Multimedia Resource Allocation in mm-Wave 5G Networks

Sandra Scott-Hayward and Emiliano Garcia-Palacios
ECIT, Queen’s University Belfast

ABSTRACT

The 5G network infrastructure is driven by the evolution of today’s most demanding applications. Already, multimedia applications such as on-demand High Definition (HD) video and Internet Protocol Television (IPTV) require Gigabit-per-second (Gbps) throughput and low delay while future technologies include Ultra HDTV and Machine-to-Machine communication. Mm-Wave technologies such as IEEE 802.15.3c and IEEE 802.11ad are ideal candidates to deliver high throughput to multiple users demanding differentiated Quality of Service (QoS). Optimization is often used as a methodology to meet throughput and delay constraints. However, traditional optimization techniques are not suited to a mixed set of multimedia applications. Particle Swarm Optimization (PSO) is shown as a promising technique in this context. Channel-Time Allocation PSO (CTA-PSO) is successfully shown here to allocate resource even in scenarios where blockage of the 60 GHz signal poses significant challenges.

1 Introduction

The volume of mobile traffic is exploding driven by a proliferation of connected devices. The high bandwidth required by multimedia applications is severely stretching the available wireless spectrum. Consumers influenced by fixed access broadband services expect almost instant file uploads/downloads and transfers between terminals (laptops, tablets and smart TVs), IPTV and HD video. The challenge for service providers is therefore to extend the capability of wireless access to maintain performance characteristics in the transition to predominantly wireless service provision. The entertainment industry will be just the first adopter to drive and inspire other futuristic scenarios in connected health, social and community interaction, and education.

The 60 GHz band, with up to four 2.16 GHz channels, supports high data rate, short range, Line-of-Sight (LOS) directional transmissions. Mm-Wave communication technologies such as IEEE 802.15.3c-2009 [1] and IEEE 802.11ad [2] are uniquely positioned due to their ability to deliver the Gigabit-per-second (Gbps) throughput that “5G” envisages.

Despite the adverse propagation characteristics of mm-Wave technologies (a result of high propagation loss due to oxygen absorption and atmospheric attenuation), the fact that waves
are confined within walls makes it ideal for personal communications: promoting security and privacy. This also promotes a more efficient frequency re-use, therefore creating very high bandwidth hubs that are ideal for the delivery of Gbps applications. A major question, however, is how to combat the impact of blockage of the LOS link by human shadowing or due to obstacles, which severely impacts the transmission?

In order to support the high volume of high quality services, optimized resource allocation is required. Optimizing resource allocation in a multi-user multimedia gigabit scenario poses interesting challenges. Most traditional optimization techniques have been used to solve convex optimization problems. However, real-time multimedia applications introduce non-convex utility functions in which the perceived quality by the user does not increase gracefully with an increase in throughput. A re-think of resource allocation optimization techniques is necessary in this context. Relatively new approaches, such as Particle Swarm Optimisation (PSO), are ideal candidates to solve non-convex problems. Our proposal compares PSO to other techniques and presents Channel-Time Allocation PSO (CTA-PSO) as a solution to optimize channel time allocation (resource) even when blockage occurs in the mm-Wave network.

2 Enabling mm-Wave technologies

IEEE 802.15.3c [1] and IEEE 802.11ad [2] provide mechanisms at the Medium Access Control (MAC) layer to allocate resource (in the form of channel access time) to multiple users. A period of contention allows for different terminals to request access time while allocated channel time slots provide dedicated data transmission time. The advantage of scheduled access is to provide a dedicated timeslot for communication thus supporting guaranteed Quality-of-Service (QoS) for an application. This guarantee is not possible when all network devices compete for bandwidth using random access techniques. The allocation process is described here and illustrated in Fig. 1.

2.1 Allocating Resource in IEEE 802.15.3c

An 802.15.3c network, or piconet, is a wireless ad-hoc data communications system in which a number of independent devices communicate with each other in a peer-to-peer fashion [1]. In the 802.15.3c protocol, medium access is controlled by a Piconet Controller (PNC). This role is usually held by the first member of the network. The superframe is initiated by a beacon from the PNC. In order to achieve the required signal range at 60 GHz, a directional beacon is employed, but in order to reach all potential neighbours, the beacon must be transmitted in all directions. A quasi-omni beacon is therefore used. The area around the PNC is divided into sectors and the PNC transmits a directional beam beacon sector-by-sector until all sectors have been covered.

2
A Contention Access Period (CAP) follows the beacon phase. The S-CAP is sectorized so in the Association S-CAP, all devices in a sector have the opportunity to contend to send association request commands and for the PNC to send immediate acknowledgments. The regular S-CAP and CAP are then available for command and data exchanges.

The CAP is a Carrier Sense Multiple Access - Collision Avoidance (CSMA/CA) phase. In the CSMA/CA protocol, a device first senses the channel prior to transmission. If the channel is free, the device transmits its frame. Otherwise, if the channel is sensed busy, the device generates a random back-off counter and defers its transmission until the back-off counter expires.

In 802.15.3c, devices compete during the CAP to transmit a request for dedicated channel time during which the application data can be transmitted. The device time allocation is called a Channel Time Allocation (CTA) and occurs during the Channel Time Allocation Period (CTAP), which follows the CAP. The CTAP operates with Time Division Multiple Access (TDMA). In this phase, a time slot/CTA is assigned to any devices that have requested an access period, provided the time is available.

In order to improve the MAC efficiency, frame aggregation and block acknowledgment are supported by the standard. These techniques are also illustrated in Fig. 1. Frame aggregation involves mapping MAC Service Data Units (MSDUs) into multiple subframe payloads. This increases the data to header ratio in the frame thus increasing throughput. A Block Ac-
knowledge (Blk-Ack) is sent for the aggregated frame. The benefit of these mechanisms is improved efficiency by reducing frame/ack overhead.

2.2 Allocating Resource in IEEE 802.11ad

Accommodating the constraints of the mm-Wave frequency band and supporting the requirements of high data rate applications, the IEEE 802.11ad protocol supplements and extends earlier versions of IEEE 802.11 MAC. It operates in a similar way to the 802.15.3c protocol with periods of contention-based and scheduled access. However, the allocation of these transmission periods differs from the 802.15.3c approach, as illustrated in Fig. 1.

Channel access time in IEEE 802.11ad is divided into beacon intervals (BIs). Each BI consists of the beacon transmission interval (BTI), association Beam-Forming training (ABFT), announcement transmission interval (ATI) and data transfer interval (DTI). During the BTI, the control/access point transmits directional beacons to each sector of the sectorized network. Request-response-based transmission allocations of scheduled service periods (SPs) and Contention Based Access Periods (CBAPs) take place in the ATI. Finally the DTI consists of a series of SPs and CBAPs during which application data transmissions occur.

The importance of the compatibility of IEEE 802.11ad with legacy 802.11 standards is clear. A transition will be possible between the lower frequency band (2.4/5 GHz) for long range communication and the higher frequency band (60 GHz) for short range communication [3]. It is this heterogeneous network style that will underpin the 5G infrastructure.

While both protocols described here combine CSMA/CA for contention-based medium access and TDMA for scheduled service, they differ in the order of allocation. In IEEE 802.15.3c, the CTA requests are sent during an initial contention-based period (CAP) and the CTAP consumes the majority of the superframe with individual timeslots (CTAs) allocated for application data transmission. In contrast, the contention-based access periods (CBAPs) and scheduled service periods (SPs) are alternated in IEEE 802.11ad to support compatibility with legacy 802.11 standards. The objective of both methods is to maximize network throughput. Optimizing resource allocation is key to maximizing network throughput.

One approach to resource allocation for multiple applications is the assignment of packets to queues where each queue has a priority level for accessing the wireless medium. For example, the packets of a real-time, low-latency application would receive the highest priority in accessing the available bandwidth. However, although this achieves prioritization between application types, it does not consider the specific (and variable) requirements of individual devices within an application type and how throughput/QoS might be optimized by appropriate scheduling of these transmissions. It is our belief that both the application type and the specific, individual application requirements must be considered for optimal resource allocation.
In our consideration of the channel time allocation optimization problem, we exploit the MAC layer structure of IEEE 802.15.3c and IEEE 802.11ad. The contention periods are used for the exchange of time-slot request-response messages. The contention-free/scheduled service periods are dedicated to application data transmission with defined QoS.

3 Optimization Techniques

Optimization involves finding the best solution to a problem. Mathematically, this involves maximization/minimization of an objective/utility function, which may be unconstrained or it may be subject to certain constraints on the variable(s) of the function.

The nature of the utility function for multimedia applications determines the type of optimization problem [4]. Classically, resource management solutions handle video as an isolated application in the network thus solving a convex optimization problem. For example, the higher the data rate at which the video can be transmitted, the better the quality perceived by the user and hence the higher the utility for the user.

In a practical network, however, mixed multimedia applications (e.g. IPTV, Voice over IP (VoIP) etc.) must be served along with video. The utility functions to describe such applications are non-convex. The non-convexity arises from the inelasticity of real-time applications, meaning that the application does not improve or degrade gracefully in response to an increase or a reduction in allocated transmission rate [5]. Rather, a reduction in data rate below a certain threshold results in a significant drop in QoS.

The challenge for 5G is therefore a combination of resolving the non-convex optimization problem (resulting from multiple applications) within the constraints of the mm-Wave network. This includes the practical consideration of the execution time of the algorithm. Framing resource allocation as a Network Utility Maximization (NUM) problem, the implementation of four popular optimization techniques are compared here for their solution potential. Techniques suitable for convex optimization problems only are included in order to illustrate the difference in attributes of each technique, specifically with respect to speed of execution of the algorithm.

The problem is defined in (1).

\[ \text{Maximize } \sum_{i=1}^{N} U_i(CTA_i) \]

Subject to \[ \sum_{i=1}^{N} CTA_i \leq CTAP \]
\[ CL_i \leq CTA_i \leq CH_i \quad \forall \quad i = 1, 2, ..., N \] (1)

NUM is the problem of maximizing the total utility, \( U_i \), of the network over the channel time allocations, \( CTA \), subject to the constraint that the sum of all CTAs should not exceed the...
network capacity (CTAP). In addition, upper and lower CTA limits, \([CL_i, CH_i]\), are set based on the desired quality constraint of the individual application. The result of the NUM will be a time-slot (CTA) allocation for each application, which takes account of both the physical transmission rate on each wireless link based on the channel condition, and the immediate quality requirements of the application.

The optimization techniques are:

1. Lagrangian Dual Decomposition Subgradient Algorithm
2. Rate Allocation Game (Nash Equilibrium via pricing mechanism)
3. Nash Bargaining Solution
4. Particle Swarm Optimization

### 3.1 Lagrangian Dual Decomposition Subgradient Algorithm

The first optimization method considered is the Lagrangian dual decomposition subgradient algorithm. This evolved from the gradient search methods used in convex optimization. With gradient search, the search direction is defined by the gradient of the function to be optimized at the current point; the gradient being the first derivative of the function. For a convex programming problem, the gradient projection controlled by the lagrange multiplier leads to convergence to the optimal solution. However, gradient search methods are inapplicable to non-differentiable convex optimization problems.

A subgradient approach to the lagrangian algorithm has been presented in [6], which enables solution of non-differentiable convex problems. An iterative algorithm is generated replacing the gradient based method. The lagrange multiplier (or price per unit rate in the resource allocation problem) is updated at each iteration. The implementation of the algorithm requires message-passing in order to communicate the lagrange multiplier from each network device to a central calculation point at each iteration. The disadvantage of message-passing is the overhead introduced to the network, which reduces application data transmission time.

### 3.2 Rate Allocation Game

Game theory is a set of mathematical tools used to analyze interactive decision processes [7]. The rate allocation game is non-cooperative in the sense that each device acts as a selfish player in the game. This enables distributed implementation. However, the use of a pricing mechanism handles the conflicting objectives of the wireless devices in the network.

The game is described by a number of devices (players), a vector of strategies/actions and a vector of payoffs based on the strategies chosen by the devices in the network.
The players are rational. This means that they take account of the consequences of their choices and of other players’ choices in order to selfishly maximize their utility. As described in [8], a rational player will use only those strategies that are best responses to some beliefs he might have about the strategies of his opponents. The requests of other devices are taken into account in a cost term describing the price of requesting additional transmission rate. If there are few devices competing for the available resource then the price is low. However, if there is high competition, then the price for requesting more resource is higher. If price increases, satisfaction decreases so an optimum is reached where no user wishes to deviate because their satisfaction will be reduced if the price goes up, which is the consequence of excess demand. The payoff reflects the overall loss/gain that the player incurs based on its selected strategy. The resulting resource allocation is a Nash Equilibrium [8]. An illustration of how the game would be implemented in a mm-Wave network is provided in Fig. 2.

![Figure 2: Implementation of the Rate Allocation Game in the hybrid MAC Framework - The request/response exchange for Transmission Opportunities (TXOPs) takes place in the CAP with the TXOPs provided in the CTAP](image)

### 3.3 Nash Bargaining Solution

The Nash Bargaining Solution (NBS) is another game-theoretic method. In this case, players cooperate to reach a fair allocation of resources. Each player has a minimum resource acceptable to it, known as the disagreement point, $d$. NBS allocates resources optimally by maximizing the Nash product, which is the product of utilities. In the channel time allocation problem, $d$
corresponds to the set of lower CTA limits, CL.

NBS is an axiomatic bargaining solution, which means that it does not require iterative bargaining among users consequently avoiding message passing. However, neither NBS nor Rate Allocation Game are applicable to non-convex functions.

3.4 Particle Swarm Optimization

Direct search methods or evolutionary techniques can be employed to solve non-convex problems. A range of evolutionary techniques have been developed in the past $\sim 40$ years, the first of which was the Genetic Algorithm (GA) developed in 1975 by John Holland and his students.

In 1995, Kennedy and Eberhart introduced their Particle Swarm Optimization (PSO) algorithm for solving global optimization problems [9]. Of the evolutionary algorithms PSO has the appeal of simplicity and evidence of good performance in a variety of application domains. PSO is based on social behaviour with multiple potential solutions of a problem generated at initialization. The solution set is called a swarm and each solution is a particle. The particles move in the problem search space searching for the optimal solution. In a similar manner to the social and cooperative behaviour of species like birds and fish, they exchange their knowledge of the search space to find the best solution by self-learning and collaboration. At each iteration, each particle adjusts its position according to its own experience and the experience of its neighbours. It is a centralized algorithm suitable for solution of non-convex problems.

A 3-dimensional PSO for a three device resource allocation problem is presented in Fig. 3. The convergence of the particles towards the global best position, $G_{best}$, is illustrated. In Fig. 3a the distribution of the swarm of 40 particles across the search space in the early PSO iterations is shown. In Fig. 3b a zoom-in on this search space highlights the connecting line indicating the progression of the global best position over a number of iterations and clustering of the particles around the final $G_{best}$, which is the resource allocation solution for the three device network.
Figure 3: PSO Convergence Illustration in a 3-D Search Space - (a) 1st (black) and 5th (grey) iterations, and (b) final (cyan) iteration with indication of progression towards $G_{best}$ (red)
3.5 Evaluation of Optimization Techniques

The attributes of each optimization technique are outlined in Table 1. The execution time is ordered in terms of increasing requirement from 1 to 4 based on an example execution of 8 devices each running a different video sequence.

The rate allocation game is limited both by its high message-passing overhead and its constraint to convex problems. Although the fastest in terms of execution time, the NBS is also limited to convex optimization problems. As indicated in Table 1, a Lagrangian relaxation approach could be explored for the non-convex resource allocation problem. However, the overhead introduced by message-passing proves a limitation of this approach.

In contrast, if the time required to reach convergence could be reduced, the benefit of centralized implementation of PSO along with its capacity to resolve non-convex optimization problems presents a potential method for solving the multimedia resource allocation problem.
<table>
<thead>
<tr>
<th>Method</th>
<th>Execution Time</th>
<th>Convex Functions</th>
<th>Non-convex Functions</th>
<th>Centralized</th>
<th>Message Passing Required</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lagrangian Dual Decomposition</td>
<td>4</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>Rate Allocation Game</td>
<td>2</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>Nash Bargaining Solution</td>
<td>1</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>Particle Swarm Optimization</td>
<td>3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
</tr>
</tbody>
</table>
4 Channel Time Allocation Particle Swarm Optimization

The results of Table 1 indicate that PSO has potential to resolve the wireless multimedia network resource allocation problem.

PSO has the appeal of simplicity and evidence of good performance in a variety of application domains. It has been demonstrated to be more computationally efficient than GA on a series of test problems [10].

With the focus of achieving a near global optimal solution in a time suitable for implementation in a real network, Channel Time Allocation Particle Swarm Optimization (CTA-PSO) has been developed [4]. CTA-PSO overcomes premature convergence by controlling exploration and exploitation in the search space. Exploration refers to the ability of the swarm to explore different regions of the search space in order to locate the global optimum. Exploitation refers to the ability of the particles to concentrate the search around a promising area of the search space in order to refine a potential solution.

The parameters in the PSO equation (2) contribute to this.

\[
V_{t+1}^i = \omega V_t^i + c_1 r_1 (P_{best}^i - CTA_t^i) + c_2 r_2 (G_{best}^i - CTA_t^i) \\
CTA_{t+1}^i = CTA_t^i + V_{t+1}^i
\]  

At each iteration, each particle’s velocity, \( V_t^i \), and channel time allocation, \( CTA_t^i \), is updated. The dimension of \( CTA_t^i \) is equal to the number of devices to be allocated resource in the network.

In (2), \( t \) is the iteration number, \( r_1 / r_2 \) are uniform random numbers, and \( c_1 / c_2 \) are learning rates/acceleration coefficients representing the weight of memory of a particle’s best position, \( P_{best}^i \), towards the memory of the swarm best position, \( G_{best}^i \). \( \omega \) is the inertia weight, controlling the contribution of the previous velocity to the velocity update.

A particle keeps track of its coordinates in the search space and aims to reach \( G_{best} \). The best solution is determined by the value of the fitness function, which in the resource allocation problem \( F \) is the utility function to be maximized.

The PSO fitness function, \( F \), for the CTA problem is described in (3).

\[
F = \begin{cases} 
\sum_{i=1}^{N} U_i (CTA_i) & \text{if} \sum_{i=1}^{N} CTA_i \leq CTAP, \\
\sum_{i=1}^{N} U_i (CTA_i) + \gamma \left( CTAP - \sum_{i=1}^{N} CTA_i \right) & \text{otherwise},
\end{cases}
\]  

where the penalty value, \( \gamma > 0 \). The penalty value accommodates the practical constraint that the sum time allocated must not exceed the available resource i.e. the \( CTAP \).

CTA-PSO monitors similarity in the swarm and achieves the global solution fast by introducing particles to increase diversity and reducing computation within the PSO. Increasing the
diversity in the swarm avoids early convergence and stagnation of the swarm at a local maximum. Based on monitoring the diversity of the swarm at intervals, a percentage of particles with the worst fitness are removed and replaced with new particles. This has the effect of injecting new energy into the PSO to break out of the local maximum and search for the global maximum.

Furthermore, in the context of dynamic implementation of CTA-PSO in a wireless network, a learning element is introduced based on neighbouring Group of Pictures (GoP) similarity. This feature takes account of the particular application type (e.g. video) and the observation that the GoP size remains approximately the same until a change in video scene takes place. As a result, rather than randomizing the swarm particles at each execution of the algorithm, knowledge of the previous best particle, $P_{best}$, positions is used. With this $P_{best}$ learning approach, CTA-PSO execution time can be further reduced.

The variation in swarm diversity is shown in Fig. 4a. The spikes in the graph illustrate the implementation of the diversity function. The higher diversity values represent greater exploration by the swarm. The base of the spike narrows (spikes 4 and 5 in Fig. 4a) indicating increased localization of the swarm exploration and can be considered as confirmation that the global optimum has been found.

The corresponding utility curve is shown in Fig. 4b illustrating the improvement in sum utility over an alternative, recognized PSO algorithm, APSO [11].
Figure 4: CTA-PSO 40 device example (a) Swarm Diversity, and (b) Convergence Behaviour
5 Solving the Challenge of 60 GHz Resource Allocation using CTA-PSO

We present an example where CTA-PSO will solve the resource allocation problem in the presence of blockage. The environment is an in-vehicle entertainment system e.g. within a train carriage where the blockage could be due to a person standing/walking in the aisle.

A real-time application such as live IPTV where a user suddenly changes to a new TV channel (IPTV Channel Change or CC IPTV) poses significant challenges when meeting delay deadlines imposed by this application (i.e. Selected Channel display deadline). The problem is aggravated by the fact that there are no IPTV frames previously buffered and that the Direct LOS is blocked (Fig. 5a - Step 2). The blockage is solved by means of the link switch relay method (IEEE 802.11ad) as shown in Fig. 5a - Step 3. An IPTV utility function is designed to reflect the delay requirement to be met and to minimize the number of IPTV frames being lost [12].

The advantages of “Link Switch” are clearly visible in comparing Figs. 5b and 5c. The user requests a Channel Change and shortly afterwards a blockage occurs for a duration of 1 second (approximately 15 MAC super-frames). The display deadline computed to meet QoS demands is 720ms (11 superframes) indicated in Figs. 5b and 5c by the “Channel should display here” legend. This value is based on the difference between the maximum QoE-linked channel change delay (2s) and the response time of the channel change request, which includes CTA-PSO and application-related timing values [12]. Without link switch there is no CTA time allocated during the blockage phase. As a result, the deadline is missed and frames are dropped. Frames are only transmitted again once the blockage is removed and CTA time is increased by PSO as shown in Fig. 5b.

Using the link switch relay method solves the problem as shown in Fig. 5c. CTA-PSO adjusts immediately to the relay introduced in the communication path. As a result, CTA time is continuously allocated despite the direct LOS blockage so that the display deadline is met with zero frame loss.

Referring back to the methods introduced in Section 3, neither NBS nor RAG would be able to solve this non-convex optimization problem. For this practical situation, the increased implementation time of the Lagrangian optimization technique would make it challenging to allocate the appropriate resource to overcome the blockage.

This case study shows a typical and realistic scenario for mm-Wave applications where not only access time has to be optimized but signal blockage also has to be overcome. The combination of CTA-PSO with the delay-sensitive utility function and link switch provides a solution to this challenge.
Figure 5: Impact of 1 s Blockage on CC IPTV - (a) Blockage Event, (b) PSO Resource Allocation without link switch, and (c) PSO Resource Allocation with link switch during the blockage phase.
6 Conclusion

Current multimedia applications (HD Video, IPTV) and future technologies for smart homes, smart vehicles and smart cities place high demands on wireless technologies. Smart resource allocation is required to achieve and maintain the QoS expected by the user.

To date, optimization techniques have focussed on single application solutions. In order to meet the future multi-user multimedia heterogeneous network demands of 5G, alternative resource allocation optimization techniques need to be explored.

In this article, we propose a reduced execution-time solution such as CTA-PSO, demonstrating its suitability for implementation in a challenging mixed multimedia wireless environment. In order to meet the evolving requirements of new applications, and converged and high-capacity networks such as 5G, alternative resource allocation techniques such as CTA-PSO must be further explored.

References


Sandra Scott-Hayward (M’13) is a Research Fellow at the Institute of Electronics, Communications and Information Technology, Queen’s University Belfast. She received a PhD from Queen’s University Belfast in 2013. She has been a Chartered Engineer since 2006 having worked as a Systems Engineer and Engineering Group Leader with Airbus. Her research interests include wireless communications and networking, resource allocation, optimization and performance analysis for next generation networks, and security in software-defined networking.

Emiliano Garcia-Palacios is a Lecturer at the Institute of Electronics, Communications and Information Technology, Queen’s University Belfast. He received a PhD from Queen’s University Belfast in 2000 and since then he has been leading research in Wireless Networks resource management. His research interests include wireless protocols, wireless resource allocation, optimization and traffic management for next generation Gigabit networks.