Load Observation and Control Model for Load Balancing with QoS in WLAN

Robil DAHER and Djamshid TAVANGARIAN

Abstract--**The usage of Load Balancing Mechanism (LBM) and Quality of Service (QoS) in WLAN Networks is currently turning into an essential process to increase WLAN performance, especially for real time applications. A model for LBM with QoS in WLAN includes three models, namely load observation, load decision, and load control model. In this paper, we introduce the Wireless Medium Busy Time (MBT) as a load metric for this model. MBT, as quantitative factor over determined time, describes the load of an Access Point (AP) more precisely than others metrics, and thus it can be used to detect the overloaded APs. A load model is discussed accordingly, and as a result, a distributed architecture is presented, where both APs and Stations (STAs) must cooperate with each other. Finally, our experiments confirm the importance of MBT as load metric, and shows that the effect of MBT variation under low bit rate on the QoS can be processed by using per-STA reservation.**

Index Terms-- **Load Balancing, Load Metric, Quality of Service, Wireless Medium Busy Time, WLAN**

I. Introduction

ADVANCED utilization and applications of WLAN require more development in the field of WLAN-Technology. A key issue of this technology development is the congestion management of mobile users at hot spots. That implies the load distribution on Access Points (APs) and also the support of real-time applications with Quality of Service (QoS).

Typically, any Load Balancing Mechanism (LBM) in a distributed system relies on an architecture model which specifies the functions of the various system components [6]. In WLAN, there are three main steps to be included in any LBM [3] [7]; these imply exploring of the load state in the WLAN, making the load distribution decision, and finally handing-off selected STAs to selected APs. Similarly, the Resource Reservation and Admission Control (RRAC) by QoS also occurs in three steps [8], resource reservation requests from STAs, deciding the reservation and the medium admission on AP, and finally informing Stations (STAs) about the decision results. In that respect, a model for load balancing with QoS can be divided into three partial models [3] [6]: load observing, load decision, and load control model.

Manuscript received February 04, 2005.

R. W. Daher is with the Institute of Computer Science, Rostock University, Albert-Einstein-Str. 21, 18059 Rostock, Germany (tel: 0049-381-498 7532, e-mail: Robil.Daher@uni-rostock.de).

D. Tavangarian is with the Institute of Computer Science, Rostock University, Albert-Einstein-Str. 21, 18059 Rostock, Germany (tel: 0049-381-498 7520, e-mail: Djamshid.Tavangarian@uni-rostock.de).

The selection of the appropriate load metrics is a key issue of the load model, which includes both load observing and load control model. Although on the "packet level" many parameters exist that may be considered as load metrics, such as link bandwidth, traffic intensity, packet error rate, Received Signal Strength (RSS), etc. nevertheless only a few of them are used by the known load balancing and RRAC solutions [2] [8] [4] [5]. The combination of LBM and RRAC in one system requires the usage of the same load metric to perform the related functions [4] [5]. For instance, the bandwidth can be considered as a load metric for LBM, and also as a resource metric for processing RRAC. Generally, bandwidth is one of the most important load metrics, which is often used in WLAN similar to the way it is used in LAN and WAN networks. However, the non-determined bit rate variations of associated links cause the wireless channel bandwidth to be variable over the time, which leads the bandwidth to be insufficient as a load metric [7]. It is worth noting that the links with low bit rate, e.g. like 1 and 2 Mbps in case of 802.11b, generate more busy time on the Wireless Medium (WM) compared with the high bit rate links, e.g. like 5.5 and 11 Mbps in case of 802.11b, despite their generation of lower data traffic [1]. Therefore, it is more useful to express the load in term of WM-Busy Time (MBT). This paper aims first of all to introduce the importance of MBT as an effective load metric for load balancing and QoS, and secondly to design a load model that provides load balancing and QoS for WLAN in infrastructure mode. The load decision model and its algorithms are beyond the scope of this paper.

In this paper, we calculate and discuss the MBT as load metric for WLAN in section II. In section III we describe our MBT-based load observing and load control model for load balancing and RRAC. After that, we demonstrate the practical measurements of the MBT for special scenarios and then evaluate the results in section IV. We conclude with a summary in section V.

II. Load Metric

A. Packet Transport Time Cost

The total Transport Process Time Cost (TPTC) of a single MAC Protocol Data Unit (MPDU) on the physical layer from the sender point of view can be divided into channel access and channel utilization time cost. The former describes the time needed to access the WM and it is dependent only on the medium Access Method (AM). The latter describes the duration of channel occupancy caused by the packet

transmission and it is dependent on the MPDU size (z_{MAC}), TxBR (r_b) and the used AM. Fig. 1 illustrates these times for DCF mode. However, the TPTC of a packet (p) can be declared as a function of these parameters, as described in (1). This equation is derived from the functions (TXTIME) of the IEEE standards 802.11a/b/g, assuming that the function "ceiling" from the function (TXTIME) is eliminated to simplify the calculation, and the control frames are transmitted at the same TxBR of the MPDU data frame:

$$TPTC(p) = t_{DT}(p) + t_{AM}(p) = a + \frac{b + z_{MAC}(p)}{r_b(p)} + t_{AM}(p) \quad (1)$$

Where: "a" and "b" are constant values in relation to stated standard and AM. "a" is expressed in µs, whereas "b" in bits. Furthermore, "b" can be declared for any layer protocol of the model ISO OSI. For instance, b_{MAC} and b_{IP} below refer to MAC-based and IP-based b-constant, respectively. Table I shows the values of these parameters under different conditions and for all standards. The time "t_{DT}" includes the data packet transport time, the channel occupancy time caused by control frames (ACK, RTS, etc.), and inter-frame spaces needed during the transport process, as long as the WM is not accessible during these inter-frame spaces (SIFS, DIFS, EIFS, etc.). "t_{DT}" is therefore considered as the WM-busy time (t_{MB}) of TPTC. The rest of TPTC is considered as t_{AM}, which is the time caused by AM that determines the nature of this time. DCF, t_{AM} includes the backoff time, and therefore considered as the WM-free time of TPTC; this time describes the time needed to access the WM as well as to complete the transport process, without causing channel occupancy. Furthermore, the WM-free time of TPTC is calculated from the sender point of view; therefore it does not affect the WM status calculation. In case of PCF is on the contrary, t_{AM} forms a part of t_{MB} and it must be calculated from the AP's point of view [1].

In the rest of this paper, we will discuss the WM from DCF perspective, unless explicitly mentioned to the contrary.

B. *Channel Utilization Efficiency*

The WM status at each time point can therefore be considered as either busy or free. The next equation (2) describes the WM-busy time (MBT) during determined time interval (T) from the AP's (WM) point of view.

$$MBT = \sum_{j=1}^{N} t_{MB}(p_j),\ MFT = T - MBT \quad (2)$$

Where p_j is the *j*th packet from all packets (N) transmitted during "T" to and from the AP, as a central node in the cell. MBT therefore indicates the AP load over T. On the other hand, the WM-free time (MFT) during T can practically not be measured because of the effect of the random backoff times as well as the random unused time gaps during T. Instead, MFT is calculated according to MBT, as explained in (2). MFT is associated with QoS of the channel during T.

The Channel Utilization Efficiency (CUE) of WM over T is calculated according to WM-busy time:

$$CUE = MBT / T,\ 0 \le CUE \le 1 \quad (3)$$

Typically, the smaller the observation time T, the smaller is the effect of the distribution of free times on the actual CUE. However, to use MBT as an effective quantitative load metric for load balancing and QoS mechanisms in comparison to other used load metrics such as bandwidth, data traffic, etc, T is selected to be 1s. CUE over 1s is definitely greater than zero, because the periodical transmittance of beacon frames.

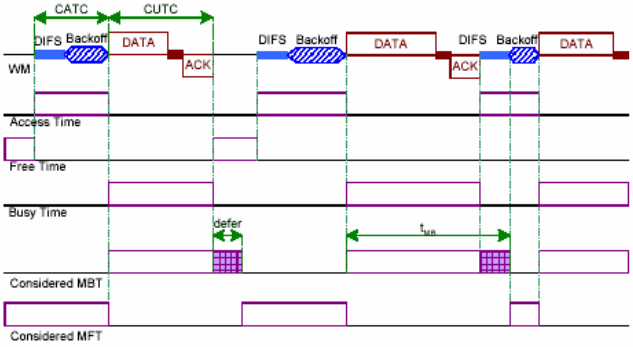

Fig. 1. Time intervals and WM-status according to DCF

TABLE I: "A" AND "B" VALUES RELATED TO IEEE STANDARD 802.11A/B/G AND ACCESS MECHANISM

	r_b [Mbps]	DCF*			DCF with CTS/RTS*			PCF*		
		a	b_{MAC}	b_{IPv4}	a	b_{MAC}	b_{IPv4}	a	b_{MAC}	b_{IPv4}
11a (OFDM)	6, 9, 12, 18, 24, 36, 48 and 54	94	428	612	162	744	928	76	588	772
11 (DSSS)	1 and 2	444	384	568	848	656	840	404	544	728
11b (DSSS)	5.5 and 11									
11g (DSSS)	1, 2	444	384	568	848	656	840	404	544	728
11g (CCK)	5.5, 11									
11g (ERP-PBCC)	5.5 and 11									
	22									
	33	446			852			406		
11g (ERP-OFDM)	6, 9, 12, 18, 24, 36, 48 and 54	112	428	612	186	744	928	84	588	772
11g (DSSS-OFDM)	6, 9, 12, 18, 24, 36, 48 and 54	480	428	612	920	744	928	440	588	772

*"a" is calculated by using the long PLCPPreamble, and "b_{MAC} and b_{IPv4}" by assuming that PBCC=0.

The WM saturation occurs when MFT has been reduced to zero, i.e. the maximum CUE (CUE=1). This case is practically impossible, because achieving this requires strong competition that increases the collision rate and thus the probability of all backoff times over T to be zero is really very small and it can be eliminated [1].

C. *Medium Busy Time as Load Metric*

Several applications, particularly the real time applications, need to transmit their packets in the WM at a constant rate per second. In a VoIP session, for example, using 20 ms frame size requires a constant packet rate of 50 pps to keep the connectivity and generally a good speech quality. If STAs using this kind of applications are known and also their transmission bit rate, then the necessary expected WM-busy time (EMBT) to be generated by the associated AP can be theoretically calculated through (2). On the contrary to (2), EMBT in a cell can be more than T, and therefore the expected CUE is greater than one. As a result, the related AP using best-effort QoS will be definitely overloaded in the reality, if this load is applied, since the expected EMBT is more than what the AP can transfer over T. Thus, EMBT is a paramount parameter to estimate the real QoS in the cell according to the expected traffic, and thus to enhance an

admission control mechanism. However, the guarantee of QoS in a cell requires that the EMBT must be clearly less than T.

The MBT calculation in (2) must consider the packet loss, bit error rate, and the retransmission rate that increase the MBT, whereas they decrease the throughput. In accordance to the relationship between RSS and TxBR, it could be noted that the lower the TxBR, the higher the bit error rate and the retransmission rate on the stated link. This effect must be considered by evaluating the AP load according to MBT. Therefore, MBT provides essential information about the AP load and its critical limits such as the overload case. On the contrary, using the throughput or traffic level to evaluate the AP load does not introduce any possibility to recognize the actual load limits of the AP, especially when STAs are transmitting at different bit rates. MBT and CUE can be measured at the AP directly by observing the transferred traffic. However, calculating the effect of packet loss and retransmission rate on the CUE can be achieved more precisely on the sender side (STA or AP) [5] [7].

III. LOAD MODEL

LBM deals only with the load of APs, and aims to distribute this load as equal as possible among APs. The distribution process is carried out through handing off one or more STAs, which are placed in an overlapped coverage area, from heavily loaded APs to lightly loaded APs. On the other hand, RRAC as a kind of Integrated Service (IntServ) QoS deals with the load of flows generated from network applications that run especially on STAs. RRAC aims to allocate certain resources to each flow, and thus reserving this allocation as long as the flow requests. Besides reservation, RRAC avoids overloading of the APs, since it continuously observes the reserved and available resources and it can therefore control the admission process appropriately. In this sense, LBM and QoS complement each other in order to provide better performance to the WLAN and its real time applications.

A. Model Architecture

The different nature of LBM and RRAC requirements leads to different possibilities of the architecture model. However, to reduce the design complexity, an architecture integrating both services must be used.

Since the handoff is a paramount process for load distribution and thus for load control model, the load model of LBM can be accordingly classified into STA-assisted and STA-controlled.

In the STA-assisted model, STAs send a measurement report including RSS, bit rate and neighbored APs to the Load Control Manager (LCG), which decides the load distribution. LCG is a component of the load decision model. A STA-assisted model describes a client/server architecture, which comprises the centralized, distributed and hybrid approach. The centralized approach implies STA/server transactions, where LCG run on a server placed in the distribution system. The distributed approach describes STA/AP transactions, where an LCG runs on every AP to manage the load locally, but it must cooperate with other LCGs in order to achieve load balancing. The hybrid approach comprises both approaches, centralized and distributed, where STA/AP/Server transactions are implied. Besides the server, each AP also runs an LCG. The APs communicate with each other through the server.

In the STA-controlled model, STAs conduct the initiation and control of load distribution, where the STAs decide by themselves when to hand off to which AP. The standard IEEE 802.11 uses this load model for LBM, where only the RSS is used to decide the handoff. Another approach implies STA/AP interaction, where the APs transmit their load level over beacon frames, and STAs therefore use this load level beside RSS to achieve effective handoff.

Resource reservation and admission control requires client/server architecture to achieve their functionalities. Similar to STA-assisted model of LBM, the load model of RRAC can also be classified into centralized, distributed and hybrid. The LCG in this model is additionally responsible for resource reservations and access admission control.

As a result, the client/server architecture model must make a trade-offs between the LBM and RRAC requirements. This architecture model requires communication between the model components (STAs and APs/server). Moreover, IEEE 802.11 does not define any method or mechanism for this task, although 802.11f can be used to carry out the transactions between APs in distributed mode. For prior study [3], however we have developed a protocol called Intelligent Management of Cells Access (IMCA) special for performing communication between STAs, APs and servers according to LBM and QoS requirements. Since then we have enhanced the previous version of IMCA with new features and functions. IMCA is a bit-based protocol and can be carried over MAC as well as UDP or even TCP over IP.

In this study we select the distributed model, where more flexibility and efficiency could be achieved in comparison to the centralized approach. To adapt the IMCA architecture to the distributed model, an IMCA-server (I-server) runs on each AP. Each I-server contains an LCG that manages both LBM and RRAC server-modules of the corresponded AP. However, the I-servers communicate with each other through the distribution system. The IMCA-client (I-client) performs the functionalities of both LBM and RRAC client-modules. In this model, two kinds of intercommunication are considered: STA-to-AP in the WM, and AP-to-AP in the distribution system. Therefore, to simplify the implementation we use IMCA over UDP/IP to carry out the LBM and RRAC functionalities.

B. Load Balancing Mechanism

In this model, STAs are classified according to their communication activity to passive and active. However, according to the coverage of APs, a STA can be local, when STA is covered only by the associated AP, and shared, when STA is covered by at least one AP besides the associated AP. From the AP's point of view, a STA can be associated, or covered when STA is associated with another covering AP. Consequently, the MBT from (2) can be expressed as function of the MBT (T_{BL}) generated by local STAs and MBT (T_{BS}) generated by shared STAs, as shown in (4).

$$MBT = T_{BL} + T_{BS} \ , \ f_d = T_{BS} / MBT \ , \ 0 \le f_d \le 1 \quad (4)$$

Where f_d is the load reducing factor that defines the AP capability to decrease its load, e.g. f_d=0.2 refers to 20% maximal possible reduction of the current AP load. However, if f_d=0, no load distribution is possible, because all associated STAs are local. In this case no LBM is affected.

The client-module of LBM runs on every STA and regularly monitors the current association and the surrounding APs. The association monitoring includes RSS and bit rate basically, whereas the monitoring of surrounding APs implies obtaining an AP-list of maximal five neighbored APs and their RSS and bit rates. Then, periodically or after a determined or critical change of one or more of the monitored parameters, the client-module sends a STA Status Message including information about the current monitored parameters or their changes to the server-module on the associated AP. The STA status messages enable each AP to recognize its neighbouring APs.

The server-module runs on every AP and monitors its load permanently. The server-module broadcasts or multicasts an AP Status Message periodically or after a critical load change to neighbored APs. An ASM contains information about the current load of the AP; other information can also be included. When the load decision model decides that one or more STAs have to handoff to another AP, it sends MOVETO Message to the selected AP to assure the handoff process. If the answer is positive, server-model sends HANDOFF-Messages to the selected STAs, which response with forced handoff [3].

The actions including STA status and AP status message belong to the load observing model, whereas the other actions including MOVETO and HANDOFF message belongs to the load control model. It is worth noting that the 802.11k standard [9] already defines some similar operations. Such operations can also be used in this model.

C. Resource Reservation and Admission Control

Due to the fact that collision is inevitable for contention-based channel access; thus, the guarantee for definite resource reservation for any flow is impossible. Therefore, this model does not offer guaranteed service, but controlled service, or soft QoS [8]. We define two types of resource reservation in this model: per-application and per-STA reservation.

Per-application reservation includes per flow and per-session reservation, where flow is unidirectional. By the initiation of per-application reservation, the client-module specifies a minimum and maximum bound for each application on the expected MBT [$emtb_{min}$, $emtb_{max}$] to be generated and sends this specification via Medium Access Message to the server-module that decides the acceptance of this reservation. After ending the application, the client-module informs the server-module via Reservation Free Message to free the reserved resources. When the resource requirements of a flow or session are changed in the runtime as a result of the variation of WM specifications such as bit rate or packet loss, the client-module must assure the reservation via Reservation Change Message by the server-module again in order to avoid disrupting other flows or sessions.

Per-STA reservation forms a kind of Channel Utilization Balancing (CUB), where the resources are reserved according to the active associated STAs. The server-module specifies for each STA a minimum and maximum bound on the expected MBT [$emtb_{min}$, $emtb_{max}$] to be generated per time interval (T) and sends this via Resource Allocation Message to the client-module that decides the acceptance of this assignment. The client-module distributes the assigned resources on the STA's flows according to specific priority rules. The generated MBT of the STA's flows must not exceed the allocated value per T. This reservation mode shifts the QoS-intelligence from the node (server/AP) to the STA/user.

This model distinguishes between two scopes of RRAC: Basis Service Set (BSS) scope and Extended Service Set (ESS) scope. In the BSS-scope, the RRAC is controlled by the server-module independent of the resource availability of other APs. This may degrade the QoS, when receiver and sender are associated with different APs. The ESS-scope solves this problem; it however requires the cooperation between APs to build a kind of virtual tunnel between STAs. The ESS-scope increases the complexity of this model, but it provides more efficiency for QoS.

IV. Measurements and Results

Our first experiments have concentrated on the importance of MBT and EMBT for LBM and RRAC according to our load model.

A. Experiments Description

For the experiments, we have used a special WLAN environment, which is based on the standard IEEE 802.11b in DCF mode, and consisted of four notebooks, one AP and three STAs (S1, S2 and S3). The AP was built upon the software "Hostap" (http://hostap.epitest.fi) and the WLAN card NetGear MA401. Linux Fedora Core 1 was used as platform for this AP. The STAs were supplied with MS-Windows XP and the WLAN cards RoamAbout from Enterasys. For traffic generation and monitoring on IP-level, the software MGEN (http://mgen.pf.itd.nrl.navy.mil) was used. However, to monitor traffic on MAC-level in relation to the number of transferred packets and data size for each bit rate, we had to modify the Hostap driver software in order to obtain the needed statistics. We have built two scenarios: in the first scenario, we have applied a variable packet rate (VPR) over a constant bit rate (CBR), and in the second scenario a constant packet rate (CPR) over a variable bit rate (VBR). The applied traffic on each link between the AP and a STA is described as a product ($n_f \times r_p \times z_{UDP}$), where n_f, r_p, and z_{UDP} refer to number of flows (half-/full-duplex), packet rate per second, and UDP payload size, respectively, as shown in Fig.2a for scenario-1 and in Fig. 2b for scenario-2.

B. Results and Discussion

Scenarion-1 presents a special case, where all STAs are transmitting at 11 Mbps. This scenario shows the effect of

MFT-gaps distribution and access collisions on the whole CUE, as shown in Fig. 2c at the time ranges "61-181" and "241-361", where the average CUE_{max}=92% in both overload cases. In LBM model, the value (MBT=0.92 T) in this case describes the upper limit of load, which the AP should not reach. In RRAC model, EMBT must be smaller than 1s, and for guaranteed QoS it must be EMBT≤0.92 T, which forms an important factor for the admission control.

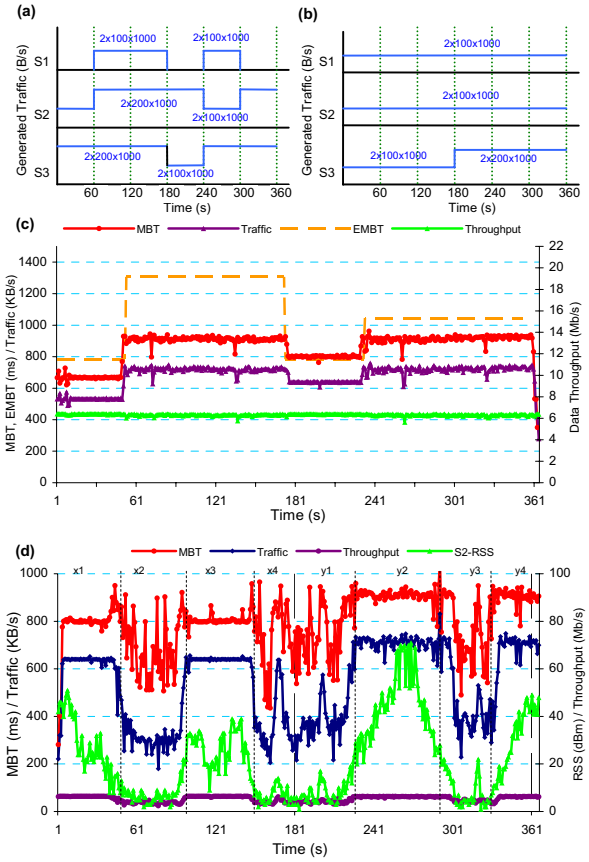

Fig. 2 MBT vs. Traffic variations for scenario 1 and 2

Scenario-2 presents the effects of VBR under low RSS on the CUE variation. S2 is used as a mobile STA to generate a VBR, which varies between 1 and 11 Mbps according to the RSS values variation between 3 and 50 dBm. The experiment for this scenario is divided into two phases: "x1 to x4" and "y1 to y4", as shown in Fig. 2d. The applied traffic was so selected that the AP in the first phase will be overloaded only if S2 will transmit at 1 or 2 Mbps, whereas the AP, in the second phase, is overloaded whatever the TxBR of S2 was. S2 transmits at different bit rates, at the sectors x2, x4, y1 and y3, particularly at 1 and 2 Mbps, according to the low RSS values, as illustrated in Fig. 2d. These sectors clarify the effects of S2 on the entire CUE, because the lower the RSS, the higher the packet loss and thus the retransmission rate, which costs relatively much time. This time in both scenarios is not measured on the AP, since the used software calculates the MBT according to the successfully transferred packets without checking the WM status, which explains the large CUE's variation (0.45 - 0.98) in these sectors. Measuring the MBT according to the WM status determines the real MBT more precisely, which will cause the registered MBT in the overloaded sectors of scenario-2 to be smoother. Thus, CUE can be used as an effective quantitative factor to determine the upper limits of AP's load independent of the TxBR variation and retransmission rate. On one hand, the traffic/throughput as load metric can provide similar information about the AP's load in comparison to MBT, when STAs and AP will transmit at CBR, which is the case of scenario-1. On the other hand, the variation of TxBR and the packet loss of S2 in scenario-2 causes the information obtained from traffic/throughput during the overload sectors to be insufficient to determine the AP's overload cases, where no upper limit can be obviously defined in this case. Consequently, since MBT expresses the real WM utilization, MBT describes the real load of this WM more effective than both traffic and throughput. However, these relatively large variations of packet rate achieved by S2 lead to the fact that per-application reservation cannot be guaranteed on the AP as long as the STA itself cannot control its packet loss under low RSS, as shown in Fig. 2d. The per-STA reservation avoids this problem by distributing the available part of the assigned resources on the locally prioritized flows.

V. Conclusion

In this paper we introduced a load observation and control model for LBM with QoS, and presented the MBT as a load metric for this model. We briefly discussed the calculation of MBT for IEEE 802.11a/b/g, and explained why the MBT can describe the AP load more precisely than other load metrics. The distributed architecture is used for this model, where each AP manages its load locally, and communicates with other APs to achieve the load balancing as well as the ESS-scope resource reservation. Our measurements confirmed the importance of MBT to detect the overload cases of APs and thus to response as fast as possible against the critical load levels. Also, the measurements showed the significance of per-STA reservation for QoS in order to avoid the effects of bit rate variations and packet loss on per-application reservation.

References

[1] IEEE Standard 802.11 (08/1999), Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications.

[2] Victor Aleo: „Load Distribution in IEEE 802.11 Cells", KTH, Royal Institute of Technology. Stockholm, March 2003.

[3] R. Daher, H. Kopp and D. Tavangarian: "Active Load Balancing in Wireless LAN hotspots", GI 2004, September 2004

[4] A. Balachandran, P. Bahl, G. M. Voelker: „Hot-Spot Congestion Relief in Public-area Wireless Networks", WMCSA, June 2002

[5] Ilenia Tinnirello and Giuseppe Bianchi:"A simulation study of load balancing algorithms in cellular packet networks", MSWiM, 2001

[6] Reinhard Riedl and Lutz Richter:"Classification of Load Distribution Algorithms", IEEE PDP, 1996

[7] Yigal Bejerano, Seung-Jae Han and Li (Erran) Li:"Fairness and Load Balancing in Wireless LANs Using Association Control", MobiCom'04, Philadelphia, Pennsylvania, USA. 2004

[8] Ming Li, B. Prabhakaran, Sathish Sathyamurthy: "On Flow Reservation and Admission Control for Distributed Scheduling Strategies in IEEE802.11 Wireless LAN", MSWiM'03, 2003.

[9] IEEE 802.11k/D0.4, Draft Supplement to Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: IEEE 802.11 WG, July 2003