Proceedings of APCC2008 copyright (c) 2008 IEICE 08 SB 0083

Admission control in wireless infrastructure networks based on the predicted percentage of delayed packets

Anselm Teh, Aruna Jayasuriya
Institute for Telecommunications Research
University of South Australia
Adelaide, Australia
Email: anselm.teh@postgrads.unisa.edu.au
aruna.jayasuriya@unisa.edu.au

Peter Pudney
Centre for Industrial and Applied Mathematics
University of South Australia
Adelaide, Australia
Email: peter.pudney@unisa.edu.au

***Abstract*—As the use of bandwidth hungry applications such as video conferencing and video on demand increases, ensuring Quality of Service (QoS) becomes increasingly important. In wireless access networks such as WiFi and WiMax, admission control is necessary to control QoS. Many admission control methods use the mean bit rate of a flow to determine if it can be allowed, however there is a lack of methods that accurately assess variable bit rate flows. We propose a simple alternative method that extends our previous work on admission control. It uses the bit rate statistics of each flow to predict the percentage of packets that are excessively delayed. A new flow is permitted only if this predicted percentage is sufficiently low. Using a simulation study we show that the proposed method can accurately predict the proportion of delayed packets and hence can control the load of the network to achieve the desired delay bounds.**

I. Introduction

The capacity of wired networking technologies is increasing rapidly, with wide availability of Gigabit Ethernet Local Area Networks (LANs) and introduction of 10 Gigabit Ethernet in recent times. With the multimedia explosion over the Internet in recent years, more and more consumers will be making use of the capacity of these high-speed networks to perform real-time communication and media sharing across large distances.

Wireless networks will always have capacity shortcomings compared to wired networks, due to the variability of the wireless medium [1]. 802.11 Wireless LANs (WLANs) are currently used by a large number of homes and businesses as local area networks and as a method to connect to the Internet. The 802.11g network has a capacity of 54 Mbps, while the up-and-coming 802.11n network can operate at up to around 160 Mbps [2]. The practical capacities of these networks are however much lower, and greatly depend on the distance from the access point, and obstacles in the way. 802.16 Metropolitan Area Networks (MANs) have also gained a lot of attention recently, and have a capacity of around 15-30 Mbps for the 802.16e standard which allows mobile stations [3],[4].

With the increased usage and focus on real-time applications, QoS has become a very important consideration in access networks, especially in the lower capacity wireless access networks. As bandwidth is so precious, a good admission control method is needed to prevent too many large flows saturating the network and decreasing the level of QoS for the rest of the users. A good admission control method should allow the network to operate as close as possible to capacity but with good Quality of Service and stability. In this paper, we describe a method for admission control that uses statistical characteristics of different flows to control the traffic admitted to a network by predicting the percentage of packets that will exceed a specified delay threshold.

The rest of the paper is organised as follows. Section II describes some existing admission control methods and their problems, while section III gives an overview of our previous work on estimating the probability of saturation. In section IV we discuss queueing related issues and our improvements to our previous work, and present simulation results and conclusion in section V and VI.

II. Background

An admission control mechanism decides which flows may be allowed into a network without the network being saturated. It uses knowledge of incoming flows and the current network situation to ensure the QoS for flows in the network. Current 802.11e and 802.16 standards allow the specification of a number of QoS parameters, such as maximum and minimum data rate, maximum and minimum service interval, and delay bound. Many types of simple admission control such as the method suggested in the 802.11e standard [5] use only the mean data rate of a flow to calculate whether a flow should be admitted or not. As many popular applications produce traffic with highly variable bit rates, e.g. video conferencing, VoIP, considering only the mean bit rate when admitting a flow does not account for the possible state of the network at any instant.

In [6] admission control is based on the 802.11e standard method. However instead of calculating admission based on mean data rate, they use time-varying data rates based on the actual usage of each flow. This may depict network usage more accurately at a particular instant, however does not accurately show the nature of the set of flows over time.

Other admission control methods rely on calculating the available bandwidth in the network based on metrics such as number of nodes, queue lengths, number of collisions, delay, channel busy time [7]. This method is not very accurate however, as metrics such as number of nodes and queue lengths do not always directly correspond to the utilisation of the network at any time and are very dependent on the type of flow being used. More direct methods such as channel busy time may come to a false conclusion due to the variable nature of the VBR flows.

Admission control methods for 802.16 networks generally use the minimum bit rate of all flows to perform admission control, then use a method to estimate the amount of bandwidth to keep as a "guard band" to allow for variable bit rates or flows performing handover into the network, which receive a higher priority [8],[9]. One method [10] calculates admission thresholds depending on the time of day. In [11], admission control is performed by monitoring available bandwidth in the 802.16e network, and comparing this value with the minimum and maximum bit rates of the incoming flow. If the available bandwidth is larger than a "required bandwidth" taken from the minimum and maximum bit rate parameters, or if there are lower priority flows whose QoS can be degraded, the flow will be admitted. These methods use QoS parameters as a guide, and do not specifically take into account the time-varying traffic characteristics of an incoming flow and its effect on the existing set of various flows.

Real time interactive flows such as video conferencing or VoIP have strict QoS bounds that must be adhered to in order to provide the best experience for the user. Too large a delay or variation in delay would cause gaps in conversation and lead to poor service for the user. The standardisation sector of the International Telecommunication Union (ITU-T) recommends in ITU-T G.114 that one way transmission delay for a voice conversation should be kept to below 150 ms for most voice conversations, however delays of 150 - 400 ms are acceptable for international connections. Delays of over 400 ms are generally unacceptable [12]. Video conferencing applications design the video stream to be viewed in sync with the audio. To do this, the video stream must have a latency similar to the audio stream accompanying it.

If we are able to predict the percentage of packets that will be delayed over a specific delay bound for any set of flows, this would be a much more useful method of determining whether a flow should be accepted or rejected in the network.

III. Estimating probability of saturation

This section will provide a quick overview of the previous work, on which the current work is based. Our method is based on the Central Limit Theorem, which states that if we have a set of independent variables, each with an arbitrary probability distribution and finite variance, then the sum will tend towards a normal distribution. Our method can be used if the network capacity is high enough that it can accommodate many flows. In our previous work [13], we found that our method was accurate if capacity was around 10 times the mean rate of a flow, or greater. This result may vary with the traffic characteristics of the flows.

Using our method, each flow entering the network that requires a specific level of service must first include in their QoS specification both a mean and variance for the flow. In both 802.11e and 802.16e systems, a station that requests a specific level of service must submit a Traffic Specification (TSPEC) or QoS Parameter List before they are accepted. Our method requires the variance parameter to be added to this list. When the QoS specifications for an incoming flow are received by the access point or the base station, we can model the generated traffic of the set of flows as a normal distribution. We then calculate the probability of reaching saturation as follows: If we model the bit rate of a single flow as a random variable B_i, given the mean bit rate μ_i and variance σ_i^2 of each flow, the probability of exceeding capacity C is

$$P[\sum B_i > C] \approx P^*[\sum B_i > C] \quad (1)$$

$$= \int_C^\infty \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} dx \quad (2)$$

where

$$\mu = \sum_i \mu_i$$

is the mean bit rate for the set of flows and

$$\sigma^2 = \sum_i \sigma_i^2$$

is the variance of the set of flows.

We can calculate the approximation using

$$P^*[\sum B_i > C] = 1 - \frac{1}{2}\left[1 + \text{erf}\left(\frac{C-\mu}{\sigma\sqrt{2}}\right)\right] \quad (3)$$

where erf is the Gauss error function

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt.$$

The probability given by equation (3) is then used as an indicator as to whether or not the flow will be accepted. If the network has strict constraints on QoS, then a one percent probability of saturation might be used as a threshold—any flow that increases the probability of saturation to one percent or more is not accepted.

The previous research found that our method of calculating the probability of saturation was a more accurate method of predicting excessive delay than compared to the suggested 802.11e that uses mean data rate of incoming flows.

IV. Queueing

In our previous work, we observed that there exists a range of scenarios that have the same probability of saturation but have different mean network delays. This is due to two reasons. First, as we are predicting only the probability of saturation, this does not tell us how much data is generated in excess of the medium capacity. Although using the probability of saturation to perform admission control is more accurate than using the mean bit rates of the flows, our previous algorithm only calculates the probability of the generated traffic exceeding the capacity of the medium. Hence it does not give an indication as to the amount of data in excess of the capacity at a given time. As a numerical example, if a set of flows

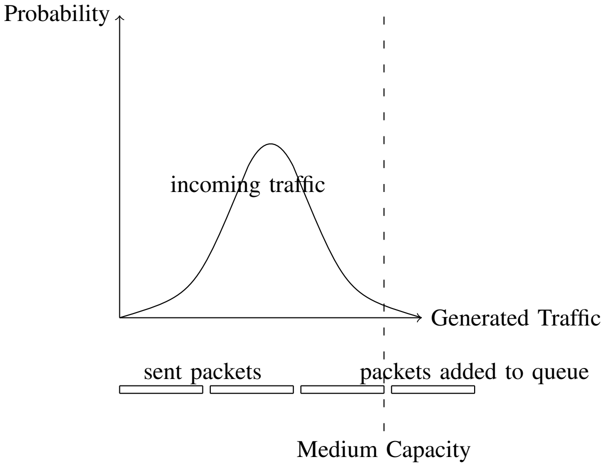

Fig. 1. Relationship between queued traffic and traffic distribution for a single interval

had a 10 percent probability of exceeding the capacity by 100 bytes in any interval, this would probably prove to be easily handled by the network, whereas a 4 percent probability of exceeding the capacity by 40 Kbytes may pose a significant problem. The second reason is a flow on effect of the amount of traffic exceeding the capacity, i.e. queuing. Once the traffic generated in an interval exceeds the capacity, this traffic needs to be added to the amount of traffic that needs to be transmitted in the next interval. Hence the total ready to be transmitted in any interval should include the traffic generated in that interval plus, any traffic that has been queued in the last interval.

As the probability of saturation can relate to a range of different delays in a network, a more accurate metric for assessing the QoS state of the network is the percentage of packets that are delayed over their specified delay threshold. For real-time applications such as video conferencing, an important QoS factor that has significant influence in the Quality of Experience (QoE) of the output is the maximum delay experienced by the flow. If we can specify a deadline and predict that only a small percentage of packets will exceed this deadline for a set of VBR flows, we can go a long way in ensuring the QoS level for our network. As shown in figure 1, if we can calculate the expected amount of generated data that exceeds the medium capacity, we can predict the growth of the queue in each interval. We describe our method to estimate the proportion of delayed packets below.

If we have a set of flows that can tolerate a maximum delay δ, then we examine the traffic in intervals of δ as this is the maximum interval at which flows will have to be served in order to ensure the delay bound is met. To examine the behaviour of the total queue size for the set of flows during the interval, we let the length of the queue at the start of the interval be q. The growth of the queue g during the interval then depends on the number of bits arriving at the queue A, and the number of bits transmitted from the queue T. In each interval, the queue size grows by $A-T$, however the queue size cannot become negative. If $A-T < -q$ then the queue goes to zero, and $g = -q$. If $A-T > -q$ then the queue does not empty, and $g = A-T$. Thus

$$g(q,A) = \begin{cases} -q & \text{if } A-T < -q \\ A-T & \text{if } A-T > -q \end{cases} \tag{4}$$

The growth of the queue can then be summarised in the following equation.

$$g(q,A) = \max(-q, A-T) \tag{5}$$

As A cannot be negative, with an incoming traffic distribution of $f(x)$ the expected growth of the queue in each interval is as follows:

$$\int_{-\infty}^{\infty} g(q,x) f(x) dx \tag{6}$$

$$= \int_{-\infty}^{T-q} -q f(x) dx + \int_{T-q}^{\infty} (x-T) f(x) dx \tag{7}$$

As the distribution of the incoming traffic for the set of flows can be described by a normal distribution as shown in our previous work [13], we can expand (7) using equation (1). Equation (7) can then be expressed as follows.

$$\begin{aligned} \Delta q = & \int_{-\infty}^{T-q} \frac{-q}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} dx \\ & + \int_{T-q}^{\infty} \frac{x-T}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} dx \\ = & \frac{1}{\sqrt{2\pi}} \Bigg[\sqrt{\frac{\pi}{2}} (T-\mu) \left(-1 - \text{erf}\left(\frac{\mu+q-T}{\sqrt{2}\sigma} \right) \right) \\ & + \sigma e^{-(\mu+q-T)^2/(2\sigma^2)} \Bigg] \end{aligned} \tag{8}$$

The expected growth of the queue, Δq, depends on the size of the queue, q. The size of the queue at which the queue is expected to stabilise is found by setting $\Delta q = 0$ in (8) and solving for q. The proportion of packets delayed over one interval in a given scenario is then q/T.

V. Simulation

To test our method of predicting the proportion of packets delayed over a specific time, we used our algorithm to calculate the expected queue size and proportion of packets delayed for a range of scenarios. The scenarios were based in an 11 Mbps 802.11e network, with varying numbers of video flows of different rates. The types of flows used in the scenarios were all MPEG4 flows, with statistics retrieved from TKN's website [14]. There were three sizes of flows used - with mean data rate / variance as follows (including all overheads):

- low rate 1.84×10^5 bps / 8.52×10^9 bps^2
- med rate 4.46×10^5 bps / 7.16×10^{10} bps^2
- high rate 1.02×10^6 bps / 3.17×10^{11} bps^2

We assume all the data flows are gamma distributed [14] and used equation (8) to predict the proportion of packets delayed excessively for all possible configurations with utilisation from 0–100% using these flows. This produced a table showing the number of high rate, medium rate, and low rate flows in each scenario, and the related utilisation and predicted proportion of delayed packets. A selection of scenarios were chosen with utilisation from between

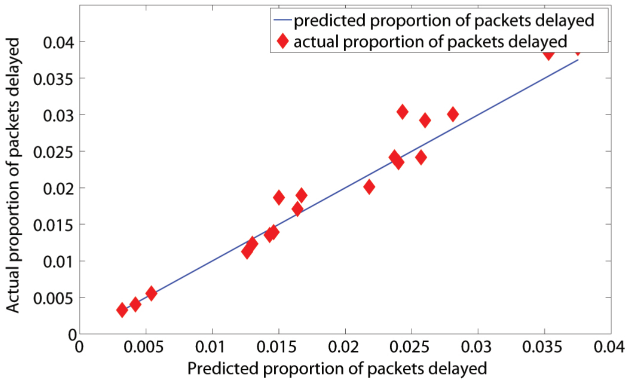

Fig. 2. Actual vs predicted proportion of packets delayed over threshold

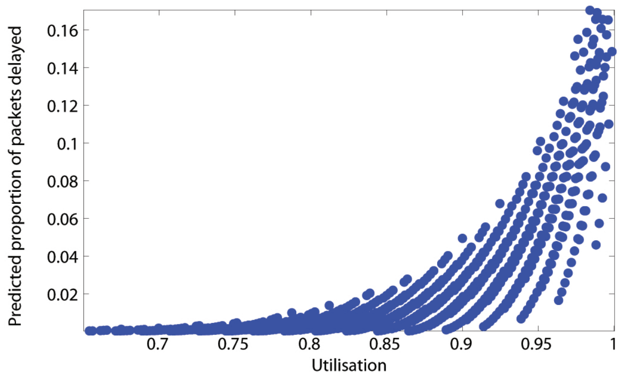

Fig. 3. Predicted proportion of packets delayed over threshold versus Utilisation

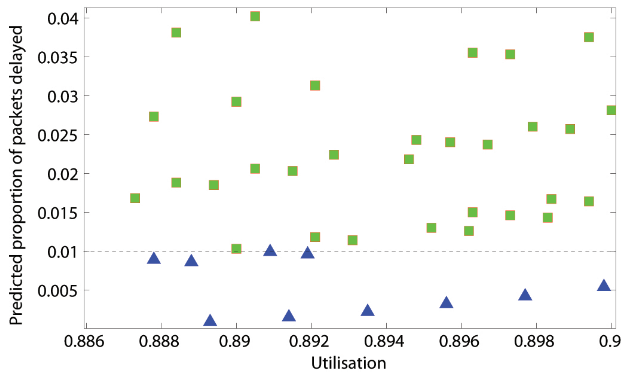

Fig. 4. Accepted and rejected scenarios - predicted proportion of packets delayed over threshold versus Utilisation

89.46 and 90%. The simulated scenarios consisted of a combination of 1 - 30 small flows, 1 - 16 medium flows, and 1 - 6 large flows. These scenarios were simulated with OPNET to compare the simulation results to the predicted results and gauge the accuracy of our method. The simulation modelled a small infrastructure network with one Quality of Service enabled Access Point (QAP) and several stations. The results were also compared to utilisation, to study its relationship with the proportion of packets delayed over the delay bound.

As mentioned earlier, the maximum one way delay for real-time flows such as video conferencing or VoIP has been suggested to be 150 ms [12] however in our simulations we specify the maximum delay to equal the source inter-arrival interval, i.e. 40 ms. This is a more realistic bound to use in the local network, as the total delay also includes such things as packetisation delay, buffering delay, and delay introduced over the backbone network. The delay in our simulation was measured from the time the generated packet arrived in the MAC queue to the time it arrived at the destination. The proportion of packets delayed over the delay bound was calculated by simply counting the number of packets with a delay over 40 ms and dividing this number by the total packets received, while utilisation was calculated as the sum of the mean data rates divided by the total capacity.

The OPNET scenarios used an 802.11e infrastructure network with 11Mbps physical layer. Only HCCA polling was used and the beaconing was turned off to simplify the simulation and review of the results. The scheduler used in HCCA was a round robin scheduler based on the simple scheduler described in annex K of the 802.11e standard [5]. The Service Interval used in scheduling was set to 40 ms, equal to the source packet interarrival interval of the MPEG flows, and the delay was calculated as the time the packet arrived at the source station MAC layer, to when it was received at the destination. Each scenario was run three times with three different seeds, and the simulation results were averaged to provide the most accurate result.

Figure 2 shows the comparison between the predicted and actual proportion of packets delayed over 40ms. The predicted proportion of delayed packets is shown on the x axis, while the actual (simulated) proportion of delayed packets is on the y axis. We can see on the graph that there is a clear relationship between the proportion of packets delayed for each scenario as predicted by our algorithm, and the simulated results. The simulated proportion of packets delayed follows the predicted proportion of packets delayed to within a small margin of error. In our simulations, the greatest error between predicted and actual proportion of packets delayed was 6.53×10^{-3}, while the average error was 1.86×10^{-3}. Using an algorithm based on equation (8), we would be able to accurately determine the proportion of packets delayed for a set of VBR (or VBR and CBR) flows.

In figure 3 we see a selection of the total graph of the predicted proportion of packets delayed vs utilisation. We can see that for the range of scenarios containing the VBR sources mentioned, there is not a simple relationship between utilisation and the predicted proportion of packets that will be delayed over a specific threshold. A single utilisation value might correspond to multiple values of proportion of delayed packets.

Figure 4 shows a comparison between the predicted proportion of packets delayed and utilisation, demarcating the line between accepted and rejected scenarios. These results are taken from OPNET simulations of 41 scenarios with utilisation from 88.7 to 90%. We can see that for the small utilisation range shown on the graph, the proportion of packets delayed (exceeding the threshold) varies greatly, from around 0-4%. This is around the percentage range that is important in controlling the QoS of real-time flows. In this group of simulations, we used a 1% proportion of total packets delayed over 40 ms as a threshold regarding the admission of a flow/scenario. In the figure, the points marked by squares correspond to the scenarios that produced over 1% of total packets delayed when simulated in OPNET, while the points marked by triangles correspond to the scenarios where the proportion of delayed packets is below 1%. Using this threshold, we can see from the figure that our algorithm predicted which scenarios would produce above and below the packet delay threshold accurately. We can also see that there is no strong relationship between utilisation and the proportion of packets delayed excessively for a network containing VBR flows. Hence, utilisation cannot be used accurately as an admission control method to tightly and accurately control the delay.

In this section we also simulated the performance of both admission control algorithms in a scenario with a mixture of small, medium and large flows. The OPNET scenario contained 15 small flows, 10 medium flows, and 6 large flows, each with an exponential on and off state time with average of 10 seconds. Before a flow was started, the flow statistics were checked and the flow was accepted or rejected based on the admission control decision. Using our method the threshold for the proportion of total packets delayed was set to 1%, while the standard admission control method rejected flows that put the total mean data rate over the medium capacity. The packets from accepted flows were scheduled using an Earliest Deadline First algorithm, similar to the method in [15]. Each scenario was simulated for 1000 seconds, and the results are listed below.

TABLE I
COMPARISON OF ADMISSION CONTROL METHODS

	Standard admission control	Predicted proportion of packets delayed
Prop packets delayed	0.1280	0.0095
Ave delay (ms)	38.9	8.1
Ave throughput (Mbps)	7.593	7.506

Table I shows that performing admission control using the method suggested in the 802.11e standard, allowed flows into the network that delayed 12.8 percent of the delivered packets over the 40ms threshold, while our proposed method kept the percentage of packets delayed below 1 percent. It can also be observed that the proposed method only reduces the throughput by 1 percent as compared to the standard method, while reducing the average delay by 80 percent.

VI. CONCLUSION

When an application requests access to a wireless network that supports QoS, the incoming application must submit a description of its resource requirements. In this paper we have shown that the variance of the required bit rate is an important parameter in characterising the resource requirement, despite receiving little attention in recent research in QoS control.

In this paper we extend our previous admission control scheme by using the mean data rate and variance of incoming flows with equation (8) to predict the proportion/percentage of packets excessively delayed. This method accurately predicts the proportion of packets delayed past a deadline, which is an important measure of QoS in all real-time flows and directly relates to the QoE of an application such as VoIP or video conferencing. We also show that utilisation itself does not have a strong relationship with the amount of delayed packets for a set of VBR flows and hence is not suited to an admission control scheme that aims to keep the number of delayed packets below a given threshold.

REFERENCES

[1] Q. Zhang, W. Zhu, and Y.-Q. Zhang, "End-to-end qos for video delivery over wireless internet," *Proceedings of the IEEE*, vol. 93, pp. 123–134, Jan. 2005.

[2] 802.11n speed. Broadcom. [Online]. Available: http://80211n.com/80211n-speed.html

[3] Wimax forum - technology. WiMAX Forum. [Online]. Available: http://www.wimaxforum.org/technology

[4] "A comparative analysis of mobile wimax development alternatives in the access network," White Paper, WiMAX Forum, pp. 1–47, May 2007. [Online]. Available: http://www.wimaxforum.org/technology/downloads/mobile_wimax_deployment_alternatives.pdf

[5] *Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements*, IEEE Std.

[6] G. Boggia, P. Camarda, L. A. Grieco, and S. Mascolo, "Feedback based bandwidth allocation with call admission control for providing delay guarantees in ieee 802.11e networks," *Computer Communications*, vol. 28, pp. 325–337, feb 2005.

[7] I. D. Chakeres and E. M. Belding-Royer, "Pac: Perceptive admission control for mobile wireless networks," *Quality of Service in Heterogeneous Wired/Wireless Networks*, pp. 18–26, Oct. 2004.

[8] S. Chandra and A. Sahoo, "An efficient call admission control for ieee 802.16 networks," *Local & Metropolitan Area Networks*, pp. 188–193, June 2007.

[9] X. Guo, W. Ma, Z. Guo, and Z. Hou, "Dynamic bandwidth reservation admission control scheme for the ieee 802.16e broadband wireless access systems," *Wireless Communications and Networking Conference*, pp. 3418–3423, Mar. 2007.

[10] F. Hou, P.-H. Ho, J. Cai, X. Shen, C.-C. Hsieh, and A.-Y. Chen, "A novel distributed connection admission control scheme for ieee 802.16 networks," *Proceedings of the 10th ACM Symposium on Modeling, analysis, and simulation of wireless and mobile systems*, pp. 173–180, Oct. 2007.

[11] Y. Ge and G. Kuo, "An efficient admission control scheme for adaptive multimedia services in ieee 802.16e networks," *Vehicular Technology Conference*, pp. 1–5, sep 2006.

[12] *One-way transmission time*, Std., 1996.

[13] A. Teh, P. Pudney, and A. Jayasuriya, "Efficient admission control based on predicted traffic characteristics," *Personal, Indoor and Mobile Radio Communications*, pp. 1–5, Sept. 2007.

[14] F. Fitzek and M. Reisslein. Mpeg-4 and h.263 video traces for network performance evaluation. TKN, Berlin University of Technology. [Online]. Available: http://www-tkn.ee.tu-berlin.de/research/trace/trace.html

[15] A. Grilo, M. Macedo, and M. Nunes, "A scheduling algorithm for qos support in ieee802.11 networks," *Wireless Communications*, vol. 10, pp. 36–43, jun 2003.