Packet Aggregation in Multi-rate Wireless LANs

Adnan Majeed and Nael B. Abu-Ghazaleh
Department of Computer Science
State University of New York at Binghamton
Binghamton, NY 13902
Email: adnan, nael@cs.binghamton.edu

***Abstract*—**
In CSMA networks, there is a significant overhead associated with each packet including header overhead and contention overhead. The high overhead problem is exacerbated because some of these overheads are required to be at the lowest data rate to ensure that all contending nodes can compete fairly. For applications where packets are small, such as Voice over IP (VOIP), this means that a majority of the available transmission time is wasted on overhead. Packet aggregation is one of the techniques that has been proposed to amortize the per-transmission overhead over multiple aggregated packets. However, existing heuristics are limited, often not considering multi-rate wireless MAC, or operation in a WLAN environment. In this paper, we first formulate the problem of optimal aggregation for a multi-rate CSMA MAC protocol and show that it is NP-Hard. We then propose two heuristics that solve the aggregation problem for multi-rate WLANs. The first, which we call Data Rate based Aggregation protocol (DRA), divides packets in the MAC queue into groups based on the data rate they are to be transmitted at. The algorithm aggregates packets in the same group and broadcasts the aggregated frame at the data rate of that group. Empirically, DRA achieves several fold increase in throughput compared to basic aggregation. DRA also achieves up to a 200% increase in the number of VoIP calls supported by a single 802.11g AP compared to using state of the art aggregation protocols. The second heuristic, which we call Data Rate based Aggregation with Selective Demotion (DRA-SD), enables cross data rate aggregation. Through preliminary evaluation, we show that selectively demoting packets can further improve performance.

I. Introduction

Header overhead and contention overhead significantly affect the performance of CSMA networks. These overheads are important; contention overhead, for example, ensures that fair and effective contention for the medium is possible. However, these overheads take up a portion of the bandwidth, reducing the bandwidth available for application traffic. For applications, such as VoIP, that generate small packets, the overheads can take up more bandwidth than the actual application traffic. Moreover, as the physical layer is improved, and new higher data rates are made available, the need for backwards compatibility often means that the packet overheads remain the same, further exacerbating the problem.

A. *Motivation – VoIP Over WiFi*

VoIP calls generate small packets and VoIP over Wifi (VoWifi) is an example of the type of applications that suffer most from the overheads of 802.11 networks. VoWifi calls have significantly increased over the past several years [1] and are predicted to continue increasing in the near future [2]. Wireless LANs (WLANs) need to be capable of meeting this exceeding demand and effectively handle multiple streams of data of varying sizes. However, current WLAN protocols are not efficient enough to do so, specially in the presence of applications, such as VoIP, that generate small packets.

Researchers show that 802.11 WLANs have poor performance, specially when used by applications that generate small packets, such as VoIP [3], [4]. Capacity models [3], [5]–[7] and measurement based works [4], [8], [9] show that 802.11b/g APs can carry a very small number of VoIP calls due to the overhead incurred by the 802.11 protocol when carrying small packets. This overhead causes VoIP calls to take up much more bandwidth at the physical layer than required by VoIP applications. For example, measurement studies show that a single 64kbps VoIP call reduces ongoing UDP traffic throughput by 900kbps [9]. This limits the number of VoIP calls that can be carried by an 802.11 AP and, in the presence of similar applications that generate small packets, affects the performance of co-existing traffic, such as video streaming traffic.

B. *Packet Aggregation*

Previous works propose combining multiple application packets into a single MAC frame, a technique called *packet aggregation* [10]–[13], to improve the performance of WLANs. Aggregation addresses the high overhead problem described above by amortizing the overhead over multiple packets.

Initial proposals for aggregation, which we call *Basic Aggregation* (BA), aggregate all packets in the MAC queue, irrespective of destination, and broadcast the aggregated frame [14], [15]. However, BA does not consider the presence of multi-rate MAC protocols. In an environment where different destinations experience different link quality, the AP must broadcast the aggregated frame at a low enough data rate to ensure that it is received by all destinations [16]. This can cause a significant loss in capacity as packets on high quality links get demoted to low transmission rates. This impact is not seen in previous studies because the evaluations assume that all links transmit at a fixed rate, the highest rate supported by the radio [14], [15]. In practice, applying aggregation without consideration to the packet data rates can cause excessive packet demotion, often leading to less efficient operation.

More recently, *Destination based Aggregation* (DA) was proposed where packets that have the same destination are aggregated and unicast to that destination [11], [17]. Previous works have used DA to improve back haul traffic in wireless mesh networks [17]. DA is also similar to the aggregation mechanism that is part of the 802.11n standard, however, to the best of our knowledge, its use for improving the performance of WLANs has not been investigated. As we show, because DA is limited to aggregating packets for a particular destination, it has limited aggregation potential. Furthermore, like BA, DA assumes that packets are transmitted in order, simplifying aggregation, but in general, missing out on more effective packet combinations that reduce the overall number of transmissions.

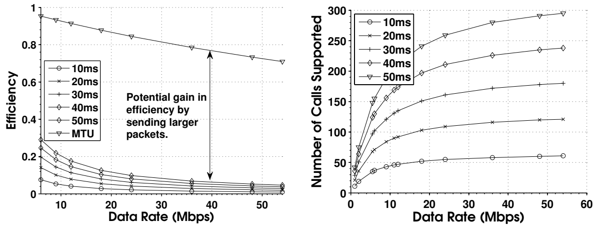
C. Towards Optimal Aggregation

In this paper we present an optimal aggregation mechanism that minimizes the total transmission time. We show that determining the optimal aggregation setting is NP-Hard and present two heuristics to solve the aggregation problem. The first heuristic disables packet demotion completely. The second heuristic allows selective packet demotion, only when it leads to a reduction in transmission time.

D. Contributions of the paper

Specifically, this paper makes the following contributions.

1) We show that, in certain cases, BA can lead to severe performance deterioration and is therefore not a feasible aggregation mechanism to be used in multirate WLANs. We also show that DA when applied to 802.11 b/g WLANs, can significantly improve downlink performance.
2) We present the notion of optimal aggregation in the context of multi-rate WLANs. We use transmission time as a metric to compare the performance of different aggregation mechanisms and to formulate the problem of optimal packet aggregation for the downlink. In the general case, and assuming packet reordering, we show that the problem is atleast NP-hard. We develop two heuristics to solve the aggregation problem efficiently.
3) Our first heuristic, Data Rate based Aggregation (DRA), aggregates packets for all links that have the same data rate and is able to transmit packets out of order. These two properties allows DRA to have a large number of opportunities to aggregate packets, and hence out perform DA, while disabling packet demotion, and hence outperform BA. We show that DRA increases WLAN capacity substantially, by upto 200% compared to DA in the presence of co-existing non-VoIP traffic.
4) Our second heuristic, Data Rate based Aggregation with Selective Demotion (DRA-SD), is a first attempt at an algorithm that allows selective demotion of packets. This demotion is allowed when it leads to a reduction in transmission time. Our evaluation of DRA-SD in basic scenarios shows that selective demotion is a promising direction of improving aggregation performance and should be pursued as a future research direction.



(a) (b)

Figure 1. Performance of VoIP using G729 codec over a 802.11g WLAN. (a) The small G729 VoIP packets cause efficiency to be very low, specially at higher data rates. The potential efficiency gain is indicated by the arrow. (b) Increasing packet sizes by increasing packetization interval improves efficiency and results in higher network capacity.

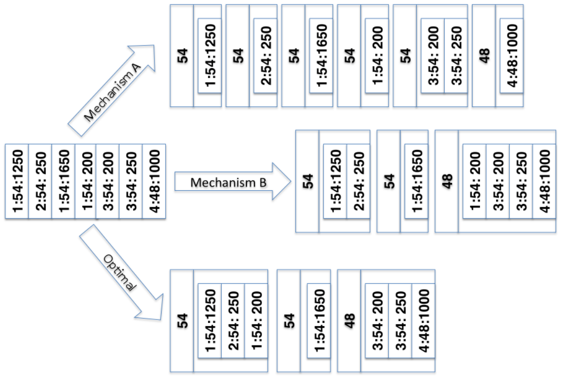
The remainder of this paper is organized as follows. Section II explains the network structure we study and shows the inefficiency of 802.11 WLANs. Section III shows by example how previously proposed aggregation mechanisms work and how an optimal aggregation protocol for a multi-rate WLAN should work; this leads to Section IV where we present a formulation of the optimal aggregation mechanism for a multi-rate WLAN and show that determining the optimal aggregation setting is NP-Hard. Section V presents the heuristics we propose to solve the aggregation problem and Section VI evaluates the performance of these heuristics. We conclude and present future work in Section VII.

II. Background and Motivation: Inefficiency of 802.11

To motivate packet aggregation, we provide a capacity analysis of an 802.11 flow and show that it is highly inefficient for applications such as VoIP that generate small packets. Without loss of generality, we assume a single WLAN cell and do not consider the impact of interference external to the cell (e.g., from nearby cells [18], [19]). The single-cell network structure consists of an Access Point (AP), connected to the wired infrastructure, and stations associated with it.

We define the capacity utilization of a flow (C) to be the portion of the medium consumed by that flow. C can be determined by three attributes: the application throughput requirement (c), the data rate at which the station communicates with the AP (R), and the efficiency of the network system (E). E determines how much of the medium is used up by communication overhead. Using the three attributes and assuming that medium capacity is not exceeded, capacity utilization is given as $C = \frac{c}{ER}$. That is, for low efficiency, the same traffic takes up a larger portion of the medium. We define efficiency, E, as the ratio of the application payload transmission time (T_p) to the total transmission time ($T_p + T_o$). T_o is the time spent in transmitting overhead and includes all possible overheads such as header overhead, contention overhead and protocol overhead.

Figure 1 shows the performance, per the model given above, of an 802.11g WLAN when carrying VoIP traffic. We modeled a G729 codec and included all overheads. These figures do



(a)

Figure 2. Performance of different aggregation mechanisms. Mechanism A: Aggregate packets for the same next hop. Mechanism B: Aggregate as many packets together as possible. Optimal: Aggregation that gives the lowest possible transmission time.

not model collisions and therefore provide upper bounds for capacity and efficiency. Figure 1(a) shows that VoIP has very low efficiency on 802.11g WLANs. G729 calls generate 10 byte packets with 10msec packetization interval. The network stack then appends 68 bytes of headers (RTP, UDP, IP and MAC). At the physical layer, the PLCP headers require a fixed $20\mu s$ to transmit. Other overheads contributing to the inefficiency include the protocol spacing overheads SIFS and DIFS, as well as the time required to transmit the ACK.

According to Figure 1(b) an 802.11g WLAN can theoretically support 61 G729 VoIP calls with 10ms packetization interval; meaning that a call with 8kbps application throughput requirement actually takes up almost 900kbps. Figure 1(a) shows that sending larger packets increases efficiency which reduces the portion of the medium taken up by each VoIP call and hence, as Figure 1(b) shows, increases network capacity.

III. An Aggregation Example

We use an example scenario to compare the aggregation mechanisms discussed in Section I and show how an optimal aggregation algorithm would work. The metric we use to compare the aggregation mechanisms is the total transmission time, T_{Tx}, as it directly maps to performance [20], [21]. Given a certain amount of data to be transmitted at particular data rates, the mechanism that has a lower value of T_{Tx} will give higher network capacity. We now show how to calculate the transmission time of a particular transmission schedule.

Consider a WLAN that consists of a single AP and the stations associated with the AP. The transmissions in the WLAN can be at one of b data rates: $R_1, R_2, \ldots, R_b$. The maximum transmission unit for the WLAN is MTU. Given a snapshot of this system with n packets and the vectors S and R_{max} defined as follows: $S = \{S_i$: Size of packet i$\}$ and $R_{max} = \{R_{max_i}$: Maximum data rate at which packet i can be transmitted$\}$ where $1 \leq i \leq n$, $0 < S_i \leq MTU$ and $R_1 \leq R_{max_i} \leq R_b$, the transmission time of these n packets is given as (notation used is explained in Table I)

$$T_{Tx} = \sum_{i=0}^{b} \left(N_{R_i} T_{o_{R_i}} + p_{R_i} T_{po_{R_i}} + \frac{s_{R_i}}{R_i} \right) \tag{1}$$

We use the scenario depicted in Figure 2 to compare the aggregation mechanisms. Figure 2 shows the MAC queue of an access point with seven packets. The packets are represented as N:D:S; N is the node ID of the next hop for the packet, D is the maximum data rate for this packet and S is the size of the packet in bytes. For this example, the MTU is set at 1700 bytes. The figure shows the aggregation setting for the three aggregation mechanisms. Mechanism A represents DA and Mechanism B represents BA. Both these mechanisms aggregate packets in order. Optimal represents the aggregation setting that gives the lowest transmission time possible; optimal aggregation is allowed to aggregate packets out of order. The total transmission times (excluding acknowledgements and protocol overheads such as SIFS, DIFS, etc.) are $849.63\mu s$, $801.67\mu s$ and $797.97\mu s$ for Mechanism A (BA), Mechanism B (DA) and Optimal, respectively.

There are a number of observations that can be made from this example. First, BA (Mechanism A) reduces the number of frames transmitted compared to DA (Mechanism B). Second, BA can increase the number of packets transmitted at lower data rates. Third, if out of order aggregation is allowed, there is more opportunity to aggregate packets. Finally, in some cases packet demotion improves performance; case in point is the last aggregated frame in the optimal setting. It was better to combine the last three packets in a frame transmitted at 48Mbps compared to generating two frames, one at 54Mbps and the other at 48Mbps. However, this should only be done when it leads to lower transmission times.

Packet demotion provides higher aggregation opportunity and so can reduce the number of frames transmitted. However, it also leads to more packets at lower data rates. Therefore, packet demotion should be done selectively. The net effect of packet demotion on total transmission time is the difference between the increase in transmission time due to transmitting more packets at lower data rates and the decrease in transmission time due to transmitting fewer frames. Section VI shows that demoting packets whenever possible, as BA does, leads to sever performance degradation.

We now present a formulation of optimal aggregation for a multi-rate WLAN based on these observations.

IV. Optimal Aggregation

In this section, we formulate the problem of optimal aggregation for multi-rate wireless LANs whose goal is to provide the optimal packet aggregation schedule to minimize overall transmission time. Specifically, the schedule consists of: (1) an assignment of packets to data rate; and (2) an aggregation schedule for the packets at each data rate identifying which packets are to be aggregated together. In general, we show that the problem is NP-hard since each packet may be considered

	Optimal Aggregation
R_i	Data rate at which packet i is transmitted
R_{max_i}	Maximum data rate at which packet i can be transmitted
N_{R_i}	Number of frames transmitted
$T_{o_{R_i}}$	Time to transmit per frame overhead
p_{R_i}	Number of application packets generated
$T_{po_{R_i}}$	Time to transmit per packet overhead
s_{R_i}	Size of application data (bits) transmitted at data rate R_i
Π	Solution space of all possible packet combinations
Π_l	A single packet combination
Π_{l_m}	Data rate of packet m in combination l
MTU	Maximum Transmission Unit

Table I
SUMMARY OF NOTATION

for transmission at any rate up to and including its maximum rate. Moreover, for a given packet to rate assignment, producing the optimal schedule for each rate is a bin-packing problem [22]. We refer to the solution space as Π, a specific packet data rate mapping as Π_l, and use Π_{l_m} to represent the data rate of packet m in mapping Π_l. The notation used in this section is summarized in Table I.

Although in general aggregation is an online problem, we consider building the schedule only for the packets that are available when the algorithm runs. The optimal aggregation problem is divided into two parts: the first part generates all possible packet combinations and the second part determines the minimum transmission time of each combination. The output is the combination that gives the overall minimum transmission time.

Considering the WLAN described in Section III, each packet mapping candidate is a vector of n elements where each element represents the data rate assigned to the corresponding packet. The optimal aggregation problem chooses the packet mapping that leads to the lowest transmission time. This can be formally stated as follows

$$\begin{aligned} &\text{Minimize} && T_{Tx} \\ &\quad\text{where} && \\ &T_{Tx} = && \sum_{i=0}^{r} \left(N_{R_i} T_{o_{R_i}} + p_{R_i} T_{po_{R_i}} + \frac{s_{R_i}}{R_i} \right) \\ &\text{subject to} && \\ & && \Pi_{l,m} \in R \quad \forall \text{m in } 1 \ldots \text{n} \\ & && \Pi_{l,m} \le R_{max_m} \quad \forall \text{m in } 1 \ldots \text{n} \end{aligned}$$

where N_{R_i} is the number of frames transmitted at data rate R_i, p_{R_i} is the number of packets that make up the N_{R_i} frames, meaning the number of packets in the packet vector for which the condition $\Pi_{l,m} = R_i$ is true. s_{R_i} is the combined size of the packets for which the condition $\Pi_{l,m} = R_i$ is true. The two constraints ensure that the packet mapping vector is legal and no packet is transmitted at a data rate higher than its maximum data rate.

We now define the second part of the problem; calculating N_{R_i} for each data rate R_i. There are a number of ways to aggregate packets to get N_{R_i}. Packets can be aggregated based on the destination, based on the next hop, based on data rates, etc. This part does not look across data rates since the first part already takes care of that. Therefore, aggregating packets according to data rates gives the maximum potential for aggregation, and hence, the fewest aggregated frames. The problem of determining N_{R_i} is presented formally as follows. Given a finite set of packets, P, with packet sizes $S_p (p \in P)$, find a partition of P in to a set F of k disjoint frames such that $v_f \le MTU$, where $f \in F$ and v_f is the size of frame f. This problem is a bin-packing problem [22]. The individual packets are items to be placed in bins; the frames represent bins. The frames have a maximum size which cannot be exceeded by packets assigned to them. We mathematically formulate the problem below where $x_{i,j}$ indicates whether packet j is part of frame i.

$$\begin{aligned} &\text{Minimize} && k \\ &\text{subject to} && \\ & && \sum_{i=1}^{k} x_{i,j} = 1 \qquad j \in P \\ & && \sum_{j=1}^{m} s_j x_{i,j} = v_i \qquad i \in F \\ & && x_{i,j} = \{0, 1\} \qquad i \in F, j \in P \\ & && v_i \ge 0 \qquad i \in F \\ & && v_i \le MTU \qquad i \in F \end{aligned}$$

This formulation means that every packet is part of exactly one frame and none of the frames has a size greater than the MTU.

Solving these two coupled components gives the optimal aggregation setting. Essentially, we need to solve an exponential number of bin packing problems.

Complexity of Optimal Aggregation: In this section we study the complexity of the optimal aggregation problem. As mentioned, the problem has two components. The first component determines the packet mapping combinations enumerating every possible rate assignment vector for the available packets. The second component determines that given a particular mapping, how to produce the optimal aggregation schedule for the packets assigned for each rate. We now show that the problem is NP-Hard.

Since each packet can be assigned to multiple data rates (those lower or equal to its maximum data rate), there are exponentially many legal mappings in general. For example, assuming a uniform distribution of data rates amongst the packets, on average each packet can be assigned $\frac{b}{2}$ data rates. This makes the number of possible mapping combinations $(\frac{b}{2})^n$ where n is the number of packets.

The second component is NP-Hard. This component calculates the optimal transmission schedule generated by aggregating packets in every data rate. This problem is equivalent to the bin-packing problem which is known to be NP-Hard [22].

Combining the two components means solving b bin-packing problems (one for each data rate) an exponential number of times (one for each of the valid packet rate mappings), making the optimal aggregation problem atleast NP-Hard.

There exist efficient approximation algorithms to solve the bin packing problem with tight bounds, such as the first-fit algorithm [22]. However, to find the combination of packets that minimizes the transmission time, all possible packet combinations have to be looked at. In this paper we look at two linear time heuristics that produce effective schedules while significantly reducing the number of packet mapping combinations to analyze.

V. Aggregation Heuristics

In this section we develop two heuristics that improve aggregation performance and are computationally feasible. The first heuristic, presented in Section V-A, completely disables packet demotion and hence only one packet combination is considered. The second heuristic, presented in Section V-B, allows selective packet demotion; the number of packet demotions considered is linear in terms of the number of data rates. We discuss implementation details of these heuristics in Section V-C.

A. Data Rate Based Aggregation (DRA)

We use observations from BA and DA to develop DRA. Aggregation improves performance by reducing the number of transmissions, while the potential loss in performance, in BA, comes from packet demotion. Based on these observations we design DRA to disable packet demotion. DRA, shown in Algorithm 1, divides packets in the AP queue into groups. Each group consists of packets that are to be transmitted at the same data rate. DRA aggregates packets from the same group together and broadcasts the aggregated frame at the data rate for that group; avoiding the loss in performance linked with aggregating across data rates. Therefore, unlike basic aggregation, DRA never performs worse than baseline operation with no aggregation. Also, since there is no packet demotion the complexity of the aggregation problem is reduced to solving b bin-packing problems. We use the first fit algorithm [22] to solve the bin-packing problems which provides larger aggregation potential by allowing packets to be aggregated out of order. DRA is an online algorithm that generates one frame at a time, every time the AP gets a transmission opportunity. Generating each frame requires at most checking each packet size once so if there are n packets in the MAC queue the algorithm is $O(n)$.

B. Data Rate based Aggregation with Selective Demotion (DRA-SD)

The example scenario in Section III revealed that in some cases aggregating across multiple data rates can lead to better performance. DRA does not allow packet demotion and therefore cannot take advantage of such aggregation opportunities. However, considering every possible packet demotion leads to exponential complexity. In this section, we present an extension of DRA that allows selective demotion of some packets; we call this heuristic Data Rate based Aggregation with Selective Demotion (DRA-SD). We use the observation that demoting a large amount of data is unlikely to yield better solutions because any gain in reducing the number of transmissions is likely to be offset by the loss of efficiency from demotion. Thus, DRA-SD focuses on demoting small frames.

Algorithm 1: Data Rate Based Aggregation

```
Input  : MAC Queue
Output: Aggregated Packet
//first packet in MAC Queue
//added to aggregated frame
$pkt$ = First Message in MAC Queue;
//Initialize aggregated frame
//data rate
$dataRate = pkt \rightarrow dataRate$;
//Initialize aggregated frame size
$aggPktSize = 0$;
while $pkt \neq null$ do
    if $(aggPktSize + pkt \rightarrow size) < MTU$ then
        $aggPktSize += pkt \rightarrow size$;
        add pkt to aggregated frame;
    end
    pkt = Next Message in Queue;
    while $pkt \rightarrow datarate \neq dataRate$ do
        pkt = Next Message in Queue;
    end
end
```

DRA-SD works as follows. DRA is run for every data rate to generate one aggregated frame per data rate. The aggregated frames generated are then merged across data rates, demoting a frame if it completely fits into an aggregated frame at a lower data rate and reduces the transmission time. The reduction in transmission time is possible if two conditions are satisfied: (1) the demotion does not increase the number of frames at the lower data rate and (2) the demoted frame is small enough so that demoting it does not cause a net increase in transmission time. This is a low-complexity implementation of packet demotion that does not consider complex packet demotion patterns which might give higher performance gains. However, as shown in Section VI-D, in certain scenarios this implementation significantly reduces delay while slightly increasing network throughput.

There are other ways of aggregating with packet demotion that can further improve network performance. DRA-SD limits demotion to taking the first frame generated by DRA on two data rates and demoting the higher data rate frame only if it completely fits into the first aggregated frame of the lower data rate. Another mechanism would be to generate all possible aggregated frames at a higher data rate and then taking the smallest frame generated and demoting this frame to a lower data rate if its individual packets can be downgraded and incorporated into different frames at the lower data rate. Such an algorithm would give a higher performance improvement compared to DRA-SD, but requires more computation. We leave the development and evaluation of more complex packet demotion algorithms for future work.

C. Implementation

In this section we discuss the implementation details of our aggregation heuristics. At the sender side the heuristics are implemented as part of the MAC transmission functionality. When a node gets a transmission opportunity it runs the aggregation protocol and transmits the aggregated frame. At the receiver side the MAC layer extracts individual packets from the aggregated frame and forwards, to higher layers, any packets that are meant for it. It should be noted here that the functionality of the 802.11 MAC protocol was not modified making these aggregation protocols compatible with unmodified 802.11 nodes.

A number of implementation decisions were made and these are discussed here.

Schedule: There are a number of ways that the aggregation schedule of DRA could be implemented. For example, round robin between the data rates. This maintains fairness between the various data rates, however, if the number of nodes in each data rate is not the same then this might lead to unfairness. Another example is weighted round robin between the data rates. Lower data rates could be given low weight and a significantly larger amount of high data rate traffic could be accommodated in the network.

The packet transmission schedule implemented in DRA is clear from Algorithm 1. The first packet in the MAC queue is picked and is the first packet added to the aggregated frame. This ensures that each data rate gets medium access proportional to the traffic generated for that data rate. We leave evaluation of alternative scheduling algorithms for future work.

Reliability: DRA, DRA-SD and BA broadcast packets to multiple recipients and without some form of reliability the delivery ratio for these algorithms can suffer. We use directed broadcast [23], where the broadcast packet is to be acknowledged by exactly one recipient. Previous implementations of aggregation have used the same approach to ensure reliability [15]. For our implementation this recipient is the destination of the first packet in the aggregated frame. Note that application level reliability can always be built on top of this mechanism.

Waiting Time: One important decision is how long a node should wait for packets before it generates an aggregated frame and transmits its. Increasing wait time allows a higher aggregation opportunity but also increases delay. In our implementation of aggregation we generate and transmit an aggregated frame whenever the AP gets a transmission opportunity. This does not add artificial delays to the traffic.

Packet Reordering: DRA and DRA-SD have the ability to reorder packets at the AP to increase the aggregation opportunity. These reordered packets are put back into their original order by a shimmer layer above the MAC layer at the receiver side. Packets are passed to the higher layer in their correct order.

VI. Performance Evaluation

In this section we evaluate the relative performance of Basic Aggregation (BA), Destination based Aggregation (DA), Data Rate based Aggregation (DRA) and Data Rate based Aggregation with Selective Demotion (DRA-SD) and compare these algorithms with the case when No Aggregation (NA) is used.

A. Experimental Setup

We use Qualnet 4.5 [24] for our simulations. The network consists of a wired and a wireless part connected through the AP. All wireless nodes are connected to the AP and all wired nodes are one hop from the AP on the wired infrastructure. The MAC protocol used is 802.11g, the nodes are static and the nodes use static routing. This setup ensures that the only factors affecting performance are the network load and aggregation performance. We simulate realistic channels using Ricean fading and use the Auto-Rate Feedback data rate selection algorithm, both of which are built into Qualnet [24].

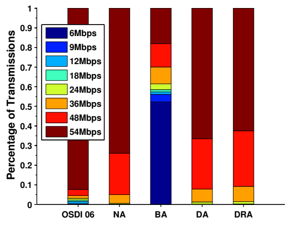
Node placements in our experiments are done at random. We use measurement results from OSDI '06 [25] to guide our node placement. Figure 3 shows the data rate distribution of transmissions from a single AP during a five minute trace at the OSDI '06 conference and also shows typical distributions for transmissions from our simulations. For traffic we use the following sources.

- *Video Streaming:* We simulate a 10Mbps HDTV stream that generates 1250 packets/second and the size of each packet is uniformly distributed between 500 bytes and 1500 bytes.
- *FTP:* We simulate an FTP download where each packet is 512bytes and the packet generation schedule is determined by FTP traces by Qualnet [24].
- *VoIP:* We simulate two way VoIP calls where talk time of either end is exponentially distributed and each end of the call takes turn transmitting packets.
- *HTTP:* We simulate HTTP connections where a user browses a server and requests pages. The HTTP connections are simulated using browsing traces.

B. Evaluation in Perfect Channel Condition

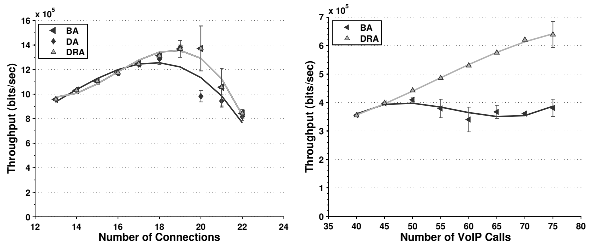
We first evaluate the relative performance of DA, DRA and BA in the presence of small sized packets. The connections are setup such that half are downlink and the other half are uplink. Each connection transmits 92byte packets every 10ms; each packet is then appended with a UDP, IP and MAC header and transmitted. For this simulation there is no fading and the data rate is fixed at 6Mbps for each link. Figure 4(a) shows the result of this experiment.

BA and DRA perform identically in this experiment. This is because when all links in a WLAN communicate at a single



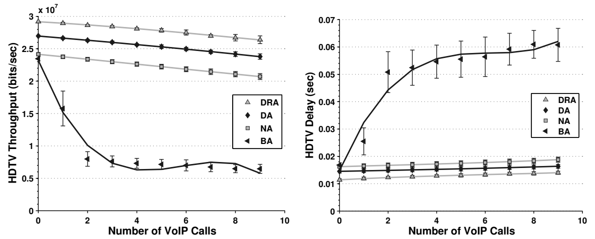
(a)

Figure 3. (a) Data rate distribution for downlink packets for an AP during OSDI '06 and simulation based downlink data rate distribution for an 802.11g WLAN.



(a) (b)

Figure 4. 95% confidence interval of throughput achievable by using the different aggregation mechanisms. (a) In the presence of a single data rate BA and DRA have identical performance. Compared to DA, DRA delays the onset of congestion by increasing efficiency, and hence reducing throughput requirement. (b) DRA and BA are compared in a multi-rate setting. BA reaches its maximum throughput at 45 downlink connections whereas DRA can sustain different data rate combination for upto 70 downlink connections, after which DRA can not sustain some data rate combinations.



(a) (b)

Figure 5. HDTV downlink stream performance as the number of VoIP calls are increased. BA has worst performance because of excessive packet demotion where as the opportunity to aggregate across destinations in DRA gives it the best performance.

fixed data rate and all packets are of the same small size then BA behaves exactly like DRA (this is the optimal solution). However, in this scenario, DA performs worse compared to both BA and DRA because it is restricted to aggregate packets with the same source. DA can support upto 19 connections where as DRA and BA support 20 connections.

The middle portion of Figure 4(a) shows the interesting performance region in the DA/DRA comparison. At low load there is sufficient capacity to carry all traffic and aggregation is not necessary. Conversely, at high loads, the queue sizes increase and DA gets enough aggregation opportunity to perform similar to DRA. The middle region, when network capacity is reached, DRA performs better by aggregating a larger number of packets than DA which is restricted to packets with the same destination.

Comparing DRA and BA requires a scenario that is not limited to using a single data rate. In this case we again use a single WLAN 802.11g setup with nodes placed with uniform distribution in the AP coverage area with the same channel and traffic setup. The link between each node and the AP is assigned a data rate depending on the distance of the node from the AP. Figure 4(b) shows the relative performance of BA and DRA. We see that for this downlink traffic setup BA throughput can match offered throughput for upto 45 connections whereas DRA matches offered throughput for upto 70 connections after which DRA throughput falls for some data rate combinations. Note that in this scenario the benefit of DRA comes purely from not demoting packets; the benefit of out of order aggregation is discussed later.

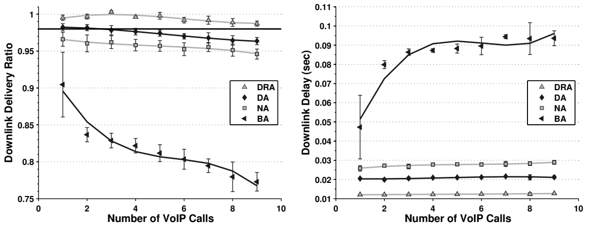
C. Accounting for Fading

In this section we evaluate the performance of DA, BA and DRA with NA for a number of traffic mixes under more realistic channel conditions.

HDTV and VoIP Traffic: We first evaluate how all aggregation mechanisms compare to NA when the WLAN carries HDTV and VoIP traffic. This is a typical traffic mix for a home wireless network. This scenario has three HDTV downlink streams and the number of VoIP calls are increased from 1 to 9. Figure 5 shows the effect on HDTV throughput and delay as the number of VoIP calls are increased. BA performs the worst while DRA has the best performance. The reason for BAs low performance is that BA aggregates as many packets as possible. If even one of these packets is at a low data rate then the whole aggregated frame is transmitted at the low data rate. This is why in Figure 3 the data rate distribution of BA has a much higher percentage of low data rate transmissions. The scenario tested here highlights the flaw in previous works that propose BA; in a multi-rate environment BA can cause performance to be worse compared to using no aggregation.

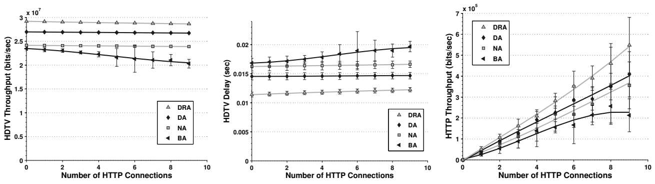
NA transmits each packet separately, using the medium very inefficiently. This causes HDTV throughput to fall and delay to increase. DA improves on the throughput and delay of NA by aggregating packets to the same destination. However, DA is able to aggregate only a few of the large HDTV packets; aggregating packets per destination therefore limits aggregation potential, limiting performance improvement. DRA has the best performance in terms of both throughput and delay. DRA is able to aggregate VoIP packets with packets from the HDTV streams and hence carry VoIP traffic with low delay and minimal effect on the HDTV performance.

Figure 6 shows the performance of the VoIP calls. The number of VoIP calls sustainable is considered as the number of calls that can be maintained while the packet loss rate is less than 2%. The uplink delay and delivery ratio of all VoIP calls is enough to maintain call quality. For downlink, both BA and NA can not sustain a single call with the three HDTV



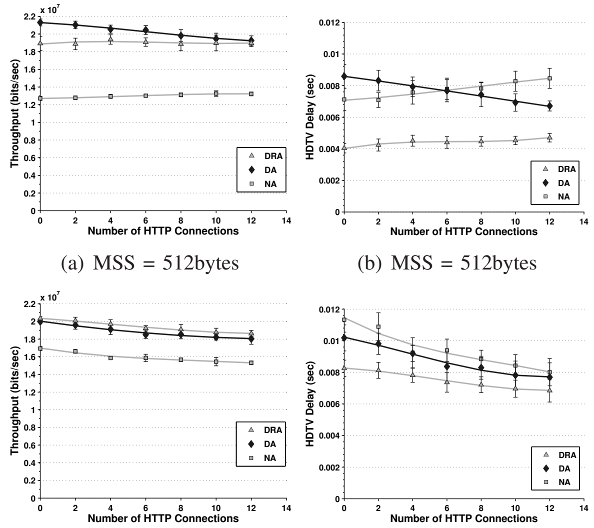
(a) (b)

Figure 6. VoIP performance in the presence of three HDTV downlink streams. (a) The horizontal line marks the point where loss rate is 2%. NA and BA can not sustain a less than 2% loss rate for a single VoIP call where as DA enables upto three VoIP calls to go on. DRA has the best performance and carries all 9 VoIP calls. (b) VoIP downlink delay performance shows that DRA reduces end-to-end delay to almost half the value of delay for DA.



(a) (b) (c)

Figure 7. HDTV and HTTP performance as the number of HTTP connections are increased in the presence of three downlink HDTV streams. (a) HDTV Throughput, (b) downlink delay for HDTV and (c) throughput of the HTTP connections.



(a) MSS = 512bytes (b) MSS = 512bytes

(c) MSS = 1024bytes (d) MSS = 1024bytes

Figure 8. FTP, HDTV and HTTP traffic. For an MSS of 512 bytes FTP throughput for NA and DRA stays constant. This is because for both NA and DRA the offered load stays the same. DRA is able to incorporate HTTP packets in the 512byte TCP packets and the HDTV packets. For 1024bytes the larger TCP packets increases end-to-end delay for HDTV. This increases number of HDTV packets in the queue and hence increases aggregation opportunity.

downlink streams running. DA can sustain three calls before the packet loss rate exceeds 2% while DRA can sustain all nine VoIP calls; a 200% improvement over DA. The relative ordering for VoIP delay performance is the same. DRA is able to cut down VoIP delay to almost half the end-to-end delay of DA.

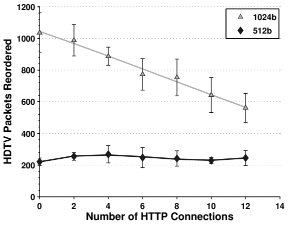
HDTV and HTTP Traffic: Next we evaluate the performance when HDTV traffic coexists with HTTP traffic. Figure 7 shows the result for this scenario. As the number of HTTP connections are increased HDTV throughput drops and delay increases. However, the relative performance order is maintained. BA performs the worst, followed by NA and then DA. DRA is able to merge HTTP traffic with the HDTV stream effectively carrying the HTTP traffic "for free". Based on these traffic mixes it is clear that BA is not an effective aggregation mechanism.

FTP, HDTV and HTTP Traffic: We now discuss a traffic mix with a downlink FTP flow, a downlink HDTV video stream and an increasing number of HTTP connections. For the FTP flow we evaluate performance by setting the TCP MSS to 512bytes and 1024 bytes. The results are shown in Figure 8. Figure 8(a) and Figure 8(c) show the network throughput for NA, DA and DRA for MSS of 512bytes and 1024 bytes, respectively. For an MSS of 512 bytes increasing number of HTTP connections does not decrease network throughput for NA and DRA. However, for NA, the extra traffic causes the HDTV stream delay to increase. In case of DA, increasing the number of HTTP connections causes the offered FTP load to decrease and hence decreases the network throughput achieved. The reduction in FTP traffic causes HDTV delay to decrease. This causes the delay "cross" in Figure 8(b). In the case of DRA, the additional HTTP traffic is incorporated with the FTP and HDTV packets, generating larger frames, therefore, HDTV delay changes very little.

Comparing the performance of MSS of 512bytes and 1024bytes also shows some interesting trends. First, increasing the MSS causes a drop in DA throughput. This is because FTP packets become too large to be aggregated and each FTP packet is transmitted separately, reducing efficiency. In case of DRA the network throughput increases when the MSS is increased. This is because larger packets means longer transmission times; this leads to more HDTV packets being queued up and hence more out of order aggregation opportunity. Figure 9 shows the significantly higher number of reordered HDTV packets because of this behavior. As the number of HTTP connections increases, the number of out of order packets drop because some of the free capacity in frames is taken up by HTTP packets. This experiment clearly shows the possible benefit of out of order aggregation. For the larger MSS, increasing the number of HTTP connections has a similar gradual decreasing effect on the throughput of all three mechanisms: NA, DA and DRA. FTP, HDTV and VoIP traffic mixes show similar results but are not shown here due to space constraints.

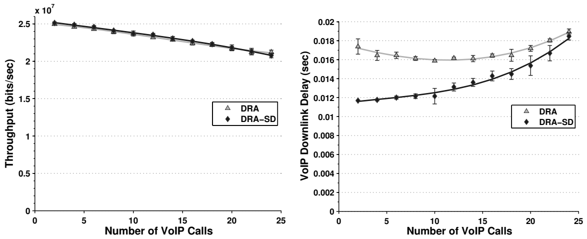
D. DRA-SD Performance Evaluation

In this section we evaluate DRA-SD performance. We use the FTP, HDTV and VoIP traffic mix. The FTP and



(a)

Figure 9. Number of HDTV packets reordered for both MSSs.



(a)

(b)

Figure 10. DRA-SD Results

HDTV connection is on a 48Mbps link and the VoIP calls are on 54Mbps links. Figure 10 shows the performance. In terms of throughput DRA-SD shows a very slight gain over DRA, however, the major difference is in delay. DRA-SD, by aggregating across data rates, is able to demote VoIP packets when it improves performance.

VII. Conclusion

In this paper we show that aggregation mechanisms proposed by previous works either lead to performance degradation in multi-rate WLANs, or have very limited performance gains. We formulate the optimal aggregation problem for a multi-rate WLAN and show that the optimal problem is atleast NP-Hard. We then propose two heuristics to solve this problem. The first heuristic, DRA, shows significant improvement in performance, compared to state of the art aggregation protocols, by aggregating packets for each data rate separately; DRA disables cross data rate aggregation and allows out of order aggregation. In certain cases, DRA shows a 200% increase in VoIP capacity of WLANs in the presence of co-existing non-VoIP traffic. The second heuristic, DRA-SD, allows limited cross data rate aggregation and shows that selective packet demotion could be used to reduce WLAN delays in certain cases.

In future work, we plan on developing and evaluating more cross data rate aggregation protocols to bridge the gap between DRA and optimal aggregation.

References

[1] T. Henderson, D. Kotz, and I. Abyzov, "The changing usage of a mature campus-wide wireless network," *Computer Networks*, vol. 52, no. 14, pp. 2690 – 2712, 2008.

[2] "471 Billion Mobile VoIP Minutes By 2015," GigaOM, July 2010. [Online]. Available: http://gigaom.com/2010/07/01/mobile-voip-forecast/

[3] D. Hole and F. Tobagi, "Capacity of an ieee 802.11b wireless lan supporting voip," in *ICC*, 2004, pp. 196 – 201.

[4] S. Shin and H. Schulzrinne, "Experimental measurement of the capacity for voip traffic in ieee 802.11 wlans," in *INFOCOM*, 2007.

[5] S. Garg and M. Kappes, "Can i add a voip call?" in *Communications, 2003. ICC '03. IEEE International Conference on*, 2003, pp. 779 – 783.

[6] L. Cai, X. Shen, J. Mark, L. Cai, and Y. Xiao, "Voice capacity analysis of wlan with unbalanced traffic," *Vehicular Technology, IEEE Transactions on*, pp. 752 – 761, 2006.

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

[8] F. Anjum, M. Elaoud, D. Famolari, A. Ghosh, R. Vaidyanathan, A. Dutta, P. Agrawal, T. Kodama, and Y. Katsube, "Voice performance in WLAN networks - an experimental study," in *GLOBECOM*, 2003.

[9] S. Garg and M. Kappes, "An experimental study of throughput for udp and voip traffic in ieee 802.11b networks," *WCNC*, vol. 3, 2003.

[10] A. Kassler, M. Castro, and P. Dely, "Voip packet aggregation based on link quality metric for multihop wireless mesh networks," *Proceedings of the Future Telecommunication Conference, Beijing, China*, 2007.

[11] M. Castro, P. Dely, J. Karlsson, and A. Kassler, "Capacity increase for voice over ip traffic through packet aggregation in wireless multihop mesh networks," in *Future Generation Communication and Networking (FGCN 2007)*, 2007, pp. 350 – 355.

[12] Y. Jeong, S. Kakumanu, C.-L. Tsao, and R. Sivakumar, "Voip over wi-fi networks: Performance analysis and acceleration algorithms," *Mobile Networks and Applications*, vol. 14, no. 4, Aug 2009.

[13] J. H. Hong and K. Sohraby, "On modeling, analysis, and optimization of packet aggregation systems," *Communications, IEEE Transactions on*, vol. 58, no. 2, pp. 660 – 668, 2010.

[14] W. Wang, S. C. Liew, and V. Li, "Solutions to performance problems in voip over a 802.11 wireless lan," *Vehicular Technology, IEEE Transactions on*, vol. 54, pp. 366–384.

[15] P. Verkaik, Y. Agarwal, R. Gupta, and A. Snoeren, "Softspeak: Making voip play fair in existing 802.11 deployments," in *Sixth USENIX Symposium on Networked Systems Design and Implementation*, 2009.

[16] S. Yun and H. Kim, "Rate diverse network coding: breaking the broadcast bottleneck," *MobiHoc*, 2010.

[17] S. Ganguly, V. Navda, K. Kim, A. Kashyap, D. Niculescu, R. Izmailov, S. Hong, and S. Das, "Performance optimizations for deploying voip services in mesh networks," *Selected Areas in Communications, IEEE Journal on*, vol. 24, no. 11, pp. 2147 – 2158, 2006.

[18] A. Mishra, V. Brik, S. Banerjee, A. Srinivasan, and W. Arbaugh, "A client-driven approach for channel management in wireless lans," *INFOCOM*, 2006.

[19] A. Mishra, S. Banerjee, and W. Arbaugh, "Weighted coloring based channel assignment for wlans," *SIGMOBILE Mobile Computing and Communications Review*, vol. 9, no. 3, Jul 2005.

[20] S. Garg and M. Kappes, "Admission control for voip traffic in ieee 802.11 networks," in *GLOBECOM*, 2003.

[21] H. Zhai, X. Chen, and Y. Fang, "A call admission and rate control scheme for multimedia support over ieee 802.11 wireless lans," *Wirel. Netw.*, vol. 12, no. 4, pp. 451–463, 2006.

[22] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York, NY, USA: W. H. Freeman & Co., 1979.

[23] P. Rogers and N. Abu-Ghazaleh, "Directed broadcast: a mac level primitive for robust network broadcast," in *WiMob'2005*, 2005.

[24] "Qualnet simulator," [online]http://www.scalable-networks.com. [Online]. Available: http://www.scalable-networks.com

[25] R. Chandra, R. Mahajan, V. Padmanabhan, and M. Zhang, "CRAWDAD data set microsoft/osdi2006 (v. 2007-05-23)," Downloaded from http://crawdad.cs.dartmouth.edu/microsoft/osdi2006, May 2007.