On Buffer Sizing for Voice in 802.11 WLANs

D. Malone, P. Clifford, D.J. Leith
Hamilton Institute, NUI Maynooth

***Abstract*— The use of 802.11 to transport delay sensitive traffic is becoming increasingly common. This raises the question of the tradeoff between buffering delay and loss in 802.11 networks. We find that there exists a sharp transition from the low-loss, low-delay regime to high-loss, high-delay operation. This transition determines the voice capacity of a WLAN and its location is largely insensitive to the buffer size used.**

I. Introduction

IEEE 802.11 technology has been enormously successful, with wireless 802.11a/b/g edge networks now very common. While data traffic (web, email, media downloads *etc*) currently constitutes the bulk of traffic in the Internet, voice applications are becoming increasingly important. Voice traffic differs fundamentally from data traffic in its sensitivity to delay and loss. This has led to substantial interest in ensuring appropriate quality of service (QoS) for voice traffic in mixed voice and data networks, including the development of the recent 802.11e standard specifically targeted at addressing QoS issues. However, the focus of published work has been largely on MAC design and operation to ensure appropriate prioritisation of delay-sensitive traffic. To our knowledge, almost no published work exists on the question of appropriate network buffer sizing for voice traffic in 802.11 WLANs.

In this paper we investigate buffer sizing for voice calls in 802.11 networks. Of course, there have been many simulation and modelling studies of 802.11 networks. While some of these studies have considered voice traffic (e.g. [1], [2], [3]), including some commenting upon the value of queueing voice separately from other traffic (e.g. [4]), to our knowledge the present paper is the first to address the question of network buffer sizing for voice traffic. At the application layer playout buffering has been considered for 802.11 (e.g. [5]), but this is a separate issue from network layer buffer sizing. In [6] it is observed that increased buffer sizing does not necessarily improve the performance of inelastic traffic.

II. Buffer sizing for voice

We consider an infrastructure mode WLAN where traffic is routed via an access point (AP). Following [7], we model a two-way voice call as a 64kbs on-off traffic stream with on and off periods exponentially distributed with mean 1.5s, subject to a minimum of 240ms. Traffic is between between a wireless client station and a device behind the AP. To account for the two-way correlated nature of voice conversations; the on/off periods of one half of a call correspond to the off/on periods of the other. We consider an 802.11b PHY with the following MAC parameters: 20 μs slot time, CW_{min} 32, DIFS 50 μs, SIFS 10 μs, long 192 μs preamble, 100 byte packets.

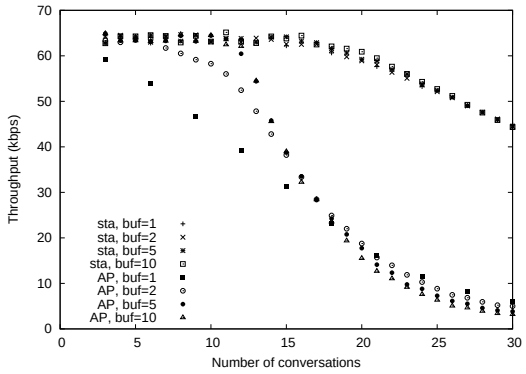

Fig. 1. Achieved throughput for AP/client voice with various buffer sizes as the no. of calls is increased. ns simulation.

Figure 1 shows the average throughput and loss per call as we increase the number of voice conversations (and so stations) in the network. Values are shown both for the aggregate client stations and the AP and results are given for buffer sizes of 1, 2, 5 and 10 packets with the buffer in the AP set to be the same size as in each of the stations. We can see immediately that the throughput achieved by the AP falls relative to that of the aggregate client stations as the number of calls is increased. This is perhaps unsurprising as the 802.11 MAC enforces per station fairness; that is, the client stations and the AP each win approximately the same number of transmission opportunities despite the fact that the AP carries n times as much traffic as each client station. The situation with on-off traffic such as voice is of course complicated by the fact that, firstly, voice traffic is relatively low rate and so need not make use of every available transmission opportunity awarded by the 802.11 MAC. Secondly, a voice conversation involves speakers approximately taking turns at talking. That is, traffic is between pairs of speakers with the on period of one speaker roughly corresponding to the off period of the other. Both of these features mitigate the contention between the wireless stations and the AP for access to the wireless channel. Hence, while a simple argument based on per station fairness would suggest that the AP throughput would be $1/(n+1)$ that of the aggregate client stations, it can be seen from Figure 1 that this is not the case[1]. This observation is not new and has been discussed elsewhere[2]. Previous work has not, however, considered the impact of buffer sizing on network behaviour.

We can see from Figure 1 that the choice of buffer size has

[1]We comment that AP throughput does scale as $1/(n+1)$ that of the client stations when the WLAN nodes are saturated, see for example [2].

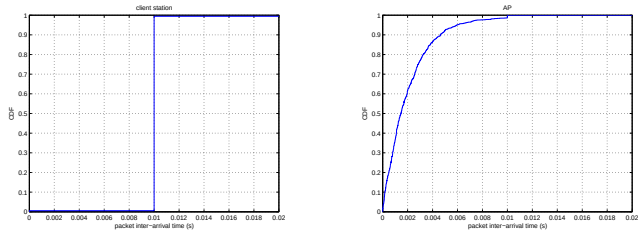

Fig. 2. Client and AP packet inter-arrival time cumulative distribution functions for the number of voice calls, $n = 10$.

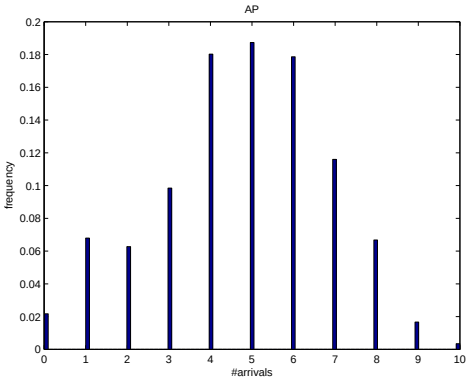

Fig. 3. Distribution of number of arrivals in a 10ms interval at an AP with 10 on/off calls in progress.

a strong impact on the throughput achieved by the AP. For very small buffers it can be seen that the AP throughput falls to around only 90% of the 64Kbs offered load from a voice call by the time two calls are active. This drop in throughput is already likely to be yield unacceptable quality of service; that is, to restrict the network voice call capacity to two calls or less. A buffer length of 10 packets improves AP throughput significantly out to about 10 conversations, thereby greatly increasing the network voice call capacity compared to the situation when very small buffers are used.

We can gain insight into this behaviour by considering the arrival processes at the client stations and AP in more detail. The arrival process at a client station consists of on-off 64Kbs traffic. During an on-period packets arrive at regular 10ms intervals; no packets are generated during an off-period. The measured cumulative distribution function of packet inter-arrival times is shown in Figure 2. Since the inter-arrival times are always greater than or equal to 10ms, stability is assured provided the mean service time at the network interface queue of a client station is less than 10ms. In contrast, the arrival process at the AP is the aggregate of n on-off arrival processes, corresponding to the n voice call halves. It may happen that packets from several calls arrive at the AP within a short time of each other and thus the inter-arrival times at the AP queue are not lower-bounded by 10ms. This is evident in the measured cumulative distribution of packet inter-arrival times shown in Figure 2. As a result, the queue sizing requirement

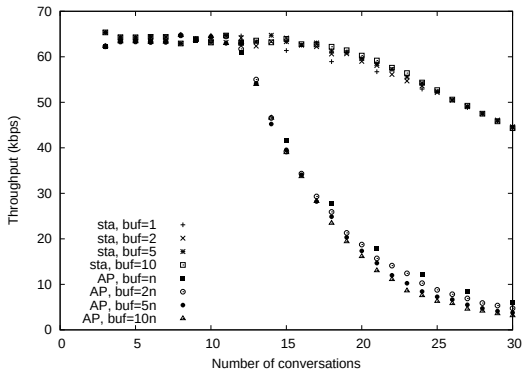

Fig. 4. Achieved throughput for AP/client voice with various buffer sizes as the number of conversations is increased. AP buffer scaled with number of calls.

at the AP differs from that at the clients.

We note, however, that since the inter-arrival times for each individual call are at least 10ms, the number of packets that can arrival at the AP during a 10ms interval is no more than n. This is a worst case bound and may occur only rarely. For example, Figure 3 shows the measured distribution of packet arrivals at the AP in a 10ms interval. This simple analysis suggests that the AP buffer size should be set equal to at least the number of calls n.

The impact of this change is demonstrated in Figure 4 where we scale the buffer of the AP to be n times the size of the stations' buffers. Figure 5 shows the corresponding delay. A number of observations can be seen immediately:

(1) The per call throughput is close to 64Kbs and the mean delay is below 10ms for up to 12 calls. For greater than 12 calls, the per call throughput falls rapidly — by 13 calls the throughput has fallen by more than 10% — and delay quickly rises. This is likely to yield an unacceptable call quality i.e. the voice call capacity of the network is therefore approximately 12 calls. This is in good agreement with a back-of-envelope calculation based on [8] which indicates a voice capacity upper limit (neglecting packet collisions and other contention overhead) of around 15 calls.

(2) There is an abrupt transition from the low-loss, low-delay regime to high-loss, high-delay operation. Below this transition, buffer sizing has little impact on throughput and delay for up to 12 calls. Above 12 calls, throughput falls below 90% of offered load for all sizes of buffer and delay rises rapidly. That is, the location of the transition is essentially independent of the level of buffer provisioning and thus network capacity is fixed at approximately 12 calls regardless of buffer size.

(3) In the high-loss, high-delay regime above 12 calls, the total delay depends strongly on buffer size. This is to be expected as in this unstable regime the buffer contains a standing queue that scales with buffer size.

(4) Smaller buffer sizes yield shorter delays in the low-throughput, high-delay regime with greater than 12 calls.

We note that across a wide range of situations including

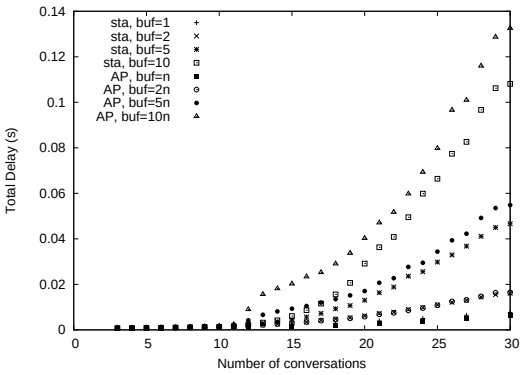

Fig. 5. Total delay (queueing delay plus MAC delay) as the no. of conversations is increased. AP buffer scaled with no. of calls.

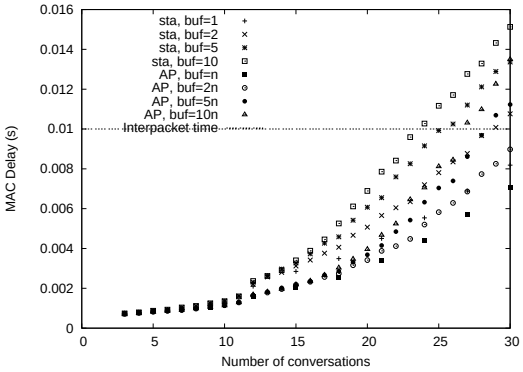

Fig. 6. MAC level delay with various buffer sizes as the number of conversations is increased.

peer-to-peer networks, infrastructure mode networks, plain 802.11, prioritised 802.11e, pure voice environments and mixed voice/data environments, we have observed similar behaviour and find that there consistently exists a sharp transition from low-loss, low-delay operation to a high-loss, high-delay regime. This transition determines the voice capacity of the network and its insensitivity to buffer sizing is surprising.

We can gain more insight into this behaviour by looking at the MAC component of the delay, see Figure 6. The MAC delay is the mean time the MAC layer spends transmitting a packet, including collisions, contention, transmission and acknowledgement. That is, the MAC delay is essentially the inter-packet service time of the network interface queue on a station. The MAC delays are naturally much shorter than the total delays. In this graph we can clearly see that increasing the buffer size results in an increase in the MAC delay when the network becomes congested[2].

The MAC delay is determined by (i) the 802.11 contention window value, which doubles each time a packet transmission fails due to a collision, (ii) the number of transmission failures due to collisions that occur before a transmission succeeds, and (iii) the length of time that the wireless channel is occupied by transmissions (countdown is halted when the channel is sensed busy). The only way that the buffer size can impact on MAC delay is by affecting one or more of these quantities. In fact, with on-off traffic such as voice, it can readily be seen that when a station's queue is not backlogged some transmission opportunities are inevitably not used as there is no packet available to send. However, once the queue becomes backlogged, the number of unused transmission opportunities must decrease. Consequently, both the frequency of packet collisions and the time that the channel is occupied by transmissions can be expected to increase. Thus, the inter-packet service time (MAC delay) of the network interface queues in the network increase as the queues becomes backlogged. Conversely, the queue backlog tends to increase as the inter-packet service time increases. Therefore the potential exists for a reinforcing feedback whereby the onset of queueing quickly leads to further queue buildup and instability. We note that this complex feedback loop coupling service rate and queueing leads to the buffer sizing task in 802.11 WLANs differing fundamentally from that in wired networks

[2]We can also see that the MAC delay increases approximately linearly as additional stations are added. This is because the 802.11 MAC mechanism is distributing the available transmission time approximately evenly among all stations.

III. Conclusions

We have considered the tradeoff between buffering and loss for voice traffic in 802.11 networks. We find that there exists a sharp transition from the low-loss, low-delay regime to high-loss, high-delay operation. This transition determines the voice capacity of a WLAN and its location is largely insensitive to the buffer size used. Interestingly, this observation indicates that recently proposed finite-load analytic models for 802.11 networks with small buffers [1] can be employed to accurately predict network capacity even when large buffers are used.

References

[1] K. Duffy, D. Malone, and D. Leith, "Modeling the 802.11 Distributed Coordination Function in non-saturated conditions," *IEEE Communications Letters*, vol. 9, no. 8, pp. 715–717, 2005.

[2] P. Clifford, K. Duffy, D. Leith, and D. Malone, "On improving voice capacity in 802.11 infrastructure networks," in *Proc. WIRELESSCOM*, 2005.

[3] N. Hegde, A. Proutiere, and J. Roberts, "Evaluating the voice capacity of 802.11 WLAN under distributed control," in *Proc. IEEE LANMAN*, September 2005.

[4] J. Yu, S. Choi, and J. Lee, "Enhancement of VoIP over IEEE 802.11 WLAN via dual queue strategy," in *Proc. IEEE ICC*, 2004.

[5] P. Bellavista, A. Corradi, and C. Giannelli, "Adaptive buffering-based on handoff prediction for wireless internet continuous services," in *Proc. HPCC*, 2005.

[6] S. Selvakennedy, "The impact of transmit buffer on EDCF with frame-bursting option for wireless networks," in *Proc. IEEE Local Computer Networks*, 2004.

[7] A. Markopoulou, F. Tobagi, and M. Karam, "Assessing the quality of voice communications over internet backbones," *IEEE Transactions on Networking*, vol. 11, no. 5, pp. 747–760, October 2003.

[8] D. Hole and F. Tobagi, "Capacity of an IEEE 802.11b wireless LAN supporting VoIP," in *International Conference on Communications*, 2004.