

Energy-efficient wireless networking for multimedia applications

(*Wireless Communications and Mobile Computing, Wiley, 2001: 1:165-184*)

Paul J.M. Havinga, Gerard J.M. Smit

University of Twente, department of Computer Science, Enschede, the Netherlands
{havinga, smit}@cs.utwente.nl

Abstract – In this paper we identify the most prominent problems of wireless multimedia networking and present several state-of-the-art solutions with a focus on energy efficiency. Three key problems in networked wireless multimedia systems are 1) the need to maintain a minimum quality of service over time-varying channels, 2) to operate with limited energy resources, and 3) to operate in a heterogeneous environment. We identify two main principles to solve these problems. The first principle is that energy efficiency should involve all layers of the system. Second, Quality of Service is an essential mechanism for mobile multimedia systems not only to give users an adequate level of service, but also as a tool to achieve an energy-efficient system. Due to the dynamic wireless environment, adaptability of the system will be a key issue in achieving this.

Keywords – energy efficiency, wireless networking, mobile computing, quality of service.

1 Introduction

Advances in technology enable portable computers to be equipped with wireless interfaces, allowing networked communication even while on the move. Whereas today's notebook computers and personal digital assistants (PDAs) are self contained (*introvert*) tomorrow's networked mobile computers are part of a greater computing infrastructure (*extrovert*). Key problems are that these portable wireless network devices need to handle multimedia traffic in a dynamic and heterogeneous wireless environment, and the need to operate with limited energy resources.

Wireless communication is much more difficult to achieve than wired communication because the surrounding environment interacts with the signal, blocking signal paths and introducing noise and echoes. As a result wireless connections have a lower quality than wired connections: lower bandwidth, less connection stability, higher error rates, and, moreover, a highly varying quality. They need to be able to operate in environments that may change drastically – in short term as well as in long term – in available resources and available services. These factors can in turn increase communication latency due to retransmissions, can give largely varying throughput, and incur a high energy consumption.

Wireless networking is a broad area, and has many applications ranging from voice communication (cellular phones) to high performance multimedia networking. In this paper we are somewhat biased towards multimedia traffic, as it is expected that the new generation of wireless networks will carry diverse types of multimedia traffic.

1.1 Key problems of wireless multimedia networking

Three key problems in wireless multimedia networking are 1) to operate with limited energy resources, 2) the need to maintain quality of service (throughput, delay, bit error rate, etc) over time-varying channels, and 3) to operate in a heterogeneous environment.

Energy-efficiency – Portable wireless devices have severe constraints on the size, the energy consumption, and the communication bandwidth. Moreover, it is expected that these devices will be multimedia oriented, and need to handle many different classes of data traffic over a limited bandwidth wireless connection, including delay sensitive, real-time traffic such as speech and video. More extensive and continuous use of network services will only aggravate this problem since

communication consumes relatively much energy. Unfortunately, the rate at which battery performance improves (in terms of available energy per unit size or weight) is fairly slow, despite the great interest generated by the booming wireless business. Aside from major breakthroughs it is doubtful that significant reduction of battery size and weight can be expected in the near future. The energy consumption these devices need for communication and computation will limit the *functionality* of the mobiles.

The way out is *energy efficiency*: doing more work with the same amount of energy. The art of low-power design used to be a narrow speciality in analog circuit design. Nowadays, it is appearing in many layers of a system.

Quality of Service – A Quality of Service model provides the basis for modern high-bandwidth and real-time multimedia applications like teleteaching and video conferencing. The notion of QoS service originally stems from communication, but because of its potential in the allocation of all scarce resources, it has found its way into other domains, e.g. operating systems [25].

Heterogeneity – In contrast to most stationary computers, mobile computers encounter heterogeneous network connections. As they leave the range of one network transceiver they switch to another. In different places they may experience different network qualities. There may be places where they can access multiple transceivers, or even may concurrently use wired access. The interface may also need to change access protocols for different networks, for example when switching from wireless LAN coverage in an office to cellular coverage in a city. This heterogeneity makes mobile computing more complex than traditional networking.

These three problem-areas that are characteristic for future mobile networking are strongly correlated. [23]. Wireless network protocols typically address network performance metrics such as throughput, efficiency, fairness and packet delay. In this paper we address the additional issue of *energy efficiency* of the wireless network protocols¹. Considerations of energy efficiency are fundamentally influenced by the trade-off between energy consumption and achievable Quality of Service. The dynamic communication and application environment is an extra challenge, which might be solved using QoS principles. The aim in mobile multimedia networking is to meet the required QoS, while minimising the required amount of energy. To deal with the dynamic variations in networking and computing resources gracefully, both the mobile computing environment and the applications that operate in such an environment need to adapt their behaviour depending on the available resources including the batteries. Current research on several aspects of wireless networks (like error control, frame-length, access scheduling) indicate that continually adapting to the current condition of the wireless link have a substantial impact on the energy-efficiency of the system [8][10][33].

1.2 Principles of energy-efficient wireless networking

In this paper we will discuss a variety of energy reduction approaches that can be used for building an energy-efficient mobile system, and show the relationship with *multimedia* and the *dynamic environment*. In this paper the following main principles are identified:

1. *Involve all layers*. Energy efficiency is an issue involving all layers of the system, its physical layer, its communication protocol stack, its system architecture (Section 2.1), its operating system, and the entire network (Section 4).
2. *Quality of Service* is an essential mechanism for mobile multimedia systems not only to give users an adequate level of service, but also as a tool to achieve an energy efficient system.

QoS support in wireless networks involves several considerations beyond those addressed in earlier work on conventional wireline networks. In traditional networks, based on fixed terminals and high-quality/high-capacity links it is feasible to provide 'hard' QoS guarantees to users. However, in a mobile environment, mobility and the need for efficient resource utilisation require the use of a 'soft' QoS model [44]. The minimum QoS requirements for multimedia applications has a wide dynamic range depending on the user's quality expectations, application usage models, and application' tolerance to degradation. Users and applications require a certain QoS level. The system then operates in such a way that it will try

¹ In general, saving energy for the base station is not really an issue, as it is part of the fixed infrastructure and typically obtains energy from a mains outlet. However, since the current trend is to have ever smaller cell sizes, and the complexity of the base station is increasing, this issue might become more important in the future mainly because of economical and thermal reasons.

to satisfy these requirements, but never gives more quality than required and necessary. Due to the dynamic wireless environment, *adaptability* of the system will be a key issue in achieving this. This implicates several mechanisms that can be used to attain a high energy efficiency, e.g.:

- *Avoid useless activity.* This is the main driving force of for instance dynamic power management (Section 2.2), link layer protocols (Section 5) and adaptive error control (Section 6). Useless activity can be caused by various factors at all levels of the system (e.g. being unnecessary in a high power operational mode, applying error control to error-resilient data, trying to transmit a video frame that is already too old). If the operations would adapt to the required QoS and current environment, then energy-efficiency can be improved.
- *Scheduled operations.* This extends power management in the sense that communication is scheduled at appropriate time such that the differences in power states are exploited as much as possible (Section 4.2, and Section 5). This has a strong relation with the QoS model since timing constraints of multimedia connections are likely to be the limiting factors in the potential energy reduction.
- *Reduce the amount of data.* This is quite obvious and is applicable in all layers of the system. This relates to the trade-off between communication and computation (Section 3). Examples are adaptive error control that adapts its error coding according to the current channel conditions and the required quality, video transmission systems that adapt the quality to the expectations of the users, the available resources, and the channel conditions. Again, QoS can be used to determine whether it is really necessary to produce the data.

1.3 Outline

In this paper we give an overview of various aspects of energy efficient wireless networking with a focus on the lower layers of network protocol stack. Although some research has been done in this area, a lot of research issues remain open. Because it is not possible to go into depth, the intention of this paper is primarily to give insight in the field of energy efficient networking.

In Section 2 we provide the fundamentals of power management. Then, in Section 3 we review the main sources of energy consumption induced by the wireless channel. In Section 4 we provide an overview of mechanisms to reduce the energy consumption needed for communication in the network protocol stack, the operating system, and by decomposition. In Section 5 and Section 6 we will delve a bit deeper into the two most developed areas to reduce energy consumption: the *MAC Layer*, and *error-control*.

2 Power management

Traditionally, energy efficiency has been focussed on low-power techniques for VLSI design. As the issue of energy efficiency becomes more pervasive, the battle to use the bare minimum of energy will be fought on multiple fronts: semiconductor technology, circuit design, design automation tools, system architecture, operating system, and application design. Energy awareness is now appearing in the mainstream digital design community affecting all aspects of the design process. Eventually, the concern for low-power design will expand from devices to modules to entire systems, including application software and even to the user.

2.1 Low power system design

Most components are currently fabricated using CMOS technology. Main reason for this bias is that CMOS technology is cost efficient and inherently consuming less power than other technologies. The dominant factor of energy consumption (85 to 90%) in CMOS is *dynamic*. A first order approximation of the dynamic energy consumption of CMOS circuitry is given by the formula:

$$P_d = C_{eff} V^2 f \quad (1)$$

where P_d is the power in Watts, C_{eff} is the effective switch capacitance, V is the supply voltage, and f is the frequency of operations. The power dissipation arises from the charging and discharging of the circuit node capacitance found on the output of every logic gate. C_{eff} combines two factors C , the capacitance being charged/discharged, and the *activity weighting* α , which is the probability that a transition occurs.

$$C_{eff} = \alpha C \quad (2)$$

At lower levels energy consumption can thus be decreased by reducing the supply voltage, reducing the capacitive load and by reducing the switching frequency.

In general, a designer tries to make a system to be optimal for a certain application and environment. The designer has to select a particular algorithm, design or use an architecture that can be used for it, and determine various parameters such as supply voltage and clock frequency. However, energy efficiency in mobile systems is not only a one-time design problem that needs to be solved during the design phase. In a mobile system, power management extends the notion of hardware/software co-design, since we have to face a *much more dynamic application and communication environment*. When the system is operational, frequent adaptations to the system are required to obtain an energy efficient system that can fulfil the requirements imposed in terms of a general QoS model. This multi-dimensional design space offers a large range of possible trade-offs.

2.2 *Dynamic power management*

The essential characteristic of energy consumption for static CMOS circuits is that quiescent portions of a system dissipate a minimal amount of energy. Dynamic power management refers to the general class of techniques that manage the performance and throughput of a system based on its computational needs within the energy constraint [6]. Dynamic power management exploits periods of *idleness* caused by system under-utilisation. Especially in mobile systems, the utilisation is not constant and power management can be used effectively. It is common practice that designers focus on worst-case conditions, peak performance requirements and peak utilisation, which, however, is in practice only fully exploited during a small fraction of their operation.

Dynamic power management is based on deactivating functional units when they are not required. The main problems involved are the cost of shutting-down and restarting a module or component. Restarting induces an increase in latency (e.g. time to restore a saved CPU state, spin-up of a disk), and possibly also an increase in energy consumption (e.g. due to higher start-up current in disks). The two main questions involved are then: 1) when to shutdown, and 2) when to wake-up.

1. The time (and thus energy) that is required to determine when a module can be shut down (the so-called *inactivity threshold*) can be assigned statically or dynamically. In a *predictive* power management strategy the threshold is adapted according to the past history of active and idle intervals.
2. The other question is *when to wake-up*, where the typical policy is to wake up in response to a certain event such as user interaction or network activity. The problem with such a demand policy is that waking up takes time, and the extra latency is not always tolerable. Again, a predictive approach, where the system initiates a wakeup in advance of the predicted end of an idle interval, often works better.

Operating modes of a wireless interface

The wireless network interface of a mobile computer consumes a significant fraction of the total power [53]. Typically, the transceiver can be in five modes; in order of increasing energy consumption, these are off, sleep, idle, receive, and transmit (see Figure 1). In transmit mode, the device is transmitting data; in receive mode, the receiver is receiving data; in idle mode, it is doing neither, but the transceiver is still powered and ready to receive or transmit; in sleep mode, the transceiver circuitry is powered down, except sometimes for a small amount of circuitry listening for incoming transmissions [35]. The difference in the amount of energy consumed in these modes is significant.

Examples:

1) The power consumption of a WaveLAN modem when transmitting is typical 1675 mW, 1425 mW when receiving, and 80 mW when in sleep mode [56]. Increasing the sleep time period of the radio thus significantly improves the energy efficiency of the wireless network. Also important to notice is that the transition times between the operating modes can be quite high. In WaveLAN a transition time from sleep to idle takes 250 μ s, and has during that period already the power consumption of the idle state [21]. Then, before the payload will be transmitted another 254 μ s is required.

2) A Bluetooth radio (Ericsson PBA 313 01/2 [2]) has similar characteristics: in sleep mode it consumes 100 mA, in idle 25 mA, in receive mode 52 mA, and in transmit mode 44 mA. From sleep to idle requires 110 ms, and from idle to transmit or receive mode typical 104 ms.

2.3 Energy efficiency

We define the *energy efficiency* e as the energy dissipation that is essentially needed to perform a certain function, divided by the actually used total energy dissipation.

$$e = \frac{\text{Essential energy dissipation for a certain function}}{\text{Actually used total energy dissipation}} \quad (3)$$

The function to be performed can be very broad: it can be a limited function like a multiply-add operation, but it can also be the complete functionality of a network protocol.

Let us for example consider a medium access control (MAC) protocol that controls access to a wireless channel. The essential energy dissipation is the energy dissipation needed to transfer a certain amount of bits over the wireless channel, and the total actually used energy dissipation also includes the overhead involved in additional packet headers, error control, etc., but also 'physical' overhead induced by e.g. a frequency hopping scheme.

Note that the energy efficiency of a certain function is independent of the actual implementation, and thus independent of the issue whether an implementation is low-power. Low-power is generally closely related to the hardware, whereas energy-efficiency relates to the algorithm using the hardware. Thus, it is possible to have two implementations of a certain function that are built with different building blocks, of which one that has been build with power-hungry components has a high energy efficiency, but dissipates more energy than the other implementation which has a lower energy efficiency, but is built with low-power components.

3 Energy consumption in mobile systems

Several researchers have studied the power consumption pattern of mobile computers. However, because they studied different platforms, their results are not always in agreement. Laptops designers use several techniques to reduce energy consumption, primarily by turning devices off after a period of no use, or by lowering the clock frequency. Lorch reported that the energy use of a typical laptop computer is dominated by the backlight of the display, the disk and the processor [36]. Ikeda et al. observed that the contribution of the CPU and memory to power consumption has been on the rise the last few years [27]. Stemm et al. [53] concluded that the network interface consumes at least the same amount of energy as the rest of the system (i.e. a Newton PDA). Further, the fraction of energy consumed for networking by these mobiles, is only likely to increase as mobiles evolve towards a thin client network computer, and the communication traffic will increase. Another source of energy consumption is due to the fact that many high-performance network protocols require that all network access be through the operating system, which adds significant overhead to both the transmission path (typically a system call and data copy) and the receive path (typically an interrupt, a system call, and a data copy). This not only causes performance problems, but also incurs a significant energy consumption. Intelligent network interfaces can relieve this problem to some extent. To address the performance problem, several *user-level communication architectures* have been developed that remove the operating system from the critical communication path [7].

To make the wireless interfaces more energy efficient, algorithms embodying energy efficient protocols must be distributed across two or more wireless end-points [32]. This implies that the focus should be on the layers of the network stack through which the mobiles interact.

Even though it is difficult to compare these results because the measurements are made for different architectures, operating systems, communication interfaces, and benchmarks, there is a common pattern: there is no primary source of energy consumption, and the energy consumed for communication increases. The energy consumption is distributed over several devices and for several operations. The conclusion is that implementing an energy efficient system should involve all the functions in the system at all layers.

3.1 System architecture

In its most abstract form, a *networked computer system* has two sources of energy drain during wireless networking [32]:

- *Communication*, due to energy spent by the wireless interface. Communication energy is, among others, dictated by the signal-to-noise ratio (SNR) requirements and the radio cell diameter.
- *Computation*, due to (signal) processing and other tasks required during communication. Computation energy is a function of the hardware and software used for tasks such as compression and forward error correction (FEC).

For long distance wireless links (macro cellular), the transmit-communication energy component dominates. However, for short distance wireless links (pico cellular) and in harsh environments where much signal processing and protocol computation may be used, the computation component can be significant or dominant [57]. Broadly speaking, minimising energy consumption is a task that will require minimising the contributions of communication and computation, making the appropriate trade-offs between the two. For example, reducing the amount of transmitted data may be beneficial. On the other hand, the computation cost (e.g. to compress the data being sent) might be high, and in the extreme it might be such that it would be better to just send the raw data.

As semiconductor technology improves, computation gets relative cheaper, whereas communication has much less advantage of the smaller feature size. Therefore, communication will get relatively more expensive. This property also holds for multimedia applications, even though these applications typically require a significant computational effort as well. For a significant part this is due to the limitations of most current hardware and operating systems that are unable to differentiate between various traffic streams [23].

3.2 Adaptability

Recent research shows that by changing the system architecture from a traditional approach to a connection oriented, reconfigurable approach gives a huge improvement of the energy efficiency of a multimedia system [37][23][24][55]. Programmability is particularly important for mobile systems because they operate in a dynamically changing environment and must be able to adapt to the new environment. For example, a mobile computer will have to deal with unpredicted network outage or should be able to switch to a different network, without changing the application. It should therefore have the *flexibility* to handle a variety of multimedia services and standards (like different video decompression schemes and security mechanisms) and the *adaptability* to accommodate the nomadic environment, required level of security, and available resources. Reconfigurable computing systems combine programmable hardware with programmable processors to capitalise on the strengths of hardware and software. While low-power solutions are already available for application specific problems, applying these solutions in a reconfigurable environment is a substantially harder problem, since programmable devices often incur significant performance and energy-consumption penalties. To reduce the energy overhead in programmable architectures, the computational granularity should be matched to the architectural granularity.

3.3 Energy consumption of a wireless channel

We will now provide the main causes of unnecessary energy consumption needed for communication over a wireless channel ([22][51]).

- First of all, for applications that have low traffic needs, the transceiver is *idling* most of the time. Measurements show that on typical applications like a web-browser or e-mail, the energy consumed while the interface is on and idle is more than the cost of actually receiving packets [53].
- Second, the typical *inactivity threshold*, which is the time before a transceiver will go in the off or standby state after a period of inactivity, causes the receiver to be in a too high energy consuming mode needlessly for a significant time.
- Third, in a typical wireless broadcast environment, the receiver has to be powered on at all times to be able to *receive messages* from the base station, resulting in significant energy consumption. The receiver subsystem typically receives all packets and forwards only the packets destined for this mobile. The access to the wireless channel is controlled by a MAC protocol. Many MAC protocols for wireless networks are basically adaptations of MAC protocols used in wired networks, and ignore

energy issues [32]. For example, random access MAC protocols such as carrier sense multiple access with collision avoidance (CSMA/CA) and 802.11 typically require the receiver to be powered on continually and monitor the channel for traffic.

- Fourth, overhead induced by the *physical layer*. This overhead can be significant and is caused by for example guard space, interfacing delay, preamble and postamble.

Examples:

WaveLAN has an overhead equivalent to approx. 58.25 bytes [56] (see Figure 2). A frequency-hopping scheme makes this effect even worse, because it requires the radio to change its frequency often, like Bluetooth. In Bluetooth [1] a group of at most seven active slave radios, is synchronised to a single master radio. Slaves may only communicate when granted permission from the master. Bluetooth radios communicate with each other using a Time Division Duplex (TDD) scheme, whereby one radio starts transmission on even slots and the other on odd slots. A multislot packet can be extended over 3 and 5 slots (see Figure 3). Due to the frequency-hopping scheme and radio requirements the overhead to transmit data can be quite significant (i.e. the effective maximal use of the channel is for a 1-slot packet (30 bytes) 38%; a 3-slot packet (185 bytes) 78%; and for a 5-slot packet (341 bytes) 87%).

- Fifth, another main contributor to overhead is due to the *transition times* between the various operating modes of the wireless radio. For example, the WaveLAN interface – that has a throughput of 2 Mbit/s –, already takes 250 μ s (or *virtually* 62.5 bytes) to make a transition from sleep to idle. An obvious conclusion is thus that efficient data transmission (in terms of bandwidth utilisation and energy consumption) can only be achieved if the amount of data transmitted is not too small. A protocol that assigns the channel per slot will cause significant overhead due to turnaround, resulting in a significant energy waste. For example, in Bluetooth, transitions are required to be very frequent, and are thus causing a lot of overhead.
- Sixth, in broadcast networks *collisions* may occur (happens mainly at high load situations). This causes the data to become useless and the energy needed to transport that data to be lost.
- Seventh, the *overhead of a protocol* also influences the energy requirements due to the amount of 'useless' control data and the required computation for protocol handling. Typical functions in the protocol stack include routing, congestion control, error control, resource reservation, scheduling, etc. The overhead can be caused by long headers (e.g. for addressing, mobility control, etc), by long trailers (e.g. for error detection and correction), and by the number of required control messages (e.g. acknowledgements).
- Finally, the high *error rate* that is typical for wireless links is another source of energy consumption. First, when the data is not correctly received the energy that was needed to transport and process that data is wasted. Secondly, energy is used for error control mechanisms. On the data link layer level error correction is generally used to reduce the impact of errors on the wireless link. The residual errors occur as burst errors covering a period of up to a few hundred milliseconds. To overcome these errors retransmission techniques or error correction techniques are used. Furthermore, energy is consumed for the calculation and transfer of redundant data packets and an error detection code (e.g. a CRC). Finally, because in wireless communication the error rate and the channel's signal-to-noise ratio (SNR) vary widely over time and space, a fixed-point error control mechanism that is designed to be able to correct errors that rarely occur, wastes energy and bandwidth. If the application is error-resilient, trying to withstand all possible errors wastes even more energy in needless error control. Another strategy is to reduce the data transmission rate or stop data transmission altogether when the channel is bad, i.e. when the probability of a dropped packet is high, so that less transmission time is wasted sending packets that will be dropped [59].

4 Energy reduction in network protocols

Energy reduction should be considered in the whole system of the mobile and through all layers of the protocol stack, including the application. Adaptability of the protocols is a key issue. The two basic principles to achieve an energy efficient system are: *avoid unnecessary actions*, and *reduce the amount of data traffic*.

4.1 Network protocol stack

We will now provide an overview of how these basic principles can be used at the layers of a typical network protocol stack.

- *Physical layer* – To allow a dynamic power management, we need to apply a radio that can be in various operating modes (like variable RF power and different sleep modes). Energy can also be saved if it is able to adapt its modulation techniques and basic error-correction schemes. The bandwidth offered by the radio also influences its energy consumption. The energy per bit transmitted or received tends to be lower at higher bit rates. A low bit-rate radio is less efficient in energy consumption for the same amount of data. However, when a mobile has to listen for a longer period for a broadcast or wake-up from the base station, then the high bit-rate radio consumes more energy than the low bit rate radio. Therefore, the low bit-rate radio should be used for the basic signalling only, and as little as possible for data transfer. This principle is for example used in HIPERLAN.

Example:

HIPERLAN [15] is the wireless LAN specified by the ETSI. Its energy saving is based on two mechanisms: a dual data rate radio, and buffering. Because HIPERLAN is based on a broadcast channel, each station needs to listen to all packets in its range. To decide whether the station is the destination of a packet, each packet is divided into a low-power low bit-rate (1.4706 Mb/s) part to transmit acknowledgement packets and the packet header, and a high power high bit-rate (23.5294 Mb/s) part to transmit the data packet itself.

To minimise the energy consumption the transmit power on the link should be minimised. Many approaches to dynamically changing the transmission power in wireless networks have been proposed. However, few of them were designed with consideration for the battery lifetime of mobile units. They are meant primarily to achieve goals like guaranteeing limits on signal to noise ratio, balancing received power levels, or maximising cell capacities [46]. The consequences of reducing transmission power are not only an increased battery lifetime, but also a lower bit error rate for neighbours (enabling higher cell capacities), and higher bit error rates for one's own transmissions. The transmission power can be chosen based on the current quality of service required and the current interference level observed. Simulations show that such an approach can yield significant energy savings, compared to other schemes that do not consider battery lifetime, such as one that attempts to always maintain a certain signal to noise ratio [46]. Note that reducing transmission power does not always reduce energy consumption, since if transmission power is so low that errors are frequent, the large number of retransmissions and/or the increased number of error correction bits necessary might cause total energy consumption to increase.

Example:

Power control is applied in GSM [47]. There are five classes of mobile stations defined, according to their peak transmitter power, rated at 20, 8, 5, 2, and 0.8 Watts. To minimise co-channel interference and to conserve power, both the mobiles and the Base Transceiver Stations operate at the lowest power level that will maintain an acceptable signal quality. The mobile station measures the signal strength or signal quality (based on the Bit Error Ratio), and passes the information to the Base Station Controller, which ultimately decides if and when the power level should be changed.

- *Link layer* – The link layer is composed out of a Medium Access Control (MAC) protocol and a logical link control protocol. In an energy efficient MAC protocol the basic objective is to minimise all actions of the network interface, i.e. minimise ‘on-time’ of the transmitter as well as the receiver. The medium access protocol can be used to dictate in advance when each wireless device may receive or transmit data. Each device is allowed to sleep when it is certain that no data will arrive for it. For example, the 802.11 LAN standard has access points that buffer data sent to the wireless devices and that periodically broadcast a beacon message indicating which mobile devices have buffered data. A different type of protocol, which does not require buffering access points, divides time into fixed-length intervals. One terminal (e.g. the base station) broadcasts a traffic schedule at the beginning of each interval that dictates when each mobile may transmit or receive data during that interval. Thus, each mobile only needs to be awake during the broadcast of the traffic schedule and may sleep until a next broadcast. Another way to reduce energy consumption is by minimising the number of transitions the wireless interface has to make. By scheduling data transfers in bulk, an

inactive terminal is allowed to doze and power off the receiver as long as the network interface is reactivated at the scheduled time to transceive the data at full speed. More about energy-efficient MAC design can be found in Section 5.

Logical Link Control. Due to the dynamic nature of wireless networks, *adaptive error control* gives significant gains in bandwidth and energy efficiency [23][52]. This avoids applying error-control overhead to connections that do not need it, and it allows to selectively match the required QoS and the conditions of the radio link. In addition to these error control adaptations, a scheduler in the base-station can also adapt its traffic scheduling to the error conditions of wireless connections to a mobile. The scheduler can try to avoid periods of bad error conditions by not scheduling non-time critical traffic during these periods.

Flow control mechanisms are needed to prevent buffer overflow, but also to discard packets that have exceeded the allowable transfer time. Multimedia applications are characterised by their various media streams. Each stream can have different quality of service requirements. Depending on the service class and QoS of a connection a different flow control can be applied so that it minimises the required bandwidth and energy consumption. For instance, in a video application it is useless to transmit images that are already outdated. It is more important to have the 'fresh' images. For such traffic the buffer is probably small, and when the connection is hindered somewhere, the oldest data will be discarded and the fresh data will be used. Flow control would needlessly spend energy for transmitting 'old' images and flow-control messages. An energy-efficient flow control adapts its control mechanism to the requirements of the connection.

- *Network layer* – The network layer takes care of routing packets in the network. Energy efficiency aspects in this layer are mainly studied in the context of ad-hoc networks.

Wireless networks can be classified as distributed (*ad-hoc*) or *centralised* networks. Essentially, the presence or lack of fixed wired infrastructure differentiates between them. Although ad-hoc networks are more flexible than centralised systems, they are less suitable for the design of low energy consuming mobiles [57]. In ad-hoc networks the data possibly has to pass multiple hops before it reaches its final destination. This leads to a waste of bandwidth as well as an increased risk of data corruption, and thus potentially higher energy consumption (due to the required error control mechanism). In a centralised system the base station can be equipped with more intelligent and sophisticated hardware, that probably has a significantly higher energy consumption than the hardware required in the mobile. Portables can then be offloaded with some functionality that will be handled by the base station.

In ad-hoc networks the mobiles cooperate to maintain information about the topology of the network in order to be able to route a packet through the network. In order to update a changed topology update messages must be exchanged through the network. The trade-off is in the gain in improved routing and the required energy and bandwidth. Several authors have presented routing algorithms that attempt to optimise routes while attempting to keep the overhead in energy consumption and bandwidth small (e.g. [41][49][50] and the references therein). Typical metrics used to determine optimal paths are shortest hop, shortest delay, link quality, and location stability. Unfortunately, some of these metrics have a negative impact because they overuse the energy resources of a small set of nodes in favour of others.

For example, consider the network illustrated in Figure 4, shortest-hop routing will route packets between 0-3, 1-4 and 2-5 via node 6, causing node 6 to die relatively early. The authors present in [49] several metrics that result in energy-efficient routes for unicast traffic:

- *Minimise energy consumed per packet.* Under light loads, the selected routes will be the shortest-hop routes. At higher loads, this metric will tend to route packets around congested areas. The main drawback is that nodes tend to have widely differing energy consumption, resulting in early death for some nodes.
- *Maximise time to network partition.* A routing algorithm should divide the work among those mobiles that will cause the network to partition in such a way that these mobiles drain their energy at equal rates.
- *Minimise variance in power levels between mobiles.* This metric ensures that all the mobiles in the network remain powered together for as long as possible.

- *Minimise cost per packet.* The paths selected should be such that mobiles with depleted energy resources do not lie on many paths.
- *Minimise maximum mobile cost.* Minimising the cost per mobile does significantly reduce the maximum mobile cost (and hence time to first mobile failure).

In broadcast traffic it is also worthwhile to know the topology of the network, since only forwarding packets to neighbours that have not received the packet can save energy. The classic approach is *flooding*, in which each mobile send a copy of its received packet to all of its neighbours. A broadcast tree approach is presented in [50], in which priority for routing packets is given to mobiles that have consumed less power and to mobiles that have more neighbours which have not already received the broadcast packets. Simulations not only show that energy is saved, but it also demonstrates very little difference in broadcast delay. In the SPIN protocols [41] mobiles negotiate with each other about the data they possess. This ensures that mobiles only transmit data when necessary.

- *Transport layer* – The transport layer provides reliable data transport, independent of the physical networks used. In the presence of a high packet error rate and periods of intermittent connectivity of wireless links, some network protocols (such as TCP) may overreact to packet losses, mistaking them for congestion. This can have significant consequence for throughput and energy efficiency. Section 6.4 will elaborate more on this issue.

As far as we know there has been no effort on mechanisms to reduce energy consumption at the higher levels of the network protocol stack (e.g. session and presentation layer).

From the previous discussion, it is clear that one should be very careful in transferring principles and protocols from the wired networks to the wireless networks. This can lead to a decrease in performance, but also in a significant increase in energy consumption.

4.2 Operating system and application layer

At the operating system and application layers various mechanisms can be used to improve energy efficiency. Examples in the operating system are hoarding and caching that can reduce the amount of data to be transmitted, and partitioning of tasks between mobiles and fixed hosts that trades energy spend for computation for energy spend in communication. Application layer topics that have addressed energy efficiency are for example database systems and video processing systems.

Hoarding and caching

One attractive way to avert the high cost (either performance, energy consumption or money) of wireless network communication is to avoid use of the network when it is expensive by predicting future access and fetching necessary data when the network is cheap. In the higher level protocols of a communication system, caching and scheduling can be used to control the transmission of messages. This works particularly well when the computer system has the ability to use various networking infrastructures (depending on the availability of the infrastructure at a certain locality), with varying and multiple network connectivity and with different characteristics and costs [11]. True prescience, of course, requires knowledge of the future. Two possible techniques, *LRU caching* and *hoarding*, are for example present in the Coda cache manager [30]. In order to support mobile computers effectively, system designers must view the network as a first-class resource, expending CPU and possibly disk resources to reduce the use of network resources during periods of poor network connectivity.

Decomposition

In a mobile multimedia system many trade-offs can be made concerning the required functionality of a certain mechanism, its actual implementation, and values of the required parameters. In a system functions can be dynamically migrated between functional modules such that an efficient configuration is obtained. Functionality can be partitioned inside a mobile system between a program running on the general-purpose CPU, dedicated hardware components (like a compressor or error correction device), and field programmable hardware devices (like FPGAs) [23].

The *networked* operation of a mobile system opens up additional opportunities for decomposition to increase energy efficiency. One opportunity is offloading computation dynamically from the mobile system, where battery energy is at a premium, to remote energy-rich servers in the wired backbone of the network. In essence, energy spent in communication is traded for computation. For example, when we

consider the transmission of an image over a wireless network, there is a trade-off between image compression, error control, communication, and energy consumption.

Partitioning of functions is an important architectural decision, which specifies where applications can run, where data can be stored, the complexity of the terminal, and the cost of the communication service [32]. The key implication for this architecture is that the runtime hardware and software environment on the mobile computer and in the network should be able to support such adaptability, and provide application developers with appropriate interfaces to control it. Software technologies such as proxies and mobile agents, and hardware technologies such as adaptive and reconfigurable computing are likely to be the key enablers.

Example:

A good example of decomposition that partitions the computational effort for video compression is described by Rabiner [42]. Due to the large amount of data needed for video traffic, efficient compression techniques are important when the video frames are transmitted over wireless channels. Motion estimation has been shown to help significantly in the compression of video sequences. However, since most motion estimation algorithms require a large amount of computation, it is undesirable to use them in power constrained applications, such as battery operated wireless video terminals. Since the location of an object in the current frame can be predicted from its location in previous frames, it is possible to optimally partition the motion estimation computation between battery operated portable devices and high powered compute servers on the wired network.

Figure 5 shows a block diagram of a wireless video terminal in a networked environment. A resource on the wired network with no power constraints can estimate the motion of the sequence based on the previous (decoded) frames and use this information to predict the motion vectors of the current frame. This can achieve a reduction in the number of operations performed at the encoder for motion estimation by over two orders of magnitude while introducing minimal degradation to the decoded video compared with full search encoder-based motion estimation.

Video processing systems

As indicated in the previous example, video processing and communication requires a significant amount of energy due to the large amounts of data to be processed. Error control is an important issue in video transmission over a wireless channel. It has been shown that ARQ and hybrid ARQ schemes can significantly improve the video transmission reliability, and can provide a much higher throughput than FEC schemes because they can adapt effectively to the varying channel conditions [34]. Adaptive source rate control schemes use a forecast of the channel condition to encode a video frame. Layered coding can be a useful tool for gaining resilience to errors. For example, in video standards such as MPEG-II the protection of I and P frames can be performed at the expense of the less important B frames. However, these techniques add redundancy thereby lowering the coding efficiency. FEC can also cause the failure mode to become more abrupt, as FEC can add more errors to a bitstream once the error correcting capability of the code has been exceeded.

To solve such problems, different coding techniques can be employed to achieve a better efficiency of the channel, and thus reducing the energy consumption.

In [54] the authors conclude that it is the loss of bitstream synchronisation, which is the primary cause of corrupted pictures. They propose to obtain synchronisation using a technique known as error resilient entropy coding, whereby all variable length datablocks always start at known positions in the bitstream. The bitstream is then re-ordered without adding redundancy such that longer blocks fill up spaces left by shorter blocks. Given bitstream synchronisation, the differential coding of the DC coefficients and motion vectors in MPEG-II cause the most visible artefacts. These appear as corrupted horizontal stripes. The authors employ a hierarchical pyramid based coding scheme to code these parameters. This technique increases the error resilience and the coding efficiency is usually improved.

In [17] a low power scalable secret key encryption scheme is proposed for secure transmission of video data. They propose a variety of techniques to reduce the power consumption. Secret key systems can be categorised into either block or stream ciphers. Stream ciphers have a better error resilience, since block ciphers will cause single bit errors in the encrypted block to propagate into multiple bit errors in the decrypted block. In addition, by dynamically scaling the encryption algorithms, the system can allocate security where it is needed within the data stream. For example, consider a differential video encoding scheme in which the initial frame is transmitted uncompressed, and then a sequence of difference frames

are transmitted. In this example the initial frame would have a higher priority than the differential frames. This is similar to the idea of priority encoding that is used to allocate additional error recovery coding for portions of the data stream that are deemed important, and reduced error correcting coding for lower priority portions of the data stream.

Information dissemination systems

Indexing mechanism as used in TDMA protocols can also be used for wireless broadcasting of data as a way of disseminating information to a massive number of users. Imielinski et al. [28][29] distinguish between two fundamental modes of providing users with information:

- *Data broadcasting*: periodic broadcasting of data on a communication channel. Accessing broadcasted data does not require uplink transmission.
- *Interactive/on-demand*: The client requests data on the uplink channel and the server responds by sending this data to the client.

In practice, a mixture of the above two methods will be used. However, even in pure on-demand mode it makes sense to batch requests for the same data and send the data once rather than cater individually to each request. To save energy, it is definitely beneficial when mobiles slip into doze mode most of the time, and come into active mode only when the data of interest is expected to arrive. The ability to wake up when data is expected is termed *selective tuning*. Imielinski et al. proposed selective tuning techniques, where clients (mobile devices) tune in periodically to the broadcast channel to download required data. When a server broadcasts the data, it also broadcasts a directory, which consists of an index indicating the time when particular records are broadcast. The index, which provides a sequence of pointers, which eventually leads to the required data, is interleaved with data. Imielinski et al. established that with these techniques, the same number of filtering requests could be served with 100 fold less energy. The savings almost doubled the working time of the mobile.

5 Energy-efficient MAC design

The medium access control protocol is responsible for driving the raw transmission facility of the physical channel, data framing, efficient sharing of the channel, and possibly some error control. The objective of an energy efficient MAC protocol design is to maximise the performance while minimising the energy consumption of the *mobile*. These requirements often conflict, and a trade-off has to be made.

On this level energy can be saved when the transceiver of the mobile can be in a low power state as much as possible, and when transmissions are successful as often as possible. The use of slotted channels allows to maintain enough synchronisation between the sender and receiver to allow a low energy waste since the nodes are then capable of predicting when packet transmissions will begin and end. Error control will be topic of the next section.

5.1 Mechanisms

There are several options for an energy-efficient MAC protocol design. In the following we outline some MAC design options.

- *Avoid unsuccessful actions of the transceiver*

Collisions and errors cause unsuccessful actions. Every time a *collision* occurs energy is wasted because the same transfer has to be repeated again after a backoff period. A protocol that does not suffer from collisions can have better throughput even under high load conditions. These protocols generally also have good energy consumption characteristics. However, if it requires the receiver to be turned on for long periods of time, the advantage diminishes. A protocol, in which a base-station broadcasts traffic control for all mobiles in range with information about when a mobile is allowed to transmit or is supposed to receive data, reduces the occurrence of collisions significantly. Collisions can only occur when new requests have to be made. New requests can be made per packet in a communication stream, per application of a mobile, or even per mobile. The trade-off between efficient use of resources and QoS determines the size to which a request applies. Note that this might waste bandwidth (but not energy) when slots are reserved for a request, but not always used.

Errors on the wireless link can be overcome by mechanisms like retransmissions or error correcting codes. In Section 6 error control will be discussed in more detail. A different strategy to reduce the

effect of errors is to avoid traffic during periods of bad error conditions. When the MAC scheduler tries to *avoid periods of bad error conditions* by not scheduling (non-time critical) traffic during these periods, energy can be saved and performance of the system is increased because only connections with a good probability of success are scheduled. This can lead to a throughput that may even exceed the average rate on the channel, due to the introduced dependence between admitted connections and channel quality [10].

- *Minimise the number of transitions*

Scheduling traffic into bursts in which a mobile can continuously transmit or receive data – possibly even bundled for different applications –, can reduce the number of transitions. Notice, however, that there is a trade-off with QoS parameters like delay and jitter.

One step further is to rearrange the scheduling of transmissions and receptions in a frame such, that a mobile can be in an operational mode for a longer consecutive period. Most MAC protocols are based on phase grouping that basically has three phases in a frame: uplink, downlink and reservation. In [21] a mechanism is presented in which there are multiple uplink and downlink phases, in which these phases are grouped per mobile in a frame. Although this has a negative effect on the capacity of the channel, it allows the mobile to turn the power off from the wireless interface for a longer period. They made this choice since in a mobile multimedia environment it is more important that connections have a certain QoS, than highest possible bandwidth. They give an example that shows that by using this method, the on time of the mobile's network interface is decreased from approx. 44% with phase grouping to approx. 24% with mobile grouping.

- *Power management*

The wireless interface should operate at the lowest energy consuming operation mode as much as possible. For example, in the 802.11 protocol [26] the mobile is allowed to turn off and the base station buffers data destined for the mobile meanwhile. The mobiles have to wake up at the same time the base station announces buffered frames for the receiver. This mechanism saves energy but also influences the QoS for the connections drastically. Such store-and-forward schemes for wireless networks not only allow a network interface to enter a sleep mode but can also perform local retransmissions not involving the higher network protocol layers. However, these schemes have the disadvantage of requiring a third party, e.g. a base station, to act as a buffering interface.

Example:

HIPERLAN does not need a dedicated base station, but any station can become a so-called forwarder. Forwarders use a forwarding mechanism to build the infrastructure. The physical size of a HIPERLAN is thus a function of the current position of all stations. Power saving is based on a contract between at least two stations. The station that wants to save power is called the power-saver, and the station that supports this is the power-supporter. Power-supporters have to queue all packets destined for one of its power-savers. Forwarders and power-supporters are not expected to be mobiles since they have to receive, buffer, and forward packets sent to one of its clients. The power-saver is active only during pre-arranged intervals. Since this interval is minimal 500 ms, it cannot be used for most time-bounded traffic [57].

Synchronisation between the mobile and the base station is beneficial for both uplink (mobile host to base station) and downlink (base station to mobile host) traffic. When the base-station and mobile are synchronised in time, the mobile can go in standby or off mode, and wake up just in time to communicate with the base-station. The premise is that the base has plenty of energy and can broadcast its beacon frequently. The application of a mobile with the least tolerable delay determines the frequency by which a mobile needs to turn its receiver on. If the wake-up call of the communication is implemented with a low-power low-performance radio, instead of the high-performance high-energy consuming radio, then the required energy can be reduced even more.

Example:

Power management is also an essential part of the Bluetooth protocol. The Bluetooth radio can be in various operational modes to support reduced power consumption for inactive radios or to provide the slave radio with increased capacity to function in other operational states. To give an idea of the power consumption for a Bluetooth radio we use the preliminary specs of the Ericsson PBA 313 01/2. Operating at 3.3 V it has a current of 40 mA in receive mode and 35 mA in transmit mode. The modes are Active, Sniff, Hold, and Park. In Active mode, the radio

actively participates, and the master periodically transmit to the slaves to maintain synchronisation. In Sniff mode, a slave radio may save power by reducing its duty cycle. The slave radio may then enter a reduced power state for those slots where it is not expected to receive a packet. Hold and Park mode lets the slave radio enter reduced power modes for extended periods of time. The consequence is that the slave radio requires more time to return to Active mode.

The principle of synchronisation between mobile and base station has been used for some time in paging systems [38]. Paging systems increase battery life by allowing the receiver to be turned off for a relatively long time, while still maintaining contact with the paging infrastructure using a well designed synchronous protocol using various forms of TDM.

Example:

A similar mechanism is applied in GSM that saves energy at the mobile station using discontinuous reception [47]. The paging channel, used by the base station to signal an incoming call, is structured such that the mobile station knows when it needs to check for a paging signal. In the time between paging signals, the mobile can go into sleep mode, where almost no power is used. GSM also uses discontinuous transmission (DTX) that takes advantage of the fact that a person speaks less than 40 percent of the time in normal conversation by turning the transmitter off during silence periods.

Although a fixed frame size can save energy because the radio can be in the lowest-power operating mode most of the time because the schedule is very deterministic, Lettieri [33] shows that there is much to be gained from variable frame length in terms of user seen throughput, effective transmission range, and transmitter power for wireless links. The latter point of view, however, was inspired by the high error rate on wireless links, where a high error rate on a large frame might not be efficient.

There are many ways in which these principles can be implemented. Several researchers (e.g. [4][9][57]) have compared various mechanisms.

6 Energy-efficient error control

Since high error rates are inevitable to the wireless environment, *energy-efficient error-control* is an important issue for mobile computing systems. This includes energy spent in the physical radio transmission process, as well as energy spent in computation, such as signal processing and error control at the transmitter and the receiver. Error control is an issue that spans several layers, from the physical layer (power control), up to the transport layer, and even applications.

6.1 The error model

Wireless networks have a much higher error rate than the normal wired networks. The errors that occur on the physical channel are caused by phenomena such as signal fading, transmission interference, and user mobility. In characterising the wireless channel, there are two variables of importance. First, there is the Bit Error Rate (BER) – a function of Signal to Noise Ratio (SNR) at the receiver -, and second the burstiness of the errors on the channel. Figure 6 presents a graphical view of packets moving through this channel [31].

This leads to two basic classes of errors: packet erasures and bit corruption errors [12][58]. Error control is applied to handle these errors. Note that even a single uncorrected bit error inside a packet will result in the loss of that packet.

6.2 Error-control alternatives

There are a large variety of error-control strategies, each with its own advantages and disadvantages in terms of latency, throughput, and energy efficiency. Basically there are two methods of dealing with errors: retransmission, also known as Automatic Repeat reQuest (ARQ), and Forward Error Correction (FEC). Hybrids of these two also exist. Adaptive error control adapts the applied error control to the observed channel conditions. In the following we will describe the consequences for the energy consumption of these alternatives.

- With *FEC* redundancy bits are attached to a packet allowing the receiver to correct errors which may occur. In principle, FEC incurs a fixed overhead for every packet, irrespective of the channel conditions. This implies a reduction of the achievable data rate and causes additional delay. When the channel is good, we still pay this overhead. Areas of applications that can benefit in particular from error-correction mechanisms are *multicast applications* [45][40].
- Using *ARQ*, feedback information is propagated in the reverse direction to inform the sender of the status of packets sent. The use of ARQ may result in an even more significant increase of delay and delay variations than FEC [48]. The retransmission requires additional buffering at the transmitter and receiver. A large penalty is paid in waiting for and carrying out the retransmission of the packet. This can be unacceptable for systems where Quality of Service (QoS) provisioning is a major concern, e.g. in wireless multimedia systems. Solutions to provide a predictable delay at the medium access control layer by reserving bandwidth for retransmission are possible [16], but spoil bandwidth.

ARQ schemes will perform well when the channel is good, since retransmissions will be rare, but perform poorly when channel conditions degrade since much effort is spent in retransmitting packets. Another often ignored side effect in ARQ schemes is that the round-trip-delay of a request-acknowledge might also cause the receiver to be waiting for the acknowledge with the receiver turned 'on', and thus wasting energy.

Classic ARQ protocols overcome errors by re-transmitting the erroneously received packet, regardless of the state of the channel. Although in this way these retransmission schemes *maximise the performance* – as soon as the channel is good again, packets are received with minimal delay – the consequence is that they expend energy. When the tolerable delay is large enough, ARQ outperforms error-correction mechanisms, since the residual error probability tends to zero in ARQ with a much better energy efficiency than FEC methods [60].

- *Hybrids* do not have to transmit with maximum FEC redundancy to deal with the worst possible channel. Under nominal channel conditions, the FEC will be sufficient, while under poor channel conditions ARQ will be used. Although more efficient than the pure categories, a hybrid system is still a rigid one since certain channel conditions are assumed.
- *Adaptive error control* allows the error-control strategy to vary as the channel conditions change. The error control can be FEC, ARQ, or a hybrid. The wireless channel quality is a function of the distance of user from base station, local and average fading conditions, interference variations, and other factors. In such a dynamic environment it is likely that any of the previous schemes is not optimal in terms of energy efficiency all the time. Adaptive error control seems likely a source of efficiency gain.

6.3 Examples of adaptive error control

Adaptive error control requires a feedback loop to allow the transmitter to adapt the error coding according to the error rate observed at the receiver. Normally, such information consists of parameters such as mean carrier-to-interference ratio (C/I), or signal-to-noise ratio (SNR), bit error statistics, and packet error rate. The feedback loop limits the responsiveness to the wireless link conditions. Information can also be gathered by purely relying on acknowledgement (ACK/NACK) information. Although quite effective for energy efficiency, adaptive error-control is mainly used to improve the throughput on a wireless link [14]. Schuler presents in [48] some considerations on the optimisation and adaptation of FEC and ARQ algorithms with focus on wireless ATM developments. The optimisation, with respect to the target bit error rate and the mapping of the wireless connection quality to the ATM QoS concept, is discussed in detail. Eckhardt and Steenkiste [13] argue and demonstrate that protocol-independent link-level local error control can achieve high communication efficiency even in a highly variable error environment, that adaptation is important to achieve this efficiency, and that inter-layer coexistence is achievable.

The choice of energy-efficient error-control strategy is a strong function of QoS parameters, channel quality, and packet size [31]. Several studies have shown that adaptive packet sizing and adaptive error control can significantly increase the throughput of a wireless LAN, using relative simple adaptation policies (e.g. [14][39]).

Adaptive error control can be applied in several ways and by several entities.

- *Avoid periods of bad channel conditions*

This mechanism is based on the principle that it is useless to transmit data when it is known that the receiver has little chance of receiving the data correctly. Chockalingam and Zorzi in [10] proposed a mechanism to avoid transmission during bad channel periods in order to reduce the number of unsuccessful transmissions. Zorzi describes in [60] an adaptive probing ARQ strategy that slows down the transmission rate when the channel is impaired without a significant loss in throughput. The protocol works normal until the transmitter detects an error due to the lack of an acknowledgement. At that time the protocol enters a probing mode in which a probing packet is transmitted regularly. A modified scheme is also analysed, which yields slightly better performance, but requires some additional complexity.

A bit more elaborated scheme is applied by Havinga and Smit [22]. They have developed a wireless communication system (network interface, a MAC protocol E²MaC, QoS mechanisms) that is dedicated to energy-efficient wireless networking for multimedia traffic. Traffic over the wireless link is scheduled by the base station based on the QoS requirements of the connections and on the current channel conditions. The scheduler tries to avoid only non-time critical traffic during bad channel periods, thereby not affecting traffic with demanding QoS.

- *Differentiate connection streams*

Since different connections do not have the same requirements concerning e.g. cell loss rate and cell transfer delay, different error-control schemes must be applied for different connection types. The error control mechanisms can be adapted to the current error condition in such a way that it minimises the energy consumption needed and still provides (just) enough fault tolerance for a certain connection. This avoids applying error control overhead to connections that do not need it, and allows the possibility to apply it selectively to match the required QoS and the conditions of the radio link. Several researchers (e.g. Havinga [22], and Lettieri [31]) have applied this principle.

The slot scheduler of E²MaC [23] groups transmissions of mobiles together as much as possible within the QoS constraints of a particular connection in order to minimise the number of operating mode transitions. The data link layer treats each connection with its own set of characteristics and requirements (flow control, error control, bandwidth and latency requirements) such that each connection will receive a satisfactory quality of service, and thus not seeks to achieve the highest performance possible. A fast feedback channel is provided that is used for both flow control and error status.

Lettieri [33] describes how energy efficiency in the wireless data link can be enhanced via adaptive frame length control in concert with adaptive error control based on hybrid FEC and ARQ. The length and error coding of the frame going over the air and the retransmission protocol are selected for each application stream based on QoS requirements and continually adapted as a function of varying radio channel conditions.

- *Application layer adaptations*

Depending on the application, the adaptation might not need to be done frequently. If, for example, the application is an error-resilient compression algorithm that when channel distortion occurs, its effects will be a gradual degradation of video quality, then the best possible quality will be maintained at all BERs ([3][54]).

All error-control techniques introduce latency, a problem that is more prominent in case of limited bandwidth. This poses the problem that low latency (for interactivity) and high reliability (for subjective quality) are fundamentally incompatible under high traffic conditions. Some multimedia applications might, however, be able to use the possibly corrupt packet. With *multiple-delivery transport service* multiple possibly corrupt but increasingly reliable versions of a packet are delivered to the receiving application [18]. The application has the option of taking advantage of the earlier arriving corrupt packet to lower the perceived latency, but eventually replaces them with the asymptotically reliable version.

In the concept of *incremental redundancy* (IR) [39], redundant data, for the purpose of error correction, is transmitted only when previously transmitted packets of information are received and acknowledged to be in error. The redundant packet is combined with the previously received (errored) information packets in order to facilitate error correction decoding. If there is a decoding failure, more redundancy is transmitted. The penalty paid for increased robustness and higher throughput is additional receiver memory and higher delay.

6.4 Local versus end-to-end error-control

The networking community has explored a wide spectrum of solutions to deal with the wireless error environment. They range from local solutions that decrease the error rate observed by upper layer protocols or applications, to transport protocol modifications and proxies inside the network that modify the behaviour of the higher level protocols [13][43].

Addressing link errors near the site of their occurrence seems intuitively attractive because they understand their particular characteristics [20] and are likely to respond more quickly to changes in environment. Performing FEC on an end-to-end basis implies codes that deal with a variety of different loss and corruption mechanisms, even on one connection. In practice this implies that different codes have to be concatenated to deal with every possible circumstance, and the resulting multiple layers of redundancy would be carried by every link with a resultant traffic and energy consumption penalty [19].

While local error-control is attractive in terms of simplicity, local error control alters the characteristics of the network, which can confuse higher layer protocols. For example, local retransmission could result in packet reordering or in large fluctuations of the round-trip time, either of which could trigger timeouts and retransmissions. In addition, end-to-end control has potentially better knowledge of the quality requirements of the connection.

The *Transport Control Protocol* (TCP) is a reliable, end-to-end, transport protocol that is widely used to support various applications. In the presence of a high packet error rate and periods of intermittent connectivity of wireless links, TCP may overreact to packet losses, mistaking them for congestion. TCP responds to all losses by invoking congestion control and avoidance algorithms. These measures result in an unnecessary reduction in the link's bandwidth utilisation and increases in energy consumption because it leads to a longer transfer time. To alleviate the effects of non-congestion-related losses on TCP performance over high-loss networks (like wireless networks), several schemes have been proposed.

These schemes choose from a variety of mechanisms to improve end-to-end throughput, such as local retransmissions, split connections and forward error correction. In [5] several schemes have been examined and compared. These schemes are classified into three categories: end-to-end protocols, where the sender is aware of the wireless link; link-layer protocols, that provide local reliability and shields the sender from wireless losses; and split-connection protocols, that break the end-to-end connection into two parts at the base station. Their results show that a reliable link-layer protocol with some knowledge of TCP provides good performance, more than using a split-connection approach. Selective acknowledgement schemes are useful, especially when the losses occur in bursts.

A different point of view is presented in [61] that analyses the energy consumption performance of various versions of TCP. They argue that in some cases the window adaptation algorithm used in TCP may in fact be more efficient for wireless channels than predicted by those early studies. The basic reason for this is the assumption that burst errors in a wireless link are long relative to the propagation delay of the connection. This assumption is likely to be true for local TCP connections. As discussed in the previous section, efficient usage of energy is achieved by avoiding periods of bad channel conditions. In fact, this is exactly what the window adaptation algorithm of TCP does.

7 Conclusions

In this paper we have identified the most prominent problems of wireless multimedia networking and presented several state-of-the-art solutions. We focussed on energy efficiency of mobile multimedia systems. Key problems of portable wireless network devices are that they need to handle multimedia traffic in a dynamic and heterogeneous wireless environment, and need to operate with limited energy resources. To achieve sufficient performance and energy efficiency, *adaptability* is important, as wireless networks are dynamic by nature. One of the key issues in the design of portable multimedia systems is to find a good balance between flexibility and high-processing power on one side, and energy-efficiency of the implementation on the other side.

We have identified the following main principles that are used in the design of energy efficient mobile multimedia systems:

1. Energy efficiency is an issue *involving all layers of the system*. We have presented an overview of energy conserving mechanisms at the lower layers of a typical network protocol stack. Much of the work in this area used a single layer view, i.e. it tried to optimise one protocol or one area of the

protocol stack. Even though there are several successful single-layer mechanisms, there remains doubt about whether approaches, which focus on a single layer in isolation, are viable. I.e. instead of trying to save energy at each separate layer, like trying to implement TCP efficiently for wireless links, applying energy saving techniques that impact all layers of the protocol stack can save more energy. This not only because it is expected that integrated solutions outperform segregated solutions, but more importantly because in many cases some level of negative interaction has been observed between schemes at different layers. Future research on this can be very useful.

2. Since system architecture, operating system, communication, energy consumption and application behaviour are closely linked, a *Quality of Service* framework is a sound basis for integrated management of all resources, including the batteries. QoS is an essential mechanism for mobile multimedia systems not only to give users an adequate level of service, but also as a tool to achieve an energy efficient system. Due to the dynamic wireless environment, *adaptability* of the system will be a key issue in achieving this.

Advances in technology enable portable computers to be equipped with wireless interfaces, allowing networked communication even while on the move. However, its energy resources available will limit the amount of functionality. Energy-efficiency will become the prominent factor in the usability of future mobile multimedia systems.

Acknowledgement

We would like to thank Lodewijk Smit for the comments on the draft version of this paper, and the anonymous reviewers, whose comments helped to improve the paper.

References

- [1] "Bluetooth Specification Version 1.0 B", http://www.bluetooth.com/developer/specification/core_10_b.pdf.
- [2] "Ericsson PBA 313 01/2 Bluetooth Radio", November 1999.
- [3] Agrawal P., Chen J-C, Kishore S., Ramanathan P., Sivalingam K.: "Battery power sensitive video processing in wireless networks", *Proceedings IEEE PIMRC'98*, Boston, September 1998.
- [4] Akyildiz I.F., McNair J., Martorell L.C., Puigjaner R., Yesha Y.: "Medium Access Control protocols for multimedia traffic in wireless networks", *IEEE Network*, pp.39-47, July/August 1999.
- [5] Balakrishnan H., et al.: "A comparison of mechanisms for improving TCP performance over wireless links", *Proceedings ACM SIGCOMM'96*, Stanford, CA, USA, August 1996.
- [6] Benini L., De Micheli G.: "Dynamic Power Management, design techniques and CAD tools", *Kluwer Academic Publishers*, ISBN 0-7923-8086-X, 1998.
- [7] Bhoedjang, R.A.F., Rühl T., Bal H.E.: "User-level network interface protocols", *Computer*, November 1998, pp. 53-60.
- [8] Chen T.-W., Krzyzanowski P., Lyu M.R., Sreenan C., Trotter: "Renegotiable Quality of Service – a new scheme for fault tolerance in wireless networks", *Proceedings FTCS'97*, 1997.
- [9] Chen, et al. "Comparison of MAC Protocols for Wireless Local Networks Based on Battery Power Consumption", *IEEE Infocom'98*, San Francisco, USA, pp. 150-157, March 1998.
- [10] Chockalingam, A., Zorzi, M.: "Energy consumption performance of a class of access protocols for mobile data networks", *VTC'98*, Ottawa, Canada, May 1998.
- [11] Ebling, M.R., Mummert, L.B., Steere D.C.: "Overcoming the Network Bottleneck in Mobile Computing", *Proceedings of the IEEE Workshop on Mobile Computing Systems and Applications*, Dec. 1994, Santa Cruz, CA.
- [12] Eckhardt D., Steenkiste P.: "Measurement and analysis of the error characteristics of an in building wireless network", *Proceedings of the SIGCOMM '96 Symposium on Communications Architectures and Protocols*, pp. 243-254, Stanford, August 1996, ACM.
- [13] Eckhardt D.A., Steenkiste P.: "Improving wireless LAN performance via adaptive local error control", *Sixth IEEE International conference on network protocols (ICNP'98)*, Austin, October 1998.
- [14] Elaoud, M, Ramanathan, P.: "Adaptive Use of Error-Correcting Codes for Real-time Communication in Wireless Networks", *proceedings IEEE Infocom'98*, pp. 548-555, March 1998.
- [15] ETSI: "High Performance Radio Local Area Network (HIPERLAN) ", *draft standard ETS 300 652*, March 1996.

- [16] Figueira, N.R., Pasquale, J.: "Remote-Queueing Multiple Access (RQMA): Providing Quality of Service for Wireless Communications", *proceedings IEEE Infocom'98*, pp. 307-314, March 1998.
- [17] Goodman J., Chandrakasan A.P.: "Low power scalable encryption for wireless systems", *Wireless Networks* 4, 1998, pp. 55-70.
- [18] Han R.Y., Messerschmitt: "Asymptotically reliable transport of multimedia/graphics over wireless channels", *Proc. Multimedia Computing and Networking*, San Jose, Jan. 29-31, 1996.
- [19] Haskell P., Messerschmitt D.G.: "In favor of an enhanced network interface for multimedia services", *IEEE Multimedia Magazine*, 1996.
- [20] Haskell P., Messerschmitt D.G.: "Some research issues in a heterogeneous terminal and transport environment for multimedia services", *Proc. COST #229 workshop on adaptive systems, Intelligent Approaches, Massively Parallel Computing and Emerging Techniques in Signal Processing and Communications*, Bayona, Spain, Oct. 1994.
- [21] Havinga P.J.M., Smit G.J.M., Bos M.: "Energy efficient adaptive wireless network design", *The Fifth Symposium on Computers and Communications (ISCC'00)*, Antibes, France, July 3-7, 2000
- [22] Havinga P.J.M., Smit G.J.M., Bos M.: "Energy efficient wireless ATM design", *ACM/Baltzer Journal on Mobile Networks and Applications (MONET), Special issue on Wireless Mobile ATM technologies, Vol. 5, No 2., 2000.*
- [23] Havinga P.J.M., "Mobile Multimedia Systems", *Ph.D. thesis University of Twente*, February 2000, ISBN 90-365-1406-1, <http://www.cs.utwente.nl/~havinga/thesis>.
- [24] Havinga P.J.M., Smit G.J.M.: "Octopus: embracing the energy efficiency of handheld multimedia computers", *proceedings fifth annual ACM/IEEE international conference on mobile computing and networking (Mobicom'99)*, pp.77-87, August 1999.
- [25] Hyden E. A., "Operating System support for Quality of Service", *Ph.D. thesis, University of Cambridge*, 1994.
- [26] IEEE, "Wireless LAN medium access control (MAC) and physical layer (PHY) Spec." P802.11VD5, *Draft Standard IEEE 802.11*, May 1996.
- [27] Ikeda T.: "ThinkPad Low-Power Evolution", *IEEE Symposium on Low Power Electronics*, October 1994.
- [28] Imielinski T., Viswanathan S., Badrinath B.R.: "Energy efficient indexing on air", *Proc. SIGMOD 1994*, pp. 25-36.
- [29] Imielinski T., Viswanathan S., Badrinath B.R.: "Data on air: Organization and Access", *IEEE Transactions on knowledge and data engineering*, Vol. 9, No. 3, May/June 1997, pp.353-371.
- [30] Kistler J.J.: "Disconnected operation in a distributed file system", PhD thesis, *Carnegie Mellon University, School of Computer Science*, 1993.
- [31] Lettieri P., Schurgers C., Srivastava M.B.: "Adaptive link layer strategies for energy efficient wireless networking", *ACM Wireless Networks*, Vol. 5, No. 5, pp.339-355, Oct. 1999.
- [32] Lettieri P., Srivastava M.B.: "Advances in wireless terminals", *IEEE Personal Communications*, pp. 6-19, February 1999
- [33] Lettieri, P., Srivastava, M.B.: "Adaptive Frame Length Control for Improving Wireless Link Throughput, Range, and Energy Efficiency", *IEEE Infocom'98*, San Francisco, USA, pp. 307-314, March 1998.
- [34] Liu H., Zarki M.: "Adaptive source rate control for real-time wireless video transmission", *Mobile Networks and Applications* 3, 1998, pp. 49-60.
- [35] Lorch, J., Smith, A. J.: "Software strategies for portable computer energy management", *IEEE Personal Communications Magazine*, 5(3):60-73, June 1998.
- [36] Lorch, J.R.: "A complete picture of the energy consumption of a portable computer", *Masters thesis, Computer Science, University of California at Berkeley*, 1995
- [37] Mangione-Smith, B. et al.: "A low power architecture for wireless multimedia systems: lessons learned from building a power hog", *proceedings of the international symposium on low power electronics and design (ISLPED) 1996*, Monterey CA, USA, pp. 23-28, August 1996.
- [38] Mangione-Smith, B.: "Low power communications protocols: paging and beyond", *Low power symposium 1995*, <http://www.icsl.ucla.edu/~billms/Publications/pagingprotocols.pdf>.
- [39] Nobelen R. van, Seshadri N., Whitehead J., Timiri S.: "An adaptive radio link protocol with enhanced data rates for GSM evolution", *IEEE Personal Communications*, pp. 54-64, February 1999.
- [40] Nonnenmacher, J., Biersack, E.W.: "Reliable multicast: where to use Forward Error Correction", *Proceedings 5th workshop on protocols for high speed networks*, pp. 134-148, Sophia Antipolis, France, Oct. 1996.
- [41] Rabiner Heinzelman W., Kulik J., Balakrishnan: "Adaptive protocols for information dissemination in wireless sensor networks", *proceedings MOBICOM '99*, Seattle, USA, pp. 174-185, August 1999.

- [42] Rabiner W., Chandrakasan A.: "Network-Driven Motion Estimation for Wireless Video Terminals", *IEEE Transactions on Circuits and Systems for Video Technologies*, Vol. 7, No. 4, August 1997, pp. 644-653.
- [43] Reiner Ludwig, Katz Randy H., "The Eifel Algorithm: Making TCP Robust Against Spurious Retransmissions", *ACM Computer Communications Review*, Vol. 30, No. 1, January 2000
- [44] Reiniger D., Izmailov R., Rajagopalan B., Ott M., Raychaudhuri D.: "Soft QoS control in the WATMnet broadband wireless system", *IEEE Personal Communications*, pp. 34-43, February 1999.
- [45] Rizzo, L.: "Effective Erasure Codes for Reliable Computer Communication Protocols", *ACM Computer Communication Review*, Vol. 27- 2, pp. 24-36, April 1997.
- [46] Rulnick J.M., Bambos N.: "Mobile power management for maximum battery life in wireless communication networks", Proc. IEEE Conf. On Computer Communications (INFOCOM'96), San Francisco, CA, March 1996, pp.443-450.
- [47] Scourias, J.: "A brief overview of GSM", University of Waterloo, <http://kbs.cs.tu-berlin.de/~jutta/gsm/js-intro.html>.
- [48] Schuler C.: "Optimization and adaptation of error control algorithms for wireless ATM", *International Journal of Wireless Information Networks*, Vol. 5, No. 2, April 1998.
- [49] Singh S., Woo M., Raghavendra C.S. "Power-aware routing in mobile ad hoc networks", *proceedings MOBICOM 98*, pp. 181-190, October 1998.
- [50] Singh S., Raghavendra C.S., J. Stephanek, "Power-aware broadcasting in mobile ad hoc networks", *technical report Oregon State University, Department of Electrical and Computer Engineering*, 1999.
- [51] Sivalingam, K.M., Chen J.C., Agrawal, P., Srivastava, M.B.: "Design and analysis of low-power access protocols for wireless and mobile ATM networks", *ACM/Baltzer Wireless Networks*, Vol. 6, No. 1, pp.73-87, Feb. 2000.
- [52] Srivastava M.: "Design and optimization of networked wireless information systems", *IEEE VLSI workshop*, April 1998.
- [53] Stemm, M, et al.: "Reducing power consumption of network interfaces in hand-held devices", *Proceedings mobile multimedia computing MoMuc-3*, Princeton, Sept 1996.
- [54] Swann R.: "Bandwidth efficient transmission of MPEG-II Video over noisy mobile links", *Signal Processing*, Vol. 12, No. 2, pp. 105-115, April 1998.
- [55] Truman T.E., Pering T., Doering R., Brodersen R.W.: "The InfoPad multimedia terminal: a portable device for wireless information access", *IEEE transactions on computers*, Vol. 47, No. 10, pp. 1073-1087, October 1998.
- [56] WaveMODEM 2.4 GHz Data Manual, Release 2, AT&T 1995.
- [57] Woesner H., Ebert J., Schläger M., Wolisz A.: "Power-saving mechanisms in emerging standards for wireless LANs: The MAC level perspective", *IEEE Personal Communications*, Vol. 5, No. 3, June 1998.
- [58] Zorzi, M., Rao, R. R.: "On the impact of burst errors on wireless ATM", *IEEE Personal Communications*, August 1999, pp.65-76.
- [59] Zorzi, M., Rao, R.R.: "Error control and energy consumption in communications for nomadic computing", *IEEE transactions on computers*, Vol. 46, pp. 279-289, March 1997.
- [60] Zorzi, M: "Performance of FEC and ARQ Error control in bursty channels under delay constraints", *VTC'98*, Ottawa, Canada, May 1998.
- [61] Zorzi M., Rao R.R.: "Is TCP energy efficient? ", *Proceedings IEEE MoMuC*, November 1999.

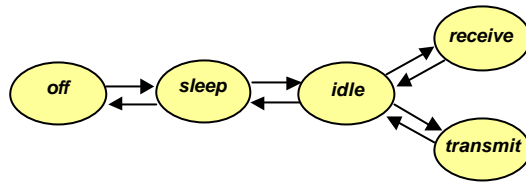


Figure 1: Typical operating modes of a wireless modem.

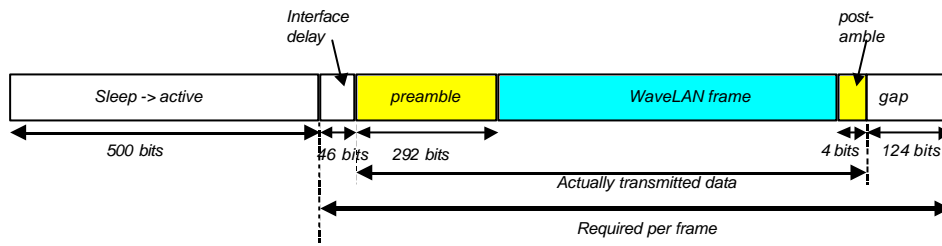


Figure 2: WaveLAN physical layer block format.

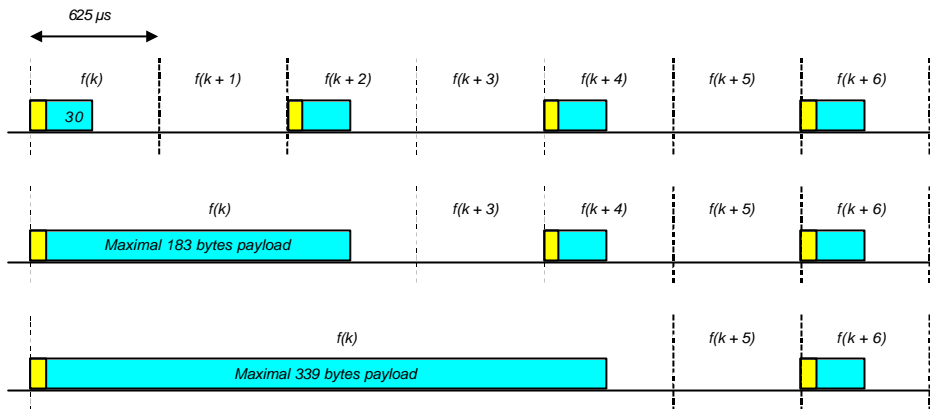


Figure 3: Bluetooth single and multislot radio packets

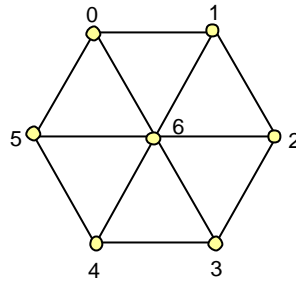


Figure 4: Ad-hoc network example.

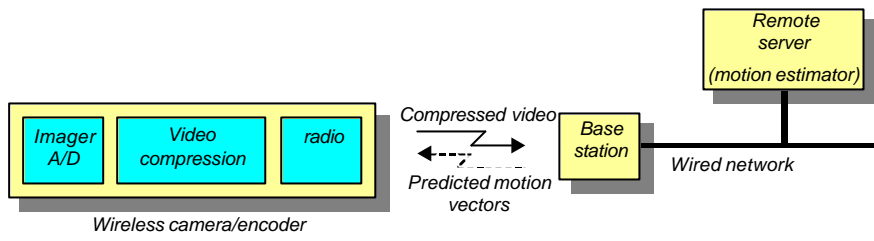


Figure 5: Partitioned video compression.

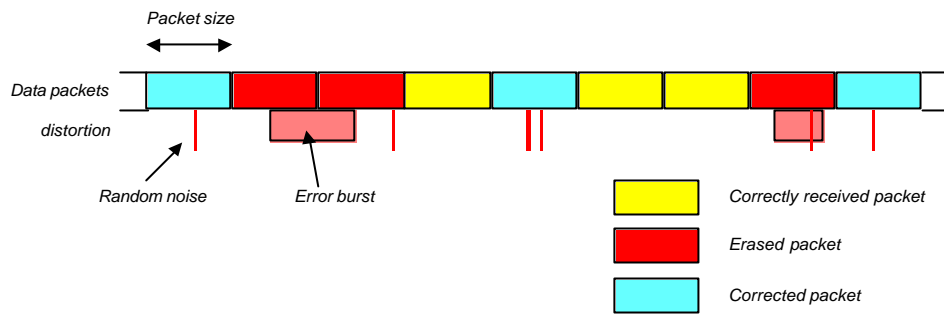


Figure 6: Error characteristics and packet erasures.