts the blocks to $n + k$ bits. The encoder also adds information on how redundancy is added, allowing the decoder, on the receiving side, to recover useful information from the codes. There are many types of error correction codes, the most well-known ones being Hamming codes, cyclic redundancy check (CRC) codes, Bose-Chaudhuri-Hochquenghem (BCH) codes, Reed-Solomon codes, and Goley codes.

Convolution coding is based on continuous operation of the encoder accepting useful data in blocks and using shift registers that generate sequences of higher rate data. Although this method is useful and convenient for error detection, error correction is much more complex than in block coding. Convolution methods employ techniques such as probabilistic decoding and approximation likelihood for error corrections.

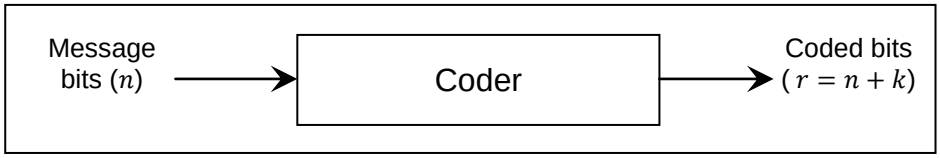

Figure 2-6 Concept of channel coding

Block and convolution coding techniques can be applied simultaneously, particularly in situations where channel errors occur in bursts. Both block and convolution coding use a technique known as interleaving, which spreads each message in a time interval and minimizes the effect of noise bursts. Interleaving sorts the data stream into a series of rows and applies coding in columns to find and eliminate errors. In this method, burst errors effectively become random errors so they can be handled as normal errors by error-correcting codes.

Once an error is detected, two main error correction methods can be applied: **automatic repeat request** (ARQ) and **forward error correction** (FEC). In ARQ, the receiver requests the transmitter to resend the part of the message that contains errors. ARQ is a powerful and effective technique, but it requires an additional feedback channel and adds delays in the data flow. In FEC, the receiver corrects the error without referring to the transmitter. It uses the additional information transmitted along with the data and employs one or more of the methods of channel coding.

In some communication systems, both random and burst errors can be severe and can occur simultaneously. In such cases, concatenation is used. Concatenation uses two types of codes, one for correcting random errors and the other for correcting burst errors.

2.2.2. ENCRYPTION

Encryption is used for protecting the transmitted information from interception or corruption by unauthorized parties. Encryption converts the original text message into an encoded form, known as cipher text. When the data is encoded (encryption) at the transmitter, the resulting cipher text appears as a random stream of symbols that does not make sense. At the receiver, the encrypted data goes through a decryption process that recovers the original information. Encryption and decryption are both controlled by secret information, called the key, known only by the transmitter and receiver. The basic structure of the encryption and decryption process is illustrated is Figure 2-7.

Data is encrypted using ciphers, which are mathematical and physical processes for encrypting data. Ciphers can be altered of modified by changing the key that generates them. In conventional cipher systems, an identical key is used on the transmitter and receiver. This makes the encryption and decryption process symmetric. There are two major types of ciphers: the block cipher and the stream cypher.

A block cipher encodes a number of bits in predetermined blocks. A typical block size is 64 bits; that is, the original data of 64 bits is encrypted to form a cipher text block of 64 bits. A stream cipher encodes each bit individually in such a way that each bit of original text is converted to the cipher text. An ideal cipher text should be completely random and unpredictable. But this is not practical since it is difficult to synchronize the transmitter and receiver keys at all times during the transmission. Instead, most ciphers use pseudorandom key streams generated by the transmitter. The receiver is based on a shared key, which is the same key as the transmitter. The use of key streams makes the transmitted information appear as a completely random signal, thus there is virtually no observable relationship between the cipher text and original text.

Pseudorandom key streams can be generated in a number of ways. A convenient method of generation is the use of linear shift registers that consist of memory elements. The memory elements are arranged as shift registers, and the memory elements of the registers shift one position to the right in each clock cycle. The output of the shift register is exclusive ORed with the original data. Since memory elements are only known to the transmitter and receiver, and the data is exclusive ORed with memory elements, an unauthorized receiver sees the stream of information as completely random and unrecoverable.

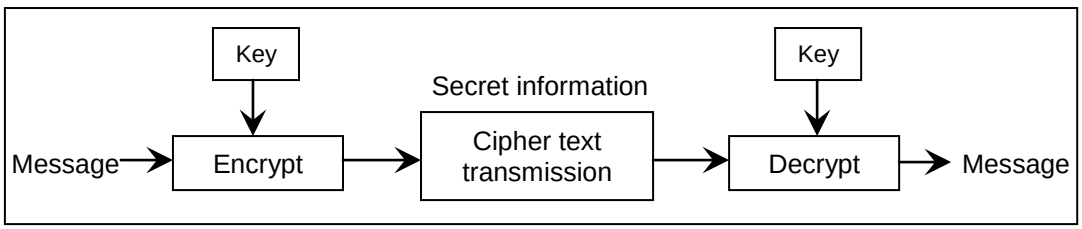

Figure 2-7 A typical encryption process

For an ideal encryption system, cryptanalysis, that is, recovery of the original text, should be unconditionally secure. However, the only known unconditionally secure encryption system is the one-time pad, in which the key

stream is completely random and unpredictable, and is used only once between the transmitter and receiver. This type of unconditionally secure system is not feasible in most application; instead, most systems aim to be computationally secure.

In computationally secure systems, the cost of breaking the encryption may be very high within the useful lifetime of the information. An attempt to break the encryption requires testing of every possible encryption key on the received blocks. This cost can be exemplified as follows: if a 56-bit key is used for encryption, all possible keys that can be used in the process will be 2^{56} keys. If, say, 10^{10} keys are tested every second, it would take about a month to test all the possible keys using a normal PC. One month is required to break the codes for a long stream of data. In general, the use of 128-bit or more keys provides long-term security within the operational speeds of today's computers.

An effective encryption system generates a cipher text so it is difficult to recover the original text if the encryption keys are not known. Ideally the cipher text should not have any observable structure. Thus the longer the key, the better the security. In many cases, minor changes in either the original text or the key can lead to larger changes in the cipher text, this is known as the **avalanche effect**.

There are many encryption algorithms and standards for wireless communication systems. The data encryption standard (DES) is an example if a symmetric encryption algorithm. The DES was originally created by IBM. It is commonly used in Internet and banking transactions, and by cable television. Recently the DES was mandated by the U.S. government for use in securing data applications not involving national security.

There are many other popular encryption techniques. Asymmetric encryption algorithms are commonly employed. Public key and private key algorithms are two variations of asymmetric encryption techniques. The security in the public key algorithm is based on the differences in the complexities of some types of inverse operations. The Rivest-Shamir-Adleman (RSA) algorithm is probably the most popular public key encryption system. This algorithm uses two or more prime numbers and complex arithmetic operations for encryption and decryption. In practice, public key algorithms are often used at the transmitter and private key algorithms are used for decryption at the receiver. Encryption techniques will be deeply discussed in chapter 4 (security).

2.3. NETWORK ESSENTIALS AND TOPOLOGIES

Networks are arrangements of hardware and software components that communicate with each other in a coordinated manner. For effective communication, the components must be mutually compatible devices. Sharing resources and exchanging information among many users and devices on a network is called networking. The most elementary network consists of two devices that are connected together to transmit information from one device to the other. Even though the network concept appears simple, a great deal of coordination and many complex technologies are required to permit communication between devices. In addition, there are many possible choices for physical connections between the network elements and the associated software.

There are various types of networks depending on the number of network elements and their spatial distribution. A **local area network** (LAN) is a system for interconnecting data communication components in a relatively confined space. LANs are most commonly contained within one or several building, as in industrial production facilities, universities, government departments, and other organizations. A networked collection of LANs is called an internetwork, as in the case of interdepartmental networks in universities and organizations. LANs can grow into **wide area networks** (WANs) that cover greater geographic distances, linking two or more separate LANs. In large, complex environments, the number of users and devices on a WAN can grow into thousands or more. For example, the Internet is a WAN internetwork that includes millions of machines and users worldwide. There are many other terms used to describe networks, such as metropolitan area networks (MANs), personal area networks (PANs), and so on, but these are basically LANs or WANs of various sizes.

The major purpose of networks is to share resources by connecting network elements, also called nodes. To be able to connect nodes, four elements are necessary: the transmission medium, the network topology, the protocols and the network operating system.

Transmission medium can be defined as the physical path between the nodes of the network that connects the nodes to each other. The physical path may be cables, fibers, RF devices, microwave devices, etc.

Topology refers to the physical layout of the devices. Topology is linked to the communication methods used between the devices and the way that resources are shared. The network topology can have a significant effect on the performance and efficiency of the network, as well as its future growth potential.

Protocols are the set of rules that are agreed to that enable communication between devices. In the simplest case, for two devices to communicate with each other, they must share a common set of rules that clearly define how they will communicate.

The operating system is the software running in the background that manages the sharing of equipment and data between the network nodes. Operating systems are important because even though two devices might share a common medium and network protocols, they still may not be able to communicate with one another unless they run appropriate software to access the network and enable communication.

2.3.1. NETWORK SOFTWARE

Devices need network software to issue the requests and responses that allow them to communicate with each other. A communication process between two devices is illustrated in Figure 2-8. In this case, communication is taking place in simplex form: device A is sending information to device B.

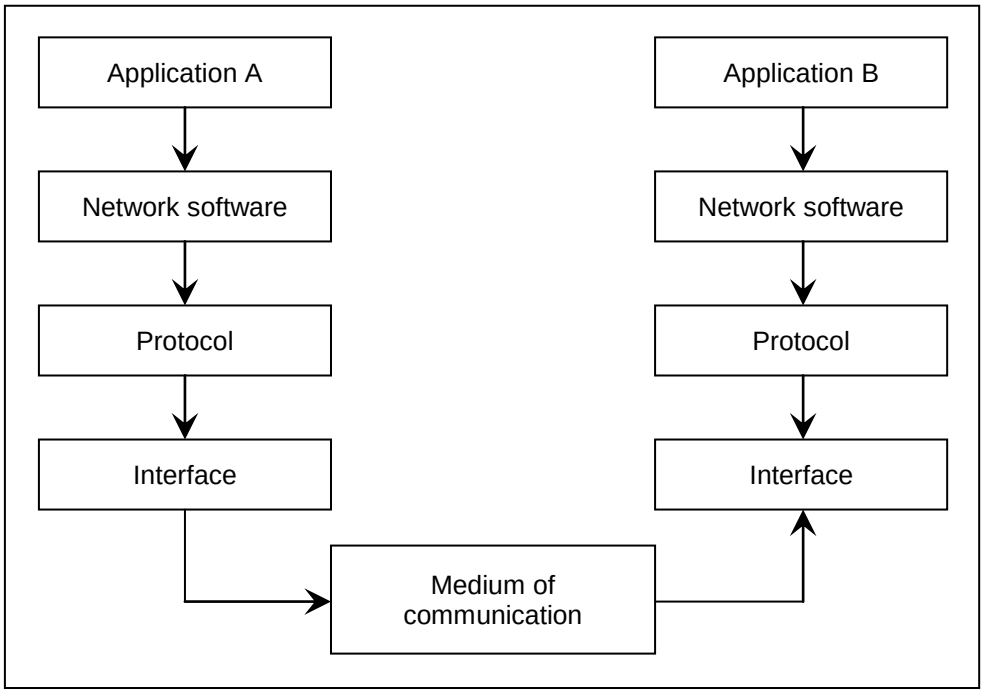

Figure 2-8 Process of communication in networked devices

In many networks, communicating devices invoke a layer of codes, which is called the **network operating system** (NOS). Network operating systems control access to network resources. Examples of common NOSs used in computers are Windows.NET, Windows XP, and Novell's NetWare.

Most network software packages come with modules for logging on and off the network. Network modules for logging on and logging off may include features such as password security, validation of user access to specific files and software, an automatic log on feature for some devices, help menus, error messages, and so on.

2.3.2. NETWORK TOPOLOGIES

Network topology refers to the physical layout of devices and supporting resources in a network and the communication methods between these devices. Topology has a significant effect on the performance of the network as well as on its future growth potential.

Topology primarily describes the patterns of connections of the network nodes. It determines the layout of the communication links between the terminal nodes, junctions, routers, repeaters, and servers. Topology algorithms are used for selecting links as well as link capacities for the associated devices on the basis of a number of factors, including transmission delays, costs associated with delays, the volume of total traffic, and future expandability of the hardware.

All devices, regardless of their topology, communicate in somewhat similar ways. They send data addressed to one or more recipients, transmit the data across the communication media, and accept and interpret the received data in a set way. In order to do all these, network elements obey some common protocols and standards shared between the involved devices.

Networks can be configured in different topologies or combinations of topologies supported by appropriate hardware and software. There are five basic types of topologies: bus, tree, ring, ad hoc, and star. Different network topologies are illustrated in Figure 2-9.

Bus topology (a) consists of a single cabling arrangement to which all nodes are connected. The message is put on the bus by one of the nodes to be received by another node or nodes. Reception of the message is acknowledged only by the addressed node or nodes. Since all the nodes receiving the information are passive when there is a message in the bus, the system may be considered to be fail-safe. Each mode can be individually installed, repaired, and disconnected without affecting the others. Faulty nodes can be easily physically disconnected or isolated by software.

Tree topology (b) is and expanded form of bus topology in which the cables branch in two directions, but offer only one transmission path between any two nodes. As in bus topology, any node can broadcast messages that can be picked up by any other node that is connected to the network.

Ring topology (c) connects each node to two other nearby nodes by point-to-point links, forming a closed loop. Transmitted messages travel from one node to the next, going around the ring in one direction. When a device in the ring receives a message signal, it either acts on it by accepting the message or passes the signal to the next device in the ring. Each node recognizes the addresses of all others nodes and has an equal opportunity to send and receive information at any time.

Token passing is a method for sending data around the ring. In token passing, a node gains an exclusive right to use the channel by grabbing a token. By using the gained token, this node has the right to access the other nodes via the ring connections. When it finishes transmitting the information, it passes the token to another node that has data ready for transmission. When the intended destination receives the information, it returns a message to the sender acknowledging that the data has arrived safely. Each node may have a bypass mechanism that enables communication to continue if it is unable to pass information or goes down.

In some networks, a centralized controller acts as the main control unit and coordinates the other nodes to transmit messages. Modern ring topologies use smart hubs that isolate the faulty devices and ensure the flow of information.

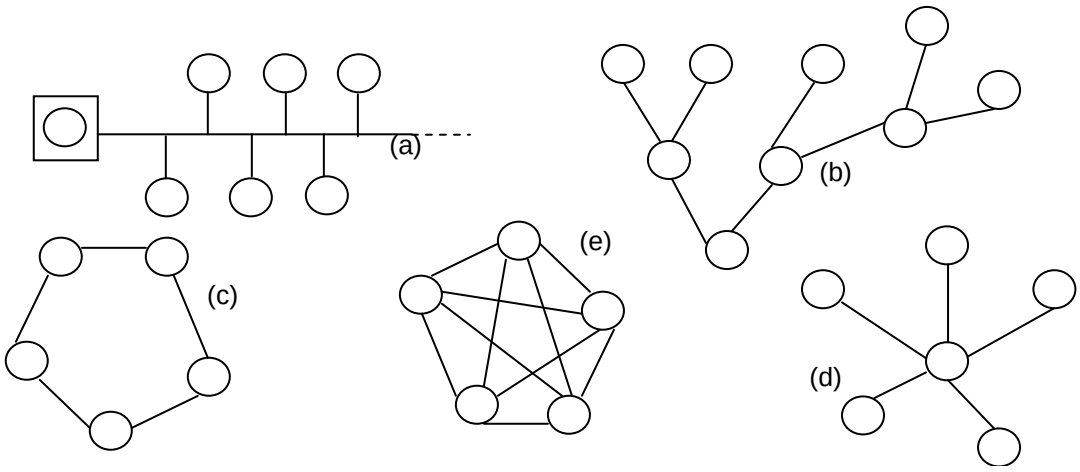

Figure 2-9 Different network topologies

In **star topology** (d), each node is connected via point-to-point link to a central control node. All routing of network traffic takes place through the central node to the outlaying nodes. If a node wants to address another node, is has to go through the central node. Star topology has simple routing algorithms that mainly consist of lookup tables containing the addresses of all nodes.

In star topology, the central control node is the most complex of all the nodes. The complexity is governed by the efficiency, size, and capacity of the network. Star topology is desirable when the bulk of the communication takes place between the central node and outlying nodes. When the communication volume is high between the outlying nodes themselves, some delays may be encountered and efficiency may decrease.

One of the benefits of star topology is that it inherently centralizes the resources. Another benefit of star topology is the relative ease of trouble-shooting. Nodes causing problems can easily be isolated by the central node without affecting the performance of other nodes. A drawback is that if the central node fails, all the devices attached to that node loose network access.

Ad hoc topology (e) is decentralized and does not rely on centralized and organized connectivity. Star, bus, tree, and ring topologies were primarily produced for wired systems, whereas ad hoc networks are more suitable for wireless networks in which a collection of autonomous devices communicate with each other. All network activity, including discovering potential nodes to communicate with, is executed by the nodes themselves. Once communication between the nodes is established, the nodes may organize themselves as one of the previous topologies.

Ad hoc networks range from small, static networks that are constrained by power sources to large, mobile, highly dynamic networks. The design of network protocols for these networks is complex. Regardless of the application, ad hoc networks need efficient distributed algorithms to determine network organization, link scheduling, and routing. However, determining viable routing paths and delivering messages in a decentralized environment where network topology fluctuates is not a well-defined problem. In wireless systems, factors such as variable wireless link quality, propagation path loss, fading, multiuser interference, power expended, and topology changes become important issues. The network should be able to adaptively alter routing paths to alleviate any of these effects.

2.3.3. INTERNETWORKING

Connecting LANs is called interworking. Many LANs can be interconnected by using repeaters, bridges, routers, and gateways. Figure 2-10 shows LAN connection devices and their levels of operation in reference to the OSI model.

Repeaters are layer 1 devices. They are located between the transmitter and the receiver and their function is to strengthen the incoming signal and retransmit it, making it suitable for long-distance operations. Repeaters operate at the physical layer level, hence they do not understand or interpret the data frames or add any new functionality. Since repeaters do not have any intelligence, they can only be used to connect networks of the same type.

Bridges operate at layer 2 and can read the destination and source addresses embedded in the frame. Therefore bridges are able to redirect the frames to their intended destinations. Bridges cannot act like interpreters between different LANs, thus they can only interconnect networks of the same type. They are unable to convert frames if different network techniques are used in the internetwork.

Bridges have limited intelligence and can act as a filter by redirecting the frames onto a segment where the intended device is connected. This is a useful property for the efficient operation of LANs, since the volume of traffic and possible collisions are reduced. Bridges increase the throughput considerably, particularly in high-speed interconnected devices on similar LANs. This is particularly true in situations where protocol conversion is not required.

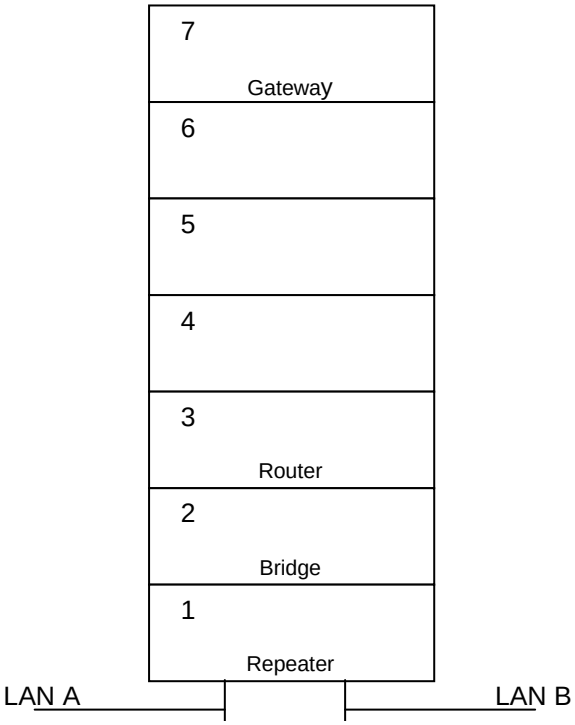

Figure 2-10 LAN connection devices and levels of operation

When the bridge receives a frame, it examines the source address and then it follows a set of rules in handling the frame. If the bridge knows the intended address, the frame is forwarded to that address. If the address of the incoming frame is not specified, the frame is ignored, resulting in the loss of information. If the address of the frame is not in the address table of the bridge, the frame is still forwarded. The bridge continuously updates the address table by adding new addresses of sending and receiving devices.

Bridges use layer 2's information to pass data to the correct destination. Some bridges, called router bridges or routers, provide extra capabilities by being able to define WAN ports or performing conversion operations between different types of networks.

Routers are level 3 network devices. They are highly intelligent and have extensive knowledge of the networks they are involved in. Routers can

determine the best route to the destination through the network, assign priorities to the information flow, and provide some limited security. Routers can link LANs that use different network protocols if they operate on a common protocol such as Transmission Control Protocol/Internet Protocol (TCP/IP).

Routers can also act as filters, as in the case of bridges, but they offer better network management capabilities. Because of these additional features, they are relatively slower. Selecting the best route between the source and destination is achieved by routing algorithms that consider complex network factors. An important network factor is the way that packets are handled. Depending on the volume of traffic, routers divide the long frames into packets, determine convenient router for the packets, and manage the timing of those packets to send them to the destination to be assembled correctly as the original frame.

Gateways operate at layer 7 and are able to connect two or more totally different networks. They can act as translators between host machines within two networks. However, since gateways can route frames as well as act on them as translators and protocol converters, they are relatively slow devices.

2.3.4. INTERNET AND INTRANET

Internet stands for **INTERconnected NETwork** and represents the global “network of networks”. The Internet is made up of a variety of computer networks that are connected through servers. Servers are computers that act as a center for connecting a particular LAN to the Internet. Servers are maintained by private and public Internet server providers (ISPs). A number of servers are connected to the Internet through gateways and routers. The complete network of the Internet is formed by the interconnection of thousands of routers. The transfer of information between computers that are connected to the Internet takes place through routers and gateways. Routers and gateways dynamically learn about network operations as they pass information from source to destination. Routers use packet switching techniques to pass information between computers.

The **TCP/IP** suite is a powerful tool used for communication on the Internet. TCO is a layer 4 protocol that provides services to applications by breaking down the messages into packets. IP provides management of packet switching.

Every computer connected to the Internet has its own discrete address – its Internet address or **IP address**. An IP address has four sets of eight but numbers separated by a period (for example, 123.231.231.123). The first two sets of numbers identify the network, the next set indentifies the subnetwork, and the last set identifies the computer in that subnetwork. IP addresses can be obtained from the ISPs, who have the rights to use block addresses obtained from an upstream registry.

IP addresses have a naming system that uses domain names rather than numbers. Specific documents on the Internet are located using a protocol called the Uniform Resource Locator (URL). Apart from the Hypertext Transfer Protocol (http), there are other protocols such as the File Transfer Protocol (ftp) and Telnet.

The Internet is used for many applications, such as electronic mail (e-mail) transfer, file locations and display, file transfers (for example, ftp), Internet searches and explorations by browsers, and so on. The **World Wide Web** (www) is the most common application of the Internet. It is based on a client/server software model where the user is the client and www is the server. The World Wide Web is a system and collection of standards for storing, retrieving, formatting, and displaying information between computers on the Internet. World Wide Web documents are written in a language called Hypertext Markup Language (html) and transferred via http.

An intranet is a private Internet-like network set up by organizations that can be accessed only by members of the organization or authorized persons and groups. An intranet is established using the same architecture and standards (for example, TCP/IP) as the Internet. Most intranets are connected to the Internet to provide authorized users with access to a wider range of resources. The connection of two or more intranets forms an **extranet**.

CHAPTER 3. PROTOCOLS AND STANDARDS

3.1. PROTOCOLS

As networks become more complex and wide reaching, the requirements for linking devices to networks and the interlinking networks themselves continue to grow. Thus the need of rules, regulations, and standards for successful connections increases proportionately. This leads to rules and standards that are accepted and practiced at national and international levels.

A **protocol** is a set of rules that are agreed to by relevant authorities to enable successful communication between devices. In the simplest case of two devices communicating with each other, they must share a common set of rules and procedures about how to communicate and exchange information. At this minimum level, such rules may include how to interpret signals, how to identify oneself and others on the network, how to initiate and end communication, how to manage the exchange of information across the networks medium, and so on. Protocols must be comprehensive enough to regulate all essential requirements of a communication system. Some of the essentials requirements are:

- Network topology: star, ring, bus, tree, or a combination.
- ISO reference model layers implemented: physical, data link, network, transport, session, presentation, and application.
- Data communication modes: simplex, half duplex, or full duplex.
- Signal type: digital, analog, or a combination.
- Data transmission mode: synchronous, asynchronous, etc.
- Data rate supported: from several bits per second to several gigabits per second, depending on the frequency of the operation and transmission medium.
- Transmission medium supported: wired, RF, optical, microwave, etc.
- Medium access control methods: carrier sense multiple access with collision detection (CSMA/CD), control token, etc.
- Data format: based on data transmission modes and individual protocol specifications.
- Error detection methods: parity, block sum check, CRC, etc.
- Error control methods: echo checking, ARQ, etc.
- Flow control methods: X-ON/X-OFF, window mechanisms, sequences, etc.

In communication systems, there are many different types of protocols addressing particular features of the communication process. Protocols are developed to enable communication in an entire network or part of a network, or in the communication devices within the network. For example, some protocols

may be developed for connections only, some for transferring messages only, and so on. The majority of protocols are developed in a hierarchy of levels or layers in the OSI reference model. The establishment of a connection between two terminals may be realized by obeying the lowest few levels of the OSI model. The transfer of a file of information to solve a specific problem would follow higher levels on the hierarchy.

3.1.1. THE OSI MODEL

Communication of devices in a network demands several carefully orchestrated activities and processes for the information to flow successfully between the sender and the receiver. The concept of networking activities and processes is as important as the configuration of the devices within the network. Several models had been proposed to create an intellectual framework within which to clarify network concepts and activities. Of all these models, none has been as successful as the OSI reference model proposed by the **International Standards Organization** (ISO). This model is commonly known as the ISO/OSI reference model, or simply the OSI model.

Because the OSI model is widely used and there are extensive hardware and software network elements complying with this model, it has become a key part of all types of communication and computer networking. An advantage of this model is that it makes many communication activities explicit by relating discrete activities and processes. The OSI model addresses all types of network concerns in an elaborate manner, starting from simple application to completely open systems, and it has taken an unrivalled position in the world of networking.

3.1.2. STRUCTURE OF THE OSI MODEL

The OSI model is configured in seven layers (Table 3-1). This layering approach helps to clarify the communication process for successful network operation. The layers are the essence of the OSI model. Essentially, networking can be broken into a series of related tasks, each of which can be conceptualized as a single entity in the communication process. This approach breaks down the complexity of the network, from the hardware supporting the network to the application software. Each individual task or activity is handled separately in each layer and issues concerning that layer can be solved independently. Once layers are developed independently, they can be interconnected to make a complete system with interrelated tasks and activities. This approach solves many problems by deconstructing the issues in each layer and breaking down overall system concerns into a series of smaller problems with possible individual solutions.

Level	Function
Process control	Application and system activities control
Presentation control	Compacting, encryption, peripheral device coding and formatting
Session control	Support of session dialog
Data transport	End-to-end control, information exchange, reliability, error control
Network control	Intranetwork operations, addressing and routing
Data link	Enables sequences to be exchanged across a single physical data link
Physical	Transmission at the physical medium

Table 3-1 ISO/OSI reference model

On top of the reference model resides the application layer, which provides a set of interfaces that permit applications such as Web browsers. At the bottom of the reference model, the physical layer resides. The physical layer is concerned with the networking medium, signals, physical connections, etc. All the activities necessary for successful network communication occur between the top and the bottom layers.

Data is typically transferred and interpreted between the transmitter and receiver, as illustrated in Figure 3-1. Layer construction follows a set process, allowing each layer on one device to behave as if it is communicating with its counterpart on the peer layers. On the transmitter side, operations occur on the way down the stack starting from the application layer and moving toward the physical layer. On the receiving side, the role is reversed as the operations move up the stack.

At the transmitter, before data passes from one layer to the next on its way down the stack, it is broken down into protocol data units (PDUs), also called packets or payloads. The PDU is a unit of information passed as a self-contained message to the peer as it moves from one layer to the next on its way down the stack. The protocol software adds its own payload information, formatting, and addressing to the PDU for delivery to its counterpart before it passes the packet to the next layer. An outgoing PDU generated by the sender in any given layer should substantially agree with the version of the PDU on the receiving side. This is important because the receiver should be able to recognize and interpret messages in the forwarded packet.

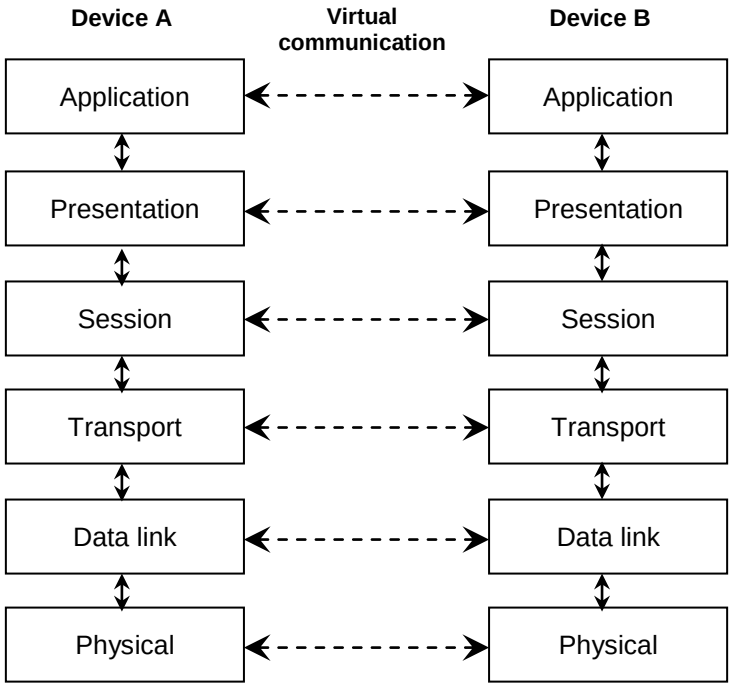

Figure 3-1 Relationship between OSI layers

When the data arrive at the receiving end, the packet travels up the stack, starting from the physical layer through to the application layer. At each layer, the software reads its specific PDU data and interprets the message. Each layer strips its specific information and passes the PDU to the next higher layer. It also performs any additional processing if required.

For any device, sender or receiver, each individual layer of the OSI model has its own set of well-defined functions, and the functions of each layer communicate and interact with the layer above and below it. Each layer concerns itself only with the information exchanged between the peers at the sender and receiver. Each layer puts what may be considered “an electronic envelope” around the data it sends down the track for transmission; conversely, it removes the electronic envelope as it travels up the stack on the receiving side.

Rigidly specified boundaries, called interfaces, separate layers in the OSI model. Any request from one layer to another must pass through the interface. Each layer is built on top of the capabilities and activities of the layer below it and acts to support the layer above. Essentially each layer provides services to the adjacent higher layer and provides a shield for the above layer from the details of lower layers. All networked communication devices are equipped with compatible protocol stacks or protocol suits. Protocol suits are a collection of software elements and services that correspond to specific layers. Network access is only possible through the use of protocols and their associated drivers.

The functionality of the individual layers of the OSI reference model is enunciated next.

Level 1 is the **physical layer**. All the details of the creating physical network connections and the regulation of transmission techniques take place at this level. The physical layer divides the transmission data into frames and gets it ready for interfacing to the network medium. Protocols at this level involve parameters such as signal voltage swings and bit durations, the type of transmission (simplex, half duplex, or full duplex), how connections are established at each end, and so on. Thus the physical layer determines the way that streams of bits are translated into signals for transmission. A compatible receiving device accepts the signals and transforms them into a stream of bits to reconstruct the information. Encoding, timing, and interpretation of signals are decided at the physical level. A typical example of a physical layer protocol and hardware is the RS-232 standard used for serial communication.

Level 2 is the **data link layer**. In this layer, outgoing messages are assembled into data frames and acknowledgment from the receivers is awaited for each message transmitted. Data frames contain the identification (ID) of the sender and the receiver, as well as the controlling information. The destination ID provides a network address for the intended recipient and the sender ID provides the return address for return messages and acknowledgements. All outgoing frames include a destination address at the link layer, and if the higher levels require it, source addresses are included as well. Data integrity is checked at this level by the issued error detecting codes and error correcting codes. There are a number of techniques for data integrity checking, such as CRC, which is a special mathematical function based on the bit pattern in the outgoing frame. Information on error checking and correcting codes is sent as a part of the frame. Possible errors are determined on the receiving end by using special mathematical functions. If the mathematically determined values of error calculations agree with the sent values, the data is assumed to have been received in the original form. When the data is received at the destination, layer 2 protocols strip the information relevant to this layer and package the raw data to pass it on to the next layer. There are many examples of hardware and software operating at level 2 of the OSI model, such as IBM’s Binary Synchronous Communications (BISYNC) protocol, X.25, and so on.

Level 3 is the **network layer**. The network layer handles addressing of messages for delivery and translates logical network addresses into their physical counterparts. Physical network addresses are known as media access control (MAC) addresses. MAC decides how to rout the transmission from the sender to the receiver. Level 3 considers important factors based on network conditions, quality of service information, the cost of alternative routes, and delivery priorities. It handles packet switching, data routing, and congestion control for the network. For the successful operation of this layer, long outgoing messages are divided into smaller packets for convenience in handling. Fragmentation and segmentation of data result in many packets, but provides

easy manageability requiring shorter transmission times. When moving data from one type of network medium to another type, the network layer handles the segmented packets and helps reassemble the data that may have been altered due to dissimilar media. Incoming packets are reassembled into messages and downsized into their original forms to pass them on to higher layers.

Level 4 is the **transport layer**. This layer manages the conveyance of data from the sender to the receiver across the network. An important task of this layer is to ensure flow control by making sure that the recipient of the transmitted data is not overwhelmed with more data than it can handle. Therefore long data payloads are segmented into portions matching the maximum packet size that is acceptable to the network medium and the recipient. The transport layer of the receiving side resequences the divided data arriving in packets into its original form. The transport layer determines the necessary parallel paths and multiplexers for routing of the packets through available channels. The transport layer is the busiest layer for end-to-end communication in the network.

Level 5 is the **session layer**. The session layer establishes system-to-system connection across the network and allows two parties to carry out ongoing communication, called a session. The exchange of messages and the transmission of data take place as long as this session continues. This layer has many functions, including setting up the session, monitoring the session identification process, providing security in data flow, providing continuity in exchanging data and messages, and terminating the session when the task is completed. The session layer also ensures synchronization of the tasks on both ends of the connection. It can place check marks in the data stream so that if communication fails at some point, only data after the most recent check mark is retransmitted. It controls logging on and off the system, verifies user identification from lookup tables, and provides billing and management issues.

Level 6 is the **presentation layer**. The presentation layer handles data formatting to make it suitable for networked communication. For outgoing messages, the presentation layer converts data into generic formats so that they can survive the rigors of transmission on the network. For incoming messages, it converts the data from its generic network representation into a format that is suitable for the requirements of the receiving device. The presentation layer performs many other tasks such as protocol conversion, handling library routines, performing data encryption and decryption, addressing character set issues, conducting compression and decompression, and providing graphic commands and code conversions.

Level 7 is the **application layer**. This is the top layer of the OSI reference model. This is the level seen by individual users, and it provides a set of interfaces for applications to gain access to network services. In the application layer, network transparency is maintained by concealing the physical distribution of resources from the user. It provides service support applications for file transfer, message handling, database management, and so on. The

application layer is able to partition complex problems among several machines in distributed process applications.

3.1.3. IEEE 802 NETWORK MODEL

The **IEEE 802 network model** is based on the OSI reference model, but it is perceived as an enhancement of the OSI model. IEEE 802 lays out a family of specifications for different types of networks, consequently there are many protocols and standards within this model. The specifications of the protocols and standards encompass and meet the requirements of a diverse range of existing networks and they are conceived to be open-ended, thus allowing the development of new types of networks. The IEEE 802 standards are the most influential networking standards in use internationally.

IEEE 802 expands the OSI reference model at the physical and data link layers. At the data link layer, two additional sublayers are included: logical link control (LLC) and media access control (MAC), as shown in Figure 3-2.

The **LLC sublayer** controls data link communication and defines the use of logical interface points, called service access points (SAPs). LLC is also responsible for error recovery in some applications. There are several modes of LLC operation; some modes require LLC to detect and recover from errors that have taken place during transmission in the selected media.

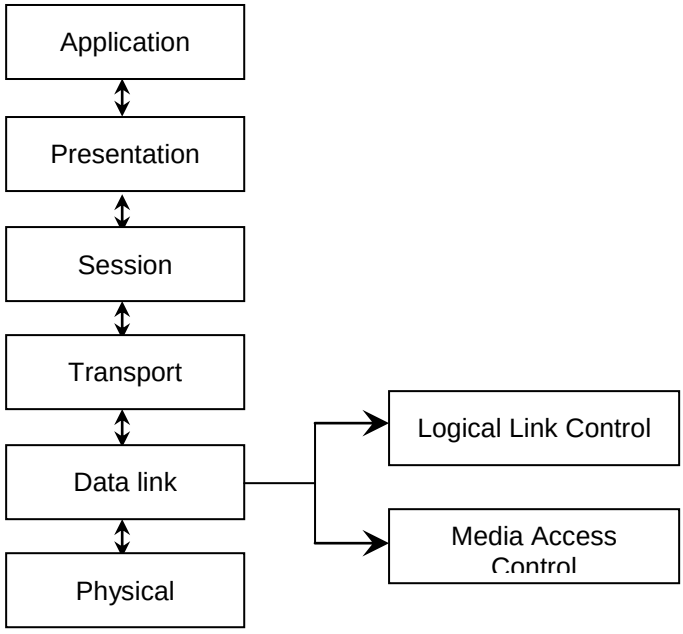

Figure 3-2 IEEE 802 with data link sublayers

The **MAC sublayer** provides shared access of multiple devices in the physical layer. MAC directly communicates with internal operations of a particular device, such as computers equipped with network interface cards

(NICs), and is responsible for ensuring error-free data transmission between the device and the network.

The IEEE 802 model concentrates on standards that describe the physical elements of a network, including network adapters, cables, connectors, signaling technologies, MAC, and so on. Most of these reside on the lower two layers of the OSI model, in the physical and data link layers. IEEE 802 is also concerned with how to manage, attach, and detach devices in and out of a networked environment.

3.1.4. TRADITIONAL MAC PROTOCOLS

We begin with a focus on contention-based MAC protocols. Contention-based MAC protocols have an advantage over contention-free scheduled MAC protocols in low data rate scenarios, where they offer lower latency characteristics and better adaption to rapid traffic variations.

3.1.4.1. Aloha and CSMA

The simplest forms of medium-access are **unslotted Aloha** and **slotted Aloha**. In the unslotted one, each node behaves independently and simply transmits a packet whenever it arrives; if a collision occurs, the packet is retransmitted after a random waiting period. The slotted version of Aloha works in a similar manner, but allows transmission only in specified synchronized slots.

Another classic MAC protocol is the **carrier sense multiple access** (CSMA) protocol. In CSMA, a node that wishes to transmit first listens to the channel to assess whether it is clear. If the channel is idle, the node proceeds to transmit. If the channel is busy, the node waits a random back-off period and tries again. CSMA with collision detection is the basic technique used in IEEE 802.3/Ethernet.

3.1.4.2. Hidden and exposed node problems

Traditional CSMA fails to avoid collisions and is inefficient in wireless networks because of two unique problems: the hidden node problem and the exposed node problems.

The **hidden node problem** is illustrated in Figure 3-3(a); here, node A is transmitting to node B. Node C, which is out of the radio range of A, will sense the channel to be idle and start packet transmission to node B too. In this case, CSMA fails to avoid the collision because A and C are hidden to each other.

The **exposed node problem** is illustrated in Figure 3-3(b). In this case, while node B is transmitting to node A, node C has a packet intended for node D. because node C is in range of B, it senses the channel to be busy and is not able to send. However, in theory, as D is outside of the range of B, and A is outside of the range of C, these two transmissions would not collide with each other. The deferred transmission by C causes bandwidth wastage.

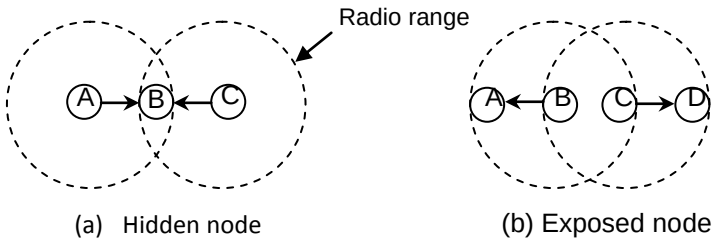

Figure 3-3 Problems with basic CSMA in wireless environments

These problems are duals of each other in a sense: in the hidden node problem packets collide because sending nodes do not know of another ongoing transmission, whereas in the exposed node problem there is a wasted opportunity to send a packet because of misleading knowledge of a non-interfering transmission. The key underlying mismatch is that it is not the transmitter that needs to sense the carrier, but the receiver. Some communication between the transmitter and receiver is needed to solve these problems.

3.1.4.3. Medium-access with collision avoidance (MACA)

The **MACA** Protocol by Karn (8) introduced the use of two control messages that can (in principle) solve the hidden and exposed node problems. The control messages are called *request to send* (RTS) and *clear to send* (CTS). The essence of the scheme is that when a node whishes to send a message, it issues an RTS packet to its intended recipient. If the recipient is able to receive the packet, it issues a CTS packet. When the sender receives the CTS, it begins to transmit the packet. When a nearby node hears an RTS addressed to another node, it inhibits its own transmission for a while, waiting for a CTS response. If a CTS is not heard, the node can begin its data transmission. If a CTS is received, regardless of whether or not an RTS is heard before, a node inhibits its own transmission for a sufficient time to allow the corresponding data communication to complete.

Under a number of idealized assumptions (for example, ignoring the possibility of RTS/CTS collisions, assuming bidirectional communication, no packet losses, no capture effect) it can be seen that the MACA scheme can solve both the hidden and the exposed node problem. Using the earlier

examples, it solves the hidden node problem because node C would have heard the CTS message and suppressed its colliding transmission. Similarly it solves the exposed node problem because, although node C hears node B's RTS, it would not receive the CTS from node A and thus can transmit its packet after a sufficient wait.

3.1.4.4. IEEE 802.11 MAC

Closely related to MACA is the widely used **IEEE 802.11 MAC standard** (9). The 802.11 device can be operated in infrastructure mode (single-hop connection to access points) or in *ad hoc* mode (multi-hope network). It also includes two mechanisms known as the distributed coordination function (DCF) and the point coordination function (PCF). The DCF is a CSMA-CA protocol (carrier sense multiple access with collision avoidance) with ACKs. A sender first checks to see if it should suppress transmission and back off because the medium is busy; if the medium is not busy, it waits a period DIFS (distributed inter-frame spacing) before transmitting. The receiver of the message sends an ACK upon successful reception after a period SIFS (short inter-frame spacing). The RTS/CTS virtual carrier sensing mechanism from MACA is employed, but only for unicast packets. Nodes which overhear RTS/CTS messages record the duration of the entire corresponding DATA-ACK exchange in their NAV (network allocation vector) and defer access during this duration. An exponential backoff is used (a) when the medium is sensed busy, (b) after each retransmission (in case an ACK is not received), and (c) after a successful transmission.

In the second mechanism, PCF, a central access point coordinates medium-access by polling the other nodes for data periodically. It is particularly useful for real-time applications because it can be used to guarantee worst-case delay bounds.

3.1.4.5. IEEE 802.15.4 MAC

The **IEEE 802.15.4 standard** is designed for use in low-rate wireless personal are networks (LR-WPAN), including embedded sensing applications (10). Most of its unique features are for a beacon-enabled mode in a star topology.

In the beacon-enabled mode for the star topology, the IEEE 802.15.4 MAC uses a superframe structure shown in Figure 3-4. A superframe is defined by a periodic beacon signal sent by the PAN coordinator. Within the superframe there is an active phrase for communication between nodes and the PAN coordinator and an inactive phase, which can be adjusted depending on the sleep duty cycle desired. The active period (CAP), and a collision-free period

(CFP) that allows for the allocation of guaranteed time slots (GTS). The presence of the collision-free period allows for reservation-based scheduled access. Nodes which communicate only on guaranteed time slots can remain asleep and need only wake-up just before their assigned GTS slots. The communication during CAP is a simple CSMA-CA algorithm, which allows for a small backoff period to reduce idle listening energy consumption (11).

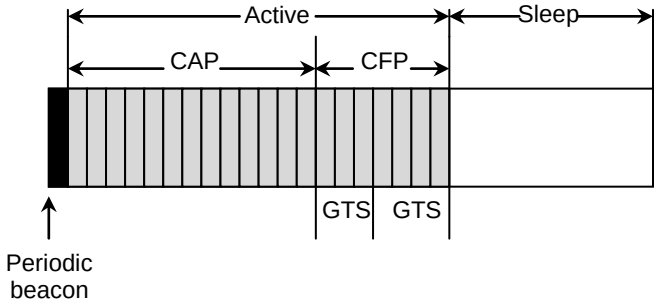

Figure 3-4 The superframe structure of IEEE 802.15.4 MAC

3.1.5. ENERGY EFFICIENCY IN MAC PROTOCOLS

Energy efficiency is obtained in MAC protocols essentially by turning off the radio to sleep mode whenever possible, to save on radio power consumption.

3.1.5.1. Power management in IEEE 802.11

There exist power management options in the infrastructure mode for 802.11. Nodes inform the access point (AP) when they wish to enter sleep mode so that any messages for them can be buffered at the AP. The nodes periodically wake-up to check for these buffered messages. Energy savings are thus provided at the expense of lower throughput and higher latency.

3.1.5.2. Power aware medium-access with signalling (PAMAS)

The **PAMAS** (power aware medium-access with signalling) (12) is an extension of the MACA technique, where the RTS/CTS signalling in carried out on a separate radio channel from the data exchange. It is one of the first power aware MAC protocols proposed for multi-hop wireless networks. In PAMAS, nodes turn off the radio (go to sleep) whenever they can neither receive nor transmit successfully. Specifically they go to sleep whenever they overhear a neighbor transmitting to another node, or if they determine through the control channel RTS/CTS signalling that one of their neighbors is receiving. The duration of the sleep mode is set to the length of the ongoing transmissions

indicated by the control signals received on the secondary channel. If a transmission is started while a node is in sleep mode, upon wake-up the node sends probe signals to determine the duration of the ongoing transmission and how long it can go back to sleep. In PAMAS, a node will only be put to sleep when it is inhibited from transmitting/receiving anyway, so that the delay/throughput performances of the network are not affected adversely. However, there can still be considerable energy wastage in the idle reception mode.

3.1.5.3. Minimizing the idle reception energy costs

While PAMAS provides ways to save energy on overhearing, further energy savings are possible by reducing idle receptions. The key challenge is to allow receivers to sleep a majority of the time, while still ensuring that a node is awake and receiving when a packet intended for it is being transmitted. Based on the methods to solve this problem, there are essentially two classes of contention-based sensor network MAC protocols.

The first approach is completely asynchronous and relies solely on the use of an additional radio or periodic low-power listening techniques to ensure that the receiver is woken up for an incoming transmission intended for it. The second approach, with many variants, uses periodic duty-cycled sleep schedules for nodes. Most often the schedules are coordinated in such a way that transmitters know in advance when their intended receiver will be awake.

3.2. STANDARDS

Standards are worthy of explaining in detail, as there may be a confusing number of standards for similar functionality devices, communication systems, and networks. The term “standard” has many definitions. The *National Standards Policy Advisory Committee* describes a standard as “a prescribed set of rules, conditions, or requirements concerning definitions of terms: classification of components; specification of materials, performance, or operations; delineation of procedures; or measurement of quantity and quality in describing materials, products, systems, services, or practices” (13).

Since the word standard is a broad term and used in many contexts, there are several kinds of standards, there are published standards (sometimes referred as paper standards) of practices and protocols that can be defined as documents describing the operations and processes to achieve unified results. Both physical and published standards play fundamental roles in shaping the efficiency of domestic and global economies. Published standards are important as they are documents that provide textual and illustrative information on what and how things should be done and are done.

Because of the national and international implications, there are many institutions and organizations that are responsible for investigating, developing, determining, and maintaining the relevant standards to support the worldwide scientific and industrial activities. However, national and international standards authorities are subject to various internal and external forces, differences of opinion, and commercial interests during the development of standards and may not always end up with ideal results in the process of standardization. This, unfortunately, can result in a confusing number of standards with different versions and interpretations describing the same process. In addition, publications grow over time, and as new technological developments occur, the standards authorities and procedures change. This can result in a multitude of standards on a specific issue.

A **standard** is intrinsically a bureaucratic process that becomes obsolete in fast moving technological environments. Standards largely benefit users (as in the case of mobile communication) and manufacturing organizations. Vendors support standards to penetrate into a market and maintain their market position by offering interoperable or interchangeable components, devices, systems, and software.

It is important to realize that standards are developed by various national and international bodies; sometimes operating totally independently, hence there may be different versions for the same subject. Some examples of national and international standards bodies are the ISO, International Electrotechnical Commission (IEC), IEEE, and American National Standards Institute (ANSI). This list can grow to hundreds of organizations worldwide. All these organizations have multiple internal departments, committees, subcommittees, and working groups to support their activities.

The evolution and expansion of wireless technology is a dynamic and ongoing process, thus there are very few standards in this area as the technology is continuing to change and develop. Nevertheless, some key developments have occurred in the standards that are directly applicable to wireless instrumentation systems. A few examples of these relevant standards are IEEE 802, high performance radio local area network (HiperLAN), PAN, Bluetooth, cellular packet radio standards, and IEEE 1451.

3.2.1. IEEE 802 STANDARDS

The IEEE 802 and ISO/OSI models were developed in collaboration of the organizations and are compatible with one another. IEEE 802 goes into much more detail on various types of networks, internetworking, high-speed networking, and network security. The network standards are numbered from 802.1 through 802.18. Each standard may have a series of standards carrying the same number but different extensions, such as 802.11b, 802.11g, etc. A complete list of 802 standards is given in Table 3-2.

The IEEE 802.11 Wireless LAN Working Group was founded in 1987 to begin standardization of spread spectrum wireless local area networks (WLANs) for use with industrial, scientific, and medical (ISM) bands. WLAN efforts of the IEEE did not gain momentum until the late 1990s, when the popularity of the Internet, combined with the wide-scale acceptance of portable devices and laptop computers, caused WLAN to become an important and rapidly growing segment of modern wireless communication. IEEE 802.11 was standardized in 1997, with the goal of providing interoperability standards for WLAN manufacturers using 11 Mbps direct sequence spread spectrum (DSSS) spreading and 2 Mbps user data rates. With an international standard now readily available for everyone, numerous manufacturers began to comply and the market began to grow rapidly. In 1999, the 802.11 high rate standard, called 802.11b, was approved, thereby providing new user data rate capabilities of 11 Mbps and 5.5 Mbps.

Standard	Name	Function
802.1	Internetworking	Routing, bridging, and Internet work communication
802.2	Logical link control	Error control and flow control over data frames
802.3	Ethernet LAN	Forms of Ethernet media and interfaces
802.4	Token bus LAN	Forms of Token bus media and interfaces
802.5	Token ring LAN	Forms of Token ring media and interfaces
802.6	Metropolitan area networks	MAN technologies, addressing and services
802.7	Broadband Advisory Group	Broadband networking media, interfaces, equipment
802.8	Fiber-optic Advisory Group	Fiber-optic media, network types, technologies
802.9	Integrated networks	Integration of voice and data traffic in a network medium
802.10	Network security	Network access control, encryption, certification, other security
802.11	Wireless networks	Wireless networks, frequency usage
802.12	High-speed networking	Variety of 100 Mbps-plus technologies
802.13	Unused	
802.14	Defunct working group	Data transfer over cable TV
802.15	Wireless personal area networks	Emerging standards for wireless PANs
802.16	Wireless metropolitan area networks	Wireless MANs
802.17	Resilient packet ring	Very high speed ring-based LANs and MANs
802.18	Wireless Advisory Group	Radio-based wireless standards

Table 3-2 IEEE 802 wireless network standards

The IEEE 802.11 (simply 802.11) standards address issues concerning wireless networks. 802.11 will continue to be developed and grow because of new wireless technologies and applications. Many manufacturers of wireless networking devices and systems have developed inexpensive, reliable wireless LANs and associated devices for domestic and industrial use that comply with the 802.11 standard. There are several versions of the 802.11 standard, the current ones being 802.1b, which specifies a bandwidth of 11 Mbps at a frequency of 2.4 GHz; 802.11a, which specifies a bandwidth of 54 Mbps at a frequency of 5 GHz; 802.11g, which specifies a high-speed wireless standard operating at speeds of up to 54 Mbps with a carrier frequency of 2.4 GHz. These 802.11 wireless LAN standards are commonly known as Wi-Fi.

3.2.2. WIRELESS ETHERNET CONCEPTS

802.11 wireless networks are viewed as an extension of Ethernet that uses electromagnetic propagation as the medium of transmission instead of electrical and optical cables. However, most 802.11 networks incorporate some wired Ethernet segments operating collaborate lively with the wireless components. The communication range of 802.11b-compatible devices is short, about 100 m, but the range of the network can extend from several meters to several hundred meters with the use of repeaters and other assisting devices, depending on the environmental factors and RF interference that exists in the area of operation.

802.11b uses a wireless access point (WAP) that serves as the center of a network configured in star topology. Workstations equipped with wireless devices such as network interface cards (NICs) can send packets to the WAP, which then redirects the packets to a destination workstation.

Wired Ethernets use CSMA/CD as the access method, but wireless networks have a special problem with this method. CSMA/CD requires that all the stations hear each other so they can identify the source that is sending data. If any two stations in the network try to send data at the same time, a collision can occur. In wired networks, the sending station will notice that a collision has occurred and it will attempt to send the data again. However, 802.11b wireless stations cannot send and receive data at the same time, so if a collision occurs, it may not be detected by the sending station. For this reason, 802.11b specifies a CSMA/CA access method in which an acknowledgement is required from the receiver for every packet that is sent and received. If there is no acknowledgment, the sending station knows the packet did not arrive safely.

The 802.11b standard specifies a transmission rate of 11 Mbps, but adverse environmental conditions may prevent transmission at this speed. Therefore transmission speeds may be decreased incrementally starting at 11 Mbps to 5.5 Mbps to 2 Mbps and finally to 1 Mbps for reliable connections. In the 802.11b standard, there is no fixed segment length because reliable communication

depends heavily on the environment and the segment length is determined by environmental conditions.

In general, an 802.11b network has a maximum range of about 100 m with no obstructions. This distance can be extended using large, high-quality antennas. However, the data rate may suffer as the distance increases or as more obstructions are present in the transmission path.

The **802.11g** and **802.11a** standards are extensions of 802.11b. Although 802.11g competes with a 802.11a, it shares many common features. 802.11g specifies a bandwidth of 54 MHz, while 802.11b specifies 11 MHz. The 802.11a standard uses an unlicensed 5 GHz portion of the spectrum, but at this time 802.11a products are more expensive to produce. 802.11g is backward compatible with 802.11b.

The 802.11a frequency bands are 5.15-5.35 GHz and 5.725-5.825 GHz, allowing at least eight simultaneous channels. In addition, another band of 255 MHz will be available for 802.1a in the United States at 5.47-5.725 GHz. 802.11b and 802.11g use an 83.5 MHz band located between 2.4 GHz and 2.4835 GHz, allowing three channels to be used simultaneously.

Of these competing standards, 802.11b appears to be most prevalent, but it has been in use longer. Of the two higher speed standards, 802.11g is backward compatible with 802.11b and therefore provides convenient bandwidth upgrade possibilities. 802.11a presents problems for upgrades to 802.11b because of its higher frequency, but it is far more reliable and flexible. It is likely that standards that provide high-speed transmission will ultimately dominate the marketplace and will be accepted by consumers as well as manufacturers for widespread use.

3.2.3. IEEE 802.16 WIRELESS METROPOLITAN AREA NETWORKS

The IEEE 802.16 wireless metropolitan area network (WMAN) standards are equivalent to HiperMAN in Europe. It can be seen as a replacement for wired or fiber-optic-based MANs. 802.16 forms the backbone of many wireless Internet service providers (WISP) operating around the globe. The base station is the fundamental component of a WMAN. The base station serves as a hub and can easily be located on buildings or transmission towers. Base stations can transmit information over long distances; typically 10 km to 15 km. Base stations are also used as bridges between the wireless world and the Internet. IEEE consortiums such as the Wi-Fi alliance try to ensure that 802.16 and 802.11 products are compatible with each other. Some WMANs use the unlicensed 2.4 GHz frequency band and orthogonal frequency division multiplexing (OFDM) techniques for data transmission.

3.2.4. CODE DIVISION MULTIPLE ACCESS-BASED STANDARDS

Code division multiple access (CDMA) is a cellular technology that competes with other technologies such as the Global System for Mobile Communications (GSM), Digital Enhanced Cordless Telecommunications (DECT), and General Packet Radio Service (GPRS). CDMA, developed by Qyalcomm and Ericsson, is a high-capacity and small cell radius cellular system. It employs spread spectrum technology and a special coding scheme.

Figure 3-5 shows how spread spectrum signal is generated. The data signal with pulse duration of T_b is XOR'ed with the code signal with pulse duration of T_c. Therefore, the bandwidth of the data signal is $\frac{1}{T_b}$ and the bandwidth of the spread spectrum signal is $\frac{1}{T_c}$. Since T_c is much smaller than T_b, the bandwidth of the spread spectrum signal is much larger than the bandwidth of the original signal. The ratio $\frac{T_b}{T_c}$ is called spreading factor or processing gain and determines to a certain extent the upper limit of the total number of users supported simultaneously by a base station (14).

The first generation of CDMA is known as CDMA One (also known as IS-95), and the second generation is CDMA2000, which is the dominant technology at the moment. CDMA2000 has many variants (1X EV, 1X EV-DO, and MC 3X) but new versions are emerging.

CDMA2000 is a wideband radio interface that offers significant advances over CDMA One with its improved performance and capacity by using turbo codes. CDMA2000 employs advanced MAC for efficient and high-speed packet data services. Its physical layer features a dedicated control channel (DCCH) and a common control channel (CCCH). The specifications of CDMA2000 also incorporate advanced multimedia quality of service (QoS) capabilities to enable scheduling and prioritization among competing services.

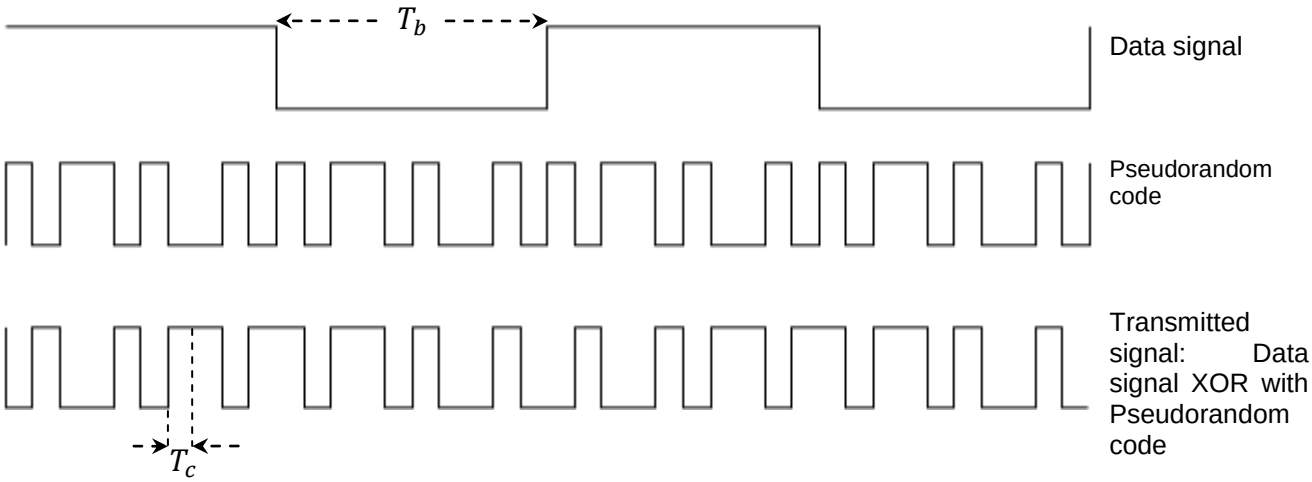

Figure 3-5 Generation of CDMA

3.2.5. TIME DIVISION MULTIPLE ACCESS-BASED STANDARDS

Time division multiple access (TDMA) standards use a single channel and divide it into a number of time slots. Each user is allowed to use only one time slot out of every few slots. Some network systems use dynamic time slot allocation to avoid wasting time slots if one side of the conversation is silent.

Time division multiple access was first implemented as the TIA-54 standard (also known as D-AMPS). TIA-54 provided three TDMA voice channels in the space of one 30 kHz analog channel. The next generation of TDMA, known as IS-136, extended the use of TDMA to the control channel. IS-136 has been adapted by ANSI and published as the TIA/EIA-136 series of standards. The UWC-136 RTT standard is based on an enhancement of the ANSI-136 TDMA standards that incorporate several others such as GSM, Enhanced Data GSM Environment (EDGE), and GPRS. It includes enhanced interfaces, capabilities, features, and services. A greater spectral efficiency is achieved with a change in the modulation scheme, new slot format, the addition of new interleaving and coding options, and other enhancements.

Time division multiple access, as defined in TIA IS-54, IS-136, and TIA/EIA-136, divides the 30 kHz cellular channel into 3-8 kbps time slots, which supports three users in strict alternation. This approach theoretically triples the capacity of cellular frequencies.

The current version of ANSI-136 supports 30 kHz channel spacing, six time slots, three calls per channel, circuit switch data rates of up to 28 kbps, and more. The UWC-136 compliant version of ANSI-136 supports 30 kHz, 200 kHz, and 1.6 MHz channel spacing, six time slots with either six or three calls per channel, and high-speed packet data rates up to 384 kbps or 2 Mbps.

3.2.6. GSM AND GPRS STANDARDS

The GSM is a digital cellular system that has found popular applications around the world. GSM networks operate in three different frequency ranges: GSM 900 operates at 900 MHz and is most common in Europe and rest of the world; GSM 1800 or DCS 1800 operates at 1800 MHz and is used in many countries (for example, France, Germany, Switzerland, United Kingdom, and Russia); and PCS 1900 or DCS 1900 operates at 1900 MHz and is used in the United States and Canada. Apart from their operational frequencies, the only differences between these systems are power levels and some minor changes in signaling. GSM standards are used for providing interfaces between various entities in GSM networks. There have been many standards introduced recently to support GSM networks. Some of these standards are GSM 04.31 (v. 8.1.0) for radio resource protocols, GSM 08.71 (v. 7.2.0) for base station system interface layers, and many others.

GSM has many features that can be used in instrumentation systems and networks, such as the short message service of GSM, which allows sending and receiving of 126-character text messages. The data speed is 9600 bps and it uses effective encryption techniques to prevent tapping and eavesdropping.

GPRS is based on GSM. It is a packet-based wireless communication service that has data rates of 56 kbps to 114 kbps. Packet-based services are more efficient than circuit-switched services since communication channels are being shared between users rather than dedicated to only one user at a time. GPRS allows several users to share the same GSM time slots via a link layer send/receive scheduling protocol.

GPRS complements Bluetooth, which effectively replaces wired connections between devices with wireless radio connections. GPRS also supports IP and X.25, a packet-based protocol that is used mainly in Europe. Hence, with GPRS, the user can access two forms of data networks, X.25 for packet-based systems and IP.

In GPRS, a TDMA packet data channel (PDCH) carries both user data and information. GPRS has a series of standards, such as GPRS-136. The main goals of these standards are to provide network architecture, radio communication technologies, and protocols for access, such as for roaming between GPRS and other networks. The GPRS-136 data model overlays the circuit-switched network nodes with packed data network nodes for service provisioning, registration, mobility management, and accounting. GPRS is an evolutionary step toward EDGE and universal mobile telephone service (UMTS).

3.2.7. OTHER WIRELESS NETWORK STANDARDS

There are many other wireless network standards such as composite CDMA/TDMA, Digital Advanced Mobile Phone System (D-AMPS; also called IS-54), DECT, EDGE, GSM EDGE Radio Access Network (GERAN), iMode, Personal Communication Service (PCS), Personal Digital Cellular (PDC), UMTS, Worldwide Interoperability for Microwave Access (WiMAX), and so on.

3.2.8. IEEE 1451 STANDARDS FOR SMART SENSOR INTERFACE

There are new standards that are emerging for hardware architecture, software, and communication of modern intelligent (smart) sensors. These standards are making a revolutionary contribution to wireless instruments, instrumentation, and networks.

In traditional instruments, sensor output, largely analog, is processed further for measurement and display purposes. Incorporating microelectronics and microprocessor technologies in both sensors and instruments has increased

their functionality. In addition, new technology has emerged where sensor signal can be directly interfaced and processed without any dedicated circuits on digital platform. With integrated communication capabilities, networks that talk directly to the sensors have emerged. Sensor networking is becoming pervasive and is causing a major shift in the measurement area.

The rapid development and emergence of smart sensors and the associated networking technologies is making smart transducers an economical and attractive solution on many measurement and control applications. However, the existence of many incompatible networks and protocols makes it very difficult to interface a wide range of sensors. In addition, a sensor customized to interface with a particular network will not necessarily work with other networks. It is clear that no particular control network is becoming the industry standard at this point in time. It appears that a variety of networks will coexist to serve specific industries. Many manufacturers are uncertain of which networks to support and they are holding back on full-scale sensor production. This condition impedes widespread adoption of smart sensors and networking technologies, despite the compelling desire to build and use them.

In view of the situation, the Instrumentation and Measurement Society of the IEEE has set up the Sensor Technology Technical Committee to organize a series of workshops to provide an open forum to exchange ideas on sensor interface issues. As a result, a series of **IEEE 1451 standards** were proposed and accepted.

The **objective** of the IEEE 1451 standard for a smart transducer interface is to define a set of common communication interfaces for connecting transducers to a microprocessor-based system, to an instrument, or to a field network in a network-independent environment. IEEE 1451 is a set of standards that defines interfaces for network-based data acquisition and control of smart sensors and transducers. The aim of the IEEE 1451 is to make it easy to create solutions using existing networking technologies, connections, and common software architectures. The standard allows application software, field network, and transducer decisions to be made independently, thus providing flexibility in choosing products and vendors that are most appropriate for a particular application.

The ultimate goal of the IEEE 1451 standards is to provide some means of achieving transducer-to-network interchangeability and transducer-to-network interoperability. To achieve this goal, these standards provide clear ways of creating measurement and control devices at the process connection level. Sensors complying with these standards are expected to have onboard information on serial numbers, calibration factors, accuracy specifications, and so on. During installation, location information can also be loaded.

The family of IEEE 1451 standards can be divided into two basic parts:

- Defining a set of hardware interfaces for connecting transducers to a microprocessor or instrumentation system.

- Defining a set of software interfaces for connecting transducers to different networks while using existing network technologies.

Understandably, the IEEE 1451 standards cover wide spectrum applications. IEEE1451 is divided into six sections – 1451.1 to 1451.6.

IEEE 1451.1 is the first section and is known as the Network Capable Application Processor (NCAP) information model. This section is concerned with the software architecture that moves the intelligence to the device level. The 1451.1 standard uses object modeling to describe the behavior of the smart transducer. It supports transducers by way of a series of transducer blocks, which can be viewed as the input/output (I/O) driver abstraction of hardware. The application software that supports the operation of smart transducers can access the transducers through the application programming interface (API). The API is a set of standardized software function routines such as “IO_Read” to request a specific type of transducer electronic data sheet (TEDS) data, and “IO_control” to set an reset sensor parameters. The software creates a flexible environment and natural modules that allow engineers to think at the level of operational real-world systems, not at the level of programming languages. This type of approach creates object-oriented technology that makes open systems possible. Object-oriented systems produce devices that are much easier to adapt to new application demands. In this way, flexibility can be achieved so that systems can be assembled, reassembled, or modified quickly and transducers can easily be reconfigured and connected to different networks.

IEEE 1451.2 is the second section and it is concerned with transducers, microprocessor communication protocols (MCP), and TEDS. The IEEE 1451.2 standard defines the transducer data and electronic interface of digital information directly from the sensor to the system, thus creating a modular type of architecture. This modular architecture allows embedding of modules in any field network in an automatic and transparent manner. This section proposes a 10-wire interface, called the transducer independent interface (TII). Two of the wires are power and ground, other lines are allocated to data in, data out, and a clock. A dedicated sensor detect line allows the NCAP to determine if a sensor is plugged in for plug-and-play operations. The interrupt line allows the Smart Transducer Interface Module (STIM) to request service. The trigger line initiates a sensor measurement and acknowledges completion of the requested action.

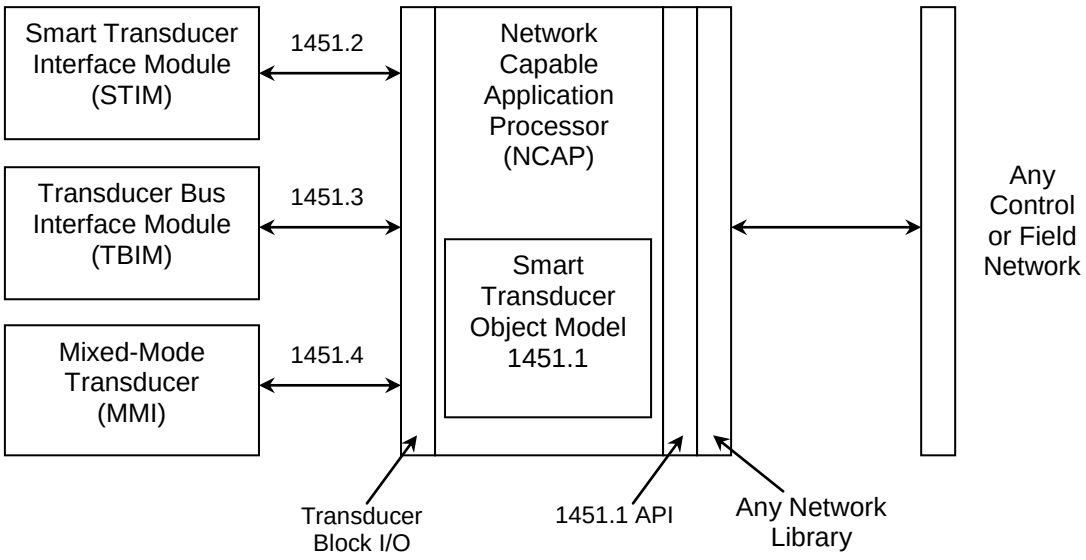

Figure 3-6 The relationship of the IEEE 1451 family of standards

IEEE 1451.3 is the third section and mainly looks after the digital communication and TEDS formats for distributed multidrop systems.

IEEE 1451.4 is the fourth section and is concerned with mixed-mode communication protocols and TEDS formats. IEEE 1451.4 establishes a universal system for the data that digital networks need to identify, characterize, interface with, and use for signals from analog sensors. This section aims to simplify the installation, creation, and maintenance of sensor networks. TEDS formats specified in the standard are self-identifiers, which are usually placed on chips embedded in sensors and actuators. Each TEDS node supplies the data a network needs to identify a device and interpret what is in its memory. TEDS formats are written in the Template Description Language described in IEEE 1451.4. The standard defines templates for commonly used devices such as accelerometers, strain gauges, microphones, and thermocouples.

IEEE 1451.5 is the fifth section and is concerned with wireless communication of sensors. This section specifies information that will enable 1451 compliant sensors and devices to communicate wirelessly. The IEEE is currently working on three different standards – 802.11, Bluetooth, and Zigbee.

IEEE 1451.6 is concerned with the information required from devices to operate on consolidated auto network (CAN) buses.

Figure 3-6 illustrates the relationships between different sections of the IEEE 1451 standard. IEEE 1451.1 deals with software. The other sections define the physical interfaces for smart transducer connectivity to networks as well as TEDS and its data format.

3.3. CONGESTION AND ERROR CONTROL

The objective of transport layer protocols is to provide reliability and other quality of service (QoS) services and guarantees for information transfer over inherently unreliable and resource-constrained networks. The following are the interrelated guarantees and services that may be needed in wireless sensor networks:

1. **Reliable delivery guarantee:** for some critical data, it may be necessary to ensure that the data arrive from origin to destination without loss.
2. **Priority delivery:** the data generated within the WSN may be of different priorities; for example, the data corresponding to an unusual event detection may have much higher priority than periodic background readings. If the network is congested, it is important to ensure that at least the high-priority data get through, even if the low-priority data have to be dropped or suppressed.
3. **Delay guarantee:** in critical applications, particularly those where the sensor data are used to initiate some form of actuation or response, the data packets generated by sensor sources may have strict requirements for delivery to the destination within a specified time.
4. **Energy-efficient delivery:** energy wastage during times of network congestion must be minimized, for instance by forcing any necessary packet drops to occur as close to the source as possible.
5. **Fairness:** different notions of fairness may be relevant, depending on the application. These range from ensuring that all nodes in the network provide equal amounts of data (for example, in a simple data-gathering application), to max-min fairness, to proportional fairness.
6. **Application-specific, data-centric quality of service:** in general, a data-centric QoS goal requires that the network as a whole provide as accurate a picture of the sensed environment as possible, given the bandwidth/energy resource constraints. In some applications, it may suffice just to ensure that some number of reports about specified events arrive at the destination in each unit time, regardless of which exact sources send the information.

In an ideal system, with large bandwidths and very low loss rates, the reliability and QoS guarantees being sought would not be difficult to provide. The task of designing transport mechanisms to provide the above functionalities is challenging because of a number of practical factors and limitations:

1. **Channel loss:** due to signal decay and multi-path fading effects, the error rates may be quite significant in WSN links. Further, these loss rates are very much a function of the exact location of the nodes as well

as the environment, and can therefore fluctuate greatly over space and time.

2. **Interference:** the error rates on wireless links are also very much a function of the level of interfering traffic in the vicinity. When traffic rates are high, packet losses can occur, even on otherwise good-quality links when the SINR drops below the successful reception threshold because of interference.
3. **Bandwidth limitation:** even if individual sensors generate data at a low rate compared with the maximum data rate available, the aggregate traffic of a large-scale network can be large enough to cause congestion in these low bandwidth networks. This happens particularly when all traffic is headed in the same destination, resulting in bottlenecks near the sink.
4. **Traffic peaks:** while the average data rate in a WSN is low most of the time, there may be a much higher peak traffic rate whenever important events are being detected. Such traffic peaks are likely to be highly correlated in both space and time – resulting in congested *hot-spots*.
5. **Node resource constraints:** the intermediate nodes in the network may also suffer from other constraints that can impact transport techniques: (a) low computational processing capabilities that preclude high-complexity approaches, (b) memory/storage constraints that limit the size of the message buffer (causing packet losses during congestion events), and, as always, (c) energy constraints that limit the amount of possible transmissions.

3.3.1. BASIC MECHANISMS AND TUNABLE PARAMETERS

Given the functionalities that need to be provided and the constraints that make this a challenging task, we must examine the various parameters that can be tuned by a WSN transport protocol.

1. **Rate control:** this refers to changing the rate at which packets are sent by a node, for instance by introducing a tunable wait time before which a packet is sent to the link layer for transmission. The rate control may be done purely at the source nodes, or at the intermediate nodes. In the latter case, there may be different rates for the route-through traffic and the traffic originating at the given mode. For implementations requiring fairness, the route-through traffic may have to be further divided on a per-child basis.
2. **Scheduling policy:** the transport buffer may be a regular first-in first-out queue (FIFO) in the simplest case. But, to support fairness, multiple FIFO queues served in a round robin fashion may be used. Similarly, to support priority and delivery guarantees, priority queues may be needed.

These may be also approximated by multiple prioritized queues to minimize complexity.

3. **Drop policy:** the queue drop policies can also have a significant impact on transport guarantees. The most basic technique is the traditional drop-tail policy employed commonly with FIFO queues whereby the packets are dropped from the tail of the queue. However, a more sophisticated policy may choose to drop low-priority packets, or even high-priority packets that cannot reach the destination before a required deadline. The drop policy may also take into account the distance travelled by a packet – it may be preferable to drop a packet near its origin as opposed to a packet that has travelled a great distance, for reasons of energy efficiency and fairness.
4. **Explicit notification:** intermediate nodes may monitor their queue lengths and when a threshold is exceeded send explicit signals to their children nodes to reduce the rates. This back pressure is typically propagated down to the sources to reduce rates. Similarly, if the queue length is sufficiently small, signals may be sent to increase rates. In other schemes, the sink itself may generate explicit notification asking nodes within the network to increase or decrease their traffic, depending upon whether application QoS needs are satisfied.
5. **Acknowledgements:** one approach to ensuring reliable transport is to use a sequence of packets and send negative acknowledgements (NACKs) hop-by-hop to trigger retransmissions for loss recovery. In some cases it may be better to use link-layer acknowledgements alone, but exercise some control over the maximum number of retransmission attempts, depending upon the priority of the packet.
6. **MAC backoff:** some of the transport mechanisms directly exert influence over link layer parameters, such as the backoff level. For instance, higher-priority packets may be given a shorter backoff interval. Another approach is to introduce an additional random backoff as a jitter to ease congestion in hot spots.
7. **Next-hop selection:** technically, forwarding decisions should be left to the network layer alone. However, sometimes the satisfaction of strict transport-level QoS objectives (such as in the case of real-time applications with strict deadlines) may requires an integrated and integrated cross-layer approach, whereby the congestion state information is used to determine the next hop.

3.3.2. CONGESTION CONTROL

Because of the low available bandwidth in WSN, congestion events are quite likely, particularly during peak traffic due to event detections. Several researchers have proposed different solutions for WSN congestion control.

3.3.2.1. Adaptive rate control (ARC)

The **adaptive rate control technique** (15) aims to provide a simple adaptive distributed rate control to ensure fairness and provide congestion control for a data-gathering tree. First, a distinction is drawn between originating traffic (for example, traffic originating at the given node) and route-through traffic (for example, traffic passing through that node that originated in upstream nodes below it on the tree).[2]

If the application rate of originating data is S, then the output originating rate is $S \cdot p$ where $p \in [0,1]$. Depending on the conditions, the rate is increased linearly by an amount α and reduced multiplicatively by a factor $\beta \in (0,1)$. Route-through traffic is given preference via a less aggressive reduction $\beta_{route} = 1.5\beta_{originate}$, and fairness is provided by ensuring that $\alpha_{originate} = \frac{\alpha_{route}}{(n+1)}$, where n is the number of descendant nodes generating the route-through traffic. Implicit acknowledgements are obtained through listening when the node on the next hop retransmits an earlier forwarded message – this signal is used to control the increase/decrease functions. Thus, while no explicit signals are sent, a congestion event would cascade down through a series of originating and route-through rate reductions, acting as an implicit back-pressure that slows the rates at all points below the location of congestion.

Another contribution of this work is to suggest the introduction of a jitter through a random initial backoff at the MAC level, to minimize contention/collisions of synchronized traffic generated by sensors sensing a common event.

3.3.2.2. Event-to-sink reliable transport (ESRT)

The **event to sink reliable transport mechanism** (16) takes a data-centric view of the problem of congestion control. It is assumed there are nodes in a given event region sending messages frequently. From an application point of view, it is assumed that there exists some threshold total reporting rate (number of reports received in a decision interval from any of the sources) that is needed for reliable event detection.

[2] The terminology of upstream/downstream can sometimes be confusing. Consider the one-way data-gathering, where the sink is the root to which all flows are directed. A node i is the parent of node j if it is the next hop from j towards the sink. We then refer to node j as being *below* i on the tree; however, we also say that node j is *upstream* from i, because its data are flowing towards the sink through i. Similarly node i is *above* j, but it is *downstream* from j.

ESRT is a closed-loop rate control technique. It is assumed that the sink node has a high-power radio that can be used to provide feedback to the sources directly. When the number of received packets is plotted as a function of the reporting frequency as in Figure 3-7 it shows a maximum at some point f_{max}, it then declines due to congestion. When the sink observes that the reliability is below threshold and the reporting rate is below f_{max}, it signals the nodes to increase their rate to meet the reliability criterion. If the reliability is above threshold, the reporting rate is reduced to save energy. In case the network is congested and the reliability is below threshold, the reporting rate is reduced aggressively. If the network provides a reliability that is within ϵ (which could be about 1%) of the reliability threshold, and no congestion is taking place, then no change in reporting rate is required. The signals from the sink are transmitted once every decision period. The sink is informed that congestion is taking place by the nodes in the network, which detect it by measuring local buffer occupancy and flag a congestion bit in packets sent towards the sink.

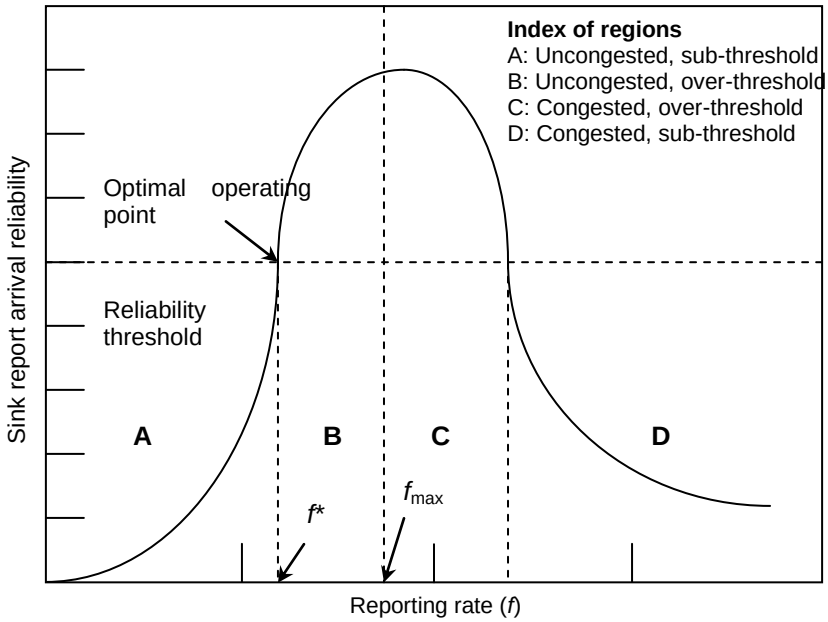

Figure 3-7 Sink report reliability curve used for congestion control in ESRT

3.3.2.3. Congestion detection and avoidance (CODA)

CODA (17) is a congestion control technique for WSN that comprises three mechanisms:

1. **Congestion detection:** buffer occupancy does not give a good indication of congestion if ARQ is not used, because the queue can potentially clear even if packets are being lost due collision. It is also possible for nodes to determine congestion by listening to the channel and to determine how busy/loaded it is; however, this can have significant energy cost. CODA therefore uses a periodic sampling technique with exponential averaging, to mitigate the impact of temporary fluctuations.

2. **Open-loop hop-by-hop back-pressure:** when congestion is detected, a suppression message (an explicit congestion notification) is propagated upstream towards the source. The suppression message is sent repeatedly. Nodes can respond to this message by dropping packets or reducing their rate. The suppression message propagates all the way upstream, or else nodes upstream will not be aware of the congestion.
3. **Closed-loop multi-source regulation:** when the congestion is persistent, it is helpful for the sink to play a role, by providing feedback for rate control. This is similar in spirit to ESRT. When the source event rate is higher than some fraction of the maximum throughput, the source may contribute to increased congestion by transmitting congestion notifications repeatedly. Therefore, is this case, the source flags a "regulate" bit to indicate to the sink that it is in this high-congestion regime. The sink then responds with periodic ACKs covering multiple event reports to regulate all sources associated with that event. If it finds the report rate to be lower than expected, it stops sending ACKs until the congestion clears, so that the sources can reduce their rate. In general, sources maintain, decrease, or increase their rate depending on how frequently they receive these ACKs.

3.3.2.4. Fair rate allocation

An explicit approach to congestion control with ensured fairness is given by Ee and Bajcsy (18). Their mechanism comprises the following three steps:

1. Determine the average transmission throughput r. One mechanism to do this is to measure the rate as the inverse of the time interval to transmit a single message. The interval is measured from the time when the transport layer sends the packet to the network layer reports that the packet has been transmitted. The rate measurement uses an exponential moving average.
2. Divide r among the number n of upstream devices (for example, on the subtree below the node on the data-gathering tree), so that the nominal per-node data packet generation rate is $r_{data} = \frac{r}{n}$. The size of the subtree is easily determined through a simple count technique. If the queue is full or overflowing, however, the rate is reduced to an even lower value.
3. Obtain the rate of the node's parent $r_{data,parent}$ through promiscuous listening or via a control message. Compare r_{data} with $r_{data,parent}$ and propagate the smaller rate upstream to the nodes in the subtree.

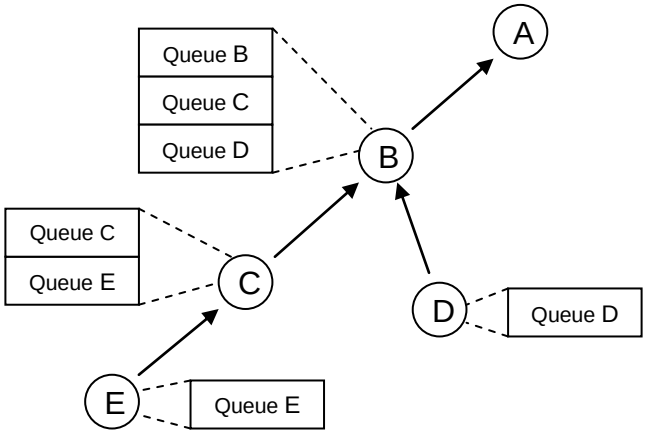

Figure 3-8 Multiple FIFO queues for fair delivery

Fairness is obtained by measuring and dividing the rate by the number of downstream nodes. To implement this, every node maintains separate FIFO queues for each child node as shown in Figure 3-8. Then, a probabilistic selection mechanism is employed to weigh the choice of packets so that the probability of choosing a queue from which to transmit a packet is proportional to the number of nodes serviced by that queue.

3.3.2.5. Fusion

The **fusion technique** (19) is actually a combination of three techniques to mitigate congestion:

1. **Hop-to-hop flow control**: this technique is very similar to the open-loop flow control in CODA.
2. **Source rate limiting**: this is similar in spirit to the ARC technique, implemented slightly differently, as follows. It is assumed that all nodes offer the same traffic load (as assumption that can be relaxed with the addition of more information collected about different node rates). Each sensor listens to the traffic its parent forwards to estimate *N*, the total number of unique sources routing through that parent. A token bucket scheme is then employed whereby a node gathers a token once it hears its parent forward *N* messages. The sensor can transmit its own generated data only if its bucket (which has a maximum token limit) contains at least one token. This approach ensures that the sensor sends its own data fairly, at the same rate as each of its descendants.
3. **Congestion-adaptive backoff control**: the length of each node's randomized backoff is determined as a function of its local congestion state. A sensor node that is congested uses a shorter backoff window, allowing it to win the contention period with high likelihood. This will help backlogged queues drain faster.

Fusion is evaluated on a testbed of 55 motes. Through an extensive experimental evaluation, the fundamental need for congestion control mechanisms is demonstrated (as otherwise efficiency and fairness are severely degraded). It is found that as loads increase or when channel variations cause reduction in bandwidth, nodes must reduce send rates so that the network avoids congestion collapse. The fusion study also shows that the hop-by-hop flow control and the backoff control with simple queue occupancy-based congestion detection, work well together under a wide range of workloads, and the rate control provides a substantial degree of fairness.

CHAPTER 4. SECURITY

4.1. SECURITY FOR WIRELESS SENSOR NETWORKS

Technology is a double-edged sword. Just as nations and societies employ technology to protect themselves, their adversaries employ technology to counter and mitigate the security afforded by these new and innovative protective measures. Consequently, steps need to be taken to ensure the security of these protective technologies. In the case of a WSN, the underlying data network, the physical sensors, and the protocols used by the WSN all need to be secured.

To exemplify this problem, consider the omnipresent threat to **distributed control systems** (DCS) (20) and **supervisory control and data acquisition systems** (SCADA) (21). These control systems are often wireless and share some similarities with WSNs. The simplest of these systems collect data associated with some metric and, based upon its value, cause some event to occur. Events include the closing or opening of railway switches, the cycling of circuit breakers, and the opening and closing of valves. Complex DCSs and SCADAs accept data from multiple sensors and using controllers, govern a wide array of devices and events in response to the measured data.

On April 23th, 2000, Vitek Boden was arrested in Queensland, Australia and was eventually found guilty of computer hacking, theft and causing significant environmental damage (22). This criminal case is of particular interest because, to date, it is the only known case where someone has employed a digital control system to deliberately and maliciously cause damage. Using a transmitter and receiver tuned to the same frequencies as the SCADA controlling Queensland's municipal water system and a laptop computer he effectively became the master controller for the municipality's water system. He then proceeded to wreak havoc upon the city's water supply by causing the release of raw sewage into local waterways and green spaces.

DCSs and SCADAs were not designed with public access in mind, consequently they lack even rudimentary security controls. Moreover, if DCSs and SCADAs were constrained by security controls, they may fail to work properly because their timing and functionality are predicted upon unfettered communications between system components.

Protocol	**Relevant attacks**
TinyOS beaconing	Bogus routing information, selective forwarding, sink-holes, Sybil, wormholes, HELLO foods
Directed diffusion and its multipath variant	Bogus routing information, selective forwarding, sink-holes, Sybil, wormholes, HELLO foods
Geographic routing (GPSR, GEAR)	Bogus routing information, selective forwarding, Sybil
Minimum cost forwarding	Bogus routing information, selective forwarding, sink-holes, wormholes, HELLO foods
Clustering based protocols (LEACH, TEEN, PEGA-SIS)	Selective forwarding, HELLO floods

Rumor routing	Bogus routing information, selective forwarding, sink-holes, Sybil, wormholes
Energy conserving topology maintenance (SPAN, GAF, CEC, AFECA)	Bogus routing information, Sybil, HELLO foods

Table 4-1 Summary of attacks against proposed sensor networks routing protocols

Similarly, the security requirements of a WSN impose costly constraints and overhead due to a sensor node's limited power supply and computational resources. In a manner not unlike the Boden case, a WSN bereft of security controls will most likely suffer the same fate as did Queensland's municipal water system.

The goal of this chapter is to present a framework for implementing security in WSNs.

4.1.1. THREATS TO A WSN

There are many vulnerabilities and threats to a WSN. They include outages due to equipment breakdown and power failures, non-deliberate damage from environmental factors, physical tampering, and information gathering. There have been identified the following threats to a WSN:

Passive information gathering: if communications between sensors or between sensors and intermediate nodes or collection points are in the clear, then an intruder with an appropriately powerful receiver and well designed antenna can passively pick off the data stream.

Subversion of a node: if a sensor node is captured it may be tampered with, electronically interrogated and perhaps compromised. Once compromised, the sensor node may disclose its cryptographic keying material and access to higher levels of communication and sensor functionality may be available to the attacker. Secure sensor nodes, therefore, must be designed to be tamper proof and should react to tampering in a *fail complete manner* where cryptographic keys and program memory are erased. Moreover, the secure sensor needs to be designed so that its emanations do not cause sensitive information to leak from the sensor.

False node: an intruder might "add" a node to a system and feed false data or block the passage of true data. Typically, a false node is a computationally robust device that impersonates a sensor node. While such problems with malicious hosts have been studied in distributed systems, as well as ad-hoc networking, the solutions proposed (group key agreements, quorums and per hop authentication) are in general too computationally demanding to work for sensors.

Node malfunction: a node in a WSN may malfunction and generate inaccurate or false data. Moreover, if the node serves as an intermediary, forwarding data on behalf of other nodes, it may drop or garble packets in transit. Detecting and culling these nodes from the WSN becomes an issue.

Node outage: if a node serves as an intermediary or collection and aggregation point, what happens if the node stops functioning? The protocols

employed by the WSN need to be robust enough to mitigate the effects of outages by providing alternate routes.

Message corruption: attacks against the integrity of a message occur when an intruder inserts them between the source and destination and modify the contents of a message.

Denial of service (DoS): a denial of service attack on a WSN may take several forms. Karlof and Wagner (23) identify several DoS attacks including: “Black hole”, “Resource exhaustion”, “Sinkholes”, “Induced routing loops”, “Wormholes”, and “Flooding” that are directed against the routing protocol employed by the WSN.

Traffic Analysis: although communications might be encrypted, an analysis of cause and effect, communications patterns and sensor activity might reveal enough information to enable an adversary to defeat or subvert the mission of WSN. Addressing and routing information transmitted in the clear often contributes to traffic analysis. We further address traffic analysis in the following subsection.

Traffic analysis is an issue that has occasionally attracted the attention of authors writing about security in networks (24). Traffic analysis is the term used for the process of inferring information about the communications of an encrypted target network.

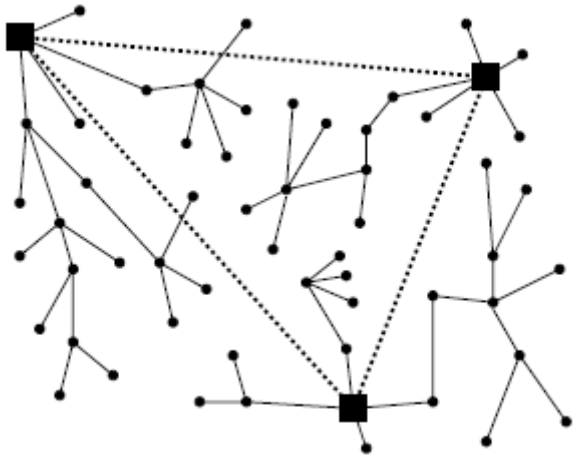

Figure 4-1 A representative sensor network architecture

Classically, traffic analysis is countered by communication systems employing traffic flow security. In this mode, the system transmits an encrypted stream continuously, encrypting idle messages when there is no valid traffic to be sent. In this way, the unauthorized listener cannot tell when the parties are actually communicating and when they are not, and is thus unable to make traffic analysis deductions. With radio links in years past, traffic flow security was often employed. Doing so did not add significant expense – it occupied a radio frequency continuously, but radio spectrum was not at a premium, nor was electric power for the transmitter a concern.

With WSNs having limited-energy nodes, the practicality of traffic flow security becomes quite problematic. Now, the cost of sending dummy traffic results directly in a reduction in the lifetime of the node. The energy required for communications is typically the dominant factor in battery lifetime of a node, so the lifetime is reduced by the ratio of dummy traffic to real traffic, which must be substantially greater than 1, to provide any useful traffic flow security protection. This will be intolerable in virtually all cases; consequently traffic flow security for WSNs is in need of further research. The effects of traffic analysis may be partially mitigated by encrypting the message header that contains addressing information, however further research is needed.

4.1.2. WSN OPERATIONAL PARADIGMS AND VULNERABILITIES

Some models of operation are simple; the sensor takes some measurement and blindly transmits the data. Other operational models are complex and include algorithms for data aggregation and data processing. In order to discuss security measures for a WSN sensibly, one must know the threats that must be defended, and equally important, those that need not be provided for. It makes no sense to even attempt to design protection against an all-powerful adversary, or even against an adversary who gets the last move in a spy-vs-spy, move-countermove game. One must select a model of the adversary's capabilities and work to that.

The following briefly describes the operational paradigms that a WSN may use. In each case, we assume the presence of a base station, or controller.

4.1.2.1. Simple Collection and Transmittal

The sensors take periodic measurements and transmit the associated data directly to the collection point. Transmission occurs either immediately following data collection or is scheduled at some periodic interval. In this paradigm each node is only concerned with its transmission to the base station, which is assumed to be within range. Thus, any notion of routing or co-operation among nodes is absent from this paradigm.

This operational paradigm is vulnerable to attacks directed against the Link Layer. DoS attacks include *jamming* the radio frequency and *collision induction*. It is also vulnerable to *spoofing* attacks in which a counterfeit data source broadcasts spurious information. If the data is considered to be sensitive and it is not encrypted, then a loss of confidentiality may occur if someone passively monitors the transmission emanating from the WSN. This paradigm and all the following are also susceptible to physical attacks – capture of a node and subsequent subversion. Such threats are countered by tamper resistant technologies, which may transmit an alert and/or self destruct when tampering is detected. Discussion of these techniques is beyond the scope of this chapter. Replay attacks in which an adversary transmits old and/or false data to nodes in the WSN can also be mounted on the six paradigms discussed here.

The primary security requirements of this paradigm are data access, authentication and data confidentiality. Confidentiality can be ensured by the use of data encryption. Symmetric key encryption methods, such as DES (25), are suitable for this paradigm. Authentication is implicitly ensured by the use of pre-deployed keys that are shared between, and unique to, the collection point and each individual sensor node (26). Each node uses its key to encrypt data before transmission; the collection point decrypts the data using the shared key corresponding to that node.

4.1.2.2. Forwarding

Sensors collect and transmit data to one or more neighboring sensors that lie on a path to the controller. In turn, the intermediate sensors forward the data to the collection point or to additional neighbors. Regardless of the length

of the path, the data eventually reaches the collection point. Unlike the first paradigm, co-operation among nodes in “routing” the data to the base station is part of this paradigm. That is, a node that receives data intended for the base station attempts to transmit the same toward the latter, instead of throwing the data away.

In addition of previous vulnerabilities mentioned in the Simple Collection and Transmittal paradigm, this method is also vulnerable to *Black Hole*, *Data Corruption* and *Resource Exhaustion* attacks. In a Black Hole attack, the sensor node that is responsible for forwarding the data drops packets instead of forwarding them. A Data Corruption attack occurs when the intermediate node modifies transient data prior to forwarding it. These attacks require that the node is subverted or that a foreign, malicious node is successfully inserted into the network. A Resource Exhaustion attack occurs when an attacker maliciously transmits an inordinate amount of data to be forwarded, consequently causing the intermediate node(s) to exhaust their power supply.

The forwarding process may span multiple sensor nodes on the path between the source node and the collection point. Thus, this paradigm requirement to minimize energy consumption complicates possible security mechanisms.

A single shared key between the collection point and a sensor node is no longer sufficient to ensure reliable transmission of a node’s data to the collection point. This is caused because a node cannot trust the next (or the previous one) to the collection point.

To resolve this problem, we can use a system (26) that utilizes pre-built headers encrypted under the intermediate node’s key. At the origin, the entire frame intended for the controller is encrypted under the sender’s key and inserted into another frame that is prepended with the pre-built header and broadcast. When the intermediate node receives it, it strips off the prepended header and re-broadcasts the frame. Then, the controller receives and decrypts it.

We can see that this solution has limited scalability in terms of number of hops; as the number of hops increases, the number of pre-built headers to prepend also increases leading to increased message size.

4.1.2.3. Receive and Process Commands

In this paradigm, sensors receive commands from a controller, either directly or via forwarding, and configure or re-configure themselves based on the commands. This ability to process commands helps in controlling the amount of data handled by the WSN.

In addition to being vulnerable to all the aforementioned attacks, the Receive and Process Commands paradigm is also vulnerable to attacks where an adversary impersonates the controller and issues spurious commands.

The previous paradigms discussed describe a *many-to-one* communication mode, designated exclusively for *unsolicited* data transmissions. In this model, the communication paradigm changes from being exclusively *many-to-one* to now include *one-to-many* communication, which means that whereas in the former, the data transmitted was intended *only* for the base station, in the latter, the data is applicable to one or more sensor nodes.

As expected, the cost of improved energy efficiency in this paradigm requires a more complicated security model. In the previous paradigm, the main security issue was to build *trust* among sensor nodes so that they could cooperate in data transmission to the collection point. This point is now extended to the controller.

Now, a sensor has to authenticate the command it received as being broadcast by the controller. This issue can be addressed by the use of *broadcast authentication* as discussed in the SPINS project (27) or by the use of shared secrets between the controller and the individual sensor nodes.

Also, a sensor has to verify the integrity of the message it receives from its neighbors; it has to know that the message was not tampered with by an intermediate node. This can be resolved by distributing encrypted identities of sensor nodes within radio range of the controller among those sensor nodes that are beyond radio range of the controller by using pre-built headers, presented in the previous paradigm.

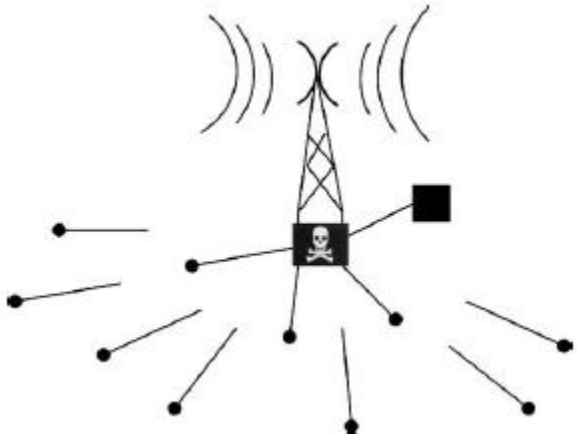

Figure 4-2 HELLO flood attack against TinyOS beaconing.

A lap-top-class adversary that can retransmit a routing update with enough power to be received by the entire network leaves many nodes stranded. They are out of normal radio range from the adversary but have chosen it as their parent.

4.1.2.4. Self-Organization

Upon deployment, the WSN self organizes, and a central controller(s) learns the network topology. Knowledge of the topology may remain at the controller or it may be shared, in whole or in part, with the nodes of the WSN. This paradigm may include the use of more powerful sensors that serve as cluster heads for small coalitions within the WSN.

This paradigm requires strong notion of routing, therefore, in addition to being vulnerable to all the previously listed attacks, this paradigm is vulnerable to attacks against the routing protocol. These attacks include *Induced Routing Loops*, *Sinkholes*, *Wormholes* and *HELLO Flooding* (23).

The previous paradigm presented, describes secure bi-directional communications between the controller and the sensor nodes. The present paradigm extends the bi-directional communication model by introducing the concept of *self-organization*. This paradigm requires that the WSN achieve organizational structure without human intervention. It consists of three primary tasks: *node discovery*, *route establishment* and *topology maintenance*.

In the node discovery process, the controller or a sensor node broadcasts a discovery message; for example, a HELLO message. In response to it, nodes unicast a message indicating their presence in the proximity of the broadcaster; for example, a HELLO-REPLY message. This message sequence is sufficient for establishing a secure single-hop WSN.

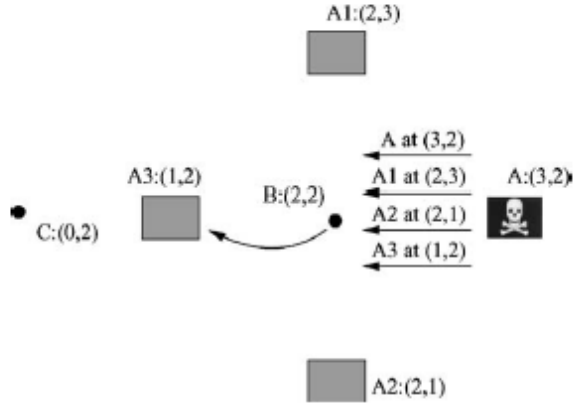

Figure 4-3 The Sybil attack against geographic routing.

Adversary A at actual location (3,2) forges location advertisements for non-existent nodes A1, A2, and A3 as well as advertising its own location. After hearing these advertisements, if B wants to send a message to destination (0,2), it will attempt to do so through A3. This transmission can be overheard and handled by the adversary A.

Following node discovery, routes between the sensor nodes and the controller must be established. In order to ensure continuous connectivity, multiple routes between a pair of nodes may be established. Most existing routing protocols for WSNs (23) are vulnerable to a host of attacks including flooding, wormhole, sinkhole and Sybil (28) attacks. Consequently, this routing protocol is extremely important and needs to be secure.

4.1.2.5. Data Aggregation

Nodes in the WSN aggregate data from downstream nodes, incorporating their own data with the incoming data. The composite data is then forwarded to a collection point.

This paradigm is particularly vulnerable to replay attacks because the *authentication* of its downstream peers becomes an issue. In the previous paradigms, the authentication of the sensor node was left to the controller, which is not an issue because controllers are robust and considerably more powerful than the sensor nodes. In this paradigm, each sensor node that utilizes data from another sensor node now cannot just forward the data as received, and therefore must ensure that the data is provided by an authorized member of the WSN.

So far, our discussion of WSNs and security has assumed that all sensor nodes transmit their data directly to the base station in either a solicited or unsolicited manner. Under this assumption, the sensors are not dependent upon the integrity or authenticity of the data. This results in hundreds or perhaps thousands of independent *data streams*. An important problem in WSNs is to control these data streams so that unnecessary data transmissions can be eliminated and the collection point can be prevented from becoming a bottleneck.

The prevailing solution to this problem is to *aggregate* or *fuse* data within the WSN and transmitting an aggregate of the data to the controller (29). The idea therefore, is to allow a sensor node to transmit its data to its neighbors, or some subset thereof. In turn, some algorithm controls which node will combine the data received from its neighbors and forward it toward the controller. This data aggregation process results in a substantial energy savings in the WSN.

4.1.2.6. Optimization: Flexibility and Adaptability

Predicated upon their own measurements and upon the values of incoming data, this paradigm requires that the sensors in the WSN make decisions. For example, a decision may be whether to perform a calculation, or if the cost is less, acquire the needed value from a peer, provided that the peer has the value and that knowledge is known in advance by the requester.

This operational paradigm shares the same security concerns and issues as does the Data Aggregation paradigm.

The WSN paradigms previously considered, focus on the data gathering and reporting functions of a WSN. The nodes in the WSN are not concerned with the *semantics* of the data they have obtained through the sensing task. The only concern is that is has to be transmitted elsewhere, possibly for further analysis.

4.2. KEY DISTRIBUTION TECHNIQUES FOR SENSOR NETWORKS

The general *key* distribution problem refers to the task of distributing secret keys between communicating parties in order to facilitate security properties such as communication secrecy and authentication.

In sensor networks, key distribution is usually combined with initial communication establishment to bootstrap a secure communication infrastructure from a collection of deployed sensor nodes. These nodes may have been pre-initialized with some secret information but would have had no prior direct contact with each other. This combined problem of key distribution and secure communications establishment is sometimes called the *bootstrapping problem*. A bootstrapping protocol must not only enable a newly deployed sensor network to initiate a secure infrastructure, but it must also allow nodes deployed at a later time to join the network securely. This is a highly challenging problem due to many limitations of sensor network hardware and software.

In this chapter, several well-known methods of key distribution will be discussed.

4.2.1. SENSOR NETWORK LIMITATIONS

The following characteristics of sensor networks complicate the design of secure protocols for sensor networks, and make the bootstrapping problem challenging.

Vulnerability of nodes to physical capture. Sensor nodes may be deployed in public or hostile locations (such as public buildings or forward battle areas) in many applications. Furthermore, the large number of nodes that are deployed implies that each sensor node must be low-cost, which makes it difficult for manufacturers to make them tamper-resistant. This exposes sensor nodes to physical attacks by an adversary. In the worst case, an adversary may

be able to undetectably take control of a sensor node and compromise the cryptographic keys.

Lack of a-priori knowledge of post-deployment configuration. If a sensor network is deployed via random scattering, the sensor network protocols cannot know beforehand which nodes will be within communication range of each other after deployment. Even if the nodes are deployed by hand, the large number of nodes involved makes it costly to pre-determine the location of every individual node. Hence, a security protocol should not assume prior knowledge of which nodes will be neighbors in a network.

Limited bandwidth and transmission power. Typical sensor network platforms have very low bandwidth. For example, the UC Berkeley Mica platform's transmitter has a bandwidth of 10 Kbps, and a packet size of about 30 bytes. Transmission reliability is often low, making the communication of large blocks of data particularly expensive.

4.2.2. THE PROBLEM OF BOOTSTRAPPING SECURITY IN SENSOR NETWORKS

A bootstrapping scheme for sensor networks needs to satisfy the following requirements:

- Deployed nodes must be able to establish secure node-to-node communication
- Additional legitimate nodes deployed at a later time can form secure connections with already-deployed nodes.
- Unauthorized nodes should not be able to gain entry into the network, either through packet injection or masquerading as a legitimate node.
- The scheme must work without prior knowledge of which nodes will come into communication range of each other after deployment.
- The computational and storage requirement of the scheme must be low, and the scheme should be robust to DoS attacks from out-of-network sources.

4.2.3. EVALUATION METRICS

Sensor networks have many characteristics that make them more vulnerable to attack than conventional computing equipment. Simply assessing a bootstrapping scheme based on its ability to provide secrecy is sufficient. Listed below are several criteria that represent desirable characteristics in a bootstrapping scheme for sensor networks.

Resilience against node capture. It is assumed the adversary can mount a physical attack on a sensor node after it is deployed and read secret information from its memory. A scheme's resilience toward node capture is calculated by estimating the fraction of total network communications that are compromised by a capture of *x* nodes *not including* the communications in which the compromised nodes are directly involved.

Resistance against node replication. Whether the adversary can insert additional hostile nodes into the network after obtaining some secret information (for example, through node capture or infiltration). This is a serious attack since the compromise of even a single node might allow an adversary to populate the network with clones of the captured node to such an extent that legitimate nodes could be outnumbered and the adversary can thus gain full control of the network.

Revocation. Whether a detected misbehaving node can be dynamically removed from the system.

Scalability. As the number of nodes in the network grows, the security characteristics mentioned above may be weakened.

Each bootstrapping protocol usually involves several steps. An *initialization* procedure is performed to initialize sensor nodes before they are deployed. After the sensor nodes are deployed, a *key setup* procedure is performed by the nodes to set up shared secret keys between some of the neighboring nodes to establish a secure link.

4.2.4. USING A SINGLE NETWORK-WIDE KEY

The simplest method of key distribution is to pre-load a single network-wide key (also known as symmetric encryption) onto all nodes before deployment. After deployment, nodes establish communications with any neighboring nodes that also possess the shared network key. This can be achieved simply by encrypting all communications in the shared network-wide key and appending *message authentication codes* (MACs) to ensure integrity.

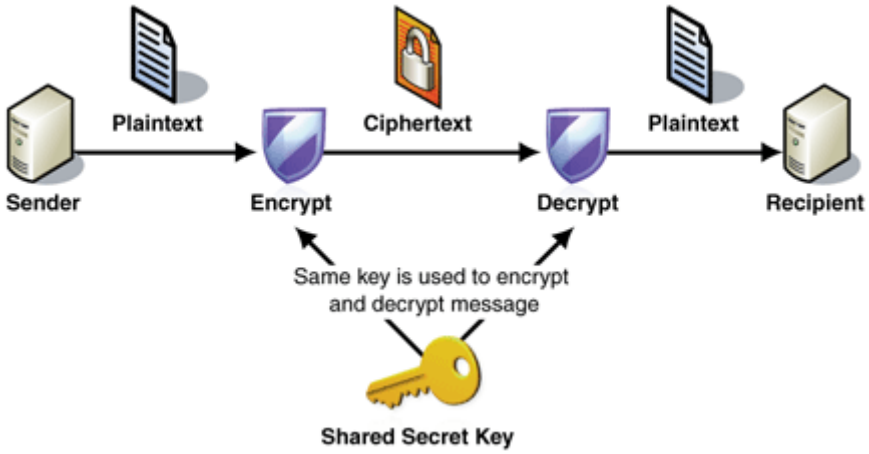

Figure 4-4 The process of symmetric encryption

The properties of the single network-wide key approach are as follows:

Minimal memory storage required. Only a single cryptographic key is needed to be stored in memory.

No additional protocol steps are necessary. The protocol works without needing to perform key discovery or key exchange. This represents savings in code size, node hardware complexity and communication energy costs.

Resistant against DoS, packet injection. The MACs guard against arbitrary packet injection by an adversary that does not know the network-wide

key *k*. replay attacks can be prevented by including the source, destination, and a timestamp in the message. A worst case denial of service attack would be to replay large numbers of packets in order to force nodes to perform many MAC verifications. However, symmetric cryptographic MAC verifications are extremely fast and this kind of DoS attack would not be very effective at preventing normal operation.

The main drawback of the network-wide key approach is that if just a single node is compromised, the entire network loses all its security properties, since the network-wide key is now known to the adversary. This makes it impractical except in two possible scenarios.

The nodes are tamper-resistant. In this case, tamper resistance is incorporated into the sensor nodes such that it becomes impractical for an adversary to attempt extraction of the network-wide key. In general, this is impractical for critical sensor systems (such as security or safety applications) since adversaries attacking these systems will have a large incentive to defeat the tamper-resistance. However, low cost tamper-resistance may actually be feasible for non-critical sensor applications such as domestic temperature monitoring.

No new nodes are ever added to the system after deployment. In this case, the sensor nodes use the network wide key to encrypt unique *link keys* which are exchanged with each of their neighbors. For example, node *A* might generate a unique key k_{AB} and send it to *B* under encryption by the network-wide key. Once the link keys are in place, all communications are encrypted using the appropriate link keys and the network-wide key is erased from the memory of the nodes. Any node which is subsequently compromised thus reveals no secret information about the rest of the network. In general such an approach is also some-what impractical since it is usually desirable to have the ability to add new nodes to the network to replace failed or exhausted nodes. A possible way to address this would be to perform a large-scale audit of all sensor nodes prior to every phase of adding new nodes. The audit would have to ensure that every node's hardware and software has not been altered maliciously. Once this has been ascertained, a new network-wide key could be distributed to the nodes to enable addition of the new nodes. However, such a comprehensive audit would probably be too costly to be practical in most applications.

4.2.5. USING ASYMMETRIC CRYPTOGRAPHY

The favored method of key distribution in most modern computer systems is via asymmetric cryptography, also known as public key methods. If sensor node hardware is able to support asymmetric cryptographic operations, then this is a potentially viable method of key distribution.

A brief outline of a possible public-key method for sensor networks is as follows. Prior to deployment, a master public/private keypair, (k_M, k_M^{-1}) is first generated. Then, for every node *A*, its public/private keypair (k_A, k_A^{-1}) is generated. This keypair is stored in node *A*'s memory along with the master public key k_M and the master key's signature on *A*'s public key. Once all the nodes are initialized in this fashion, they are ready for deployment.

One the nodes have been deployed, they perform key exchange. Nodes exchange their respective public keys and master key signatures. Each node's public key is verified as legitimate by verifying the master key's signature using the master public key which is known to every node in the network. Once the public key of a node has been received, a symmetric link key can be generated and sent to it, encrypted by its public key. Upon reception of the session key, key establishment is complete and the two nodes can communicate using the symmetric link key.

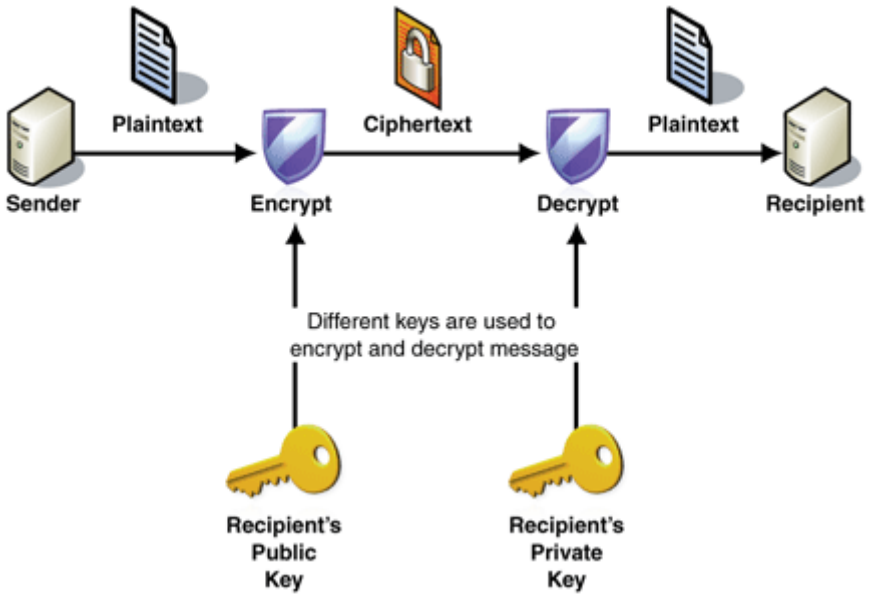

Figure 4-5 The process of asymmetric encryption

The properties of this approach are the next ones:

Perfectly resilient against node capture. Capture of any number of nodes does not expose any additional communications in the network, since these nodes will have no knowledge of any secret link keys besides the one that they are actively using.

Possible to revoke known compromised keypairs. Revocation can be performed by broadcasting the entire revoked keypair, signed by the master key. Nodes receiving the broadcast can authenticate it as coming from the central authority and ignore any future communications purporting to originate from the revoked keypair. If a large number of keypairs are to be revoked, the master keypair itself could be updated by broadcasting an updated new master key signed by the old master key, then unicasting the new master key's signature on each of the legitimate public keys (the unicast is encrypted in the respective nodes' public keys). The revoked keypairs will not be signed by the new master key and will thus be rejected as invalid by all nodes in the network.

Fully scalable. Signature schemes function just as effectively regardless of the number of nodes in the network.

However, using asymmetric cryptography has its disadvantages:

Dependence on asymmetric key cryptographic hardware or software. All known asymmetric cryptographic schemes involve computationally intensive mathematical functions, for example, modular exponentiation of large numbers. Basic sensor node CPU hardware, however,

is extremely limited, often lacking even an integer multiply instruction. In order to implement asymmetric cryptography on sensor nodes, it is necessary to either implement dedicated cryptographic hardware on a sensor node, thus increasing its hardware cost, or encode the mathematical functions in software, thus reading the amount of code space and memory for sensing functions. Given that asymmetric cryptography is only used for key setup upon node deployment, this represents a tiny fraction of the sensor node's lifetime. Significantly increasing the cost of a node is thus difficult to justify.

Vulnerability to denial-of-service. Asymmetric cryptographic operations involve significant amounts of computation and it can take up to minutes of intense processing for a sensor node to complete verification of a single signature. Thus, the nodes are vulnerable to a battery exhaustion denial of service attack where they are continuously flooded with illegal signatures. The sensor nodes will attempt to verify each signature, consuming valuable battery power in the process, while not being able to establish connections with the legitimate nodes. Since this denial-of-service can only occur during the key establishment phase, the sensor network is only vulnerable when it is newly deployed or when additional nodes are being added to the network. Increased monitoring of the site for adversarial transmissions during those times could alleviate the DoS problem. However, such DoS attacks make it infeasible for asymmetric cryptography to be used in sensor networks deployed in large or unmonitored areas such as battlefields.

No resistance against node replication. Once a node is captured, its keypair can be used to set up links with every single one of the nodes in the networks, effectively making the node "omnipresent". This could potentially put it in a position to subvert the routing infrastructure of alter sensor network operation. Such an attack is not difficult to prevent with some added countermeasures. For example, each time a node forms a connection with another node, they could both report the event to a base station encrypted using their master public key. Any node that has an exceptionally high degree could then be immediately revoked.

4.2.6. USING PAIRWISE KEYS

In this approach, every node in the sensor network shares a unique symmetric key with every other node in the network. Hence, in a network of n nodes, there are a total of ${}^{n}C_{2}$ unique keys. Every node stores $n-1$ keys, one for each of the other nodes in the network.

After deployment, nodes must perform key discovery to verify the identity of the node that they are communicating with. This can be accomplished with a challenge/response protocol.

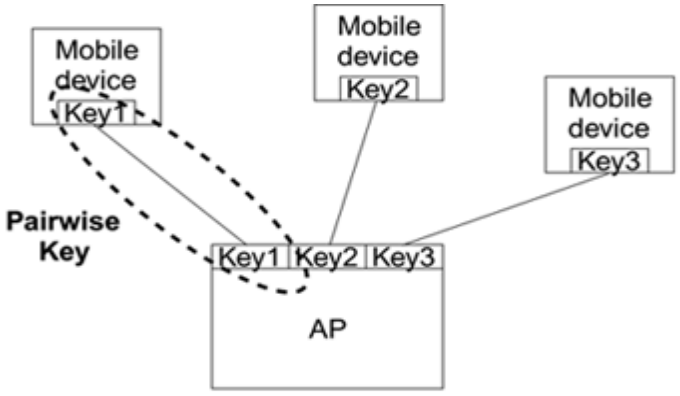

Figure 4-6 Pairwise Key

The properties of this approach are the next:

Perfect resilience to node capture. Similar to the asymmetric cryptographic scheme, any node that is captured reveals no information about the communications being performed in any other part of the network. Its pairwise keys could be used to perform a node replication attack throughout the network, but this could be countered using the same method as described for asymmetric cryptography in the previous section.

Compromised keys can be revoked. If a node is detected to be compromised, its entire set of $n-1$ pairwise keys is simply broadcast to the network. No authentication is necessary. Any node that hears a key in its set of pairwise keys broadcast in the open immediately stops using it. This effectively cuts off the revoked node from the network.

Only uses symmetric cryptography. The pairwise keys scheme achieves many of the benefits of using asymmetric cryptography without needing dedicated hardware of software to before the more complex asymmetric cryptographic primitives. This not only makes the network less vulnerable to energy- sapping denial of service attacks.

The main problem with the pairwise keys scheme is poor scalability. The number of keys that must be stored in each node is proportional to the total number of nodes in the network. With an 80 bit key, a hundred node network will require 1 kB of storage on each node for keys alone. This means it will probably be prohibitively expensive to scale this scheme up to thousands of nodes.

4.3. SECURITY IN SENSOR NETWORKS: WATERMARKING TECHNIQUES

WSN are distributed embedded systems where each unit is equipped with a certain amount of computation, communication, storage, and sensing resources. In addition each node may have control over one or more actuators and input/output devices such as displays. A variety of applications for sensor networks are envisioned, starting from nano-scale device networks to interplanetary scale distributed systems. In many senses, WSN are a unique type of systems which have unique technical and operational challenges.

Among these, security and privacy are most often mentioned as the key prerequisite for actual deployment of sensor networks.

There are at least three major reasons why security and privacy in WSN is such an important topic. The first one is that sensor networks are intrinsically more susceptible to attacks. They are often deployed in uncontrolled and sometimes even hostile environments. Wireless communication on a large scale can be easily observed and interfered with. WSN nodes are both complex component systems with numerous weak points from a security point of view. In addition, there are several constrained in terms of energy and therefore extensive on-line security checking is not viable. Finally, sensors can be manipulated even without interfering with the electronic subsystem of the node and actuators can pose strong safety and hazard concerns.

The second argument that emphasizes the role of security in WSN is the importance of protecting typical applications. WSN cannot only have data about one or more users, but can also contain a great deal of information about their past and even future actions. In addition, they may contain significant amounts of information about a user's physiological and even psychological profiles. Furthermore, once the sensors are equipped with actuators both the sensors and the environment can be impacted in a variety of ways.

The third reason for security in WSN is, in a sense, the most scientific and engineering based reason. WSN require new concepts and a new way of thinking with respect to security, privacy, digital rights management, and usage measurement. The Internet was a great facilitator of computer and communication security on a large scale. Note that the Internet itself created opportunities for new types of attacks such as DoS and intrusion detection. It also created new conceptual techniques on how to defend the Internet infrastructure. For example, Honeypots are now widely used to obtain information about the behavior and thinking of an attacker (30). It is easy to see that WSN will further accentuate these trends. For example, denial of sleep attacks will be brought to a new level of importance.

In this section it is explained how to develop and evaluate watermarking schemes for the protection of the data and information in sensor networks.

4.3.1. MOBILITY AND SECURITY

In this section, the interplay between mobility and security is discussed. More specifically, focused on security and mobility with respect to multi-hop wireless networks. This area of research is still in the very early phases of its development and only a few research results have been reported. But a very rapid growth is expected in this direction in the near future.

It is expected that a large percentage of wireless networks will be mobile. There are two main reasons for this prediction. The fist one is that in many applications one can achieve significantly higher performance if mobility is provided and exploited. For example, sensor nodes may move closer to the phenomenon or event of interests. The second reason is even more compelling: sensor networks associated with individual users, cars, trains, airplanes and other transportation vehicles are intrinsically mobile.

It is not clear whether mobility makes security in wireless sensor networks easier or more difficult to be achieved. From one point of view, it

makes it easier because one can leverage on conducting specific security tasks on the nodes of interest which are in favorable locations. From the other point of view, it makes it more difficult due to the dynamic structure of the topology and potential introduction and departure of any given node.

Until recently, mobility received relatively little attention in wireless ad-hoc networks. The main reason for this is that the majority of standard tasks in wireless ad-hoc networks are easier to address in the static scenario. Even more importantly, there experimental data that would enable realistic modeling of mobility does not exist. Essentially, all current models are of random statistical nature (31).

4.3.2. WATERMARKING

The notion of intellectual property protection and specifically watermarking has been widely studied for items such as text, audio/video (32), and circuit designs. Specifically, watermarking techniques have been proposed for two domains: static artifacts and functional artifacts.

Static artifacts (33) consist of only syntactic components which are not altered during their use, for example, images and audio. Watermarks can also be placed in graphical objects such as 3D graphics and animation. The essential property of all watermarking techniques for static artifacts is that they leverage the imperfection of human perception. The main objectives of watermarking techniques for static artifacts include requirements for global placement of the watermark in the artifact, resiliency against removal, and suitability for rapid detection.

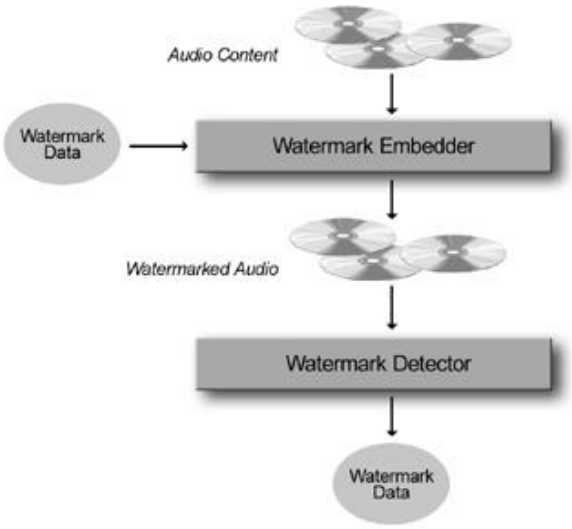

Figure 4-7 Watermarking in audio

Watermarking for functional artifacts, such as software and integrated circuits design have also been proposed. The common denominator for functional artifacts is that they must fully preserve their functional specifications and therefore cannot leverage on the principles for watermarking static artifacts. Functional artifacts can be specified and therefore watermarked at several levels of abstraction such as system level designs, FPGA designs (34), at the

behavioral and logic synthesis levels, and the physical design level. These approaches leverage on the fact that for a given optimization problem, a large number of similar quality solutions exist which can be selected from in order to have certain characteristics which math a designer's signature. More complex watermarking protocols, such as multiple watermarks (34), fragile watermarks, publicly detectable watermarks and software watermarking, have also been proposed. Techniques have also been developed for watermarking of DSP algorithms, sequential circuits, sequential functions, and analog designs.

Additionally, other techniques for intellectual property protection such as fingerprinting, obfuscation, reverse engineering, and forensic engineering have been proposed.

In sensor networks, watermarking and other intellectual property protection techniques can be applied at a variety of levels. The design of the sensor nodes and the software used in the network can be protected using functional techniques. Additionally, both static and functional watermarking can be applied on the data collected from the network depending on the types of sensors and actuators deployed (for example, video, audio, measured data).

4.3.2.1. Real-time watermarking

Real time watermarking aims to authenticate data which is collected by a sensor network. The first watermarking technique for cryptologically watermarking data and information acquired by a WSN has been developed by Feng and Potkonjak (35).

The key idea of their technique is to impose additional constraints to the system during the sensing data acquisition and/or sensor data processing phases. Constraints that correspond to the encrypted embedded signature are selected is such a way that they provide favorable tradeoffs between the accuracy of the sensing process and the strength of the proof of authorship. The first set of techniques embeds the signature into the process of sensing data. The crucial idea is to modulate by imposing additional constraints on the parameters which define the sensor relationship with the physical world. Options for these parameters include the location and orientation on sensor, time management (for example, frequency and phase of intervals between consecutive data capturing), resolution, and intentional addition of obstacles and use of actuators. In particular, an attractive alternative is to impose constraints on intrinsic properties (for example, sensitivity, compression laws) of a particular sensor; therefore the measured data will have certain unique characteristics that are strongly correlated with the signature of the author/owner.

The second technique is to embed a signature during data processing either in the sensor or control data. There are at least three degrees of freedom that can be exploited: error minimization procedures, physical world model building, and solving computationally intractable problems. In the first scenario, there are usually a large number of solutions that have similar levels of error. The task is to choose one that maintains the maximal consistency is measured data and also contains a strong strength of the signature. Typical examples of this type of tasks are location discovery and tracking. In the second scenario, they add additional constraints during the model building of the physical world.

In the final scenario, they are dealing with NP-complete problems, and therefore it is impossible to find the provably optimal solution. Therefore, the goal is to find a high quality solution that also has convincing strength of the signature.

4.3.2.2. Generic procedure

There exist numerous types of sensor networks and they can be used for many different purposes. Their goal is to watermark all data provided by a sensor network generically regardless of the type of data the network is collecting or what the purpose of the network is.

There are two types of data being produced by a sensor network: raw sensor data and processed application data. The first type, sensor data, is the original data the sensor network captures or measures. It may or may not be what the user or the network desires. However, the second type, processed data, is the output of the network to the user. The distinction of these two types of data provides insight into where watermarking can take place: (a) during the process of sensing data (original data capturing); (b) during the process of processing the original data. Therefore, they call these two processes watermarking in sensing data and watermarking in processing data.

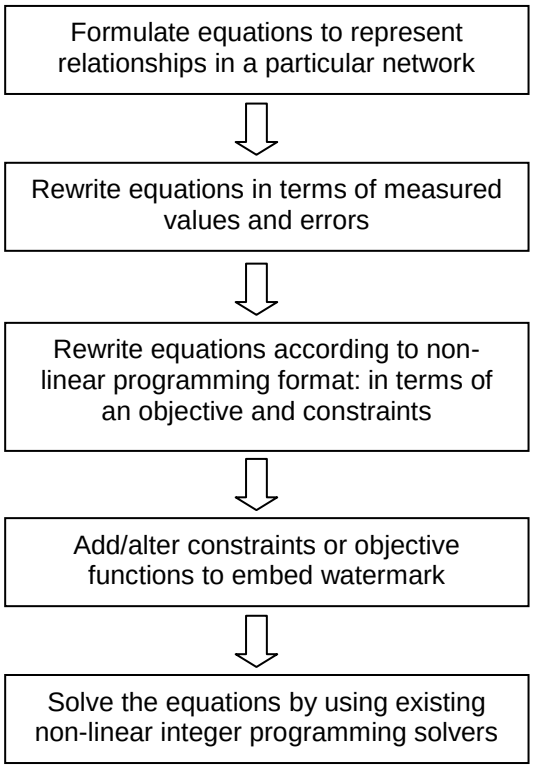

Figure 4-8 General procedure for embedding a watermark

An important question to ask is how is the original raw data being processed in order to generate the processed application data. In this case,

Feng and Potkonjak enquired the technique of non-linear programming. The general procedure can be summarized as Figure 4-8.

They first represent all the relationships that exist in the network using equations. Since everything is measured there always exists some degree of error. Realizing this, they replace the variables with the summation of a reasonable estimate and some error value. Their next goal is to minimize the errors in the equations, and achieve the closest possible estimates to the true values. This can be achieved by using effective non-linear programming solvers.

CHAPTER 5. NEW TECHNOLOGIES AND MATERIALS

5.1. MATERIALS

Materials used in electronics can play active or passive role. Very often one material can play both roles.

5.1.1. PASSIVE MATERIALS

Passive materials could be described as materials which are only used to provide either mechanical structure or electrical connection (36). The following Tables 5-1 and 5-2 summarize some examples of the physical properties of several materials (37) that determine their use in applications (38). Some of these materials can be used as active as well as passive materials, mainly silicon and gallium arsenide.

	Si	GaAs	SiO_2	Si_3N_4	GaN
Density [kg/m^3]	2,330	5,316	2,200	3,100	6,150
Melting point [°C]	1,414	1,238	1,600	–	2,500
Thermal conductivity [W/m/K]	168	47	6.5, 11	19	130
Dielectric constant	11.7	12	4.5, 4.3	7.5	4
Young's modulus [GPa]	190	–	380	380	–
Forbidden gap (300 °C)	1.12	1.427	–	–	3.2

Table 5-1 Physical properties of non-metallic materials

	Al	Au	Cr	Ti
Density [kg/m^3]	2,699	19,320	7,194	4,508
Melting point [°C]	660	1,064	1,875	1,660
Thermal conductivity [W/m/K]	236	319	97	22
Work function [eV]	4.3	5.1	4.5	4.3
Young's modulus [GPa]	70	78	279	40

Table 5-2 Physical properties of metallic materials (often used in the passive role)

5.1.2. ACTIVE MATERIALS

These materials are essential to the sensing process used in various types of micro sensors (36), such as photosensitive, piezoelectric, magneto resistive and chemo resistive films.

Nowadays, a wide range of functional materials are currently used in micro sensors and these often take the form of thin or thick films and play an active role in the sensing system. Some of them can be deposited using IC-

compatible deposition techniques (CVD or LPCVD) but others need special techniques such as electrochemical deposition as in the case of conducting polymers (38). The properties of some active materials are given in Table 5-3.

	Density [kg/m^3]	Melting point [°C]	Electrical conductivity [10^3 S/cm]	Thermal conductivity [W/m/K]
Thermal				
Pt	21,470	1,769	9×10^4	72
Radiation				
Ge	5,323	937	3×10^{-4}	67
Mechanical				
Quartz AT-cut	1,544	1,880	5.1	2.8
Magnetic				
Fe-pure	7,874	1,535	10^5	449
Chemical				
SnO_2	6,950	1,360	low	–

Table 5-3 Some properties of active metals

5.1.3. SILICON

Silicon makes up 27% of the Earth's crust by weight. Elemental silicon is not found in nature, but occurs in compounds like oxides and silicates. Silicon is prepared by heating silica and carbon in an electric furnace, using carbon electrodes. Though silicon is under normal conditions a relatively inert element, it still reacts with halogens and dilute alkalis, but most acids (except for some hyper-reactive combinations of nitric acid and hydrofluoric acid) do not affect it (39).

Silicon is abundant, relatively inexpensive and exhibits a number of physical properties which are useful for sensor application (40).

However, a major problem with silicon is that many of its characteristics are temperature dependent. Silicon does not display the piezoelectric effect and it is not ferromagnetic. Silicon also has no efficient photo- or electro-luminescent properties with the exception of porous or some nanocrystalline forms of silicon, which have not yet found any industrial applications in this field. This is the reason for its limited role in the field of optoelectronic active sources. Although silicon does not display the desired effect, it is possible to deposit layers of materials with the desired properties on the silicon substrate (41).

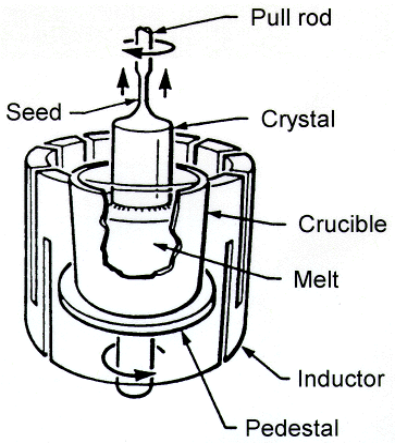

Figure 5-1 A schematic of rotary electromagnetic stirring used for Czochralski growth of semiconductor single crystals.

5.1.3.1. Single-crystalline silicon

Single-crystalline silicon is the most widely used semiconducting material. It is a brittle material, yielding catastrophically rather than deforming plastically (40). This material may be produced with high purity and quality (containing very few structural defects).

The silicon is cleaned by zonal melting (removing many impurities). Single crystals of silicon are then mostly prepared by cooling the melt using the Czochralski method, shown in Figure 5-1.

5.1.3.2. Polysilicon

Polycrystalline layers may be formed by vacuum deposition onto an oxidized silicon wafer with an oxide thickness of about 0.1 μm. Polysilicon structures may be doped with boron or other elements by ion implantation or other techniques to reach the required conductivity. Even if the boron concentration is very high, the resistivity of the polysilicon layers is always higher than that of a single-crystalline material. The resistance coefficient of the resistance may be changed over a wide range, positive or negative, through selective doping. The temperature sensitivity and the resistance of undoped polysilicon is substantially higher than that of single-crystalline silicon. For some specific doping concentrations, the resistance may become insensitive to temperature variation.

Polysilicon resistors are capable of reaching as high level of long-term stability as can be expected from resistors in single-crystalline silicon, since surface effects play only a secondary role in the device characteristics (40).

5.1.4. OTHER SEMICONDUCTORS

There is a wide range of compound semiconductors that combine atoms from columns III and V, II and VI or IV and VI of the periodic table. The importance of compound semiconductors is the possibility of combining semiconductors from the same family (for instance III/V) to prepare heterostructures with unique properties. Only two compound semiconductors are presented in this chapter: GaAs, the most important and widely used compound semiconductor and InSb because of its use in magnetic sensors.

5.1.4.1. Gallium arsenide (GaAs)

GaAs is a compound semiconductor combining group III and V elements from the same row as the traditional group IV semiconductor, germanium. GaAs has a density of 5,314.4 kg/cm^{-3} at room temperature and crystallizes into the zinc blended structure. A shift in valence charge from gallium to arsenic atoms produces a mixed ionic/covalent bond compared to the covalent bond of germanium and silicon, which increases the melting point (1,260 °C) but decreases hardness.

The most significant attribute is the electronic band structure of GaAs, which determines the major electrical and optical properties. Firstly, the optical absorption and luminescence across the band gap do not require the participation of momentum conserving phonons. This means that efficient luminescence is achievable for GaAs unlike Si and Ge. Secondly, the effective mass of electrons is substantially lower for GaAs than for Si (compare 0.3 m_0 for Si to 0.0067 m_0 for GaAs), so faster electronic devices are achievable in GaAs. Thirdly, since the forbidden energy gap for GaAs (1.42 eV) is higher than that for Si (1.08 eV) superior device isolation is potentially available for GaAs.

GaAs is dominantly used in heterostructures combining other ternary compound semiconductors with wider band gaps such as AlGaAs, or lower band gaps such as InGaAs.

5.1.4.2. Indium antimonide (InSb)

InSb is useful for magnetic sensing devices such as Hall effect sensors and magnetic resistors. InSb magnetoresistors are used as magnetic position sensors in automotive applications such as crankshaft and camshaft sensor for engine control. The sensitivity of magnetoresistors is proportional to the square of the electron mobility. Thus, the very large room temperature electron mobility of InSb is an advantage for these sensors. The narrow energy gap (0.18 eV) makes the intrinsic electron density high. Since the device operating temperature may be 200 °C in some applications, InSb is normally n-type doped to stabilize the electron density. InSb is also used for infrared imaging.

5.1.5. PLASTICS

Plastics are synthetic materials. They are made from monomers which consist of one chemical unit. The long chains of repeating units (ethylene) form polymers (polyethylene). In the same way, for example, polystyrene is formed from styrene monomers. Polymers consist of carbon atoms in combination with only seven elements –hydrogen (H), nitrogen (N), oxygen (O), fluorine (F), silicon (Si), sulfur (S) and chlorine (Cl). The combinations of these elements create thousands of various plastics.

The combination of the atoms must correspond to the rules of joining them with other atoms. Each atom has a limited capacity of chemical bonds if the compound should be stable. Polymers are also used as detectors or radiation, chemical sensors and other sensing applications.

5.1.5.1. Thermoplastics

Heavier molecules are created by adding more carbon and hydrogen to a chain, the step increase being 14 (one carbon + two hydrogens). For example: ethane gas (C_2H_6) is heavier than methane gas (it contains and additional carbon and two additional hydrogens, and its molecular weight is 30). Pentane C_5H_{12} is too heavy to be a gas and it is a liquid at room temperature. Further additions of CH_2 groups make progressively heavier liquids until $C_{10}H_{38}$ – this is

not a liquid but is a solid – paraffin wax. If we reach a molecular weight of 1,402 ($C_{100}H_{202}$), the material is tough and is called a low molecular weight polyethylene – the simplest of the thermoplastics. Further addition of CH_2 groups increases the toughness of the material and we get medium and high-molecular weight polyethylene (40). Polyethylene – the simplest polymer – is reasonably transparent and is used for fabrication of infrared windows and lenses.

The long chains are formed by heat, pressure and by using catalysts. This process is called polymerization. The chain length (molecular weight) determines many properties of a plastic – toughness, creep resistance, stress-crack resistance, melt temperature, melt viscosity, difficulty of processing, etc. these polymers are called thermoplastic polymers (heat-moldable).

If we pack the chains closer to one another we get denser polyethylene. These plastics have crystal structures. Crystallized areas are stiffer and stronger. These polymers are more difficult to process because they have higher and sharper melting temperatures. The crystalline thermoplastics abruptly transform into low-viscosity liquids, while amorphous thermoplastics soften gradually. For example, amorphous polymers include polystyrene, polycarbonate, polysulfone, etc. While crystalline plastics include polyethylene, polypropylene, nylon, acetal, etc.

5.1.5.2. Thermosets

Thermosets are another type of plastic. The polymerization – curing – is performed in two steps: (a) material manufacturing and (b) molding.

For example, phenolic compounds are liquefied under pressure during the molding process and a cross-linking reaction between molecular chains takes place. After it has been molded, a thermoset plastic has all its molecules interconnected with strong chemical bonds, which are not reversible by heating.

Thermoset plastics resist higher temperatures and provide greater dimensional stability. Thermoplastics offer higher impact strength than thermosets. They are also easier to be processed and allow more complex designs.

The useful thermoplastics in sensor-related applications are: alkyd, alkyl, epoxy, phenolic and monomers.

A *Copolymer* is a polymer formed in a polymerization reaction with two different monomers.

Plastics are electrical insulators, but often we require them to behave as conductors. In order to make them conductive we may either use lamination of the metal foil, metallization (for example, for shielding purposes) or we can mix plastics with conductive additives (graphite, metal fibers).

Piezoelectric plastics are made from PVF_2 and PVDF (crystalline materials). Initially, they do not have piezoelectric properties and they must be processed using high voltages or by corona discharge. These plastic films are used in some applications instead of ceramics, because they have better flexibility, stability against mechanical stress and they can be formed into any desirable shape (40).

5.1.6. METALS

Ferromagnetic metals (steel, iron, manganese, nickel and some alloys) are used in magnetic sensors. Ferromagnetic metals are also used for magnetic shielding. Non-ferromagnetic metals such as copper, aluminum and certain alloys such as some stainless steels have relative permeability close to 1.

When selecting a metal for the sensor design, we must take into the account not only the physical properties but also its mechanical processing. For example, copper has excellent thermal and electrical properties, but it is difficult to machine. An alternative compromise in this case is very often aluminum.

5.1.7. CERAMICS

Ceramics are crystalline materials which are very useful in sensor fabrication. The main common properties are structural strength, thermal stability, low weight, resistance to many chemicals, ability to bond with other materials and excellent electrical insulating properties. Another advantage of ceramics is that they mostly do not react with oxygen and thus do not create oxides.

Several metal carbides and nitrides belong to ceramics. Boron carbides and nitrides and aluminum nitrides (which have excellent heat transfer) are most commonly used. Silicon carbide has a high dielectric constant, so that it is ideal for designing capacitive sensors. Ceramics are usually hard, therefore they require special processing techniques. Various shapes of ceramic substrates are fabricated by scribing, machining, and drilling with the use of computer-controlled CO_2 lasers.

Ceramics for the sensor substrates are available from many manufacturers in thicknesses ranging from 0.1-10 mm (40).

5.1.8. GLASS

Glass is an amorphous solid material made by fusing usually silica with basic oxide. Although its atoms do not form a crystalline structure, its atomic arrangement is rather dense. Glass is a transparent material available in many colors. It is hard and resistant to most chemicals (except hydrofluoric acid). A lot of glasses are based on silicate and are composed of three major components – silica (SiO_2), lime ($CaCO_3$) and sodium carbonate (Na_2CO_3).

Non-silicate glasses include phosphate glass (resistant to hydrofluoric acid), heat absorbing glass (made with FeO), glass based on oxides of aluminum, vanadium, germanium and other types of metal. For example, borosilicate glass is massively resistant to thermal shocks due to its low thermal expansion and is used for the fabrication of optical mirrors. Lead-alkali glass (leas glass) contains lead monoxide (PbO), which increases the index of reflection and it is a better electrical insulator. It is used for the construction of optical windows, prisms and as a shield against nuclear radiation.

Light-sensitive glass forms another group. Photo chromatic glass darkens when exposed to ultraviolet radiation and clears when the UV is

removed and/or the glass is heated. The photo chromatic material may keep its color (at room temperature) from a few minutes to a week or longer (40).

5.2. SILICON PLANAR IC TECHNOLOGY

Micro sensor processing has similar requirements as the current microelectronic technology. The basic processing steps in silicon planar IC (integrated circuit) technology often form the basis of micro sensor technology. Conventional silicon planar IC technologies are subsequently modified to include some additional processing steps.

The monolithic fabrication processes can be divided into two basic types known as bipolar and CMOS. MOS is one of the most common IC technologies presently used in micro sensors (42).

The silicon planar IC fabrication generally involves all of the following processes:

- Crystal growth and epitaxy
- Oxidation and film deposition
- Diffusion or implantation of dopants
- Lithography and etching
- Metallization and wire bonding
- Testing and encapsulation

The existence of an oxide is very important: SiO_2, the preparation of which is simple, and which is suitable for lithography and possesses high electrical resistivity. Therefore, most ICs and micro sensors are produced whenever possible using silicon rather than gallium arsenide or other semiconductors. For the majority of IC structures it is necessary to grow thin epitaxial layers, which have better crystallographic quality in comparison with the bulk material (43).

5.2.1. THE SUBSTRATE: CRYSTAL GROWTH

There are two main techniques for bulk silicon crystal growth: Czochralski crystal pulling and floating zone process. Czochralski growth is used for the growth of larger diameter crystals (300 mm diameter are already industrially used) and for doped Si crystals.

The advantage of the float zone crystals is purity not only with respect to dopants but also as far as non-doping impurities such as carbon, oxygen, heavy metals and others are concerned. By applying multi-pass zone melting, the purity can even be enhanced. This method is not suitable for the growth of doped crystals. The largest diameter, which can be grown by this method, is about 100 mm, because of the stability problems arising from the melted zone under gravity conditions.

5.2.2. OXIDATION

The oxide layer is formed by placing the wafers into a furnace containing oxygen at 1,100 °C. Oxygen reacts with silicon and diffuses through the growing SiO2 layer (38).

The oxide films, which are used to prevent re-doping of areas with different types of doping materials, also form an electrically insulated region on the semiconductor device and are used for surface passivation.

5.2.3. DIFFUSION AND ION IMPLANTATION

MOS transistors are generally made from conducting or semi-insulating silicon wafers with layers that have been doped with n- or p-type materials (38). Controlled amounts of dopants are inserted into the wafer by thermal diffusion, ion implantation or during epitaxial growth.

The procedure of thermal diffusion of n-type materials is following. The wafers are placed in a furnace and an inert gas containing the required dopant (for example, AsH_3 or PH_3[3]) is passed over them. The p-type diffusion can be achieved by passing an inert gas carrying, for instance, B_2H_6.

An alternative method to thermal doping is ion implantation. The charged ions of the desired dopant are accelerated to energies to the range of 10 to 1,000 keV and are fired at the surface. The technique is now commonly used by penetrating As, P and B to a depth of 0.5, 1 and 2 μm at 1,000 keV in silicon.

Ion implantation at 10 keV gives very little yield on the ion source on this energy is used only very rarely, minimum reasonable energy is about 40-50 keV. On the opposite side of the spectrum, typical implantation does not allow energy higher than 200 keV. Besides the elements listed, BF^{2+} is also very common. After ion implantation the material should be annealed.

Doped layers can be prepared directly by epitaxial growth. By this way it is possible to dope layers homogeneously or with a defined doping profile and at the exact doping levels.

5.2.4. LITHOGRAPHY AND ETCHING

Lithography is an image transfer process of a geometric pattern from a mask onto a thin layer of material, called a resist. It is used in traditional planar processes, but it also is the principal mechanism for pattern definition in micromachining. The resist stands for a radiation-sensitive material. Firstly, a resist is usually either spin coated or sprayed onto the silicon wafer. The next procedure is to place the mask onto the resist correctly. Secondly, in optical lithography, ultraviolet (UV) radiation is used to change the solubility of the photo resist in a given solvent. The positive photo resist becomes more soluble

[3] AsH_3 and PH_3 are very toxic, in fact AsH_3 is the most toxic gas ever used in planar technology and they are typically replaced with less harmful ways of placing the same element into the silicon substrate. Arsenic is implanted from a solid source and phosphorus is doped in a furnace using $POCl_3$. B is also doped using a solid source.

after exposure to the UV light. The negative photo resist becomes less soluble due to a polymerization process.

A photo resist may also be used as a template for patterning material deposited after lithography. The resist is subsequently etched away, and the material deposited on the resist is “lifted off”.

5.2.5. DEPOSITION OF MATERIALS

The forming of the gate electrode is a typical example of deposition. The silane is pyrolyzed to produce the polysilicon layer. The polysilicon layer can be doped by the use of dopant gases during its formation or by diffusion or ion implantation. The higher reliability of the polysilicon as compared to the aluminum is the reason it is used in the construction of the gate. The polysilicon gate serves as a mask against source and drain implant.

5.2.6. METALLIZATION AND WIRE BONDING

At the end of processing the MOSFET, another oxide layer is formed and the window for metallization is exposed to lithography. The metal is then deposited by either physical vapor deposition – evaporation, chemical vapor deposition, or sputtering (≈ 1 μm). The metal, mostly aluminum (Al) or gold (Au), forms the ohmic contacts to the source, drain and gate electrode.

The final wafer is then diced up. A saw or diamond scribing is used and the IC is mounted into the package. The ultrasonic welding of thin Al or Au wire or ribbons makes up the electrical connections between the pads (ohmic contacts) and package terminals. An interesting fact is that the crucial influence on the reliability of the whole IC depends on the wire-bonding (38).

5.2.7. PASSIVATION AND ENCAPSULATION

The final IC chip must be protected from the atmosphere. The sensing areas are often covered up by the photo resist or by a silicon nitride layer. Silicon nitrate can be deposited by LPCVD or CVD and acts as a firm barrier against water.

The film thickness for IC is usually limited to less than 0.2 μm, because thicker layers cause thermally-induced stresses.

The last step is the encapsulation of the IC in the following manner, for example, sealed in a plastic resin or hermetically sealed in a metal case. This process is highly desirable, because it protects the silicon device from the surrounding environment. In the case of MEMS, isolating the silicon structure from the atmosphere is not always required. The atmosphere could transmit the measured quantity (38).

5.3. DEPOSITION TECHNOLOGIES

One of the basic steps in MEMS[4] processing is to deposit thing and thick films of material, which provides the sensing surface with the required properties. For example, sensitivity to thermal radiation is given by coating with nichrome. Thick films are used to construct pressure sensor (44) or microphones, where the membranes have to be produced. The following processes allow the fabrication of films to have a thickness anywhere between a few nanometers and about 100 micrometers. The film can be locally etched using lithography and wet chemical etching processes. Dry physical etching and laser processing can also be used (45).

There are two kinds of deposition processes: (a) chemical deposition and (b) physical deposition.

The first one is based on the creation of solid materials directly by chemical reactions with gas and/or liquid compounds or from the substrate material. In the second one, the deposited material is physically placed onto the substrate, with no traditional chemical reaction, which forms the material on the substrate (46).

5.3.1. CHEMICAL REACTIONS

5.3.1.1. Chemical vapor deposition (CVD)

The substrate is placed inside the reactor, into which a number of gases are introduced. The basic principle is that a chemical reaction takes place between the source gases. This reaction creates a solid material, which is deposited on free surfaces inside the reactor.

The CVD system is used to process thin films with good uniformity. This technology allows a variety of materials to be deposited, although some of them are less popular, because of hazardous by-products formed during the process. This is, however, the technology which is preferred by industry nowadays.

One of the simplified forms of CVD process is illustrated in Figure 5-2. The substrates or wafers are positioned on a stationary or rotating table whose temperature is elevated up to the required level by heating elements. There are three reasons for this: (a) oxides are decomposed and evaporated from the wafer surface, (b) the surface is smoothed, often at the atomic scale, and (c) the source gases are thermally decomposed, which is necessary for the layer growth. The top cover of the chamber has an inlet for the carrier gas, which can be added with various precursors and dopants. These additives, while being carried over the heated surface of the substrate, form a layer. The gas mixture flows from the distribution cone over the top surface of the wafers and exits though the exhaust gas outlets.

[4] MEMS is an abbreviation for “micro electro-mechanical system”. These devices feature the integration of mechanical elements, sensors, actuators and operating electronics on a common silicon substrate with the use of microfabrication technology.

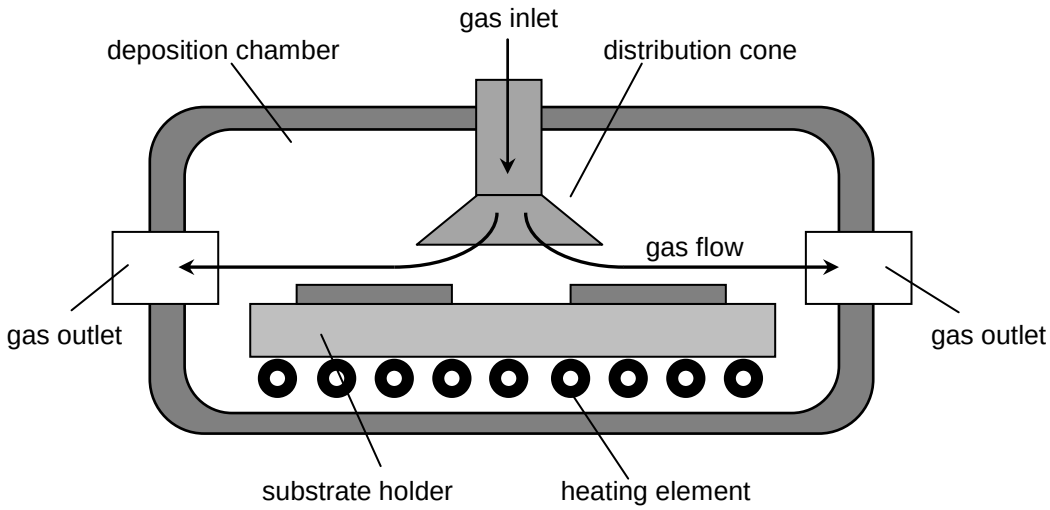

Figure 5-2 Simplified structure of a reactor chamber

5.3.1.2. CVD epitaxy

Epitaxial growth is used not only for layer deposition, but also because it is a way of preparing heterostructures of the highest quality from different materials. This technology is a special type of CVD process. It consists of depositing atoms of the desired material onto the substrate with the same crystallographic characteristic as the substrate. This concerns mainly crystallographic structures and orientation of the axes. In particular, it is possible to produce almost perfect single-crystalline films on single-crystalline substrates, if the lattice constants f the two materials are very close to each other. Film grown on a polycrystalline or amorphous substrate will also be amorphous or polycrystalline.

The most important epitaxial growth is vapor phase epitaxy (VPE). In this process, a number of gases are introduced into an induction-heated reactor where only the substrate is heated. The temperature of the substrate typically must be high because of oxide decomposition and evaporation, smoothing of the surface at an atomic level as well as for the decomposition of precursors.

The technology is primarily used for deposition of silicon and is widely used for producing silicon-on-insulator (SOI) substrates. The advantage of epitaxy is the high growth rate of material – it allows the formation of films with a thickness ranging from ≈ 1 μm to > 100 μm. Some processes require high temperature exposure of the substrate; others do not demand significant heating of the substrate. Some processes can be used to perform selective deposition, depending on the surface of the substrate (43). The typical VPE reactor is shown in the schema in Figure 5-3.

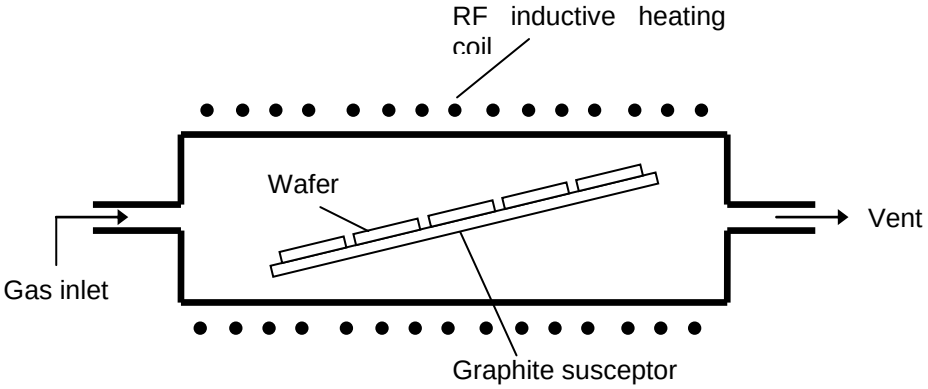

Figure 5-3 Typical "cold-wall" vapor phase epitaxial reactor

5.3.1.3. Electrodeposition (electroplating)

This process is restricted to electrically conductive materials. The process is used to make films of metals such as copper, gold and nickel (the thickness ≈ 1 μm to 100 μm). The deposition is best controlled when used with an external electrical potential, but it requires electrical contact to the substrate. The typical set-up for electroplating is shown in Figure 5-4.

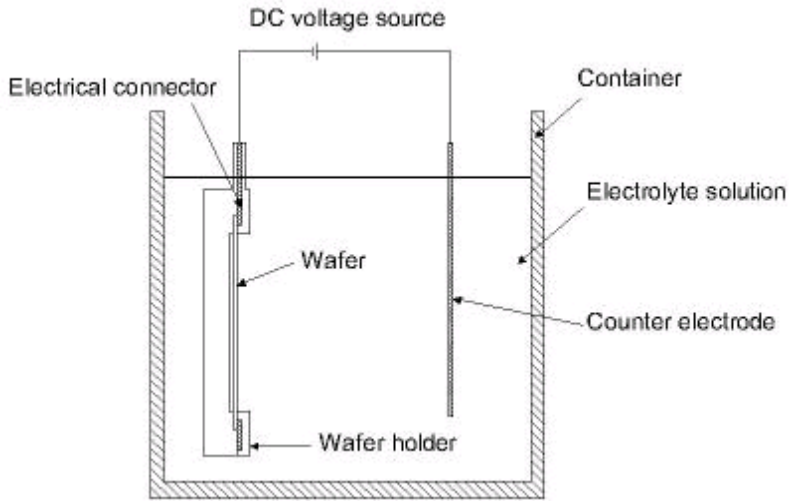

Figure 5-4 Typical set-up for electrodeposition

The substrate is paced in liquid solution (electrolyte) and an electrical potential is applied between a conducting area on the substrate and a counter electrode (usually platinum) in the liquid. The chemical process forms a layer of material on the substrate. Various types of gases are very often generated at the counter electrode (47).

Electroless plating does not require any external electrical potential and contact with the substrate during processing. These processes use chemical solutions in which deposition takes place spontaneously on any surface. The

disadvantage of this fabrication is that it is more difficult to control the film thickness and uniformity (46).

5.3.2. PHYSICAL REACTIONS

5.3.2.1. Physical vapor deposition (PVD)

PVD comprises technologies for deposition of metallic as well as dielectric films. It is more common than CVD for deposition of metals (lower process risk, cheaper), even if the quality of the films is inferior – higher resistivity for metals, more defects and traps for insulators. The choice of deposition method is in many cases arbitrary and depends on which technology is available for the specific material at a given moment. Two main techniques for PVD are evaporation and sputtering.

5.3.2.2. Evaporation

The main principle of this PVD technique is that metal can be converted into gaseous form and the deposited on the surface of the sample. The substrate is placed inside the vacuum chamber (usually 10^{-6} to 10^{-7} Torr). The block (source) of the material to be deposited is also located in the chamber. It is heated so that it evaporates. For some materials (Cr, Ti) sublimation temperatures are lower than the melting temperatures. The vacuum is required to allow the molecules to evaporate and move freely in the chamber. They subsequently condense on all surfaces. All evaporation technologies use this principle but the methods differ in the way the source material is heated and evaporated.

The two most popular evaporation technologies are e-beam and resistive evaporation. In e-beam evaporation, an electron beam is focused on the surface of the source material and causes local heating and subsequent evaporation. In resistive evaporation, a tungsten boat, containing the source material, is heated electrically.

The film thickness is determined by the evaporation time and the vapor pressure of the metal. The evaporation, just like all vacuum deposition processes, produces layers with large residual stresses and therefore these techniques are mostly used for depositing thin layers.

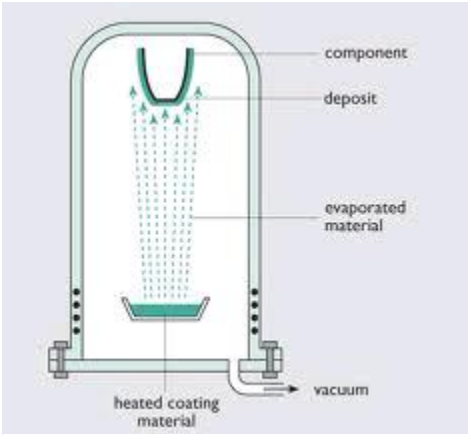

Figure 5-5 The evaporation - deposition of thin material film in a vacuum chamber

Very often it is important to keep the substrate at an elevated temperature, in order to evaporate moisture from its surface as well to remove some oxides or other surface impurities. The resistive evaporation is shown in Figure 5-5.

5.3.2.3. Sputtering

The source material in this PVD technology is subjected to a lower temperature compared to evaporation. The substrate is placed in a vacuum chamber (about 2.10^{-6} to 5.10^{-6} Torr) with the source material (denoted as the target or the cathode) and an inert gas (for example, argon or helium) of low pressure. A gas plasma is ignited using an AC or DC high voltage power source. The gas becomes ionized. The target-cathode is connected to this voltage. The sample wafer is attached to the anode at some distance from the cathode. In some cases, when the non-conductive substrate is used, the wafer doesn't need to be connected to the electrode and it is sufficient to put the wafer between the anode and cathode. The ions are accelerated against the target. The kinetic energy of the bombarding ions is sufficiently high to free some atoms from the target surface. The source material, now in a vapor form, condenses on all surfaces including the substrate (46). This principle of sputtering is common for all sputtering technologies. The different are typically in the method of ion bombardment of the target (40).

The advantage of this technology is better uniformity, especially if a magnetic field is introduced into the chamber. The field allows improved flow of atoms towards the anode. Since this method does not require a high target temperature, theoretically any material, including and organic material, can be sputtered. The process of sputtering can be extended to the sputtering of more than one target at the same time (co-sputtering), for example, sputtering nichrome (Ni and Cr) (43).

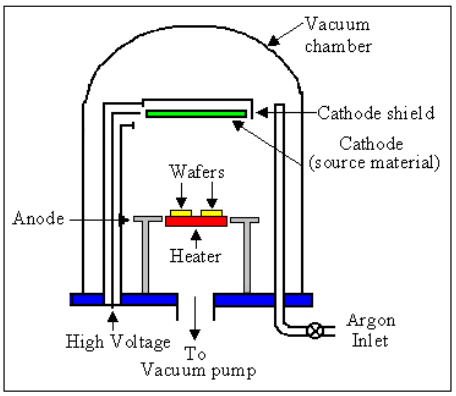

Figure 5-6 The sputtering process in a vacuum chamber

The sputtering process is shown in Figure 5-6.

5.3.2.4. Casting

In the casting process, the material to be deposited is dissolved in a volatile liquid solvent. When the solvent is evaporated, a thin layer of the material remains on the substrate (46). The differences in this process are in the way the material is transported onto the substrate. The most widely used are transmission by spraying and spinning.

The thinness of the casting layer on the substrate depends on the solubility of the deposited material and can be in the range from a single monolayer of molecules/atoms (adhesion promotion) to tens of micrometers (0.1 to 50 μm). The control of the film thickness depends on exact conditions, but the thickness can be uniform within +/- 10% in a wide range (40).

This process if often used for polymer, polyimide and other organic materials. The casting method is also used for transferring the photoresists to the substrate in the photolithography process and is an integral part of the photolithographic technique (43). This technique is often used for fabrication of humidity and chemical sensors.

5.3.2.5. Spray coating

Thermal spray coating is used for metal deposition. The coating material is fed into the flame where it melts. The melted is atomized by a high-velocity stream of air or other gas. When the stream reaches the target, atoms are bonded to the surface. This technology may replace traditional plating which often has serious pollution problems as the electrolytes contain toxic chemicals such as cyanide.

Low-temperature spray is used to deposit paints. One of the applications is in the production of thermal radiation sensors. The deposition layer (the surface of the sensor) is processed by covering it with a coating having a high infrared emissivity. The coating must be very thin to have a good thermal conductivity and a very small thermal capacity. However, the available organic materials have low thermal conductivity and cannot be effectively deposited with thickness less than 10 µm. this characteristic influences the sensor response (43).

5.3.2.6. Screen printing

This process has been used for many years as a cheap way to make hybrid circuits in electronics. The simplified explanation of the technique consists of the preparation of an ink paste using suitable organic solvents. The paste is then squeezed though a fine gauze mask and forms a 25 to 100 µm film in the desired areas. The film is then dried by heat treatment to form a conductive layer. The lateral resolution is only about 100 µm but the printing cost makes this technique commercially viable for low volume low-cost electronic circuits (38).

For example, platinum electrodes have been printed and used in electrochemical and bioelectrochemical sensors.

5.3.2.7. Laser ablation

Laser ablation in general is removing material from the surface using a laser beam. Localized heating causes the material to evaporate. This technology is also used for film deposition: in this case the target is heated by a pulsed laser and the substrate is positioned in the ablation plasma plume. This processed is made in vacuum or in the low pressure background gases. Laser ablation makes it possible to deposit complex materials (including organic) from the target to the substrate. High deposition rates are achievable; however the thickness is usually not perfectly even.

CONCLUSIONS

Wireless sensor networks must be designed to meet a number of challenging requirements including extended lifetime in the face of energy constraints, robustness, scalability, and autonomous operation. The many design concepts and protocols described in the preceding chapters address these challenges in different aspects of network operation. While much additional work remains to be done to realize the potential of tomorrow's systems, even these partial solutions offer reason for optimism.

Perhaps the most important lesson to take away from the studies described in this thesis is that the fundamental challenges must be tackled by making appropriate design choices and optimizations across multiple layers.

We can observe that the deployment of a sensor network can have a significant impact on its operational performance and therefore requires careful planning and design. The fundamental objective is to ensure that the network will have the desired connectivity and application-specific coverage properties during its operational lifetime. Particularly for small-medium-scale deployments, where there are equipment cost constraints and a well-specified set of desired sensor locations, structured placements and desirable. In other applications involving large-scale deployments of thousands of inexpensive nodes, such as surveillance of remote environments, a random scattering of nodes may be most flexible and convenient option.

Consider energy efficiency, which is perhaps the most fundamental concern due to limited battery resources. The most significant source of energy consumption in many applications is radio communication. At deployment time, energy efficiency concerns can inform the selection of an appropriate mixture of heterogeneous nodes and their placement. Localization and synchronization techniques can be performed with low communication overheads for energy efficiency. At the physical/link layers, parameters such as the choice of a modulation scheme, transmit power settings, packet size, and error control techniques can provide energy savings. Medium access techniques for sensor network use sleep modes to minimize idle radio energy consumption. Topology control techniques suitable for over-deployed networks also put redundant nodes to sleep until they are needed to provide coverage and connectivity. At the network layer, routing techniques can be designed to incorporate energy-awareness and in-network compression to minimize energy usage. Combining all these techniques, it may well be possible to make sensor network deployment with battery-operated devices last for several years.

In wireless sensor networks, medium-access protocols provide a dual functionality, not only providing arbitration for access to the channel as in traditional MAC protocols, but also providing energy efficiency by putting the radio to sleep during periods of communication inactivity. Energy efficiency is a key concern for sensor network MAC protocols. As we see, significant energy savings are possible by avoiding idle listening. While there have been proposals to use a secondary low-power wake-up radio to achieve this, the simple low-power listening/preamble sampling technique provides effectively the same benefit.

Transport-layer mechanisms are needed in wireless sensor networks to provide guarantees on reliable, low-latency, energy-efficient, fair delivery of

information. Several challenges must be overcome in order to provide these guarantees: channel loss, interference, bandwidth limitations, bursty traffic, and node resource constraints. For reliable data delivery, unlike the traditional Internet, end-to-end TCP-based mechanisms are not appropriate, due to the high loss rate on links.

The unreliability of wireless links, with large spatio-temporal variations in quality due to multi-path fading effects, poses another fundamental challenge. Link layer solutions include the implementation of link quality monitoring along with power control, blacklisting, and ARQ techniques. The use of link quality metrics as well as diversity-based multi-path and cooperative routing techniques also provides robustness at the network layer. Finally, rate control and priority scheduling techniques enable reliable communication of critical information though the network.

Scalability concerns are addressed in protocols at all layers by emphasizing distributed and hierarchical algorithms with localized interactions. We have seen that many of these techniques are also designed to be inherently self-configuring and adaptive to changes in the environment, to meet the need of autonomous, unattended operation.

BIBLIOGRAPHY

1. **W.J. Kaiser, K. Bult, A. Burstein, D. Chang, et al.** *"Wireless Integrated Microsensors", Technical Digest of the 1996 Solid State Sensor and Actuator Workshop.* June 1996.

2. **J. Hill, R. Szewcyk, A. Woo, D. Culler, S. Hollar, and K. Pister.** *"System Architecture Directions for Networked Sensors", Proceedings of the 9th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS).* November 2000.

3. **A. Mainwaring, J. Polastre, R. Swecyk, D. Culler, and J. Anderson.** *"Wireless Sensor Networks for Habitat Monitoring", Proceedings of the First ACM International Workshop on Wireless Sensor Networks and Applications (WSNA) .* Atlanta, Georgia : s.n., September 2002.

4. **IPTO, DARPA.** *"Sensit: Sensor Information Technology Program".*

5. **N. Xu, S. Rangwala, K. Chintalapudi, D. Ganesan, A. Broad, R. Govindan, and D. Estrin.** *"A Wireless Sensor Network for Structural Monitoring", Proceedings of ACM Conference on Embedded Networked Sensor Systems (SensSys).* November 2004.

6. *"It's Time for Sensors to Go Wireless" .* **Manges, W.** s.l. : Sensors Magazine, April 1999.

7. **IEEE 802.15.4, IEEE Standard for Information Technology-Part 15.4.** *Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low Ride Wireless Personal Area Networks (LR-WPANS).* 2003.

8. **Karn, P.** *MACA: A New Channel Access Method for Packet Radio.* September 1990.

9. **IEEE 802.11.** *IEEE standards for Information Technology - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications.* 1999.

10. **IEEE 802.15.4.** *IEEE Standard for Information Technology - Part 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low Rate Wireless Personal Area Networks (LR-WPANS).* 2003.

11. **G. Lu, B. Krishnamachari, C. Raghavendra.** *Performance Evaluation of the IEEE 802.15.4 MAC for Low-Rate Low-Power Wireless Networks.* April 2004.

12. **S. Singh, C.S. Raghavendra.** *PAMAS - Power Aware Multi-Access Protocol with Signalling for Ad Hoc Networks.* October 1998.

13. **National Standards Policy Advisory Committee.** *National Policy on Standards for the United States and a Recommended Implementation Plan.* Washington, D.C. : s.n., December 1978.

14. **Dubendorf, Vern A.** *Wireless Data Technologies.* s.l. : John Wiley & Sons, Ltd., 2003.

15. **D.Culler, A. Woo.** *A Transmission Control Scheme for Media Access in Sensor Networks, Proceedings of ACM Mobicom.* July 2001.

16. **Y. Sankarasubramaniam, O.B. Akan, I.F. Akyildiz.** *ERST: Event-to-Sink Reliable Transport in Wireless Sensor Networks, Proceedings of ACM MobiHoc.* June 2003.

17. **C.-Y. Wan, A. Campbell, L. Krishnamurthy.** *PSFQ: A Reliable Transport Protocol for Wireless Sensor Networks, Proceedings of ACM SenSys.* November 2003.

18. **C.-T. Ee, R.Bajcsy.** *Congestion Control and Fairness for Many-to-One Routing in Sensor Networks, Proceedings of ACM SenSys.* November 2004.

19. **B. Hull, K. Jamieson, H. Balakrishnan.** *Techniques for Mitigating Congestion in Sensor Networks, Proceedings of ACM SenSys.* November 2004.

20. **Lewis, Robert.** *Modelling Distributed Control Systems Using IEC 61499.* UK : s.n., February 2001.

21. **Boyer, Stuart A.** *SACDA: Supervisory Control and Data Acquisition.* NC : s.n., January 1999.

22. *Cybercrime's New Foe.* **Wilson, Tony.** Australia : The Gold Coast Bulletin, page 14, October 2002.

23. *Secure Routing in Sensor Networks: Attacks and Countermeasures, pages 293-315.* **C. Karlof, D. Wagner.** Berkeley : Ad Hoc Networks, May 2003, Vol. 1.

24. **X. Fu, B. Graham, R. Bettati, W. Zhao.** *On Countermeasures to Traffic Analysis Attack.* 2003.

25. **National Institute of Standards and Technology.** *FIPS 46-2; Data Encryption Standard.* December 1993.

26. *Secure Sensor Networks for Perimeter Protection.* **S. Avancha, J. Undercoffer, A. Joshi, J. Pinkston.** s.l. : Special Issue Computer Networks on Wireless Sensor Networks, 2003.

27. *SPINS: Security Protocols for Sensor Networks.* **A. Perrig, R. Szewczyk, V. Wen, D. Culler, J. D. Tygar.** s.l. : Wireless Networks Journal (WINET), pages 521-534, September 2002.

28. *The Sybil Attack.* **Douceur, J. R.** s.l. : Proc. IPTPS '02, March 2002.

29. **B. Krishnamachari, D. Estrin, S. B. Wicker.** *Modeling Data-Centric Routing in Wireless Sensor Networks.* s.l. : Technical Report CENG 02-14, Dept. of Computer Engineering, USC, 2002.

30. **Spitzner, L.** *Honeypots: Tracking Hackers.* 2002.

31. *Drm: doesn't really mean digital copyright management.* **Camp., L.J.** s.l. : ACM Computer and Communications Security, pages 78-87, 2002.

32. *Perceptual watermarks for digital images and video.* **R. Wolfgang, C.I. Podilchuck, E. Delp.** s.l. : International Conference on Security and Watermarking of Multimedia Contents, pages 219-222, 1996, Vol. 3657.

33. *Multimedia watermarking techniques.* **F. Hartung, M. Kutter.** 87, 1987, Vols. Proceedings of the IEEE, pages 1079-1107.

34. *Robust FPGA intellectual property protection through multiple small watermarks.* **J. Lach, W.H. Mangione-Smith, M. Potkonjak.** s.l. : Design Automation Conference, pages 831-836, 1999.

35. *Real-time watermarking techniques for sensor networks.* **J. Feng, M. Potkonjak.** s.l. : SPIE Security and Watermarking of Multimedia Contents, pages 391-402, 2003.

36. **Culshaw, B.** *Smart Structures and Materials.* 1996.

37. **Sze, S.M.** *Semiconductor Devices Physics and Technology.* NY : s.n., 1985.

38. **Gardner, W.J.** *Microsensors - Principles and Applications.* NY : s.n., 1994.

39. **P.T. Moseley, A.J. Crocker.** *Sensor Materials.* 1996.

40. **Fraden, Jacob.** *Modern Sensor Handbook - Physics, Designs and Applications.* NY : s.n., 1996.

41. Okmetic Oyj, Finland (high-quality silicon wafers). *http://www.okmetic.com.* [Online]

42. **H. Baltes, O. Brand.** *CMOS-based microsensors.* 2001.

43. MEMS Exchange. *http://www.mems-exchange.org.* [Online]

44. Micromachined Sensors of a Global Market. *http://www.sensonor.com.* [Online]

45. Institute of Microelectronics (IME), National University of Singapore. *http://www.ime.a-star.edu.sg/.* [Online]

46. MEMSnet, Corporation for National Research Initiatives. *http://www.memsnet.org.* [Online]

47. *Electroplating and characterization of cobalt-nickel-iron and nickel-iron for magnetic microsystems application.* **F.E. Rasmussen, J.T. Ravnkilde, P.T. TANG, O. Hansem, S. Bouwstra.** 2001.