
Peer reviewed version

Link to published version (if available):
10.1109/VETECS.2006.1683027

Link to publication record in Explore Bristol Research

PDF-document

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
http://www.bristol.ac.uk/pure/about/ebr-terms.html
Non-Linear Optimization of IEEE802.11e Super-frame Configuration

Russell Haines, Tim Lewis, Justin Coon, Neil Fanning
Telecommunications Research Laboratory,
Toshiba Research Europe Ltd.,
32, Queen Square, Bristol, BS1 4ND, UK

Abstract — As the potential data-rates of wireless local area networks (WLANs) continue to rise, the ability of such systems to support a rich set of applications increases. The centralized control functions in the IEEE802.11 family of standards have been developed to enable both data-oriented (browsing, email, etc.) traffic and quality of service (QoS) sensitive traffic to coexist. Balancing the demands of the two types of traffic has, to date, been achieved by algorithms based on experimental, heuristic data. In this paper we present a non-linear optimization theory-based approach for deriving optimum configurations with the IEEE802.11e centralized control functions in mind. The optimization algorithm itself (the “Barrier Method”) is well-known; the challenge in problems such as this is in the formation of the utility function and its constraints, so these are explored in detail. Finally, we show the advantages of this approach over discrete look-up based approaches.

Keywords — Optimization, WLAN, 802.11, HCF, PCF, CFP, CP.

I. INTRODUCTION

Wireless local area networks (WLAN), such as those defined by the IEEE802.11 family of standards, are increasingly being used to support rich multimedia audio/visual (A/V) applications. Such applications have very demanding quality of service (QoS) needs such as guaranteed delay and delay-variance, as well as, typically, higher bandwidth requirements. The best-effort contention-based Distributed Coordination Function (DCF) of the IEEE802.11 standard [1] struggles to support such traffic. The centralized Point Coordination Function (PCF) of the original IEEE802.11 standard and the enhanced Hybrid Coordination Function (HCF) in the IEEE802.11e standard [2] both introduce centralized coordination to allow QoS-sensitive traffic to coexist alongside contention-based data exchanges. However, there is a relative paucity of research on the centralized control functions of IEEE802.11 when compared to the vast body of work available on the DCF and ad hoc networking, and this contribution partially redresses this imbalance.

Centralized coordination imposes a time-based repeating super-frame onto the medium (as illustrated in Fig. 1), characterized by the transmission of a broadcast beacon, followed by a contention-free (pollled) period (CFP) and then a contention-based access period (CP). The duration of the super-frame (i.e. the beacon and CFP repetition rate) and the relative size of the CFP to the rest of the super-frame, typically termed $\text{CFP}_{\text{REF}}$ and $\text{CFP}_{\text{MAX}}$, are both configurable by the centralized controller entity located at the Access Point (AP).

The configuration of these parameters determines the success of a given WLAN deployment from the perspective of the polled traffic, the contention-based traffic or both. A badly configured system will fail to deliver the performance that the end-user has the right to expect, irrespective of the headline data rate of the product.

A self-adaptive scheme has been proposed and studied [3]. This scheme selects parameters from pre-defined look-up tables indexed by a quantized number of active polled stations and stepped values for the maximum allowable delay of the applications. The values populating the look-up tables are derived through experimental simulation results, and do not take into account the minimum CFP and CP sizes mandated by the standard [1].

A more flexible and adaptable approach would allow an continuous optimized set of super-frame parameters to be derived; an approach with a more theoretical basis would permit greater confidence in the optimal nature of the values being employed than is possible with experimental results.

The mathematical technique proposed as a candidate solution in this paper is that of non-linear optimization. The various methods within non-linear optimization theory optimize (as the name suggests) any number of variable parameters to provide a stable system solution. Non-linear optimization has been applied to various problem domains within communications, including wireless sensor network access [4] and deriving training sequences for orthogonal frequency division multiplexing (OFDM) systems [5, 6]. We use the barrier method [7] in this work. The success of non-linear optimization approaches is dependant on how well the objective and constraint functions model the behavior under study.

Fig. 1 Super-frame Structure

0-7803-9392-9/06/$20.00 (c) 2006 IEEE
II. Utilisation Model and Objective Function

The goal of configuring the super-frame is to try to satisfy both the polled and contending traffic flows. If just one traffic flow were given free rein, this would be at the expense of the other. It is important to guarantee the polled (QoS-sensitive) traffic, but not if it starves the contention traffic. Our approach is to maximize the utilisation of the two phases simultaneously within a number of constraints, such that the two phases’ utilizations are traded-off against each other.

To begin with, then, the utilisation expressions of the two phases must be developed; i.e. the efficiency of the allocation of air-time to the different phases. The goal is to develop expressions that indicate how far from the ideal each component is.

Certain assumptions are required even at this high level starting point in order to simplify the mathematics that follows. These include:

- bit-errors and interference neglected
- no hidden-terminal or capture effect
- no collisions
- no stations are power-saving
- terminals are fully backlogged

The efficiency of each phase can be further decomposed into two factors—the inherent inefficiency in each individual exchange (which scales linearly with the number of exchanges) and the phase inefficiency, comprising any unused-airtime wastage at the end of the phase. We seek to derive, then, two expressions, one for the utilisation of the polled traffic and one for the contending data traffic.

In the case of the polled traffic it is easy to compute how much of the bandwidth is being wasted and aim to minimize that. Each polled exchange incurs the standard inter-frame space penalties, specifically two SIFS periods, as shown in Fig. 2. Hence for a polled exchange, the overhead is simply twice the SIFS duration of the PHY in question.

\[ C_a = 2 \times SIFS \quad (1) \]

The second factor is the wastage in the CFP caused by it being configured to any size not divisible exactly by the frame exchange duration (in practice the central controller can terminate the CFP early and make this “wasted” period available to the CP).

Hence, the wastage incurred within the polling period, comprising the wastage per polled-exchange plus whatever surplus remains at the end of the CFP, can be expressed as:

\[ V(N_p) = \left(1 - \frac{N_p (C_b - C_a)}{xy}\right) \quad (2) \]

Where \( C_b \) is the entire polled exchange duration (ms) and \( C_a \) is the polled exchange overhead from (1), and \( x \) and \( y \) are CFP\textsubscript{MAX} and CFP\textsubscript{REP} respectively.

The number of polled terminals, \( N_p \), is a parameter that the AP can reasonably be expected to know as all stations must associate with the AP if polling service is required.

For the contending traffic it is more straightforward to consider and maximize the percentage of the bandwidth that was actually used to transmit useful data. During contention, stations must wait for the DIFS period of silence on the medium (with the 802.11b physical layer, this is 50\( \mu \text{s} \), compared to the SIFS of 10\( \mu \text{s} \)). Having reached the end of DIFS, the station backs off for a random number of slots (each of 20\( \mu \text{s} \) duration in 11b) drawn from the range [0, CW], where CW begins at 31 (11b again) and can increase as a binary exponential up to the limit 1023. If the medium goes busy during the contention window (the slot-count-down) then the STA will suspend the count down, wait for the medium to go idle for DIFS again, then resume counting from where it left off.

One of the aforementioned assumptions was “no collisions”, and this assumption can be used to simplify the “truncated binary exponential back-off” mechanism by freezing CW at 31, and simply taking a mean CW value of 15.5 (albeit a non-integer value) for every contention. If every contention is assumed to win without any other terminal transmitting during the CW phase (although in reality the probability of seeing another terminal transmit is going to increase with the number of terminals present) then a single DIFS per contention can be assumed.

A final simplification is for the calculations to consider a single “standardized” payload size for the contention-based traffic, corresponding to the mean. The first part of the utilisation expression is the implicit utilisation of a particular exchange, as illustrated in Fig. 3.
This gives an overhead of:

\[ M_s = \text{DIFS} + \text{Backoff} + \text{SIFS} + \text{ACKFrame} \quad (3) \]

Then we can incorporate the waste at the end of each CP. The effective number of contending stations will depend on the traffic level and the total number of contending stations \( N_c \). If we know the approximate packet rate of this traffic, \( P_r \), the effective number of concurrently sending stations will be \( y \times P_r \times N_c \). This results in the utilisation for the contention period:

\[ L(N_c) = 1 - \frac{y \times P_r \times N_c (H_s - M_s) + y (1 - x)}{y (1 - x)} \]

Which simplifies to:

\[ L(N_c) = \frac{P_r \times N_c (M_s - H_s)}{1 - x} \quad (4) \]

Where \( H_s \) is the entire standardized contended exchange duration (in ms), \( M_s \) is the standardized contended exchange overhead (ms) from (3), \( N_c \) is the number of contending stations, \( y \) is CPF\(_{\text{REP}}\) and \( x \) is CPF\(_{\text{MAX}}\).

This expression must be constrained by the frame-generation rate of the traffic, otherwise this becomes almost a “self optimizing” model that will always fill the CP to capacity. We can use the utilisation functions \( L \) and \( V \) in the following objective function:

\[ f_s(x, y) = (1 - L(N_c))^2 + (V(N_p))^2 \quad (5) \]

We use the \( 1 - L(N_c) \) term since higher values of \( L \) correspond to good performance (in contrast high values of \( V \) indicate poorer performance), and square both terms to ensure that both are positive and differentiable over the whole domain of interest. Plugging in the expressions for \( L \) given in (4) and \( V \) from (2) and simplifying gives:

\[ f_s(x, y) = \left( 1 - \frac{P_r \times N_c (M_s - H_s)}{1 - x} \right) + \left( 1 - \frac{N_c (C_a - C_s)}{xy} \right)^2 \quad (6) \]

The possible solutions are constrained in a number of ways. CPF\(_{\text{MAX}}\) is a ratio of two time periods so must be positive and less than one. CPF\(_{\text{REP}}\) is bounded by the worst case polling frequency (“delay”, \( D \)) specified by the application. Additionally, both the CPF and CP are subject to minimum duration constraints (“CPF\(_{\text{MIN}}\)” and “CPF\(_{\text{MIN}}\)” respectively) according to the standard [1]. The CPF has to be at least big enough to contain one polled exchange comprising the largest payload possible in each direction, plus a Beacon and a CF-End. The CP has to be large enough to contain an acknowledged exchange of the largest payload possible.

Mathematically, the problem reduces to an optimization problem over two variables, \( x \) and \( y \): Minimize \( f_s(x, y) \) from equation (6), subject to the set of constraints:

\[ \begin{align*}
\text{CPF}_{\text{MIN}} - xy & \leq 0 \\
\text{CPF}_{\text{MIN}} - (1 - x) y & \leq 0 \\
0 & \leq x \leq 1 \\
0 & \leq y \leq D
\end{align*} \quad (7) \]

III. NON-LINEAR VECTOR OPTIMIZATION OF MODEL

The barrier method [7] can be used to solve this optimization problem. However, it is desirable to first reformulate the objective function as a function of a single variable that is a vector. For this reformulation, let \( z = (x, y) \), and define the length two unit vectors as \( e_1 = (1, 0)^T \) and \( e_2 = (0, 1)^T \). With these definitions in mind, the objective function can be rewritten as follows:

\[ f_0(z) = \left( 1 - \frac{\alpha}{1 - e_1^T z} \right)^2 + \left( 1 - \frac{\beta}{e_2^T E z} \right)^2 \quad (8) \]

Here \( \alpha = p_r N_c (M_s - H_s) \), \( \beta = N_p (C_b - C_a) \) and \( E = e_2 e_2^T \). In vector notation the constraints can be restated as follows:

\[ \begin{align*}
\text{CPF}_{\text{MIN}} - z^T E z & \leq 0 & 1^{\text{st}} \text{constraint} \\
\text{CPF}_{\text{MIN}} - e_1^T z + z^T E z & \leq 0 & 2^{\text{nd}} \text{constraint} \\
e_1^T z - 1 & \leq 0 & 3^{\text{rd}} \text{constraint}, \text{upper bound} \\
e_2^T z & \leq 0 & 3^{\text{rd}} \text{constraint}, \text{lower bound} \\
e_1^T z - D & \leq 0 & 4^{\text{th}} \text{constraint}, \text{upper bound} \\
e_2^T z & \leq 0 & 4^{\text{th}} \text{constraint}, \text{lower bound}
\end{align*} \]

The constants are determined by the physical layer under consideration and the characteristics of the traffic flows.

\[ \begin{align*}
M_s & \text{ standardized data exchange overhead, ms} \\
H_s & \text{ standardized data exchange, ms} \\
C_a & \text{ polled exchange overhead, ms} \\
C_b & \text{ polled exchange duration, ms} \\
N_c & \text{ number of data stations} \\
\text{CPF}_{\text{MIN}} & \text{CPF}_{\text{MIN}}, \text{ ms} \\
\text{CPF}_{\text{MIN}} & \text{CPF}_{\text{MIN}}, \text{ ms} \\
D & \in \{ \ldots \} \text{ polling rates under consideration} \\
N_p & \in \{ \ldots \} \text{ numbers of polling stations considered} \\
P_r & \text{ packet generation rate for conflicting traffic}
\end{align*} \]

This problem can now be solved using standard convex optimization techniques such as the barrier method [7], however, the objective function is not convex so feasible starting points must be determined to pick the appropriate
local minima (there are never more than three such points and the three initial values we use will always locate the desired one). By examining the inequality constraints of the original problem it is possible to find feasible starting points $x_0$ and $y_0$ that can be used to initialize the barrier method. Observe the following two inequalities:

- $CFP_{\text{min}} \leq xy$
- $CP_{\text{min}} \leq (1 - x)y$

These are obtained by rearranging the first two inequalities of the original problem statement. Solving the second inequality for $xy$ enables the composite inequality to be written as: $CFP_{\text{min}} \leq xy \leq y - CP_{\text{min}}$

Thus, for a given $y = y_0$, a feasible $x = x_0$ can be taken from the interval:

$$x_0 \in \left[\frac{CFP_{\text{min}}}{y}, 1 - \frac{CP_{\text{min}}}{y}\right] \quad (9)$$

and the following feasible starting point constraint must be met.

$$CFP_{\text{min}} > y_0 - CP_{\text{min}} \quad (10)$$

**IV. APPLICATION OF MODEL**

The assumptions and parameters used in the aforementioned static self-adaptive scheme [3] can be adopted by this model to give some concrete values. These parameters include an 802.11b MAC/PHY configuration (affecting IFS times and the like), with ten data stations contending for access in the CP. These parameters are given in Table 1 along with the resulting concrete values for the constants in the expressions developed in the previous sections.

Feasible starting points can then be considered - the starting point constraint in (10) can be met for these values when, for example, $CFP_{\text{min}} = 39.922$, $CP_{\text{min}} = 21.404$ and $y_0 = 48$.

The barrier method was used to find optimal values of $x$ and $y$ for the different combinations of $D$ and $N_p$ given above. As mentioned previously the objective is not in general convex, for certain values of the parameters $D$ and $N_p$ it can have up to three local minima. The particular minimum that the algorithm converges on depends on the initial values, and is particularly sensitive to the value of the $x$ component. We keep the initial $y$ value constant and close to its maximum of $D$. The three local minima were discovered using the following set of initial $x$ values:

1. $1.2 * (CP_{\text{min}}) / y$
2. $0.5 * (1 - CP_{\text{min}} - CP_{\text{min}}) / y$
3. $0.8 * (1 - CP_{\text{min}} - CP_{\text{min}}) / y$

The first of these is a point near the lower end of the feasible set, the second a point in the middle and the third a point towards the top end of the feasible set for $x$. For many values of $D$ and $N_p$, all of these local minima were found to be identical, indicating that the local minimum is a global minimum. In the case where more than one local minima was found the objective function was evaluated at each and the true minimum chosen. The minimum values obtained are illustrated in Fig. 4 and Fig. 5, and listed in Table 2.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slot</td>
<td>0.02ms</td>
<td>PIFS</td>
<td>0.03ms</td>
</tr>
<tr>
<td>SIFS</td>
<td>0.01ms</td>
<td>DIFS</td>
<td>0.05ms</td>
</tr>
<tr>
<td>MAC Header</td>
<td>28 bytes</td>
<td>PLCP/</td>
<td>0.192ms</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Header/Preamble</td>
<td></td>
</tr>
<tr>
<td>Mean data</td>
<td>1000 bytes</td>
<td>Mean data rate</td>
<td>7.5 fps</td>
</tr>
<tr>
<td>MSDU</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A/V MSDU</td>
<td>200 bytes</td>
<td>A/V data rate</td>
<td>64 kbps,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1s on, 1.35s off</td>
</tr>
<tr>
<td>Data rate</td>
<td>2Mbps</td>
<td>Control rate</td>
<td>1Mbps</td>
</tr>
<tr>
<td>Beacon</td>
<td>160 bytes</td>
<td>ACK</td>
<td>14 bytes</td>
</tr>
<tr>
<td>M</td>
<td>0.674 ms</td>
<td>H</td>
<td>4.978 ms</td>
</tr>
<tr>
<td>C</td>
<td>0.02 ms</td>
<td>C</td>
<td>2.228 ms</td>
</tr>
<tr>
<td>N</td>
<td>10 stations</td>
<td>N_p</td>
<td>{2, 4, ..., 20}</td>
</tr>
<tr>
<td>CFP_{\text{min}}</td>
<td>39.922 ms</td>
<td>CP_{\text{min}}</td>
<td>21.404 ms</td>
</tr>
<tr>
<td>P</td>
<td>0.0075 s^{-1}</td>
<td>D</td>
<td>475, 87.5, 100, 112.5, ..., 200 ms</td>
</tr>
</tbody>
</table>

![Fig. 4 CFP_{\text{MAX}} Optimization Results](image1)

![Fig. 5 CFP_{\text{REF}} Optimization Results](image2)
Table 2: Optimum \(\text{CFP}_{\text{MAX}}\) and \(\text{CFP}_{\text{REP}}\) values

<table>
<thead>
<tr>
<th>(N_p)</th>
<th>(T_f = 200,\text{ms})</th>
<th>(T_f = 150,\text{ms})</th>
<th>(T_f = 100,\text{ms})</th>
<th>(T_f = 75,\text{ms})</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.293 173</td>
<td>0.487 126</td>
<td>0.399 100</td>
<td>0.532 75</td>
</tr>
<tr>
<td>4</td>
<td>0.246 184</td>
<td>0.528 130</td>
<td>0.502 91</td>
<td>0.532 75</td>
</tr>
<tr>
<td>6</td>
<td>0.242 181</td>
<td>0.569 135</td>
<td>0.520 90</td>
<td>0.532 75</td>
</tr>
<tr>
<td>8</td>
<td>0.244 182</td>
<td>0.412 125</td>
<td>0.538 89</td>
<td>0.532 75</td>
</tr>
<tr>
<td>10</td>
<td>0.241 184</td>
<td>0.353 132</td>
<td>0.519 90</td>
<td>0.532 75</td>
</tr>
<tr>
<td>12</td>
<td>0.453 168</td>
<td>0.443 126</td>
<td>0.510 78</td>
<td>0.532 75</td>
</tr>
<tr>
<td>14</td>
<td>0.380 165</td>
<td>0.604 139</td>
<td>0.504 85</td>
<td>0.532 75</td>
</tr>
<tr>
<td>16</td>
<td>0.520 162</td>
<td>0.507 92</td>
<td>0.399 100</td>
<td>0.532 75</td>
</tr>
<tr>
<td>18</td>
<td>0.346 148</td>
<td>0.506 108</td>
<td>0.399 100</td>
<td>0.532 75</td>
</tr>
<tr>
<td>20</td>
<td>0.385 152</td>
<td>0.482 125</td>
<td>0.399 100</td>
<td>0.532 75</td>
</tr>
</tbody>
</table>

The optimum values of \(\text{CFP}_{\text{MAX}}\) are fairly variable, especially for larger values of \(D\) and the smaller values of \(N_p\). This variability seems to occur mainly when the objective is most flat: in that it does not vary much over a wide range of \(\text{CFP}_{\text{MAX}}\) values. This means the instability happens in exactly the situations where choosing a precise value of \(\text{CFP}_{\text{MAX}}\) is least important. The \(\text{CFP}_{\text{REP}}\) optima tend to be close to the maximum \(D\), especially for smaller \(D\) where the constraints do not permit much variation anyway. For larger \(D\), the optimum values are significantly smaller than \(D\), this is in line with the fact that there is much more potential to fit the polled and contention periods within a smaller repetition time.

The OPNET™ simulation tool has been used to model a centrally coordinated IEEE802.11 system with the parameters given in Table 1. Simulations were run for a variety of \(\text{CFP}_{\text{REP}}\) values, providing throughput, delay and delay jitter values in each case for both polled and contending traffic. By examining these results, it is possible to identify several points of interest, such as the smallest \(\text{CFP}_{\text{MAX}}\) value at which required throughput for polled traffic can be achieved, the largest \(\text{CFP}_{\text{MAX}}\) value at which the required throughput threshold for contending traffic can be maintained, and the resulting delay characteristics in each case. The \(\text{CFP}_{\text{MAX}}\) values predicted by the optimization process fall within these upper and lower bounds for \(\text{CFP}_{\text{MAX}}\), justifying the values obtained.

By means of comparison against the benchmark for these preliminary results, Table 3 gives the upper and lower bounds from the simulation alongside the values predicted by both the optimization method and the benchmark results [3]. Whilst the comparative benchmark results consistently sit toward the upper end of the range established by these simulation results, the values derived by the optimization approach are predominantly in the middle of the observed range of possible \(\text{CFP}_{\text{MAX}}\) values.

Table 3: Derived \(\text{CFP}_{\text{MAX}}\) vs. Simulation Results (12 Station case)

<table>
<thead>
<tr>
<th>Delay (ms)</th>
<th>Simulation Lower</th>
<th>Simulation Upper</th>
<th>Benchmark [3]</th>
<th>Optimization (Table 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>75</td>
<td>0.55</td>
<td>0.75</td>
<td>0.7</td>
<td>0.532</td>
</tr>
<tr>
<td>100</td>
<td>0.45</td>
<td>0.55</td>
<td>0.7</td>
<td>0.51</td>
</tr>
<tr>
<td>150</td>
<td>0.3</td>
<td>0.45</td>
<td>0.6</td>
<td>0.443</td>
</tr>
<tr>
<td>200</td>
<td>0.3</td>
<td>0.5</td>
<td>0.6</td>
<td>0.453</td>
</tr>
</tbody>
</table>

V. CONCLUSIONS

We have presented an application of non-linear optimization that fine-tunes the structure of the super-frame in centralized WLAN applications. This is of particular interest because the latest WLAN data rates make supporting A/V traffic attractive, but centralized control functions are then required to meet the strict QoS requirements.

We have focused on the objective function and the constraints that bound it, along with the operations required to reach the standard form amenable to optimization by the Barrier Method.

The previously published work in this area [3] has the significant limitation that the minimum sizes of the CFP and CP are not taken into account, which would severely hinder its usage in a real implementation. This work offers a more viable solution for that reason alone. In addition to this, our preliminary results show agreement between simulated system behavior and the optimal values predicted by the non-linear optimization approach.

The optimized super-frame configurations are applicable for a range of delivery delay requirements and numbers of stations, and are fixed for a particular physical layer and for given application traffic characteristics; both of these aspects are either known or can be bounded at design time.

In terms of future work, further simulation to consider situations outside of the stated assumptions (e.g. with more realistic collisions and back-offs, or when terminals are not fully backlogged) would be valuable in justifying the superframe configurations under more realistic conditions. The utility function can be further developed to incorporate a bias term to favor either CFP or CP, depending on the priority of the system. The heterogeneity and time-varying nature of the application traffic profiles must also be considered. Lastly, more detailed comparisons of the performance of this approach against the performance of other published works are required, along with complexity and performance analysis of the dynamic re-optimization of revised utility functions when the parameters change.

REFERENCES