Dynamic Resource Reservation and QoS Management in IEEE 802.11e Networks

Jürgen Wolf **Stephan Heckmüller** **Bernd E. Wolfinger**

Dept. of Computer Science, TKRN, University of Hamburg
Vogt-Kölln-Str. 30, 22527 Hamburg, Germany

E-mail: {`jwolf`|`0heckmue`|`wolfinger`}`@informatik.uni-hamburg.de`

Keywords: Quality of Service, Resource Management, Real-Time Communications, Performance Evaluation, IEEE 802.11e

Abstract

Quality of service (QoS) guarantees are becoming increasingly relevant for mobile networks, too, as they have to support, e.g., more and more audio and video applications with real-time requirements. Therefore, recent extensions of WLANs, namely IEEE 802.11e, are introducing new mechanisms to enable the provisioning of QoS guarantees, in particular by means of resource reservation.

This paper suggests to diminish the inefficiencies of the resource reservation approach as taken by IEEE 802.11e. The solution advocated is to embed the authors' earlier abstract approach for a dynamic resource redistribution into 802.11e based WLANs. By means of analytical and simulation models it is demonstrated that the extension of 802.11e based networks by a dynamic resource redistribution can lead to a significant increase in overall network utilization, though QoS requirements can still be guaranteed.

1 INTRODUCTION

In nowadays communication networks the provisioning of quality of service (QoS) to communicating applications and end-users has become an important factor. This is due to the strongly increasing number of applications with real-time requirements.

Some networks try to satisfy QoS requirements simply by over-provisioning of resources; in mobile networks resource over-provisioning is typically infeasible because in these networks the available overall bandwidth (BW) is quite limited. As alternatives to resource over-provisioning one could try to achieve QoS guarantees by means of a priori reservation of resources or by means of prioritizing data units, e.g. the packets transmitted, in combination with access control for high-priority traffic. Recent enhancements of the wireless LAN (WLAN) standard, namely the IEEE 802.11e draft [18], also take the approach of reserving resources in order to satisfy QoS requirements in mobile networks.

Static reservation of resources may be quite inefficient, however, if significant variations of load exist during the life-time of a connection for which resources have been reserved. We elaborated a general approach which allows one to dynamically redistribute reserved resources in local broadcast networks in a load-adaptive manner [36]. The present paper suggests to integrate this earlier abstract approach for resource redistribution into WLANs with IEEE 802.11e architecture and thus achieve some dynamic QoS management. In this study, we focus on deterministic QoS provisioning, assuming resources statically allocated to real-time streams. In particular, we model real-world scenarios where network characteristics are mostly constant over time.

We will provide a course evaluation, based on mathematical (analytical) models, of the proposed method for QoS management. More detailed simulation models, based on the *ns-2* simulator, allow us to confirm the results of the analytical models, which indicate that the overall packet throughput as well as the overall utilization of bandwidth within WLANs can be improved significantly by applying our approach. In order to increase the validity of the simulation results the simulation models not only have been carefully verified but also highly realistic load models, directly based on real traces of video traffic, have been used in the simulation experiments.

The paper is structured as follows: Section 2 presents the state-of-the-art in load-adaptive resource management. In Section 3 we introduce our approach for dynamic resource redistribution; in particular, we will cover the aspects of load estimation and of solving the challenging trade-off between efficiency and actuality in the exchange of state information. The basic characteristics of IEEE 802.11e are described in Section 4, focusing on the mechanisms for QoS support and including our approach to embed dynamic resource redistribution into IEEE 802.11e. The analytical models to quantify the expected gain in throughput and the additional end-to-end delay, both resulting from the dynamic resource redistribution, are part of Section 5. Section 6, under realistic application scenarios, comprises the simulation results for an in-depth evaluation of the IEEE 802.11e extension as suggested by us. Section 7 concludes with a short summary of lessons learned.

2 RELATED WORK

Various aspects are noteworthy regarding ***QoS management based on dynamic resource reservation***. First of all, the *reservation of resources* is an important means to achieve quality of service guarantees in computer and communication networks [8]. In continuous media communications with real-time requirements, such as audio/video communications, it is typical to reserve the resources, as they are expected to be required later on, during connection setup of a stream, cf. IntServ [7]. Some real-time protocols and services, such as CMTP/CMTS [28], allow that resources may be re-allocated by means of setting up new streams dynamically. Other approaches assume that variations of load within single stations are smoothed out to a great extent by means of multiplexing a larger number of streams and doing the reservation for the complete overlay of those streams [26].

Up to now, only few proposals exist which – as the one we presented in [36] – advocate for a dynamic freeing and recalling of bandwidth between stations. Earlier approaches [4, 21] did not provide a redistribution of the unused prioritized bandwidth. Evidently, bandwidth reservation has already been studied in the literature [39, 16].

Estimation of load has been an important topic since more than one decade [20, 32]. Besides the idea of basing load estimation directly on the utilization observed [12] it has also been suggested to directly observe arrival times, thus estimating arrival rate, and service-time requirements, thus estimating mean service-times. Geometric weighting [30] seems to be the dominating approach for calculating the estimates. In load sampling besides the periodic sampling which we suggest for simplicity, more complex measurement procedures have been proposed, e.g. based on so-called random sampling [10]. Measurements, as required in our paper, refer to observations of primary load rather than of secondary load in the sense of [37].

As an alternative to provisioning deterministic *QoS guarantees*, as we strive for in our approach, stochastic QoS guarantees are possible [9]. Then, evidently, no longer a reservation for the worst case would be required but QoS requirements could become violated temporarily, unless applications have the capability to adapt to network changes [6]. For applications tolerating transient fluctuations in the QoS, different approaches for resource management exist, in particular, in the context of wireless networks [14, 23].

IEEE 802.11 WLAN [19] is originally designed for best-effort (BE) services. Thus various ***QoS enhancements*** have been advocated. Most of them introduce prioritized QoS by means of modified contention-based access mechanisms; one approach introduces different backoff increment functions, interframe spaces, and variable maximum frame lengths, according to the corresponding priority level [3]. Other proposals specify mechanisms enhancing or partially replacing the distributed coordination function [34, 33]. The point coordination function is extended and investigated only by few research works [22].

The upcoming standard 802.11e [18] introduces various medium access modes at MAC level. Most investigations concentrate on the analysis of contention based modes, cf. [24, 29]. Since error rates in wireless LANs are orders of magnitude higher than in wired LANs, enhancement schemes regarding forward error correction are evaluated as well [11].

Resource management architectures for wireless networks have been an important research topic, too. Approaches exist covering adaptive resource management architectures providing algorithms for reservation, advance reservation, and resource adaptation in mobile computing environments [5]. This work is focused on the coordination between the different layers of the network. Furthermore, architectures to support expedited forwarding service, as defined in DiffServ, over WLANs are proposed [17]. Other research work addresses methods to handle networks with variable bandwidth, where applications declare their range of required QoS at connection setup time [27].

3 DYNAMIC QoS-RESOURCE MANAGEMENT

In this section we summarize briefly our approach of dynamic resource management which we presented in more detail in [36]. We focus on the resource reservation approach and its usage in local broadcast networks, especially networks – as assumed in this study – with wireless communication infrastructure. In particular, we investigate ways of how communication resources which are already reserved and allocated to communicating end-users or end-systems can be dynamically handed over to other end-users or end-systems in case that they are not needed by their original "owner" for some time.

The approach we suggest assumes that the problem of reserving and statically allocating resources to communicating entities has already been solved [9]. We also suppose that the resource reservation has been established for a rather long-term time interval, e.g. duration of an audio/video stream, typically in the order of minutes. In our approach, each owner of resources determines whether its communication load, i.e. the amount of time-critical data waiting for transmission at the owner, justifies the continued reservation of all the resources as allocated to the owner. If a sufficiently large amount of resources is observed to be temporarily free, the owner of the resources will pass some of these resources to its "neighbors". In order to respect real-time requirements for its data to be sent each owner will continue to observe the arrivals to its transmission queue. If its local load is increasing again, the owner informs its neighbors that they are no longer allowed to make use of the resources which were offered for public access at an earlier instant. The freeing of resources and their reclaiming could lead to oscillations in the resource redistribution. Therefore, our approach for making resource allocation depending on the actual communication load, is based on load estimators, which produce some kind of smoothed estimates for the load as it is generated locally over time. Moreover, we introduce systems of load thresholds which, only when being crossed by the load estimates of an owner of resources, leads to a broadcast of messages for freeing or reclaiming resources.

3.1 Assumptions and Underlying Model for Resource Redistribution

In the following we present the basic proceeding which we suppose throughout this paper for resource reservation and redistribution between the stations of a broadcast network.

Regarding ***resource reservation*** we assume:

- n stations $STA_1, \dots, STA_n$ communicating via a local broadcast network (Ethernet, WLAN,...)
- communication load of the stations consisting of time-critical transmissions and non-time-critical data transfers, cf. the real-time (RT_Q) and non-real-time queues (nRT_Q) in Figure 1
- in each station with real-time communication requirements there exists at least one owner of resources, where reserved data rate (allocated "bandwidth") is the resource considered by way of example; scenarios could be that the owner reflects the complete station or an owner could model a single end-user
- resources are supposed to be reserved before starting a new stream with real-time requirements and are available to its owner until it releases connection after end of the stream; the amount of bandwidth reserved is such that the owner's QoS requirements can be guaranteed.

Relating to ***temporary resource redistribution*** we assume that an owner of resources can pass resources to other stations temporarily, whenever these resources are not required locally for some time. We introduce a message of type

- FREE_BW $(STA, \Delta d)$ with which a station STA hands over a bandwidth (respectively data rate) of $\Delta d\ bit/s$ to other stations, and
- RECALL_BW $(STA, \Delta d)$ to recall bandwidth Δd to its owner station STA. With receipt of a RECALL_BW message the recalled bandwidth has to be released immediately.

The policy of how to distribute the freed resources among the other entities can be constituted arbitrarily, maybe distributed algorithm or central assignment, and may depend on the individual network.

Approach to ***estimate the level of load*** resulting from time-critical transmission requests:

We assume periodic measurements available at instants $t_i := t_0 + i \cdot \Delta t$, for $i \geq 1$, where the measured samples ρ_i characterize the load of time-critical transmission requests which have been generated by the source during the interval $T_i := [t_{i-1}, t_i)$. Evidently, $\rho_i := d_i/(r \cdot \Delta t)$, where d_i denotes the amount of time-critical data generated during T_i and r denotes the data rate as it was allocated to its owner at connection setup time. This sequence of samples (ρ_i) may be used at each instant t_i to calculate an estimate $\hat{\rho}(t_i)$ of the actual level of load.

3.2 Traffic Load Estimation

In general, arbitrary functions are supposable to calculate an estimation of the current workload based on the individual load measurements. Nevertheless, so far, we focus on methods based on *geometric* and *arithmetic* weighting. These functions are well understood, easy to calculate, and can be parameterized in an appropriate manner.

Here, ***geometric weighting*** is defined as follows:

$$\hat{\rho}(t_i) = \alpha\rho_i + (1-\alpha) \cdot \hat{\rho}(t_{i-1}), \quad i \geq 1$$

where $\alpha \in (0, 1]$ and $\hat{\rho}(t_0)$ to be initialized, e.g. $\hat{\rho}(t_0) := 0$. The α factor indicates how fast the estimator changes with a strongly different new sample. The geometric weighting is also known as the *exponential weighted moving average* (EWMA) [30]. As an abbreviation we denote by G_α geometric weighting with parameter $\alpha \in (0, 1]$.

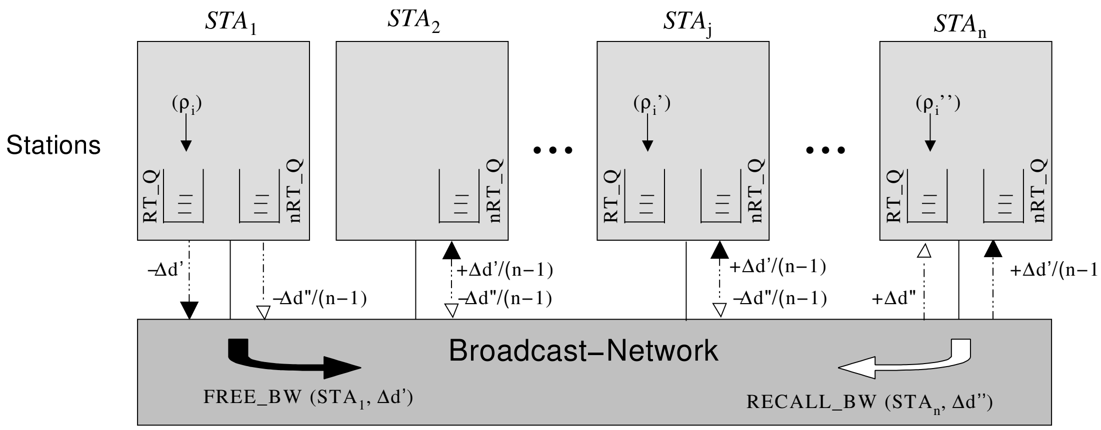

Figure 1. Redistribution of bandwidth: the depicted example scenario illustrates the freeing of $\Delta d'$ by station STA_1 and recalling of $\Delta d''$ by STA_n, here with resources distributed equally.

The ***arithmetic weighting*** estimator is defined as:

$$\hat{\rho}(t_i) = C_0 \cdot \sum_{j=0}^{w-1} \frac{w-j}{w} \rho_{i-j}, \quad i \geq 1,$$

where $w \in \{1, 2, \ldots\}$ and ρ_k to be initialized for all $2 - w \leq k \leq 0$. $C_0 = \frac{2}{w+1}$ denotes a normalization constant. This estimator is equivalent to a sliding window with window size w and weights increasing in a linear way when the samples are more recent. Arithmetic weighting uses $\frac{k}{w}$ as weights, where k is the position in the window of size w with the most recent sample taking the position $k = w$. The size of w indicates how many past samples including the present sample must be kept to estimate the current load. As an abbreviation, A_w denotes arithmetic weighting with a window size of $w \in \mathbb{N}$.

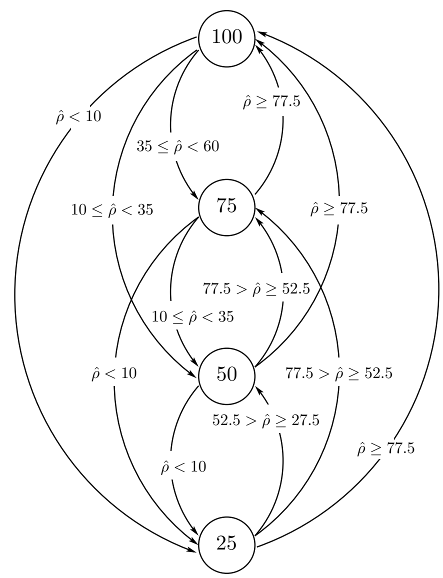

Figure 2. An example 4-state threshold system parameterized according to $E_4^{0.15,0.25}$; states and transitions are labeled in percent.

3.3 System of Load Thresholds

In this section we introduce a system of load thresholds, which determines when to free or claim back bandwidth based on current load estimation as discussed above. The main tasks of such a system are on the one hand to assure real-time requirements to be respected and secondly to avert oscillations caused by small variations of the estimator's value. After giving the general definition, we introduce reasonable constraints which should be taken into account when constructing specific threshold systems by means of actually parameterizing the general model.

Definition 1 *We define an n-***state threshold system** *as tuple $TS(S, \vartheta)$. The n-tuple $S = (S_1, S_2, \ldots, S_n) \in [0,1]^n$, with $S_i < S_j$ for all $1 \leq i < j \leq n$, denotes the set of states while the state-transitions are defined by the $n \times (n-1)$-matrix*

$$\vartheta = \begin{pmatrix} \vartheta_{1,2} & \cdots & \vartheta_{1,n} \\ \vdots & \ddots & \vdots \\ \vartheta_{n,2} & \cdots & \vartheta_{n,n} \end{pmatrix} \in (0,1]^n \times (0,1]^{n-1}$$

with $i < j \Rightarrow \vartheta_{k,i} < \vartheta_{k,j} \quad \forall k$. With given state S_i and estimated load $\hat{\rho}$ the next state is given by

$$\boxed{r(S_i, \hat{\rho}) := S_{\max\limits_{j \in \{2,..,n\}} \{\vartheta_{i,j} \leq \hat{\rho}\ ,\ 1\}}}$$

Interpretation of ϑ and S_i:

- Up-thresholds: For $i < j$, each entry $\vartheta_{i,j}$, determines the up-threshold which has to be crossed upwards (or at least reached) by the estimator $\hat{\rho}$, in order to change the system's state from S_i to S_j.

- Down-thresholds: For $i \geq j$, each entry $\vartheta_{i,j}$, determines the down-threshold which has to be crossed downwards by the estimator $\hat{\rho}$, in order to change the system's state from S_i to S_{j-1}.
- If the system's state is S_i this means that a proportion of $S_i \cdot r$ of the originally reserved data rate r is actually available to this station.

If neither any up-threshold nor any down-threshold is crossed, the system remains in the current state. Transitions are passed through periodically after each multiple of Δt. Figure 2 shows a specific 4-state threshold system − actually parameterized according to $E_4^{0.15,0.25}$ (see below) − by way of example.

Definition 1 is kept little restrictive in order not to limit its universality. Nevertheless, concerning our study we assume reasonable constraints for the rest of the paper:

C1 $S_n = 1$ (or 100%), otherwise full capacity can not be reached

C2 down-thresholds *in* S_i are strictly less than up-thresholds *to* S_i,
formally $1 \leq k < i \Rightarrow \vartheta_{i,i} < \vartheta_{k,i}$

C3 up-thresholds *in* S_i are greater than up-thresholds *to* S_i,
formally $1 \leq k < i \Rightarrow \vartheta_{i,i+1} > \vartheta_{k,i}$

C4 down-thresholds *in* S_i are less than down-thresholds *to* S_i,
formally $n \geq k \geq i+1 \Rightarrow \vartheta_{i,i} < \vartheta_{k,i+1}$

Nota bene: If transitions $S_i \rightarrow S_j$ occurs with $i > j$ or $i < j$ the control messages FREE_BW and RECALL_BW respectively are sent.

In an earlier work we demonstrated that threshold systems with equal distances between their thresholds are indeed a suitable way to parameterize the systems [36]. Therefore we still evaluate these kind of parameterized systems. Below we establish an abbreviation $E_n^{b,m}$ for the threshold systems we investigate in this study; thus we do not always have to write out the complete matrices. We define

$$E_n^{b,m} := TS(S, \vartheta)$$

with

$$S_i := 1 - (n-i) \cdot \frac{(1-m)}{n-1}$$

$$\vartheta_{i,j} := \begin{cases} S_{j-1} - b & \text{for } i \geq j \\ S_j - \frac{3}{2} \cdot b & \text{for } i < j \end{cases}$$

whereby n denotes the number of states. S_1 and S_n are set to m and 1 respectively and all states between are interpolated linearly. The spacing between down-threshold and destination state is denoted by b; the spacing between up-threshold and state is set to $\frac{3}{2}b$. By way of example, Figure 2 shows a 4-state threshold system parameterized according to $E_4^{0.15,0.25}$. Evidently, not every possible parameterization is reasonable. We discuss further specific parameterizations in more detail in Section 5 and 6.

4 ESTABLISH RESOURCE MANAGEMENT IN WLANs USING QoS SUPPORT MECHANISMS OF 802.11e

Enhancements to the IEEE 802.11 [19] medium access control (MAC) are currently under development and will establish the basis for the 802.11e extension of the original 802.11 standard (the latter also called the *legacy* 802.11). 802.11e introduces the *hybrid coordination function* for QoS support [18], which defines the two medium access mechanisms *contention-based channel access* and *controlled channel access*. At first, in this section we briefly describe the mechanisms provided for QoS support within an 802.11e network. In the next step, we present the enhancements in order to support adaptive resource management using the methods as described in Section 3.

4.1 QoS Support Mechanisms of 802.11e

At first, we roughly illustrate the principles of the 802.11 medium access control and highlight the fundamental QoS limitations of legacy 802.11. Then we discuss those parts of the new 802.11e extension which are fundamental in the context of the resource management. For further information we refer to summaries [24] and the draft itself [18].

The ***802.11 medium access control*** is based on *carrier sense multiple access* with *collision avoidance*. Different *interframe spaces* (IFSs) are provided in order to reduce the probability of collisions. Nevertheless each transmitted *MAC protocol data unit* requires an immediately transmitted acknowledgement (ACK). Fragmentation is optional but recommended for long MAC service data units. Furthermore, the MAC consists of a *request-to-send/clear-to-send* (RTS/CTS) mechanism in order to ease the "hidden station problem", cf. [24].

The legacy 802.11 [19] provides the *point coordination function* (PCF) to support some kind of QoS. Within a *contention-free period* (CFP) the *point coordinator* – typically the access point (AP) – can poll other stations or send data frames to them. The end of the CFP is announced with a CF-end control frame. During the subsequent *contention period* (CP) the stations compete for the medium access with equal priorities using the *distributed coordination function* (DCF). A CFP and the following CP form a *superframe* and alternate periodically. Each superframe starts with a beacon frame broadcast by the AP, regardless whether or not the optional PCF is used. Each beacon announces the time of the arrival of the next beacon frame, the so called *target beacon transmission time* (TBTT).

There exist a couple of ***QoS limitations of legacy 802.11***. The DCF supports only best-effort services, no QoS guarantees are possible at all. Also the PCF mode has three main problems the QoS performance suffers from [22]:

- All data exchange between two stations in the same *basic service set* (BSS), i.e. a collection of stations using the same radio frequency, goes through the AP, thus approximately twice as much bandwidth is needed.

- Beacon delays are unpredictable since the beacon frame can only be sent if the medium is sensed idle (for at least the duration of a PCF IFS (PIFS)). Admittedly sending stations are not further restricted in order to respect the TBTT, thus the beacon frame can be delayed up to about $4.9ms$ [25].

- No mechanisms exist to control the transmission time of a polled station. Once polled a station decides the length of its frame for its own – up to the maximum frame length of 2304 byte (resp. 2312 byte encrypted). This behavior makes it difficult to provide transmission resources to several stations.

Since the legacy 802.11 was originally designed for best-effort services, but the need for providing QoS increases steadily, various approaches for enhancements have been proposed [29]. Refering to this, the upcoming standard 802.11e covers the issues of service differentiation with several ***extensions to support QoS in 802.11e***. For that purpose a new mechanism is specified, namely the *hybrid coordination function* (HCF), which differentiates various medium access modes at MAC level. Either the nodes compete for channel access during CP using the *enhanced DCF* (EDCF), offering prioritized QoS; alternatively they wait for being polled by the *hybrid coordinator* (HC) using *HCF controlled channel access* (HCCA), offering strict parameterized QoS. Another feature is the *direct link protocol*, allowing any entity within one quality BSS to communicate directly with any other.

In this paper we focus on strict parameterized quality of service, that's why we omit the details of the EDCF access rules here, refering to further reading [18, 24]. 802.11e still defines a superframe periodic over time (cf. Figure 3), starting with a beacon frame and an optional CFP followed by the CP. Contrary to the legacy version, stations can be polled both during CFP and CP in order to allow the HC greater latitude to provide its services. Moreover, HCCA has highest priority so that the HC does not need to compete if a station shall be polled. Another important feature is the concept of *transmission opportunity* (TXOP), reflecting the time interval a station is allowed to occupy for its transmissions – including fragments, ACKs, and, if needed, RTS/CTS frames. The length of an EDCF-TXOP interval – assigned for winning an EDCF contention – relates to the corresponding access class, at which the TBTT must not be crossed, so that the problem of beacon delay is solved. The TXOPs are constituted centrally by the HC.

Figure 3 exemplifies how channel access could take place over time. One superframe consisting of CFP and CP is shown, illustrating the different channel access modes: STA_1, which has data enqueued for STA_2, is polled by HC twice, once during the CFP and again in the CP. Further on, HC and after that STA_2 wins the competition for access according to the EDCF rules. All transmissions including ACKs, regardless if polled or won in contention, must be finished within their TXOP and before TBTT.

4.2 802.11e Resource Management Enhancement

Our resource management approach reflects an abstract concept, thus it does not dictate at all which layers or interfaces to use for establishing the necessary methods and services, including data rate provisioning, load estimation, control message exchange and distribution of freed capacities.

In the following we are going to map the abstract concept onto 802.11e networks. The main issues we take into account in this mapping are the desire for an uncomplicated implementation, usability and transparency from end-users' view, unchanged usage of the 802.11e standard as such, consideration of the network infrastructure as well as a fair but easy distribution of resources. These goals led us to the following concept:

The ***measurement component*** is located at the UDP input queue, estimating the load at the corresponding interface periodically over time, since, in this study, we make use of one typical application, the transmission of video-streams, typically sending using UDP. To provide transparency, the data rate, as part of the negotiated QoS, is allocated at UDP layer. Thus a function has to be provided mapping the request to traffic specification (TSPEC) parameters at MAC level. This transformation function can be adapted by the HC to individual network characteristics, e.g. to expected above-average error rate – if known accurately timed.

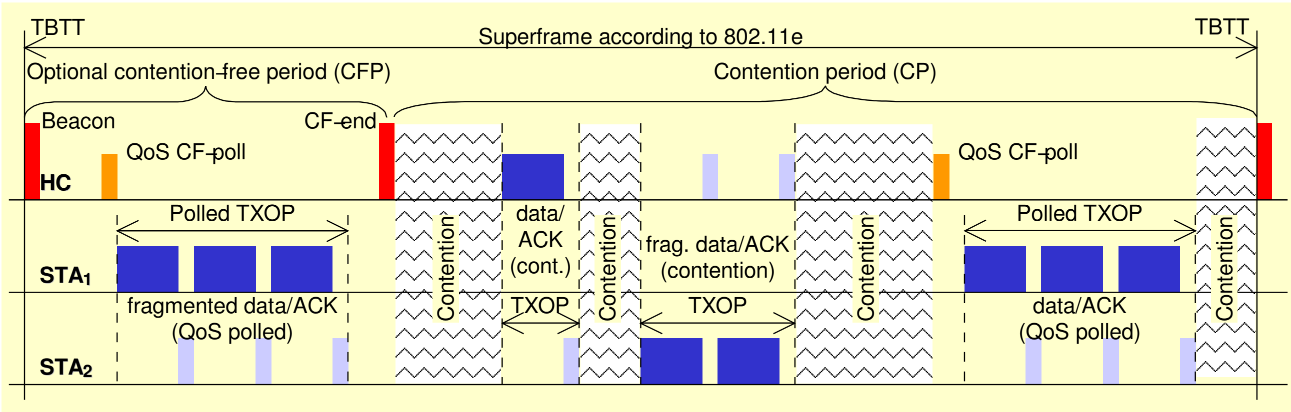

Figure 3. Example of a superframe according to 802.11e including the optional contention-free period; the optional RTS/CTS signaling is omitted.

Evidently, arbitrary interfaces are suggestable for QoS negotiation and dynamic redistribution; thus, the ***exchange of control messages*** is organized at transport layer, too. As these control messages are responsible for proper resource management, they are sent reliably and with highest priority – we use UDP whereby a confirmation is immediately sent back. The information provided consists of source node, stream id, and the percentage of the static allocated data rate claimed for the owner. Obviously, it is possible to simultanously establish the adaptive redistribution for different queues at different layers.

Static resource reservation can be applied to 802.11e networks using the HCCA method. Thus, effectively, the control of the redistribution is done at MAC level – handled by the HC. Each successfully registered entity gets assigned a computed duration of polled-TXOP. The simple round-robin scheduler as proposed by IEEE [18] is adjusted so that real-time stations are polled – even when they signal that their queues are currently empty – until their assigned TXOPs become zero. Thus, the negotiated service is provided – just in the moment generated data can then be dispatched instantly. There is no need to change MAC level implementation but within the HC.

The ***distribution of the remaining resources*** is organized conditioned by the underlying policy. Each time when the HC receives FREE_BW or RECALL_BW messages the resources are reassigned for the corresponding stream according to its claim. Our policy integrated for this study, is to add free bandwidth to the CP, so it is fairly distributed since all backoff entities can compete for access according to their priorities.

Nota bene: When migrating to a different MAC, only the low-level implementation at the HC has to be adapted.

5 ANALYTICAL EVALUATION

By means of dynamic resource management the overall network utilization typically can be improved quite significantly. However, in order to set suitable parameters for an individual application it is essential to know their impacts on the provided QoS, in particular regarding the additional delays that can occur. Therefore, it is desirable to make calculations available in order to adjust the parameterization according to the characteristics and delay sensibility of an individual application. Below, we firstly depict how an upper bound can be derived from the parameterization; in Section 5.2 we discuss by way of example – in an implementation independent way – the resource management's impact on the utilization achieved for the reserved transmission capacity, concerning different estimators and state models.

In this section, we still want to abstract from an actual network configuration or even an individual implementation. Therefore, we can assume that requests to change the real-time streams' capacity are translated into action instantly – at instances of time in multiples of Δt. Furthermore, we suppose that data can be sent with any size, so that no capacity is wasted and no delay is induced by oversized frames that do not fit into slots of remaining transmission capacity. Mentionable, we also neglect here at first the signaling communication overhead needed for controling the sharing of bandwidth.

5.1 Upper Bound of Delays

In the past, we have already theoretically analyzed the impact of the threshold systems combined with weight functions as introduced previously. In particular, the maximum delay τ_i that can occur for data received during interval T_i can be upper bounded, which we summarize in the following – for details cf. [36].

At first, that instant $t' \in T_i$ is determined at which data is delayed longestly. The delay is computed then as difference between the instant t'', that is the instant when all data waiting in the sending queue at instant t' has just been sent, and t'.

The backlog delay within T_i is maximal for the instant $t' := t_{i-1} + \frac{\rho_i}{r}$. Assuming the worst case that data is generated with the maximum possible data rate r during the first part of the interval, t' is the instant when the last data unit within T_i has been added to the sending buffer. In order to determine t'', with S_i denoting the portion of r allocated during T_i by the threshold system, let B_i denote the amount of backlogged data buffered after interval T_i:

$$B_i = \max(0, B_{i-1} + (\rho_i - S_i) \cdot r \cdot \Delta t) \qquad \text{for } i > 0$$

Algorithm 1 DELAY(x, i)

Require: Amount of data x and index i of interval; S_i denotes portion of r allocated within T_i.
Ensure: Time needed to send amount of data x starting at interval T_i;

```
x_r ← x − S_i · r · Δt            /* data not sent during T_i */
if x_r > 0 then
    return 1 + DELAY(x_r, i + 1)      /* start sending remainder in T_{i+1} */
else
    return x / (S_i · r)              /* time to send x in T_i */
end if
```

Furthermore, DELAY(x, i), as described by Algorithm 1, computes the maximum time needed to send the amount of data x where sending starts at the beginning of interval T_i. With

$$t'' = t_{i-1} + \text{DELAY}(B_{i-1} + \rho_i \cdot r \cdot \Delta t, i)$$

it follows

$$\tau_i \leq t'' - t' = \text{DELAY}(B_{i-1} + \rho_i \cdot r \cdot \Delta t, i) - \frac{\rho_i}{r}.$$

Applying the "rapid boost" load scenario as given by

$$\rho_i = \begin{cases} 0 & \text{for } i = 0 \\ 1 & \text{for } i \geq 1 \end{cases} \qquad (1)$$

the maximum delays shown by Table 1 can be computed. It can be seen that even strict real-time requirements can be solved by selecting adequate estimators and models. To give an example, by using a geometric estimator with $\alpha = 0.3$ combined with Model $E_2^{0.35,0.5}$, the maximum backlog delay does not exceed Δt.

5.2 Utilization of Reserved Transmission Capacity

We study the behavior of the dynamic resource management for several different real-time streams represented by real traces. As representatives for real-time streams with in fact variable bit rates, which we consider in this paper, we investigate video streams. We study traces generated from MPEG-2 and MPEG-4 streams as well as video sequences coded according to the ITU Standard H.263. For the latter some of the resource redistribution‘s possible impacts on the perceived video quality are pointed out in [35]. In this section we single out one typical example.

Table 1. Maximum backlog delay in multiples of Δt in the load scenario given by Equation (1) for different threshold systems.

	$G_{0.1}$	$G_{0.3}$	A_{10}	A_{20}
Model $E_2^{0.35,0.5}$	3.5	1	1.5	3
Model $E_4^{0.15,0.25}$	6.75	2.25	3	5.5
Model $E_9^{0.1,0.2}$	6.7	2.2	3.1	5.5

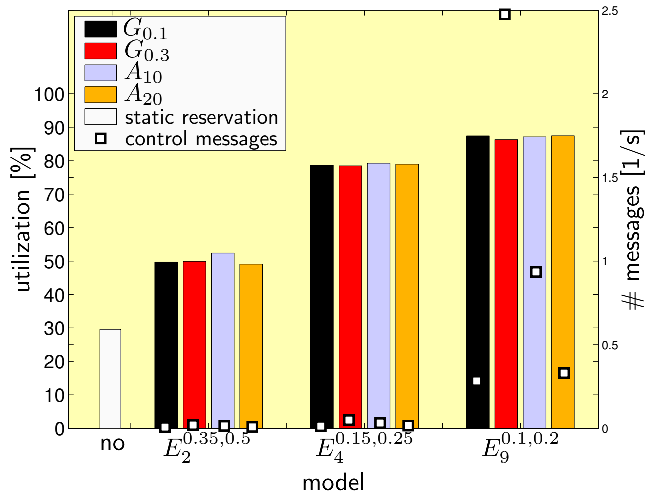

Figure 4. Overall bandwidth available for usage within the network and mean message rate for signaling purposes

The input MPEG-4 trace chosen is a mix of the two MPEG-4 traces obtained from [38]. The traces are that of the movies "Jurassic Park" and "Mr. Bean" which we superimpose. We study the reactivity of some transition models using arithmetic and geometric estimators with different weights with this typical stream of total size 573.8 MB, lasting 59.37 minutes. The size of the sample intervals is $\Delta t = 40\ ms$ (corresponding to 25 frames per sec.) with maximum measured data within one interval of $d_{max} = 178.41\ kbit$ which induces an a priori reserved capacity of $r = 4.4\ Mbit/s$ in order to fulfill the precondition $\rho_i \leq 1\ \forall i$.

The chosen state models are parameterized according to $E_2^{0.35,0.5}$, $E_4^{0.15,0.25}$, and $E_9^{0.1,0.2}$ in order to provide models with different granularities. Figure 4 shows the network capacity put at other stations' disposal and the number of FREE/RECALL messages exchanged by the

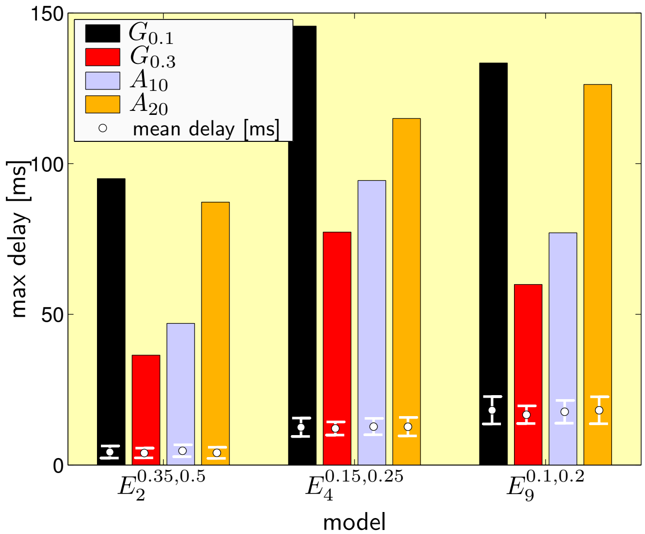

Figure 5. Maximum and mean delay (including 95% confidence intervals)

system. Figure 5 depicts the maximum and mean time the transmission of data was delayed for.

In all experiments transition model $E_9^{0.1,0.2}$ achieves the minimal capacity loss. This is due to the fact that variable-bit-rate applications produce variations and more detailed models can better adapt the resource allocation to the actual requirements. However, the allocated bandwidth of $E_9^{0.1,0.2}$ is often smaller than the value required by the input trace; the models with 2 and 4 states respectively, fit this requirement better. Therefore, a more detailed model achieves better performance in terms of lost bandwidth, but also introduces more delay if the trace fluctuations are rapid and also more communication overhead is generated.

It can be seen, that significant enhancements under realistic load scenarios are achieved by the threshold systems. Further experiments, in particular with smoothed traces, show similar results. Though offered load sometimes exceeds the presumed upper bound of $\rho_i \leq 1$, neither mean nor maximum delay differ significantly, although the number of exchanged messages rises.

6 EXPERIMENTAL RESULTS WITH NS2

In addition to the QoS support mechanisms of the upcoming 802.11e standard, we presented in Section 4 an enhancement scheme for adaptive resource management. We mentioned, that modifications at MAC level are limited to the AP. To evaluate the described scheme, we use the event-driven network simulator *ns-2* [1]. Based on the *CMU-Monarch Project's* wireless and mobility extension to *ns-2* [13, 15], we built up a bug-fixed and verified version, with methods and functionality implemented according to the current 802.11e draft [18].

6.1 Goals and Metrics

The behavior and impacts of the resource redistribution are well understood at an abstract level. However, the actual quantities are connected with the actual network used and its resource management integration. Therefore, the main goal of our simulation studies has been to show that, not only on abstract level, but also when taking into account important aspects of its concrete realization the resource redistribution leads to encouraging results for the selected WLAN architecture.

Changes against the abstract analysis (cf. last section) are that now the control messages which are sent via the network, are indeed counted for the real-time stream's resources. Furthermore, data rate and frames can not be divided into any size, but are bounded by the rules of the corresponding layers. Anyway, the transformation between requests for resources based on measurements of primary load and its realization at MAC level (secondary load) impacts the overall behavior.

Basically, one main issue is to demonstrate that the dynamic resource management can be applied beneficially to WLANs beyond an abstract level. We show that the behavior within a concrete realization meets the expectations inspired by the analytical evaluation. In order to get still better understanding regarding the parameterization, we investigate the way gained transmission capacity is affected by different estimator and state model parameterizations. Of similar significance are the parameters' impacts on the QoS. Thus we evaluate the delays between sending and receiving real-time entities and compare them to the analytically obtained results, cf. Section 5. By means of an example configuration we demonstrate that transmission of best-effort data is indeed improved by the resource management.

The metrics we use in the evaluation are *utilization* of static reserved capacity, *delay* of real-time traffic, best-effort *throughput*, and *bandwidth consumption by signaling procedures*.

We study the ***utilization*** of the static reserved data rate, whereby we highlight the share of the bandwidth allocated statically to real-time streams, which is made available for usage within the overall network, in particular comprising the real-time data. Values of utilizition are given as percentage, reflecting the share of the time (i.e. TXOP) reserved for that stream at MAC level used by the real-time stream itself or handed over to be competed for.

Since many real-time applications discard data received too late, ***delays*** are one important metric for the provided QoS. Values given are measured end-to-end at the UDP level for streams with data rate reserved statically.

One means to highlight the noticable improvements is to depict the overall best-effort ***throughput***. To be able to compare graphs from different levels of load, we use normalized throughput, i.e. the amount of best-effort data actually transmitted divided by the offered load – considered at UDP level.

The ***bandwidth consumption by control messages*** is given in percentage either at UDP or at MAC level. At UDP level the amount of data queued for signaling is divided by the overall offered load. At MAC level the time consumed for sending the message, including the confirmation, is divided by the total length TXOPs claimed for the corresponding stream.

6.2 Simulation Model and Scenarios

In our simulations modeling an 11 $Mbit/s$ wireless LAN, the topology consists of one base station acting as HC and several wireless stations, which are located such that all of them can communicate pairwise. We assume that there is no mobility in the system and only a negligible medium error rate during the periods covered by the experiments. The traffic simulated is composed of real-time and best-effort data. The stations handling real-time queues allocate their data rate statically in advance. These real-time stations are polled by the HC according to the negotiated QoS.

All best-effort traffic is sent with equal priorities as we do not aim to evaluate the impacts of service differentiation using prioritized communications; we refer to other sources [29]. The best-effort stations send data according to miscellaneous kinds of traffic, e.g. traffic is generated similar to variable bit rate audio or video encoders (cf. load models in [37]); packet sizes are taken from normal distributions with mean 400 byte, standard deviations 20 to 80 byte, and inter-packet arrivals of 25 and 40 ms, coming up to flows with an average bit rate of 80 and 128$kbit/s$. We set up the number of best-effort flows such that the WLAN operates at full capacity.

Our reporting starts after a short period of static resource allocation and exchange of network control information (like ARP, etc.), so that this period does not affect the results. We use parameter settings – if available – as recommended by IEEE. Unfortunately, there do not exist any recommendations of the superframe size. Thus, we constituted different values which appeared reasonable, whereby we performed most of the simulations with superframe size 40 ms, which seems to be a proper value for the WLAN communication with the given constraints. The settings in detail are summarized in Table 2.

6.3 Experimental Studies

We run the simulations with different real-time streams represented by real traces. We study several different streams encoded according to MPEG-2, MPEG-4, and ITU Standard H.263. Since all experiments lead to analogous results we single out one typical example, namely the mix of the movies "Jurassic Park" and "Mr. Bean", cf. Section 5.2. In an earlier work we demonstrated that choosing equal distances between adjacent states are one proper way to parameterize the threshold systems [36]. Thus, we make use for the simulations of the models $E_n^{b,m}$ (cf. Section 3.3).

Table 2. Parameters of simulated WLAN

Parameter	Value
time slot	$9\,\mu s$
SIFS,PIFS,DIFS	$16, 25, 34\,\mu s$
superframe	$40\,ms$
bit rate (channel)	$11\,Mbit/s$
size of control msg (UDP data)	$64\,bit$

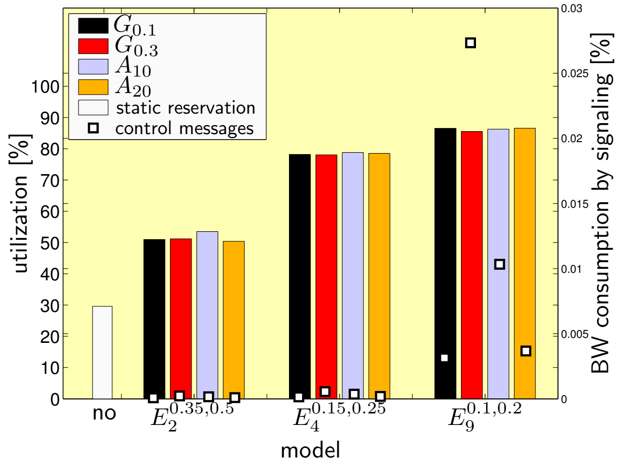

Figure 6. *ns-2*: Overall utilization at MAC level and BW consumed by signaling messages

- Initially we compare the behavior of the WLAN to the idealistic analysis on abstract level. For that we evaluate the impacts of the models $E_2^{0.35,0.5}$, $E_4^{0.15,0.25}$, and $E_9^{0.1,0.2}$ within a scenario consisting of 8 nodes. The data rate r reserved for the real-time stream is set to $r = 4.4$ $Mbit/s$ in order to fulfill the precondition $\rho_i \leq 1$ $\forall i$. The stations altogether offer 8096 $kbit/s$ of best-effort traffic.

  Analogously to Section 5, Figure 6 shows the share of the reserved TXOPs at MAC level used for sending the real-time data or provided for usage within the overall network. Figure 7 shows the maximum and mean delays – including the 95% confidence intervals – as well for different models and estimators as for the case of static reservation. The utilizations appear roughly identical to the calculations done at the abstract level. Evaluating the end-to-end delays, it is indeed quite astonishing how well the parameters' impacts still comply with the values obtained at the abstract level; only model $E_4^{0.15,0.25}$ shows some dif-

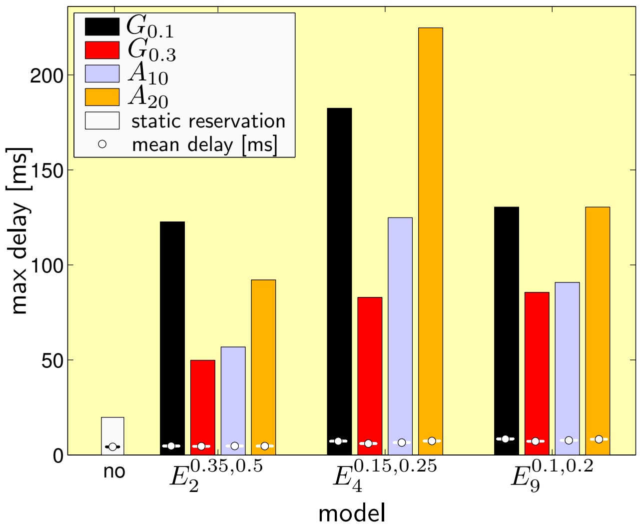

Figure 7. *ns-2*: Maximum and mean delay (including 95% confidence intervals)

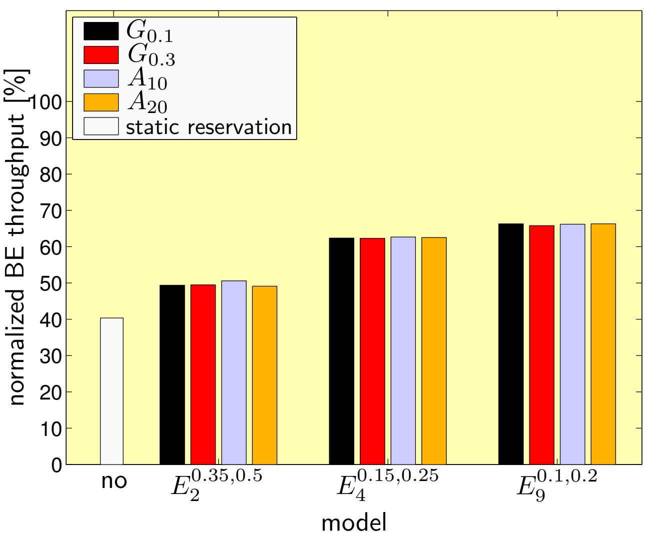

Figure 8. *ns-2*: Normalized best-effort throughput

ferences. As well, the fraction of the data rate used for signaling is negligible as predicted.

Resources not required by real-time flows are fairly distributed among all other stations by adding the remaining TXOPs to the CP, so that the throughput of best-effort traffic increases if the WLAN operates at full capacity. Figure 8 depicts the normalized throughput at UDP level for different models and estimators. All models cause significant improvements to the best-effort traffic. While using static reservation the WLAN can transmit only about 40% of the offered UDP network load, the models $E_4^{0.15,0.25}$ and $E_9^{0.1,0.2}$, however, make it possible to transmit about 62 and 66% respectively. Evidently, the strong

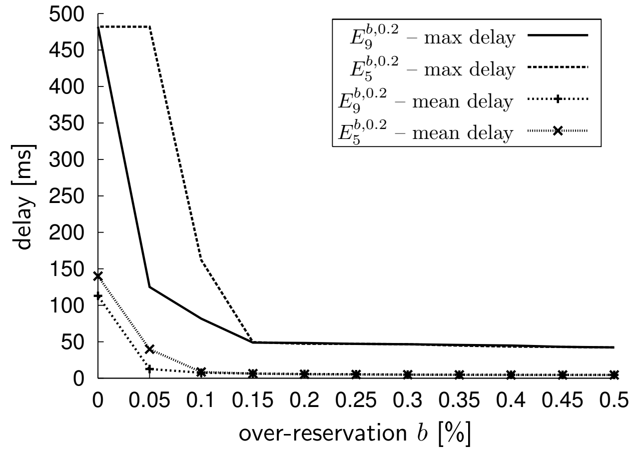

Figure 9. Maximum and mean delays

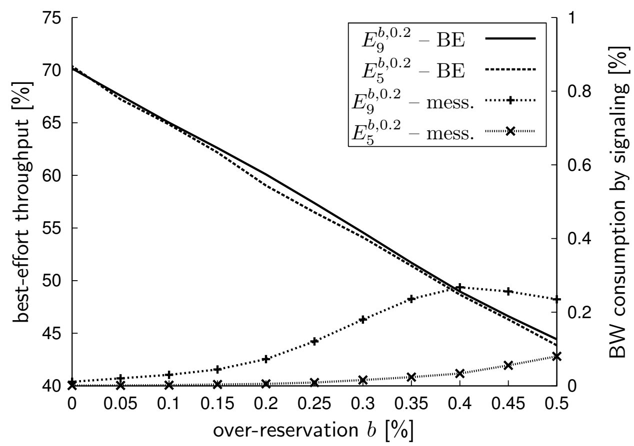

Figure 10. Best-effort throughput and share of reserved data rate consumed by control messages

increase for this configuration is caused due to the fact, that the resources for this high quality stream represent a large share of the overall capacity (about 47%).

- For the second series we model a scenario consisting of 10 different nodes, still offering 8096 $kbit/s$ of best-effort traffic. We vary the amount of desired over-reservation, represented by the parameter b, for models with 5 and 9 states and a minimum allocation of 20%. The load estimate $\hat{\rho}$ is calculated according to $G_{0.3}$. The delay, maximum and mean, as connected with b is depicted by Figure 9. Evidently, a minimum over-reservation has to be instantiated in order to obtain suitable delays. Figure 10 depicts the normalized overall throughput of the best-effort queues and the share of the reserved UDP data rate that is consumed by exchange of control messages. Considering the trade-off between throughput improvement and decreasing QoS various reasonable values can be

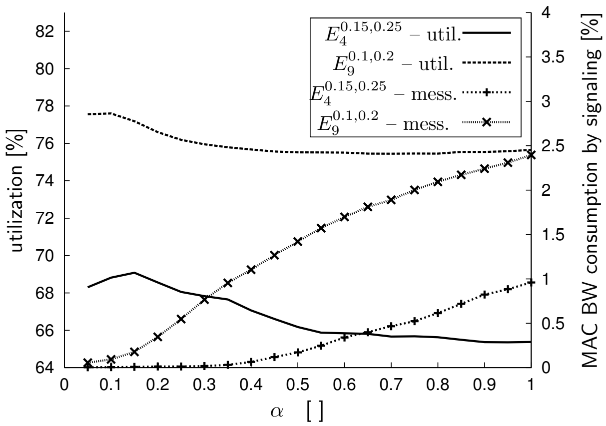

Figure 11. Utilization and fraction of BW at MAC level consumed by control messages

chosen, depending on the corresponding application, in order to improve the overall network capacity.

- For the last series, we choose video streams with lower bit rate in order to highlight the effects of the signaling procedures; the example we single out is a MPEG-4 stream of the movie "Toy Story" [2, 31] with maximum measured data within one interval of $d_{max} = 19.3$ *kbit* which induces an a priori reserved capacity of $r = 471$ *kbit/s*. Figure 11 depicts the impacts of the estimator's parameter α on the potential utilization of the reserved bandwidth for $E_4^{0.15,0.25}$ and $E_9^{0.1,0.2}$. It can be seen, that the increasing geometric weight α leads to an increase in the exchange of signaling information. However, improper choice of the parameterization can affect the improvements in a negative way. Nevertheless, the capacities for the overall network are significantly improved, as compared to the static reservation inducing a poor utilization of about 33% for the reserved part.

Nota bene: We also performed series at which real-time data, accumulated due to lack of freed transmission opportunities, was sent using contention-based channel access. This can lead, depending on the corresponding priorities and the overall load, to very small delays.

It is worth mentioning that all obtained results regarding the real-time flows, remain unaffected by changes of best-effort load or the number of communicating nodes. As the whole controlling and timing that affects real-time queues works self-sufficient, scalability is no issue – of course, except for the initial reservation procedures and the redistribution using contention-based access mechanisms.

7 CONCLUSIONS

Satisfying QoS requirements by means of resource reservations is a promising approach for (wireless) local-area networks, because the expenditure for such reservations is relatively small in LANs, which is unlike to very large WAN configurations where resource reservation may lack scalability. Nevertheless, static resource reservation may lead to an inefficient utilization of resources if significant load variations occur during the lifetime of connections for which a fixed amount of resources has been reserved statically.

We have demonstrated in this paper that efficiency of static resource reservation, as suggested by IEEE project 802 for the standard 802.11e, can be ameliorated considerably by means of dynamic resource redistribution. In particular, we have indicated how reserved resources can be redistributed in WLANs with IEEE 802.11e architecture in a highly efficient manner. We have elaborated analytical and simulation models and have used these models to quantitatively assess the performance of WLANs with load-adaptive resource redistribution. Thorough model evaluations have proven that – despite the very significant gains in bandwidth utilization thanks to dynamic resource redistribution – we are still able to fulfill QoS requirements as they would be typical for multimedia applications, e.g. for applications based on video communications with quite restrictive real-time requirements.

REFERENCES

[1] ns-2, http://www.isi.edu/nsnam/ns, Access date: Feb. 2005.

[2] Video Traces for Network Performance Evaluation, Available under http://trace.eas.asu.edu/, Access date: Mar. 2005.

[3] I. Aad and C. Castelluccia. Differentiation mechanisms for IEEE 802.11. Anchorage - AK, USA, 2001.

[4] P. Anelli and G. Le Grand. Differentiated Services over Shared Media. In *IWQoS*, pages 288–293, 2001.

[5] V. Bharghavan, K.-W. Lee, S. Lu, S. Ha, J. R. Li, and D. Dwyer. The TIMELY Adaptive Resource Management Architecture. *IEEE Personal Communications Magazine*, 5(4), August 1998.

[6] C. Bouras and A. Gkamas. Multimedia Transmission with Adaptive QoS based on Real-Time Protocols. *Int. Journal of Comm. Systems*, 16(3):225–248, April 2003.

[7] R. Braden, D. Clark, and S. Shenker. Integrated Services in the Internet Architecture: An Overview. RFC 1633, IETF, June 1994.

[8] M. Cardei, I. Cardei, and D.-Z. Du, editors. *Resource Management in Wireless Networking*. Springer, 2005.

[9] H. J. Chao and X. Guo. *Quality of Service Control in High-Speed Networks*. J. Wiley, 2002.

[10] B.-Y. Choi, J. Park, and Z.-L. Zhang. Adaptive Random Sampling for Load Change Detection. *Performance Evaluation Review*, 30(1):272–273, 2002.

[11] S. Choi. Emerging IEEE 802.11e WLAN for Quality-of-Service (QoS) Provisioning. *SK Telecom Telecommunications Review*, 12(6):894–906, December 2002.

[12] D. D. Clark and W. Fang. Explicit Allocation of Best-Effort Packet Delivery Service. *IEEE/ACM Trans. on Networking*, 6(4):362–373, August 1998.

[13] CMU-Monarch Project. The CMU Monarch Project's Wireless and Mobility Extensions to NS, http://www.monarch.cs.cmu.edu/cmu-ns.html, Access date: Feb. 2005, August 1999.

[14] M. El-Kadi, S. Olariu, and H. Abdel-Wahab. A Rate-Based Borrowing Scheme for QoS Provisioning in Multimedia Wireless Networks. *IEEE Trans. on Parallel and Distributed Systems*, 13(2):156–166, February 2002.

[15] P. Garg, R. Doshi, R. Greene, M. Baker, M. Malek, and X. Cheng. Using IEEE 802.11e MAC for QoS over Wireless. In *Proc. of the IEEE International Performance Computing and Communications Conference (IPCCC 2003)*, Phoenix, Arizona, 2003.

[16] A. Ghanwani, W. Pace, V. Srinivasan, A. Smith, and M. Seaman. A Framework for Integrated Services Over Shared and Switched IEEE 802 LAN Technologies. RFC 2816, IETF, May 2000.

[17] F. Harivelo, G. Le Grand, P. Anelli, J. Wolf, and B. E. Wolfinger. Expedited Forwarding for WiFi. In *Proc. of ISWCS'04, 1st Int. Symp. on Wireless Communication Systems*, Mauritius, September 2004.

[18] IEEE 802.11e/D12.0. Draft Supplement to Part 11: Wireless Medium Access Control (MAC) and physical layer (PHY) specifications: Medium Access Control (MAC) Enhancements for Quality of Service (QoS), IEEE 802.11 WG, November 2004. IEEE.

[19] IEEE Std. 802.11-1999. Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications, Reference number ISO/IEC 8802-11:1999(E), IEEE Std 802.11, 1999 edition, 1999. IEEE.

[20] C. R. Kalmanek, H. Kanakia, and S. Keshav. Rate Controlled Servers for Very High-Speed Networks. In *Proc. of the Conference on Global Communications (GLOBECOM)*, pages 12–20, 1990.

[21] G. Le Grand. *Qualité de service dans des environments Internet mobile*. PhD thesis, Université P. et M. Curie, LIP 6, Paris, July 2001.

[22] A. Lindgren, A. Almquist, and O. Schelén. Quality of Service Schemes for IEEE 802.11 Wireless LANs - An Evaluation. *In Special Issue of the Journal on Special Topics in Mobile Networking and Applications (MONET) on Performance Evaluation of QoS Architectures in Mobile Networks*, 8(3):223–235, June 2003.

[23] A. Malla, M. El-Kadi, S. Olariu, and P. Todorova. A Fair Resource Allocation Protocol for Multimedia Wireless Networks. *IEEE Trans. on Parallel and Distributed Systems*, 14(1):63–71, January 2003.

[24] S. Mangold, S. Choi, G. R. Hiertz, O. Klein, and B. Walke. Analysis of IEEE 802.11e for QoS Support in Wireless LANs. *IEEE Wireless Communications*, 10(6):40–50, December 2003.

[25] S. Mangold, S. Choi, P. May, O. Klein, G. Hiertz, and L. Stibor. IEEE 802.11e Wireless LAN for Quality of Service. In *Proc. of European Wireless*, pages 32–39, February 2002.

[26] L. Massoulié and J. Roberts. Bandwidth Sharing: Objectives and Algorithms. *IEEE/ACM Trans. on Networking*, 10(3):320–328, 2002.

[27] M. Mirhakkak, N. L. Schult, and D. Thomson. Dynamic Bandwidth Management and Adaptive Applications for a Variable Bandwidth Wireless Environment. *IEEE Journal of Selected Areas in Communications (JSAC)*, 19(10):1984–1997, October 2001.

[28] M. Moran and B. Wolfinger. Design of a Continuous Media Transport Service and Protocol. Technical Report TR-92-019, Int. Computer Science Institute, Berkeley, CA, April 1992.

[29] Q. Ni, L. Romdhani, and T. Turletti. A Survey of QoS Enhancements for IEEE 802.11 Wireless LAN. *Journal of Wireless Communications and Mobile Computing (JWCMC)*, 4(5):547–566, 2004.

[30] E. P. Rathgeb. Modelling and Performance Comparison of Policing Mechanisms for ATM Networks. *IEEE Journal On Selected Areas In Comm.*, 9(3):325–334, 1991.

[31] P. Seeling, M. Reisslein, and B. Kulapala. Network Performance Evaluation Using Frame Size and Quality Traces of Single-Layer and Two-Layer Video: A Tutorial. *IEEE Communications Surveys and Tutorials*, 6(2):58–78, 2004.

[32] A. Shalkh, J. Rexford, and K. G. Shin. Load-Sensitive Routing of Long-Lived IP Flows. In *Proc. of ACM SIGCOMM '99*, pages 215–226, Cambridge, USA, 1999.

[33] J. Sobrinho and A. Krishnakumar. Real-time Traffic over the IEEE 802.11 medium access control layer. *Bell Labs Technical Journal*, pages 172–187, 1996.

[34] A. Veres, A. Campbell, M. Barry, and L. Sun. Supporting Service Differentiation in Wireless Packet Networks using Distributed Control. *IEEE Journal of Selected Areas in Communications (JSAC), Special Issue on Mobility and Resource Management in Next-Generation Wireless Systems*, 19(10):2094–2104, 2001.

[35] J. Wolf. Network Resource Management for Real-Time Streams within a Multimedia Document Server Architecture. In *Proc. of 49. Int. Wissenschaftliches Kolloquium (IWK)*, volume 2, pages 297–302, Ilmenau, September 2004.

[36] B. E. Wolfinger, J. Wolf, and G. Le Grand. Improving Node Behaviour in a QoS Control Environment by Means of Load-dependent Resource Redistributions in LANs. *Int. Journal of Comm. Systems*, 18(4):373–394, 2005.

[37] B. E. Wolfinger, M. Zaddach, K. D. Heidtmann, and G. Bai. Analytical Modeling of Primary and Secondary Load as Induced by Video Applications using UDP/IP. *Computer Commun.*, 25(11/12):1094–1102, 2002.

[38] A. Wolisz. MPEG-4 and H.263 Video Traces for Network Performance Evaluation, Available under http://www-tkn.ee.tu-berlin.de/research/trace/trace.html, 2000.

[39] R. Yavatkar, D.Hoffman, Y. Bernet, F. Baker, and M. Speer. SBM (Subnet Bandwidth Manager): A Protocol for RSVP-based Admission Control over IEEE 802-Style Networks. RFC 2814, IETF, May 2000.