Cross-Layer Optimization for Opportunistic Multi-MAC Aggregation

Tariq Elkourdi, Amith Chincholi, Tan Le, Alpaslan Demir
InterDigital Communications LLC
Melville, NY, USA
E-mail: firstname.lastname@interdigital.com

Abstract—In this paper, we describe a RAT (radio access technology) agnostic approach referred to as OMMA (Opportunistic Multi-MAC Aggregation) that can aggregate multiple RATs operating over independent spectral bands. Aggregation and/or traffic shaping is performed above the air interface protocol stacks but below the IP layer. We investigate the problem of minimizing the average packet latency (the sum of queuing delay and serving delay) under certain constraints and propose an algorithm to compute optimal packet distribution ratio across RATs to be implemented by OMMA. Moreover, we statistically characterize the reordering delay for the OMMA system and propose an algorithm to compute a modified optimal packet distribution ratio which minimizes the packet end-to-end delay, i.e., the sum of average packet latency and average packet reordering delay. Finally, some numerical results are presented to study the average packet latency, average packet reordering delay and average packet end-to-end delay for different system parameters.

Keywords—Multi-radio systems, packet scheduling, traffic shaping strategies.

I. INTRODUCTION

The concept of multi-radio access technology has been widely studied both in the industry and academia due to the proliferation of diverse radio access technologies [1]-[4]. With the availability of a multitude of radio access technologies and their relatively low cost, the idea of using multiple radios simultaneously has gained a lot of momentum as a viable approach for higher throughput, diversity and security. For instance, it was shown in [2] that exploiting multi-radio diversity via opportunistic scheduling significantly increases spectrum usage. Multi-radio transmission diversity was also shown to resolve some critical implementation challenges that throttle the throughput of TCP flows over wireless channels [4].

This paper describes a system and method that can distribute IP packets across two or more RATs operating over independent spectral bands. This paper will show that the IP throughput can be significantly enhanced when compared to the case when a single RAT is used. This is achieved by introducing an aggregation module called opportunistic multi-MAC aggregation (OMMA) which resides above the air interface protocol stacks but just below the IP layer and is common to all RATs in the device.

In a typical IEEE 802.11 Wi-Fi system, the access point (AP) and associated stations (STAs) typically use a single flavor of the IEEE 802.11 system (11a/b/g/n) over a specific ISM band (2.4GHz or 5GHz) when communicating with each other. Instead, a system that implements OMMA may have a multi-band, and multi-RAT AP and STAs that support simultaneous operation over the 2.4GHz ISM band using the IEEE 802.11n protocol, 5.0GHz ISM band using the IEEE 802.11ac protocol stack and an aggregation of multiple TVWS bands (512-698MHz) using the IEEE 802.11af proposal.

The OMMA scheduler within the OMMA layer is responsible for traffic shaping of IP packets across RATs. It can operate in one of two possible modes: i) selection mode, in which all packets are forwarded to a single RAT based on a predefined rule; and ii) multiplexing mode in which the main stream of packets is dynamically split into sub-streams each of which is forwarded to a corresponding RAT. The splitting is based on the rate of traffic generation of the main stream and the channel quality statistics such as medium access delay, frame error rate, average data rate and RSSI. More precisely, these statistics are collected by each RAT and fed back to OMMA to update the portion of main stream traffic that is forwarded to each RAT.

Analysis of the optimal band and power allocation that maximizes the total multi-RAT system capacity can be found in [3]. The main contribution of this paper is to investigate the problem of optimally distributing IP packets across RATs based on minimizing the average packet latency, which
consists of the average queuing delay and the average propagation and transmission (serving) delay, under certain constraints. Additionally, this paper provides a statistical characterization of average packet reordering delay and provides an optimal traffic shaping rule to minimize the average packet end-to-end delay, which consists of the average packet latency and the average packet reordering delay.

II. SYSTEM MODEL

We consider a wireless system that consists of a network terminal (NT), e.g., Wi-Fi access point, and a number of user terminals (UTs), e.g., Wi-Fi stations (STAs). Both NT and UT have the capability of supporting multiple RATs where all RATs operate on different spectral bands. Bands are orthogonal so signals on different bands do not interfere with each other. Transmitting terminals operate in OMMA multiplexing mode. For simplicity of analysis, we study the case of a single NT and a single UT in the downlink. The uplink scenario and extension to multiple STAs by modeling medium access delays are omitted for brevity. At the NT, a stream of packets is generated in a Poisson fashion with an average packet arrival rate \( \lambda \) (packets/unit time). The main stream is then split into \( K \) sub-streams each of which is assigned to a corresponding transmit buffer in each RAT. The \( i \)th sub-stream, \( i \in \{1, \ldots, K\} \), is therefore also Poisson with an average arrival rate \( \lambda_i \) (packets/unit time) where

\[
\lambda = \sum_{i=1}^{K} \lambda_i. \tag{1}
\]

We model our multi-RAT system in each terminal device as \( K \) parallel M/M/1 queues \cite{7} to account for the fact that transmit buffers can possibly have different serving rates. Packets are transmitted over channel \( i \) in a first-come-first-served (FCFS) basis and are received successfully at serving rate \( R_i \) (packets/unit time). To assure system stability we impose the constraint

\[
\sum_{i=1}^{K} R_i > \lambda. \tag{2}
\]

Rate \( R_i \) reflects the quality of channel \( i \) and is fed back from each RAT to the transmitting OMMA layer to update its traffic shaping strategy. At the receiver, the sub-streams are combined in the receiving OMMA layer.

III. PROBLEM STATEMENT AND PERFORMANCE ANALYSIS

Based on the system model described above, we seek to answer the question of how to distribute packets across RATs such that the average end-to-end delay is minimized. Placing all packets in the transmit buffer of the RAT with the lowest latency may increase the average packet queuing delay. On the other hand, dispersing them across all RATs may decrease the average packet queuing delay and serving delay, however it may result in out-of-order reception of packets at the receiver due to differences in link latencies. This can cause longer queuing delays at the receiver to rearrange packets before sending them up to the IP layer. In the following, we analyze the packet latency and reordering delay and propose an optimal operational point that minimizes the sum of both delays, i.e., end-to-end delay.

A. Average Packet Latency

In this section, we study the problem of finding the optimal average arrival rate \( \lambda_i \) for RAT, \( i \in \{1, \ldots, K\} \) such that the average system packet latency \( \tau_s \) is minimized. The latency for each packet in queue \( i \) is given as \cite{7}

\[
\tau_i = \frac{1}{R_i - x_i}, \tag{3}
\]

where

\[
0 \leq x_i \leq R_i. \tag{4}
\]

Therefore, the average system packet latency, i.e., the objective function, can be expressed as

\[
\tau_L(x_1, \ldots, x_K) = \sum_{i=1}^{K} \tau_i P_i = \frac{\sum_{i=1}^{K} x_i}{R_i - x_i}, \tag{5}
\]

where \( P_i = \frac{x_i}{x} \) is the probability of sending packets to RAT \( i \). The optimization problem can be formulated as follows

\[
\text{minimize}_{x_1, \ldots, x_K} \tau_L(x_1, \ldots, x_K), \tag{6a}
\]

\[
\text{subject to } x = \sum_{i=1}^{K} x_i, \tag{6b}
\]

\[
x_i \geq 0, \tag{6c}
\]

\[
x_i \leq R_i \quad \forall i. \tag{6d}
\]

\( \tau_L(x_1, \ldots, x_K) \) is convex over the supportable rate region \([x - \sum_{i=1}^{K} R_i]^+ \leq x_i \leq \min\{x, R_i\}\) defined by (2) and (4) and hence the Lagrangian approach can be used. The Lagrangian is given as

\[
L(x_1, \ldots, x_K, \lambda, \lambda_1, \ldots, \lambda_K, \hat{\lambda}_1, \ldots, \hat{\lambda}_K) = \tau_L(x_1, \ldots, x_K) + \lambda \left( x - \sum_{i=1}^{K} x_i \right) - \sum_{i=1}^{K} \lambda_i x_i + \sum_{i=1}^{K} \hat{\lambda}_i (x_i - R_i), \tag{7}
\]

where \( \lambda, \lambda_i \) and \( \hat{\lambda}_i \) are the Lagrange multipliers. By taking the partial derivatives of the Lagrangian with respect to \( x_i \) for \( i \in \{1, \ldots, K\} \) and making them equal to zero, we obtain

\[
\frac{d}{dx_i} L(x_1, \ldots, x_K) = \frac{R_i}{(R_i - x_i)^2} - \lambda - \lambda_i + \hat{\lambda}_i = 0, \forall i \in \{1, \ldots, K\}, \tag{8}
\]

where \( x_i^* \) is the optimal arrival rate for RAT \( i \). Solving for \( x_i^* \), we finally obtain (8).
The updated Lagrange multipliers $\lambda^{k+1}, \lambda_i^{k+1}$ and $\bar{\lambda}_i^{k+1}$ are calculated according to

$$\lambda^{k+1} = \left[ \lambda^k + \gamma \left(\sum_{i=1}^{K} x_i - R \right) \right]^+, \forall i \quad (9)$$

$$\lambda_i^{k+1} = \left[ \lambda_i^k - \gamma_i (x_i^k) \right]^+, \forall i \quad (10)$$

$$\bar{\lambda}_i^{k+1} = \left[ \bar{\lambda}_i^k + \gamma_i (x_i - R_i) \right]^+, \forall i \quad (11)$$

where $\gamma, \gamma_i$ and $\bar{\gamma}_i$ are the step sizes, $R = \sum_{i=1}^{K} R_i$ and $[x]^+ = \max\{x, 0\}$. For the simple case of $K=2$, since $x_1^k + x_2^k = x$, if $R_2$ is set to 0 we have $x_1^k = R_1$, which implies that all packet will be forwarded to RAT 1. Moreover, if $R_1 = R_2$, then $x_1^k = x_2^k = \frac{x}{2}$ which implies that if both channels have equal performance then packets are distributed equally. We propose the following algorithm to compute $x_i^*$. 

**Algorithm I: Minimum Average Packet Latency**

1. Choose feasible initial multipliers $\lambda^{(k)}, \lambda_i^{(k)}, \bar{\lambda}_i^{(k)}$ with initial index $k = 0$ and $\tau = \inf \{\tau_1, \tau_2, \ldots, \tau_K\}$.
2. Calculate $x_i^{(k)}$ according to (8);
3. if $(x_i^{(k)} < 0)$ or $(x_i^{(k)} > R_i)$
   - Set $k \leftarrow k+1$ and update $\lambda^{(k)}, \lambda_i^{(k)}, \bar{\lambda}_i^{(k)}$ according to (9), (10), (11) then go to Step 2;
4. else
   - if $\tau_1 (x_1^{(k)}, \ldots, x_K^{(k)}) < \tau$
     - Set $x_i^* = x_i^{(k)}$ and $\tau = \tau_1 (x_1^{(k)}, \ldots, x_K^{(k)})$;
     - Set $k \leftarrow k+1$ and update $\lambda^{(k)}, \lambda_i^{(k)}, \bar{\lambda}_i^{(k)}$ according to (9), (10), (11) then go to Step 2;
   - else
     - Stop;
   end if
end if

This algorithm converges to the optimal value $x_i^*$ as given by (8).

**B. Average Packet Reordering Delay**

In this section, we calculate the reordering delay based on the solution (8). Assume $N$ packets are placed in the transmit queue and indexed sequentially to indicate each packet’s position relative to all other packets inside the queue. A packet of an arbitrary index $n \in \{1, \ldots, N\}$ is transmitted over RAT$h$, $h \in \{1, \ldots, K\}$, with a probability $P_h$, where $\sum_{h=1}^{K} P_h = 1$. This packet will experience a certain delay and upon reception its position will be offset by $\varepsilon \in \{-N + 1, \ldots, -2, N - 1\}$, as shown in Fig. 2. We say that a packet is received in the correct order if $\varepsilon = 0$ and is out-of-order otherwise. A key observation here is that reordering delay happens either when packets with indices lower than $n$ are received late or when packets with indices higher than $n$ are received early. This can happen in one of the following cases: i) Packet $n$ is transmitted over the link with the higher latency whereas $\delta$ subsequent packets are transmitted over the lower latency link as shown in Fig. 3-(a); ii) Packet $n$ is transmitted over the link with lower latency whereas $\delta$ preceding packets are transmitted over the higher latency link as shown in Fig. 3-(b). For instance, consider the scenario where packets 1, 2, 3, 4 and 5 are ready for transmission over a two-RAT system RAT 1 and RAT 2, where RAT 1’s link is faster than RAT 2’s link. Assume packet 1 is scheduled for transmission over RAT 1 whereas all (or some) of the subsequent packets, that have an index that belongs to the set $\{2, 3, \ldots, 1+\delta\}$ are assigned to the RAT 2 (case (ii)). At the receiver side, packets may be received out of their original order (e.g., 2, 3, 4, 5, 1) causing a reordering delay. In this case the reordering delay that packet 1 experiences is $\frac{\delta}{R_2}$.

We calculate the average packet reordering delay based on the probability mass distribution of the offset $\varepsilon$. For simplicity, assume a two-RAT system, i.e., RAT 1 and RAT 2. Packets sent over RAT 1 and RAT 2 will experience a delay $d_1 = \frac{1}{R_1}$ and $d_2 = \frac{1}{R_2}$, respectively. Without loss of generality, assume that $d_2 > d_1$, then the delay difference is $d = d_2 - d_1$. Therefore, a packet on RAT 2 will arrive after

$$\delta = \begin{cases} 
|dx_1| & \text{if } x_1 \geq x_2 \\
|dx_2| & \text{otherwise}
\end{cases} \quad (12)$$

Throughout this paper, packet of position index $n$ is referred to as packet $n$. 

---

Fig. 2. Packet of position index $n$ is received with an ordering offset $\varepsilon$. 

---

*(Throughout this paper, packet of position index $n$ is referred to as packet $n$.*

---
A subset of $\delta$ packets

A subset of $\delta$ packets

OMMA

RAT1 (low latency)

RAT 2 (high latency)

OMMA

RAT1 (low latency)

RAT 2 (high latency)

Fig. 3. Packet reordering cases: (a) Packet $n$ is transmitted on the link with higher latency whereas $\delta$ subsequent packets are transmitted on the lower latency link; (b) Packet $n$ is transmitted on the link with lower latency whereas $\delta$ preceding packets are transmitted on the higher latency link.

of its subsequent packets transmitted on RAT$_1$. Note that if packet $n$ is scheduled to be transmitted over the faster link then the packets after packet $n$ and before packet $n + \delta$ in the transmitter’s main queue will not cause a reordering issue regardless of their RAT assignment. Conversely, if packet $n$ is scheduled to be transmitted over the slower link then packets before packet $n$ and after packet $n + \delta$ in the transmitter’s main queue will not cause reordering issues regardless of their RAT assignment as shown in Fig. 3.

For case (i), the conditional probability mass function of $\mathcal{E}$ can be written as $\Pr [\mathcal{E} = \varepsilon |$ packet $n$ is transmitted on RAT$_2] = \left( \frac{\delta}{\varepsilon} \right) \left( \frac{\varepsilon}{\varepsilon} \right)^{\delta - \varepsilon}$, and similarly for case (ii), we have $\Pr [\mathcal{E} = - \varepsilon |$ packet $n$ is transmitted on RAT$_1] = \left( \frac{\delta}{|\varepsilon|} \right) \left( \frac{|\varepsilon|}{|\varepsilon|} \right)^{\delta - |\varepsilon|}$. Note that the average serving time for $\varepsilon$ packets transmitted on RAT$_1$ is $\frac{\varepsilon}{R_1}$ and the average serving time for a single packet transmitted on RAT$_2$ is $\frac{1}{R_2}$.

Therefore, the average packet reordering delay can be calculated using the law of total probability as

$$
\tau_R = \sum_{\varepsilon=1}^{\delta} \mathcal{E} \left( \frac{\delta}{\varepsilon} \right) \left( \frac{\varepsilon}{\varepsilon} \right)^{\delta - \varepsilon + 1} + \sum_{\varepsilon=-\delta}^{-1} \frac{1}{R_2} \left( \frac{\delta}{|\varepsilon|} \right) \left( \frac{|\varepsilon|}{|\varepsilon|} \right)^{\delta - |\varepsilon| + 1}
$$

Arrival rate for RAT $i$ that minimizes (13) is denoted as $\hat{x}_i^*$. 

C. Average End-to-End Delay

We recall that the average packet end-to-end delay $\tau_E$ is the sum of the average packet latency $\tau_L$ and the average reordering delay $\tau_R$ experienced at the OMMA layer. Therefore, an optimal design aims to minimize that sum rather than just one of its components. The optimal packet arrival rate for RAT $i$ that minimizes the average packet end-to-end delay, i.e., $\tau_E = \tau_L + \tau_R$, can be formulated as follows:

$$
\hat{x}_i^* = \arg \min_{x_i} \tau_L + \tau_R, \forall i
$$

subject to $x = \sum_{i=1}^{K} \hat{x}_i^*$

$$
\hat{x}_i^* \geq 0, \forall i
$$

We propose the following algorithm based on the Golden Section Search technique [5] to solve for (14) given (8) and (12).

Algorithm II: Minimum Average End-to-End Delay

1. Initialize $x_{\min} = a, x_{\text{opt}} = b, x_{\max} = c$, where $a$ is the minimum supportable rate, $b$ is the optimal split $x_{\text{opt}}$ of Algorithm I and $c$ is the maximum supportable rate.

2. if $x_{\max} - x_{\text{opt}} > x_{\text{opt}} - x_{\min}$

   Set $d = x_{\text{opt}} - \omega (x_{\max} - x_{\text{opt}})$

   else Set $d = x_{\text{opt}} - \omega (x_{\text{opt}} - x_{\min})$

end

3. if $|x_{\max} - x_{\min}| < \rho |(x_{\text{opt}} - d)|$

   Set $\hat{x}^* = x_{\max} + x_{\min}$

   else Stop the algorithm

end

4. if $\tau_E (\hat{x}_i^*) < \tau_E (b)$

   if $x_{\max} - x_{\text{opt}} > x_{\text{opt}} - x_{\min}$

      Set $x_{\min} = b, x_{\text{opt}} = d$

   else $x_{\text{opt}} = d,$ $x_{\min} = c$

end

5. else if $x_{\max} - x_{\text{opt}} > x_{\text{opt}} - x_{\min}$

   Set $x_{\min} = d$

else $x_{\max} = d$

end

6. end if and go to step 3

Where $\omega$ and $\rho$ are constants. This algorithm will iteratively converge to $\hat{x}_i^*$. The speed of convergence depends on the initial value (8) and the step size dictated by (13).

IV. NUMERICAL RESULTS

In this section, we present some numerical results to study the average packet latency, average packet reordering delay and average packet end-to-end delay for different system parameters. Fig. 4 shows the arrival rates $x_1^*$ and $x_2^*$ for $R_1 = 4, 6, 8$. In general, the optimal minimum latency solution aims to balance the load on each RAT by forwarding a portion of the total load that is proportional to the RAT’s serving rate. For instance, for a fixed $R_2$, as $R_1$ increases, $x_1^*$ increases and $x_2^*$ decreases correspondingly. It is also seen that when $R_1 = R_2 = 8$, the optimal minimum latency solution splits the total load equally between the two RATs. This is due to the fact that an equal split minimizes queuing time for each packet when serving rates are equal.

Fig. 5 plots the minimum packet latency arrival rate $x_1^*$, minimum packet reordering delay arrival rate $\hat{x}_1^*$ and minimum end-to-end delay arrival rate $\hat{x}_1^*$ vs. overall main stream arrival rate $x$ for $R_1 = 8$ and $R_2 = 4$. As it can be seen, the minimum latency is achieved via selection mode by
using RAT 1 for light load ($x<2.2$) and via multiplexing mode by exploiting RAT 1 and RAT 2 simultaneously for heavier loads ($x>2.2$). On the other hand, minimum reordering delay can be achieved via selection mode using RAT $i$, $i \in \{1,2\}$ for $x \leq R_i$. For $x > R_i$, traffic is split via multiplexing mode by forwarding $R_i$ portion of the traffic to RAT $i$ and $(x-R_i)$ portion to RAT $j$, $j \in \{1,2\}$ and $j \neq i$. As for the minimum end-to-end delay, we notice that the optimal portions are different from that of the other two approaches, and tend to converge asymptotically to the minimum latency solution as load increases.

Fig. 6 plots the latency $\tau_L$, reordering delay $\tau_R$ and end-to-end delay $\tau_E$ versus $x_i$ for the case of $x_{1}=5$, $R_{L}=6$, $R_{G}=5$ and supportable rates $x_{1} \in [0,5]$ and $x_{2} \in [0,5]$. It can be seen that $x_{1}^{*}$ that achieves the minimum latency splits traffic almost in half between the two RATs. This solution encourages the exploitation of both RATs simultaneously which in turn minimizes the amount of average packet queuing time as compared to forwarding all traffic to a single RAT. The minimum reordering delay solution $x_{1}^{*}$, however, suggest forwarding all traffic to one RAT or the other so that all traffic is transmitted and received in sequence. As it is seen from the figure, the two solutions, i.e., the one that minimizes latency and the one that minimizes reordering delay, need not coincide with each other. A joint optimal split $x_{1}^{*}$ exists, which minimizes the sum delay.

V. CONCLUDING REMARKS

In this paper, we proposed a system and method (OMMA) to distribute IP packets across two or more RATs operating over independent spectral bands. This paper showed that OMMA layer provides a means of aggregating and exploiting multiple bands to enhance overall system performance. We derived the optimal packet distribution strategy based on minimizing the average packet latency under certain constraints. We presented a statistical characterization of the reordering delay experienced by packets at the receiver’s OMMA layer and realized that the optimal packet distribution strategy resulted in a large reordering delay. To address this, we proposed a jointly optimal packet distribution algorithm which minimizes the packet end-to-end delay. We finally presented numerical results to study the average packet latency, average packet reordering delay and average packet end-to-end delay.

REFERENCES


