Frame Aggregation in Central Controlled 802.11 WLANs: the Latency vs. Throughput Trade-off

Jose Saldana, *Member, IEEE,* José Ruiz-Mas, and José Almodóvar

Abstract—Frame aggregation is widely used in 802.11 WLANs in order to provide a significant throughput improvement. However, the latency increase that comes as a counterpart can reduce the quality experienced by the users of applications with real-time constraints. This letter explores the throughput vs. latency trade-off in the context of central controlled solutions (e.g. SDWN-based). First, a scenario with a single Access Point is used to illustrate the problem, and to propose two possible solutions. Then, a centralized algorithm that dynamically (de)activates aggregation is tested in a scenario with a number of APs. The results show that aggregation parameters can be tuned in order to keep latency in low levels, with a low throughput penalty.

Index Terms—802.11, Wireless LAN, frame aggregation, real-time services

I. INTRODUCTION

IN the last years, IEEE 802.11 is being massively deployed in home scenarios, connecting the user devices to the home gateway. In many cases, a number of Access Points (AP) are installed in order to improve the coverage in the house. 802.11 hotspots are also widely used in airports, malls, business centers, campuses, etc. Different vendors have developed WLAN solutions based on a coordinated set of APs, bringing coverage to an area of interest. In addition, new SDWN solutions in which a set of APs is coordinated by a central controller have been proposed [1], [2], bringing fine-grain control of the network parameters, thus providing smart radio resource management features as dynamic channel allocation, load balancing, power control and seamless handover.

Services of different nature may coexist in these scenarios: while some of the users are downloading a video, browsing the web or using an app, others may be running services with real-time constraints as e.g. VoIP, video conferencing or online games (many titles have been ported to smartphones or tablets, and some others have been designed for them). Therefore, a distinction between two kinds of services can be established: on the one hand, bandwidth-demanding services that want to get as much throughput as possible; on the other hand, low-latency services, which may use very little bandwidth.

In order to meet the growing traffic demand in these scenarios, a set of improvements are being included in the different 802.11 amendments. One of them is frame aggregation, which was first introduced in 802.11n as a way to improve the

J. Saldana and J. Ruiz-Mas are with CeNITEQ Group - Aragon Institute of Engineering Research (I3A), EINA, University of Zaragoza, Spain. Email: {jsaldana, jruiz}@unizar.es. J. Almodóvar is with Networks Department, TNO, The Netherlands. Email: jose.almodovarchico@tno.nl

efficiency at MAC level. In normal operation, media access control is used before sending each frame. When frame aggregation is employed, a large frame can be assembled by concatenating multiple small ones: therefore, a number of sub-frames can be sent together, i.e. the channel sensing and the exponential backoff algorithm will only be run once.

Two frame aggregation mechanisms are considered in the standard: AMSDU (Aggregated MAC Service Data Unit aggregation) and AMPDU (Aggregated MAC Protocol Data Unit aggregation). In the former, a single FCS (Frame Check Sequence) is included, which protects the integrity of the whole multi-frame. Therefore, if some bits are corrupted, all the information has to be retransmitted. In contrast, in AMPDU, each sub-frame carries its own FCS, which permits individual retransmissions of the corrupted sub-frames. Therefore, the latter method has become more successful: in fact, in 802.11ac, every frame must have an AMPDU format, and very big frames (up to 1 MB) can be sent.

Frame aggregation can provide significant improvements in terms of efficiency: in one of the first works on the topic [3], simulations showed a significant efficiency increase, and a better performance of AMPDU, which was emphasized when the packet-error rate was high. In [4], an optimization method was proposed to dynamically adjust the number of AMPDUs, according to the sub-frame size, the maximum aggregation level allowed and the instantaneous channel bit-error-rate.

However, as a counterpart of the throughput increase, when a wireless station (STA) gains access to the medium and transmits a long frame including a number of sub-frames, the rest of the STAs connected to the AP have to wait until the end of the transmission. This delay can be significant, especially if the STA is far from the AP and it is transmitting at a low rate. This results in delay, and also in jitter, as some frames will wait more time than others. All in all, a trade-off appears: frame aggregation is positive for bandwidth-demanding services, but it can negatively affect services with real-time constraints.

The contribution of the present paper is the detailed study of this trade-off, and the proposal of three solutions that can be employed (separately or jointly), in order to achieve a good balance: *a)* prioritization [5] for real-time services; *b)* setting a limit on the AMPDU size; *c)* employing a real-time central coordination of frame aggregation in all the APs, depending on the traffic demand, the presence of real-time flows, etc. To the best of our knowledge, this is the first study proposing a centralized coordination and control of 802.11 frame aggregation functionalities.

The next facts should be taken into account in order to illustrate the convenience of developing solutions for this problem: *a)* the growing popularity of SDWN solutions in which the

TABLE I
SIMULATION PARAMETERS

ns3 version	ns-3.26	UDP pkt size	60 bytes
Mobility model	Random Waypoint	UDP rate	50 pps
Walking speed	1.5 m/s with pause time 2 s	TCP pkt size	1,500 bytes
Channel model	YANS	TCP variant	New Reno
Channels	36 to 128 (20 MHz channels)	Simulation time	300 s (600 s for Fig. 2)
WiFi phy std	80211n_5GHZ	RTS/CTS	Not enabled
WiFi rate control model	Idealwifi Manager	Inter-AP distance	100 m (30 m to the border)
Propagation model	Friis loss model Constant speed	Short guard	Not enabled

wireless parameters of a number of APs can be coordinated by a central controller [1], [2]; *b)* the high availability of devices capable of frame aggregation, i.e. implementing 802.11n and subsequent versions; *c)* the extensive use of services with real-time constraints in these scenarios.

II. TESTS AND RESULTS

In this section we will first present the trade-off under study, and the results obtained when applying different solutions. All the results have been obtained with ns-3 simulator[1]. The simulation parameters are detailed in Table I.

A. The Throughput vs. Latency Trade-off

The trade-off under study is illustrated in Fig. 1: a single AP is shared by $2N$ users: N are using VoIP (simulated with UDP upload flows) and N are downloading a file with TCP. The users move at walking speed, following a Random Waypoint Model in a 60x60 meters square.

It can be observed that the throughput obtained by the TCP users is much higher if aggregation is employed, as it could be expected [3]. However, as a counterpart, a meaning increase of the one-way delay is observed, which can rise up to 25 ms. This delay is caused by the waiting of the VoIP packets while the TCP AMPDUs are transmitted. It can be considered significant if we take into account the requirements of next generation networks, which *"will need to be able to support a roundtrip latency of about 1 ms"* [6]. Obviously, adding a big amount of latency in the first hop of the communication is not desirable.

B. Prioritization

EDCA (Enhanced Distributed Channel Access) prioritization, introduced by IEEE 802.11e, is based on the use of four queues with different values of the contention window. It relies on the Differentiated Service CodePoint (DSCP) value in the Differentiated Service (DS) field of the IP header.

Fig. 1. A single AP shared by $2N$ users: a) total throughput of the N TCP users; b) average one-way latency experienced by the N VoIP users. 95% confidence intervals are included. The packet loss rate for VoIP (not shown) is below 0.2%.

Fig. 2. A single AP shared by a number of users, using EDCA priorities: a) total throughput of the TCP users; b) average one-way latency experienced by the VoIP users. The packet loss rate for VoIP is below 0.1% in all cases.

---

[1]The ns3 source code developed for the simulations, and the scripts for obtaining each graph, are available at https://github.com/Wi5/ns3_802.11_frame_aggregation
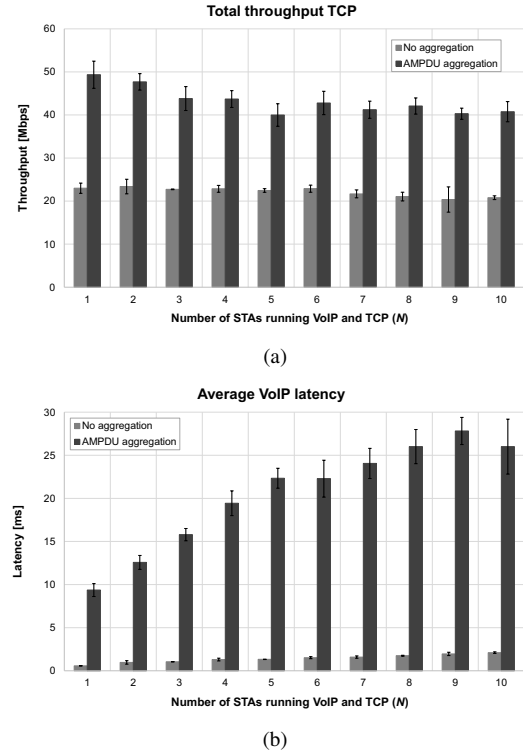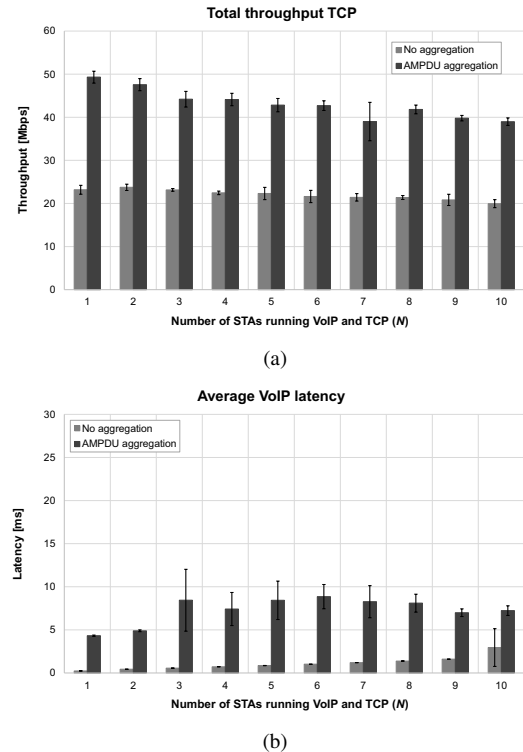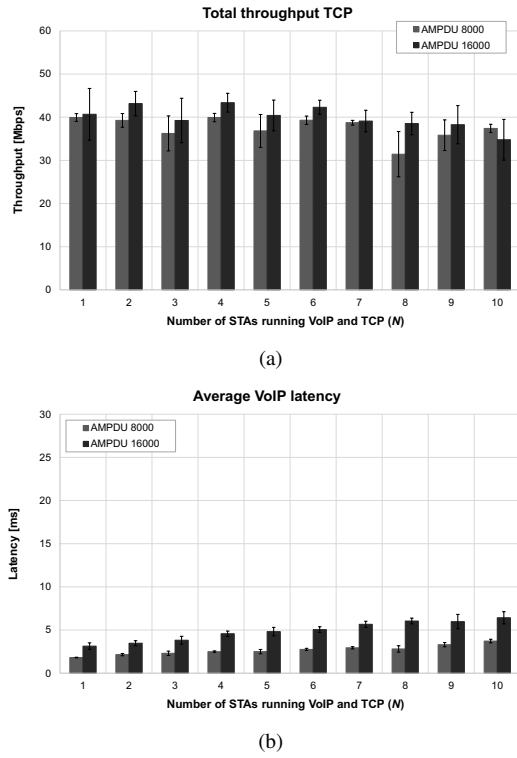
(a)

(b)

Fig. 3. A single AP shared by a number of users, using a limit of 8 kB and 16 kB for the size of the AMPDU: a) total throughput of the TCP users; b) average one-way latency experienced by the VoIP users. The packet loss rate for VoIP is below 0.4% in all cases.

In our simulation scenario, we have assigned *Voice* (VO) priority to the VoIP flows and *Best Effort* (BE) to the TCP ones, obtaining the results shown in Fig. 2. It can be observed that the results in terms of throughput of the TCP flows are quite similar to the previous ones (see Fig. 1(a)). At the same time, the latency gets reduced in both cases, as the VoIP packets have a higher priority: if aggregation is not enabled, the latency is again low; if it is enabled, the latency is smaller than before, especially when the number of flows grows. However, its value is significant (8-9 ms) in some cases. Another limitation of this solution is that it will only work if the required value of the DS field is set, which not always happens in real networks.

C. Setting a limit on the AMPDU size

Another possibility is to set a limit for the maximum AMPDU size, instead of using the default limit of 65,535 bytes defined in 802.11n. In our case, two different values have been used: 8 kB, which allows AMPDUs including 5 packets of 1,500 bytes; and 16 kB, which allows the inclusion of 10 packets into a single AMPDU. The results are shown in Fig. 3.

If we compare Fig. 3(b) with Fig. 1(b), a significant reduction of the latency experienced by the VoIP applications can be observed, which is now always below 7 ms. The counterpart is a reduction of the throughput, which is between 80 and 95% of the amount obtained with the default limit (compare Fig. 3(a) and Fig. 1(a), *AMPDU Aggregation* bars). It can be seen that the higher the maximum size of the AMPDU, the better the throughput, and the worse the latency.
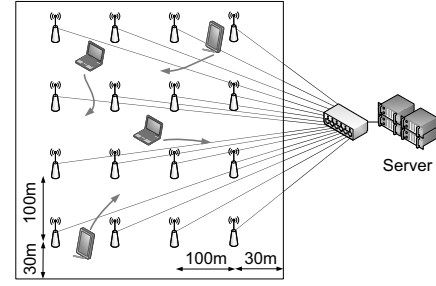

Fig. 4. Test Scenario with 16 APs.

This parameter can be tuned in order to modify the tradeoff: if a higher delay can be tolerated by the applications, we can obtain a better throughput. However, the difference in throughput is smaller than the difference in latency, i.e. with a throughput reduction between 10-20%, the latency can be reduced by a factor between 4 and 8. The cause of this fact is the asymptotic behavior of the efficiency increase when aggregating (see e.g. the results presented in [3]): there is a moment where the increase of the number of aggregated frames only provides a small throughput improvement, but the added delay grows linearly.

Finally, it should be remarked that this solution presents an advantage with respect to prioritization: it is easier to deploy, i.e. it just requires an adjustment of the AMPDU maximum size parameter in the APs.

D. Central coordination

The solutions already presented (prioritization and setting a limit on the AMPDU size) have been illustrated using a scenario with a single AP. However, in order to test the central coordination of aggregation, a bigger scenario including 16 APs (in a 4x4 array) has been implemented in ns3 (see Fig. 4). As it happens in [1], [2], we assume that the APs are coordinated by a central controller (e.g. based on SDWN) able to tune different network parameters.

As before, $2N$ users walk through the scenario: N are using VoIP and N are downloading a file with TCP. Each AP uses a different channel, and a fast handover mechanism has been included in the simulation environment: as soon as a STA gets disconnected, it looks for the closest AP and requests association. EDCA priorities are used.

The central controller runs an algorithm that dynamically disables aggregation in an AP if a VoIP flow appears, and enables it when the VoIP user leaves. Therefore, VoIP users will always *see* a non-aggregating AP, whereas TCP users will receive non-aggregated frames in some moments, thus experiencing some throughput penalty.

It should be noted that some knowledge about the nature of the flows present in the network is required in order to run this algorithm. Although this does not represent a problem in the simulations (all the information is available), in a real network it would require the use of a traffic detection mechanism. This is nowadays feasible, as solutions based on machine learning can detect the presence of real-time flows just using the packet size and inter-packet time [7], thus maintaining user's privacy.
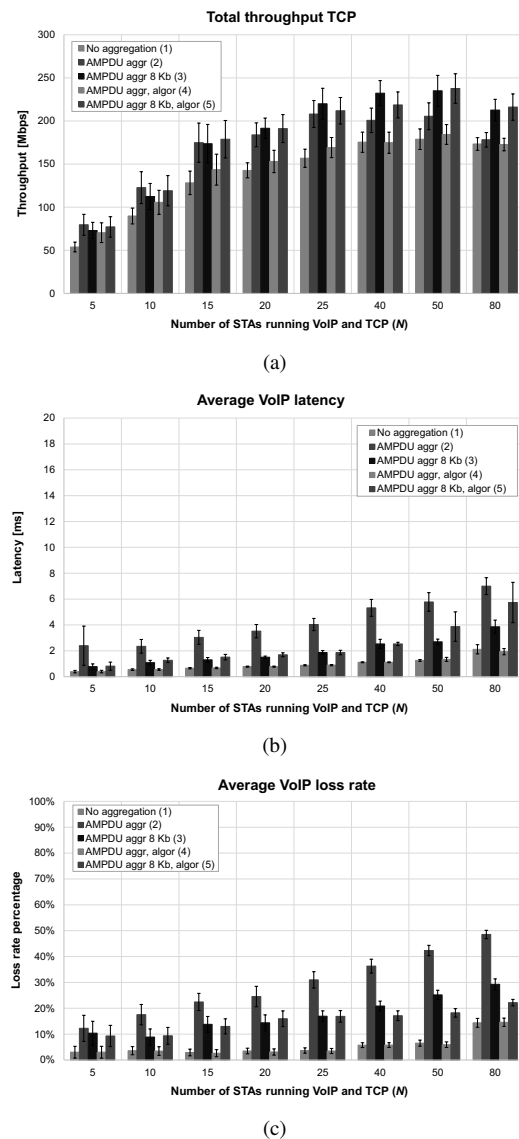
Fig. 5. A network including 16 APs shared by a number of users, using different options for controlling the size of the AMPDU: a) total throughput of the TCP users; b) average one-way latency experienced by the VoIP users; c) average loss rate experienced by the VoIP users.

The results obtained in this scenario are shown in Fig. 5, where the *No aggregation* case (1) is compared with four different ones: (2) AMPDU aggregation is always active (with the default limit of 65 kB); (3) aggregation with a fixed limit of 8 kB; (4) aggregation, with the algorithm deactivating it when a VoIP STA is associated to an AP; and finally, in (5) the algorithm is used but, instead of deactivating the aggregation, a maximum size of 8 kB is set when a VoIP flow is present in an AP, in order to limit the added delay.

It can be observed that *No aggregation* (1) is always the worst option in terms of throughput, but it maintains a low value of latency and packet loss. Regarding AMPDU aggregation (2), the throughput is improved but at the cost of an increased latency and packet loss rate. If the AMPDU limit is set to 8 kB (3), the latency gets reduced. In addition, although the throughput gets reduced when the number of users is low, it even gets improved for 20 and more users.

This can be caused by the difference in terms of packet loss rate.

If the algorithm controlling aggregation is running (4), a throughput penalty between 3 and 19% with respect to (2) is translated into very low values of latency (a reduction factor between 3 and 6) and packet loss (factor between 3 and 9). Finally, the combination of the algorithm and the limit of 8 kB (5) gives very similar results to (3): the throughput penalty is minimum, and the VoIP users will experience a similar delay and packet loss rate.

The proposed algorithms can be useful in order to maintain the VoIP QoE [8] in acceptable values, as delay and packet loss can be kept low, thus providing a good Mean Opinion Score.

The decision of which solution is better on each case will first be determined by the availability of information about the services present on each AP. Nevertheless, the network manager can always tune the maximum AMPDU size, in order to limit the maximum delay and packet loss rate that the users will experience. All in all, a penalty in terms of throughput can be expected as a counterpart of the improvement of the quality of real-time services, but this trade-off can be tuned in order to get a good balance between the users of both services.

III. CONCLUSION

This paper has studied the trade-off between the throughput improvement and the extra latency that come as a consequence of frame aggregation in 802.11 WLANs. Three solutions, based on prioritization, limiting the maximum AMPDU size and using a central coordination of the aggregation have been studied and discussed. A tradeoff appears: if the maximum delay of real-time services has to be reduced, a penalty in the throughput of TCP connections will appear. As future work, smarter algorithms can be developed in order to dynamically adapt the aggregation parameters as a function of the instantaneous values of the latency. In addition, the influence of other options could be explored, as e.g. RTS/CTS and 802.11ac.

REFERENCES

[1] J. Schulz-Zander, L. Suresh, N. Sarrar, A. Feldmann, T. Hühn, R. Merz, "Programmatic orchestration of wifi networks," in USENIX Annual Technical Conference (USENIX ATC 14), pp. 347-358, Jun 2014.

[2] L. Sequeira, J. L. de la Cruz, J. Ruiz-Mas, J. Saldana, J. Fernandez-Navajas, J. L. Almodovar, "Building a SDN Enterprise WLAN Based On Virtual APs," IEEE Communications Letters, vol.21, no.2, pp. 374-377, Feb. 2017.

[3] B. Ginzburg, A. Kesselman, "Performance analysis of A-MPDU and A-MSDU aggregation in IEEE 802.11n," Sarnoff Symposium, 2007 IEEE, vol., no., pp.1,5, April 30 2007-May 2 2007.

[4] J. Liu, M. Yao, Z. Qiu, "Enhanced Two-Level Frame Aggregation with Optimized Aggregation Level for IEEE 802.11n WLANs," in IEEE Communications Letters, vol.19, no.12, pp.2254-2257, Dec. 2015.

[5] R. Moraes, P. Portugal, F. Vasques, R.F. Custódio, "Assessment of the IEEE 802.11 e EDCA protocol limitations when dealing with real-time communication," EURASIP Journal on Wireless Communications and Networking, 2010, 29.

[6] J. G. Andrews *et al.*, "What Will 5G Be?," in IEEE Journal on Selected Areas in Communications, vol. 32, no. 6, pp. 1065-1082, June 2014.

[7] T. T. T. Nguyen, G. Armitage, P. Branch, S. Zander, "Timely and Continuous Machine-Learning-Based Classification for Interactive IP Traffic," in Networking, IEEE/ACM Transactions on, vol.20, no.6, pp.1880-1894, Dec. 2012.

[8] The E-model, a computational model for use in transmission planning, ITU-T Recommendation G.107, March 2003.