

QoS Scheduling for Energy-Efficient Wireless Communication

P.J.M. Havinga, G.J.M. Smit

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

Abstract

In this paper we present a QoS scheduler that assigns the bandwidth over the wireless channel such that the amount of energy spend by the mobile is minimized, while maintaining the Quality of Service of the connections. Energy efficiency is an important issue for mobile computers since they must rely on their batteries. We have designed and implemented an energy-efficient architecture and MAC protocol for wireless multimedia traffic. The scheduling is based on two mechanisms, 1) a short term transmission frame scheduling that concatenates uplink and downlink traffic of one mobile, and 2) a long-term scheduling, that tries to collect traffic as much as possible within the QoS requirements of the connections. The result is that the transceiver can be in a low-power operating mode for an extended period of time and that the number of operating mode transitions is reduced.

1. Introduction

The energy consumption of portable computers like PDAs and laptops is a limiting factor in the amount of functionality that can be placed in these devices. The wireless network interface of a mobile computer consumes a significant fraction of the total energy consumed by a mobile computer. More extensive and continuous use of network services will aggravate this problem. Energy efficiency can be improved at various layers of the communication protocol stack. However, even today, research is still focused on performance and (low power) circuit design. There has been substantial research in the hardware aspects of mobile communications energy-efficiency, such as low-power electronics, power-down modes, and energy efficient modulation. Due to fundamental physical limitations though, progress towards further energy-efficiency will become mostly an architectural and software-level issue.

The context of this paper is data link-level communication protocols for wireless networks which provide multimedia services to mobile users. Portable devices have severe constraints on the size, the energy consumption, and the communication bandwidth available, and are required to handle many classes of data transfer over a limited bandwidth wireless connection, including

delay sensitive, real-time traffic such as speech and video. Our approach is driven by two major factors. The first factor is that the design should be *energy-efficient* since the mobiles typically have limited energy capacity. The second factor is that it should provide support for multiple traffic types, with appropriate Quality of Service levels for each type. The aim is to meet the required QoS, while minimising the required amount of energy.

In previous work [5] we have presented a MAC scheduling principle for a TDMA system that reduces significantly the energy consumption that is needed for the mobile to communicate, albeit not providing the most efficient bandwidth utilisation. The scheduling principle (called *mobile grouping*) is based on a reordering of the transmission frame such that mobiles can operate in a low power operating state as long as possible. We believe that this is a valid choice since in a mobile multimedia environment it is more important that connections have a certain QoS, than pure raw bandwidth. Sufficient performance and energy efficiency have become the predominant platform requirements for battery-powered mobile multimedia computing devices [7]. In this paper we extend the mobile grouping principle by trying to schedule the traffic extending several transmission frames in order to be able to transmit as much as possible in bursts.

2. Energy dissipation in wireless communication

A significant part of the power consumption needed for wireless communication is due to the wireless interface, the transceiver. Typically, the transceiver can be in five modes (see Figure 1); in order of increasing energy consumption these are: off, sleep, idle, receive, and transmit. 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 circuit is still powered and ready to receive or transmit; in sleep mode, the transceiver circuitry is powered down, although in some implementations a small amount of circuitry is still listening to incoming transmissions.

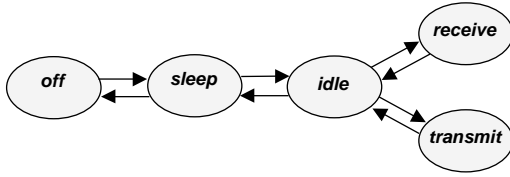


Figure 1: Typical operating modes of a wireless modem.

A wide variety of reasons for unessential energy consumption exist. These reasons include the overhead of the protocol, the high error rate on wireless channels, the inactivity threshold time after which a radio will enter a low energy consuming operation mode, the need to receive messages, the occurrence of collisions, and the turnaround time between various operating modes. The data link layer, and in particular the Medium Access Control protocol, can alleviate these problems significantly. In [6] the causes of unessential energy consumption and main principles of energy-efficient MAC protocol design have been explored in more detail. In this paper we concentrate on ways to reduce the effect of turnaround time, and on minimising the time a radio needs to be in high power mode (idle, receive, transmit).

There are basically three effects that contribute to the required energy for a transition from sleep to transmission or reception:

1. the required time and energy to change the operating mode from sleep to idle.
2. the required time and energy the interface has to be either in idle, receive and transmit mode, but is not transmitting or receiving actual data. This is the overhead required to initiate and terminate the actual transmission. This time includes the required gap (guard time), interfacing delay, preamble, and the postamble.
3. the required time and energy to switch to the sleep mode after transmission or reception.

We assume the wireless physical header and trailer to be a fact that cannot be changed or improved with a MAC protocol, although the protocol can try to minimise the number of times that these are required.

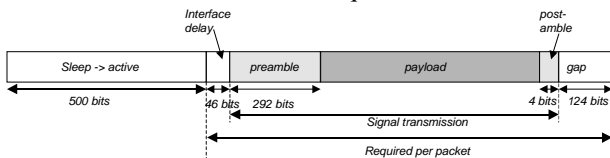


Figure 2: WaveLAN physical layer block format.

The overhead introduced in the physical layer can be significant, e.g. for WaveLAN [12] it can be up to *virtual* 58.25 bytes (i.e. the time normally 58.25 bytes could have been sent) (for guard space (in which the silence level is measured), interfacing delay (required to synchronise to

the internal slotsync moments), preamble and postamble, see Figure 2). Moreover, with this interface that has a throughput of 2 Mbit/s, a transition time from sleep to idle of 250 μ s already takes *virtually* 62.5 bytes (500 bits). The power consumption of the WaveLAN modem when transmitting is typical 1675 mW, 1425 mW when receiving, and 80 mW when in sleep mode (according to the specs [12]). Minimising the on time of the radio will thus significantly improve the energy efficiency of wireless communication.

This also shows that efficient data transmission (in terms of channel utilisation and energy consumption) can only be achieved if the amount of data that is transmitted in one burst is not too small. Scheduling traffic into bursts in which a mobile can continuously transmit or receive data – possibly bundled for different connections –, can reduce the number of transitions. Notice, however, that there is a trade-off with QoS parameters like delay and jitter.

3. Energy-efficient traffic scheduling

In this study we concentrate on *Time Division Multiple Access* (TDMA) schemes where a base-station coordinates access to one or more channels for mobiles in its cell. We have implemented a highly adaptive network interface and a MAC protocol (E^2 MaC). We use a WaveLAN modem as the physical layer. We consider an office environment in which the cells are small and have the size of one or several rooms. The system provides support for diverse traffic types and QoS while achieving a good energy efficiency of the wireless interface of the mobile.

3.1. Overview

In E^2 MaC a channel is divided into *slots*, and these slots are grouped into *frames*. The payload transmitted in one physical block is referred to as *packet*, and can span multiple slots. Mobiles can reserve resources (slots) for connections. A QoS manager (typically located on the base-station) receives transmission requests from the mobiles. The key to providing QoS for these connections will be the scheduling algorithm that assigns the bandwidth.

The *QoS manager* establishes, maintains and releases wireless connections between the base-station and the mobile and also provides support for handover and mobility services. Multimedia networking requires at least a certain minimum QoS and bandwidth allocation for satisfactory application performance. This minimum QoS requirement has a wide dynamic range depending on the user's quality expectations, application usage modes, and application's tolerance to degradation. In addition, some applications can gracefully adapt to sporadic network congestion while still providing acceptable performance.

The *slot scheduler*, which is a part of the QoS manager, assigns bandwidth for connections. A schedule is broadcast to all mobiles so that they know when they should transmit or receive data. This schedule is called *Traffic Control Slot (TCS)*. The slot scheduler is designed to preserve the admitted connections as much as possible within the negotiated connection QoS parameters. In the scheduling of our MAC protocol we apply two mechanisms to reduce the energy consumption.

- *Mobile grouping strategy*. This strategy is based on the principle that a mobile has a concatenated uplink and downlink phase, and that the transceiver will enter a low power operating mode for the remaining time. Although this strategy 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. We have made this choice since in a mobile multimedia environment it is more important that connections have a certain QoS, than highest possible bandwidth. Notice that mobile grouping is applied within a single transmission frame.
- *Schedule traffic in bursts*. This mechanism is applied on a larger time-frame than mobile grouping. It schedules traffic several frames in advance, based on the buffer status information at both the mobile and the base station, and also based on the characteristics of the connections. The objective is to group traffic from one mobile as much as possible within the QoS requirements of all its current connections.

Both mechanisms will be described in more detail in the following sections.

3.2. QoS manager

The notion of QoS over a wireless link has been the focus of much recent research (e.g. [1][3][6][9]), and several scheduling algorithms have been proposed (e.g. [4][8][11]). The *QoS manager* establishes, maintains and releases wireless connections between the base-station and the mobile and also provides support for handover and mobility services. Applications contact the QoS manager when setting up a connection. The QoS manager will inform the applications when they should adapt their data streams when the QoS of a connection has changed significantly. Figure 3 conceptually illustrates the role of adaptive applications in the QoS model.

The application requests a new connection for a certain *Service Class* that defines the media type (e.g. video, audio, data), interactivity model (e.g. multimedia browsing, videoconference), and various QoS traffic parameters (e.g. required bandwidth, allowable cell loss ratio). The service classes allow multimedia sessions to transparently adapt the quality of the connection when the

available resources change marginally without the need to further specify details and without explicit renegotiations.

Network resource allocation is done in two phases. First, the QoS manager checks the availability of resources on the base-stations coverage area at connection setup. The necessary resources are estimated based on the required service. The new connection is accepted if sufficient resources are estimated to be available for the connection to operate within the service contract without affecting the service of other ongoing connections. Otherwise, the connection is refused. Second, while the connection is in progress, dynamic bandwidth allocation is performed to match the requirements of interactive traffic and the available resources. When the available bandwidth changes (because congestion occurs, or the error conditions change drastically), the QoS manager reallocates bandwidth among connections to maintain the service of all ongoing connections within their service contracts.

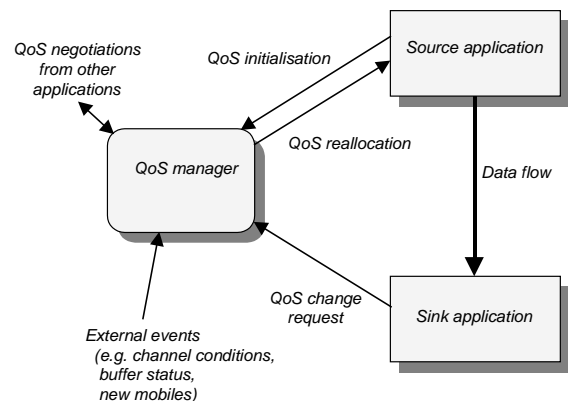


Figure 3: The service model for adaptive applications.

The resulting allocation improves the satisfaction of under-satisfied connections, while maintaining the overall satisfaction of other connections as high as possible. In [9] a bandwidth reallocation algorithm is described that fits well to the QoS model used by the QoS manager.

3.3. Slot scheduler

The *slot scheduler* on the base-station assigns bandwidth and determines the required error coding for each individual connection. The QoS manager provides the service contracts used.

For a proper slot assignment, the slot scheduler needs to know the current state of each connection. For the downlink direction, the scheduler acquires this information directly by monitoring the corresponding queues in the base station. For the uplink direction, this information can be obtained through the implementation of a dedicated protocol, which can be a *polling* scheme or a *contention* scheme. In E²MaC we use a combination of both polling

and contention. In E²MaC all packets include the buffer status of the connection queues. Thus, when there are enough uplink connections (either normal data packets or control connection packets), then the slot scheduler will receive the connection queue status frequently. If this is not sufficient (for example because a connection queue receives more data than anticipated), then an unreserved (contention) slot can be used to transfer a recent buffer status update to the slot scheduler using a control packet.

The slot scheduler maintains two tables: a *request table* and a *slot schedule table*. The request table maintains several aspects of the current connections handled by the base station (like the connection type, the connection queue size and status, the error state of the channel with mobile, the assigned bandwidth, the requested reliability). The slot schedule table reflects the assigned number of slots to connections, and the error coding to be applied. This table is essentially broadcast as Traffic Control Slot (TCS) to the mobiles.

These two tables are used by the QoS manager and slot scheduler to assign bandwidth to connections. Since these entities are implemented as software modules on the base-station, their implementation can be adapted easily to other scheduling policies if needed.

A schedule is broadcast to all mobiles so that they know when they should transmit or receive data. In composing this traffic control, the slot scheduler takes into account: the state of the downlink and uplink queues, and the radio link conditions per connection. The slot scheduler is designed to preserve the admitted connections as much as possible within the negotiated connection QoS parameters.

3.4. Mobile grouping strategy

The slot scheduler schedules all traffic according to the QoS requirements and tries to minimise the number of transitions the mobile has to make. It schedules the traffic of a mobile such that all downlink and uplink connections are grouped into packets taking into account the limitations imposed by the QoS of the connections. In general there are three phases: uplink phase, downlink phase, and reservation phase. In the downlink phase the base station transmits data to the mobiles, and in the uplink phase the mobiles transmit data to the base station. In the reservation phase mobiles can request new connections. We refer to this mechanism as *phase grouping*.

In our protocol we have in principle similar phases, but these are not grouped together in a frame according to the phase, but are grouped together according to the mobile involved. In our protocol we thus group the uplink and downlink phase of *one mobile*. We refer to this mechanism as *mobile grouping*.

Figure 4 shows the two grouping strategies. In mobile grouping the uplink and downlink packets for a mobile are

grouped sequentially (if possible) so that the mobile can power down longer and make minimal transitions between power modes. As indicated above, the power consumption of the WaveLAN modem when transmitting is typical 1675 mW, 1425 mW when receiving, and 80 mW when in sleep mode [12]. Increasing the sleep time period of the radio thus significantly improves the energy efficiency of the wireless network. Moreover, due to the large power-transition times, this mechanism might give the mobile enough time to enter a power-down mode at all. This is shown in the Figure 4 where Mobile 2 with phase grouping cannot enter sleep mode after reception of the downlink packet, but is forced to idling¹. Because the operating modes of phase grouping for a mobile are spread in the frame, the power-mode transition times T_{sleep} to enter sleep mode, and $T_{wake-up}$ to wake from sleep mode limits the time a mobile can stay in sleep mode.

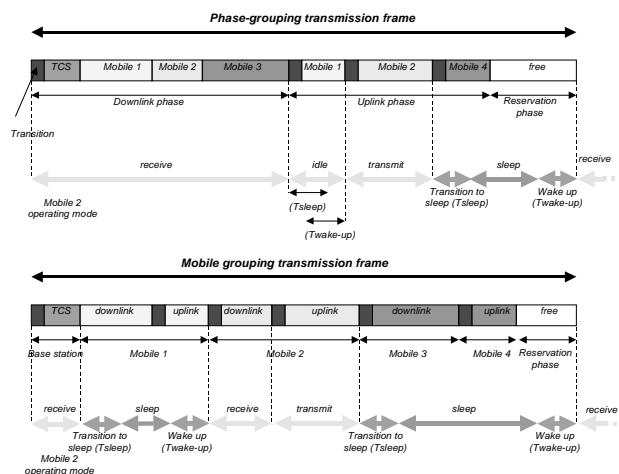


Figure 4: Grouping strategies.

Notice that in the mobile grouping strategy there is more transition overhead (i.e. one transition per mobile) since the base station does not transmit its data to the mobiles in one packet during the downlink phase of phase grouping. The transition overhead consists of guard space (gap), interfacing delay, preamble, and postamble (see for an example Figure 2). The transition overhead involved with each transmission packet is the reason that the available bandwidth of mobile grouping is less than the available bandwidth in phase grouping. However, since the traffic of a mobile is grouped, the mobile can enter a low-power mode (sleep) for a longer time. In fact, with phase grouping, the mobile is in general forced to receive the complete downlink packet, and will ignore the data not destined for the mobile. The consequences of using mobile

¹ A power-optimised network interface could stop receiving the downlink packet after it has received data for mobile 2, and thus also enter sleep mode.

grouping on the channel efficiency and the energy consumption is analysed in [5].

3.5. Burst scheduling

The QoS manager and slot scheduler at the base-station receives transmission requests from the mobiles. Each mobile can have multiple unidirectional connections with different Quality of Service requirements. We have applied the service categories that have been defined under ATM: constant bit rate (CBR), real time VBR, non-real time VBR, unspecified bit rate (UBR), and available bit rate (ABR).

The QoS manager has knowledge of the connection types and queue status of all connections in its cell. It collects information on the number of packets in the connection queues of both the base-station and the mobiles. This queue status is updated frequently. The QoS manager translates the requirements of the active connections to deadlines using the various parameters that describe the connection. This basically produces two types of connections: real-time connections, and non real-time connections. Real-time connections are described as a transmission of an amount of data with some size and with a certain frequency, CBR and VBR have such real-time connections. These parameters have in general a mean and a maximum value. Non real-time connections have no stringent delivery requirements, UBR and ABR traffic is thus treated non-real time.

The scheduler operates in two phases. Initially, the deadlines are scheduled according the Earliest Deadline First (EDF). This schedule spans a relative long period of several frames (which is in our first implementation over 100 frames, which implies a period of one second). This 'long-term' schedule incorporates both the actual queue status as the worst-case requirements of the connections (expected traffic). This schedule is already suitable as it adheres the QoS requirements of the active connections. Then, in the second optimisation phase, the scheduler tries to group the connections of the mobiles as much as possible. Connections of one mobile are grouped together incorporating both uplink and downlink connections. Reordering of traffic is applied if the number of transitions is reduced, and the QoS requirements of the connections are still met.

4. Conclusion

In this paper we have presented a QoS scheduling mechanism for a MAC protocol that is aimed for a high energy efficiency. In the scheduling we apply two strategies. Firstly, uplink and downlink traffic of a mobile is concatenated in a transmission frame. Secondly, traffic of a mobile is scheduled in large bursts, taking into account the limitations imposed by the QoS of the connections and

the queues in the mobile and the base station. Reordering of the traffic is performed on a larger time scale, in our implementation traffic is scheduled for 100 transmission frames in advance. This strategy reduces the number of operating mode transitions between transmitting, receiving, idle, and sleep, and maximises the possible sleep period of the transceiver.

References

- [1] 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.
- [2] 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.
- [3] Choi S., Shin K.G.: "A cellular wireless local area network with QoS guarantees for heterogeneous traffic", *Mobile networks and applications* 3, pp. 89-100, 1998.
- [4] Colombo G., Lenzini L., Mingozzi E., Cornaglia B., Santaniello R.: "Performance evaluation of PRADOS: a scheduling algorithm for traffic integration in a wireless ATM network", *Proceedings of the fifth annual ACM/IEEE international conference on mobile computing and networking (MobiCom'99)*, pp. 143-150, August 1999.
- [5] Havinga P.J.M., Smit G.J.M.: "Energy-efficient TDMA medium access control protocol scheduling", *proceedings Asian International Mobile Computing Conference (AMOC 2000)*, Nov. 2000.
- [6] 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.*
- [7] McKenna D.: "Mobile platform benchmarks, a methodology for evaluating mobile computing devices", *Transmeta Corporation*, January 2000, <http://www.transmeta.com>.
- [8] Moorman, J.R., Lockwood J.W.: "Multiclass priority fair queuing for hybrid wired/wireless quality of service support", *Proceedings of the second ACM international workshop on Wireless Mobile Multimedia (WoWMoM'99)*, pp. 43-50, August 1999.
- [9] 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.
- [10] Shakkottai S., Srikant R.: "Scheduling real-time traffic with deadlines over a wireless channel", *Proceedings of the second ACM international workshop on Wireless Mobile Multimedia (WoWMoM'99)*, pp. 35-42, August 1999.
- [11] Su W., Gerla M.: "Bandwidth allocation strategies for wireless ATM networks using predictive reservation", *IEEE Globecom '97*, 1997.
- [12] WaveMODEM 2.4 GHz Data Manual, Release 2, AT&T 1995.