Heterogeneous Wireless Network Selection: Load Balancing and Multicast Scenario

Svetlana Boudko
Norsk Regnesentral
Oslo, Norway
svetlana.boudko@nr.no

Wolfgang Leister
Norsk Regnesentral
Oslo, Norway
wolfgang.leister@nr.no

Stein Gjessing
University of Oslo
Norway
steing@ifi.uio.no

Abstract—**The increasing demand for real-time multimedia streaming from mobile users makes important deployment of network selection in wireless networks. Coexistence of various wireless access networks and ability of mobile terminals to switch between them make an optimal selection of serving mobile networks for groups of mobile clients a challenging problem. Since scalability can easily become a bottleneck in large-scale networks, we study the decision-making process and selection of the data that needs to be exchanged between different network components. In this paper, we present two decentralized solutions to this problem that we compare and evaluate in the OMNet++ simulation environment.**

Keywords—Wireless networking; mobile network selection; decentralized algorithms.

I. Introduction

This article extends the work presented by Boudko et al. [1, 2] and studies load balancing and multicast communication over heterogeneous mobile networks. This article is also an extention of [3, 4].

Availability of various wireless network technologies and continuous development of mobile devices and services lead to complex and highly dynamic networking and challenge resource limitations of wireless access networks. According to a recent forecast [5], monthly data consumption for wireless networks will increase over 15 times in the years between 2011 and 2016. In 2016, the demand for mobile bandwidth will exceed the average capacity by about 32 %. Despite constantly increasing demand, the range of frequencies is the same. Consequently, during peak demands when the bandwidth becomes an insufficient resource the consumer is likely to experience degradations in the form of reduced service, slow service, or even no service.

To avoid some of these negative effects of a network that is challenged by resource limitations we need to consider the resource allocation problem from a different angle, including collaboration between mobile user nodes and networks to improve the overall utilization of resources. Referring to wireless access networks, the ability to be connected to several network technologies simultaneously offers new possibilities to formulate effective strategies for network selection.

The network selection problem inspired by the “always best connected” concept was mostly focused on the definition of metrics to address the best end user quality of service for each single consumer, neglecting the impact to the other consumers in the network. Contrary to this, we use metrics that express quality of service for all users collectively and evaluate benefits for the system components from complexly applied network selection. In our problem formulation, we take into account that *1*) a large number of mobile devices can operate simultaneously inside an area with overlapping coverage of various mobile networks; *2*) some of the networks can experience degradation of service while some of them can accommodate more users; and *3*) some groups of these devices can listen to the same feeds from the same Internet locations while being connected to different access points. We consider the network selection problem to use for multi-user environments with possible multicast configurations that allows the network to perform load balancing, improve the users’ overall QoS, and increase the networks’ throughput.

Being originally introduced for use in the wired Internet, multicast is an efficient method for point-to-multipoint communications, which reduces drastically the traffic load when the same content is sent to a large group of users. The 3rd Generation Partnership Project (3GPP) and its successor 3GPP2 defined the *Multimedia Broadcast and Multicast Service* (MBMS) and the *Broadcast and Multicast Service* (BCMCS) [6], respectively. The Long-Term Evolution (LTE) project introduced *LTE Broadcast*, also denoted as *evolved Multimedia Broadcast Multicast Service* (eMBMS) [7]. Different types of applications like video conferencing, file distribution, live multimedia streaming, IPTV can benefit from deploying multicast networking. It is also advantageous in cases of the flash crowd phenomenon when the popularity of a certain item increases rapidly over a short period of time. The *LTE whitepaper* [8] shows that already from three to five subscribers in one cell site achieve break-even of cost between unicast and multicast.

However, the complexity of managing multicast networks makes the deployment of multicast even more challenging in wireless environments when mobility issues have to be considered. Also, the notion of a link interface for a wireless multicast channel differs from that for a wired network. Multicast management in wireless heterogeneous networking also involves mobile network selection for a group of clients in addition to construction of multicast trees like in conventional multicast protocols.

In this paper, we consider a solution for the network selection problem for heterogeneous mobile networking as a part of multicast group management. Previously, we have proposed a method that provides an optimal network selection for a given network topology, network conditions and user preferences assuming that all needed information can be collected from the network and is available for a central decision-making unit that

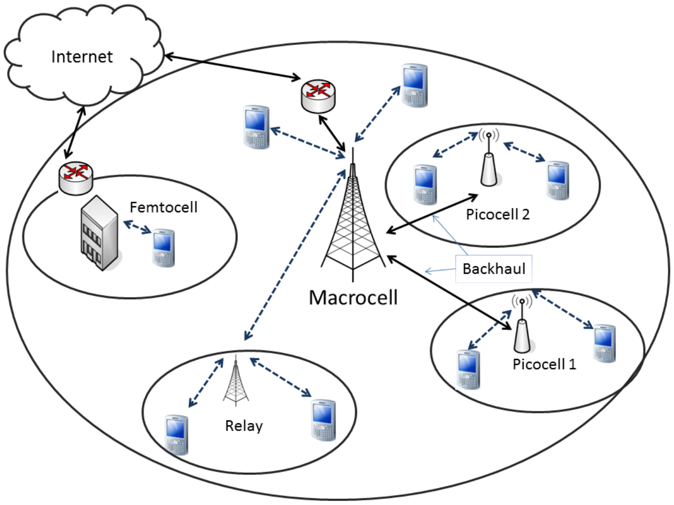


Figure 1. LTE Advanced Heterogeneous Network Architecture.

computes the assignment of mobile multicast groups of nodes to networks [1, 2]. In this paper, we refine the aforementioned method and apply it as an upper bound for evaluating methods that use only limited information shared among the decision makers.

The work proposed in this paper includes the following contributions: *1*) We propose two approaches that allow network selection in a decentralized manner with only limited information shared among the decision makers. *2*) Through extensive simulations, we study how different sets of information available to decision makers influence the performance of the system. *3*) We discuss how the evaluated approaches can be combined and propose ideas for further improvements.

The remainder of the paper is organized as follows. After presenting an overview of related work in Section II, we discuss two representative scenarios for load balancing and for multicast networking in Section III. Load balancing problem in heterogeneous wireless network is addressed as follows. We present the problem as a team-decision problem in Section IV, and outline suitable algorithms. The simulation set-up and the results of simulations are presented and analyzed in Section V Multicast transmission in heterogeneous wireless network is covered in the following sections. We present the problem formulation and outline an optimal solution to the problem in Section VI. The system components are considered in Section VII. Two decentralized solutions are presented in Section VIII. Simulation results are given in Section X. We discuss future work and conclude in Section XI.

II. Related Work

To the best of our knowledge, the research field regarding resource allocation and selection of a network in heterogeneous wireless networks from a perspective of load balancing and multicast delivery is not well explored. In what concerns network selection for mobile multicast groups, four research areas can be considered as related work *1*) handoff management; *2*) network selection in wireless networks; *3*) multicast in wireless networks; *4*) LTE-Advanced Heterogeneous Networks.

A. Handoff Management in Mobile Networks

Prediction-based techniques have been suggested in several studies [9–12] aiming to reduce handoff delays.

To represent the movement behavior of a mobile user, Paramvir et al. [9] propose a two-level user mobility model consisting of a local level and a global level. A hierarchical location prediction algorithm is proposed based on an approximate pattern-matching algorithm implemented in the global level and Kalman filtering techniques implemented in the local level.

Akyildiz and Wang [10] propose a mobility model that uses historical records and stochastic behavior of mobile users to predict their future position. The model is built upon a framework of user mobility profiles (UMP). In the proposed prediction algorithms, many factors are taken into consideration including velocity and direction of mobile users, historical records, stochastic model of cell residence time and path characteristics. The authors claim that these algorithms predict more accurately than previous schemes. However, the complexity of the algorithms make them impractical for mobile applications.

In two studies, Tseng et al. [11] and Choi et al. [12] propose using cross-layer information to perform layer-3 handoff in parallel with or prior to the layer-2 handoff. However, these schemes can lead to false alarms and cause unnecessary MIP registrations. Ray et al. [13] conclude that deciding upon the ideal choice and timing of cross-layer triggers in order to reduce layer-3 latency is still an open problem.

Vertical handoff is the handoff between the networks of different wireless technologies and has been addressed in several studies [14–20]. While horizontal handoffs are typically triggered when the received signal strength (RSS) of the serving access router drops below a certain threshold the vertical handoff can be initiated due to other reasons such as user preferences or network conditions including coverage, bandwidth, cost and power consumption. The decision process is therefore more complex for vertical handoffs than for horizontal ones.

While some authors only use RSS as an input parameter for the handoff decision process [10, 21], others combine the use of RSS with bandwidth information [22–24]. Using cost functions has been proposed earlier [14–16]. Nasser et al. [25] propose a cost function that depends of the cost of service, security, power consumption, network conditions and network performance. However, in their evaluation, all weights except the bandwidth weight are set to zero. This renders their cost function to a function of one parameter: bandwidth.

Algorithms based on fuzzy logic or artificial neural networks in combination with multiple criteria [17–19] suffer from high handover delay because of their complexity and the training process. Unfortunately, the authors of these algorithms did not provide throughput results.

Recently, some studies [26, 27], proposed solutions for group vertical handoffs in heterogeneous environments. These studies consider scenarios when many mobile users send handover requests almost at the same time. In these scenarios, the influence of multiple users is important to consider for optimal network selection. The solutions presented in both studies require a centralized approach to be adopted to implement the proposed schemes. The obvious drawback of this approach is a poor scalability of these solutions.

B. Admission control and network selection in wireless networks

Ormond and Murphy [28] propose a network selection strategy that explores a number of possible utility functions. The solution is user-centric, and an interplay between different users and networks is not considered. Ormond and Murphy conclude that the impact of multiple users operating in the same region needs to be further examined.

Gluhak et al. [29] consider the problem of selecting the optimal bearer paths for multicast services with groups of heterogeneous receivers. The proposed algorithm selects the bearer path based on different optimization goals. However, Gluhak et al. address the problem only for the ideal static multicast case without taking into account users crossing different cells. In their work, multicast membership does not change during the duration of a service, and multicast groups are not built with consideration of users' movements. In our opinion, this is not a realistic case for wireless networks.

Jang et al. [30] present a mechanism for efficient network resource usage in a mobile multicast scenario. This mechanism is developed for heterogeneous networks and implements network selection based on network and terminal characteristics and QoS. However, in the proposed mechanism, the network selection is performed purely based on terminal's preferences, the network perspective is not considered, and the solution does not optimize the utilization of network resources.

Tragos et al. [31] propose a generic admission control algorithm that allows network selection for 4G heterogeneous wireless networks. The algorithm aims to provide maximum utilization of the network, prevent overloading situations and ensure best QoS. However, implementation of the algorithm requires the presence of a centralized entity.

Khan et al. [32] present a game theoretic solution for resource allocation and call admission in wireless networks using cooperative games. The main goal is to increase the utilization of the available bandwidth and to reduce the call blocking. The solution is applicable to wireless network scenarios where networks are willing to cooperate. Kalai-Smorodinsky Bargaining Solution is used to solve the cooperative game. The authors also propose the request distribution algorithm that allows to allocate the request to several different networks and split the requested bandwidth between these networks. Simular to Tragos et al. [31], the implementation requires also a centralized entity that is responsible to handle bargins between the participating networks.

In our analysis, we recognize that several problems are not yet addressed, or where the currently available solutions need to be improved. Decentralized algorithms that rely on information only partly shared between the decision-makers need to be implemented and evaluated in multi-user scenarios. These considerations motivate us to look at distributed and computationally efficient methods of network selection in heterogeneous mobile environments.

C. Network Selection in Wireless Networks

Ormond and Murphy [28] propose a network selection approach that uses a number of possible utility functions. Their solution is user-centric and does not present any multicast scenario. An interplay between different users and networks is not considered either. Ormond and Murphy conclude that the impact of multiple users operating in the same region needs to be further examined.

Gluhak et al. [29] consider the problem of selecting the optimal bearer paths for multicast services with groups of heterogeneous receivers. The proposed algorithm selects the bearer path based on different optimization goals. However, Gluhak et al. address the problem only for the ideal static multicast case without taking into account users crossing different cells. In addition, it requires that the knowledge of the conditions in wireless networks and preferences of receivers is fully shared. In their work, multicast membership does not change during the duration of a service, and multicast groups are not built with consideration of users' movements. In our opinion, this is not a realistic case for wireless networks. Also, the proposed selection algorithm is built upon a rule according to which the receivers are partitioned into two sets: the receivers for which only one network is available versus the receivers for which several networks are available. The impact of the users inside the second group, as a result of this partitioning, is neglected.

Yang and Chen [33] propose a bandwidth-efficient multicast algorithm for heterogeneous wireless networks that is formulated as an Integer Linear Programming problem that is solved using Lagrangian relaxation [34]. The algorithm deals only with constructing optimal shortest path trees for multicast groups. In this approach, important parameters, such as cost of service or the user's velocity, are not considered.

Jang et al. [30] present a mechanism for efficient network resource usage in a mobile multicast scenario. This mechanism is developed for heterogeneous networks and implements network selection based on network and terminal characteristics and Quality of Service (QoS). However, in the proposed mechanism, the network selection is performed purely based on terminal's preferences, the network perspective is not considered, and the solution does not optimize the utilization of network resources.

Hou et al. [35] propose a cooperative multicast scheduling scheme for multimedia services in IEEE 802.16 based wireless metropolitan area networks (WMAN). The scheduling is considered for one base station that further re-sends the data to multiple subscriber stations. These are grouped into different multicast groups and the users are assigned to the groups. The authors consider two approaches to select multicast groups for services: the random selection and the channel state aware selection. The process is controlled by the base station and limited to one network technology. No network heterogeneity is considered.

D. Multicast in Wireless Networks

The Multicast Mobility (multimob) working group [36] focuses its activity on supporting multicast in a mobile environment. The main goals of the group are to work out mechanisms for supporting multicast source mobility and mechanisms that optimize multicast traffic during a handover. The group also documents the configuration of IGMPv3/MLDv2 in mobile environments. In this sense, they extend the IGMPv3/MLDv2 protocols for implementation in the mobile domain and improve *Proxy Mobile IPv6* to handle multicast efficiently. However, they do not consider any modifications across different access networks.

The Long-Term Evolution (LTE) project introduces evolved Multimedia Broadcast Multicast Service (eMBMS) [7]. This standard covers technically the terminal, radio, core network, and user service aspects that provide a point-to-multipoint service for transmitting data from a single source to multiple recipients. The performance is improved due to higher and more flexible LTE bit rates, single frequency network (SFN) operations, and carrier configuration flexibility. The eMBMS Service Layer offers a Streaming- and a Download Delivery Method and is enhanced with video codec for higher resolutions and frame rates and forward error correction (FEC), and the radio network with procedures to ensure MBMS reception in a multifrequency LTE network. eMBMS also allows LTE network and backhaul offloads.

E. LTE-Advanced: Heterogeneous Networks

Though several improvements were introduced in the LTE-Advanced standard [37], the homogeneous networks with only macrocells deployments will not be able to cope with future mobile traffic. A step towards optimization of performance in wireless networks is done by LTE-Advanced [38] through enhancements in network topology. The LTE-Advanced proposed implementation of heterogeneous networks (HetNets) topologies that combine utilization of both macrocells and small cells, the latter including micro, pico, femtocells and relays, each having different transmit power and access rules for user devices.

1) Macrocells: Macrocell is an outdoor base station and the main base station in the cell. The transmitted power is about 45 dBm. Macro cells are connected with each other through backhaul that are usually built upon a wired infrastructure. In some cases, e.g., for rural areas, wireless links can also be used.

2) Micro and Picocells: These cells are, usually, an outdoor low cost base stations with open access and a small coverage.They are connected with the macro cell using a backhaul link. The transmitted power is about 35 dBm. The range is about two kilometers wide for microcell and about 200 meters for picocell.

3) Femtocells: Femtocells are indoor base stations either with open or limited access and low transmitted power that is less than 23 dBm. Though these cells are positioned as an alternative to micro and picocells, their coordination with the macrocell still is not fully achieved in current deployments.

4) Relays: Relay stations receive, demodulate and retransmit the signals between base stations and mobile users. They can decode the data and provide error correction. Relays are used to increase throughput and to extend coverage of cellular networks. Relays do not need wired connection to the base station; therefore, the backhaul costs can be saved.

In the HetNet network model, macrocells provide full coverage for a wide area and small cells cover some areas with extra traffic demand. It is a useful network architectural feature since the bandwidth demand is not uniform across the area and users and traffic are often concentrated in particular areas. Another important benefit of using small cells inside of a macrocell is to improve coverage in places where coverage of the macrocell is not sufficient, e.g., in cell edges. Since deploying extra macro cells in these areas results in additional interferences, the deployment of lower power picos is a better solution and can give a cost reduction. A typical LTE-Advanced HetNet scenario with a macro base station and several small cells is illustrated in Figure 1.

For the purpose of this paper, we evaluate network selection solutions considering both the LTE-Advanced HetNet network model and a general heterogeneous mobile network system with overlapping of several wireless technologies, e.g., Wi-Fi and cellular networks.

In our analysis, we recognize that the presented previous work has not addressed several important aspects related to the network selection for mobile multicast groups. We need to study how the users' movements influence the optimal selection of members for multicast groups and how the information needed for network selection is exchanged between the decision makers.

III. Scenario

A. Load Balancing Scenario

We consider a network selection scenario for a group of users in a hotspot area like a crowded city center, a public transportation node or an exhibition site where a coverage of several base stations or access points from different networks is possible. We assume a substantial overlap in coverage of these stations. The networks implement different network technologies. We also consider a situation with multiple overlapping IEEE 802.22 wireless regional area networks where self-coexistence is allowed. A representative scenario of such networking is illustrated in Figure 2. These user terminals are capable of connecting to several access networks, and vertical handoffs between different networks are technically possible. The terminals periodically receive beacon signals from base stations or access points of the available access networks that are typically broadcast once per second.

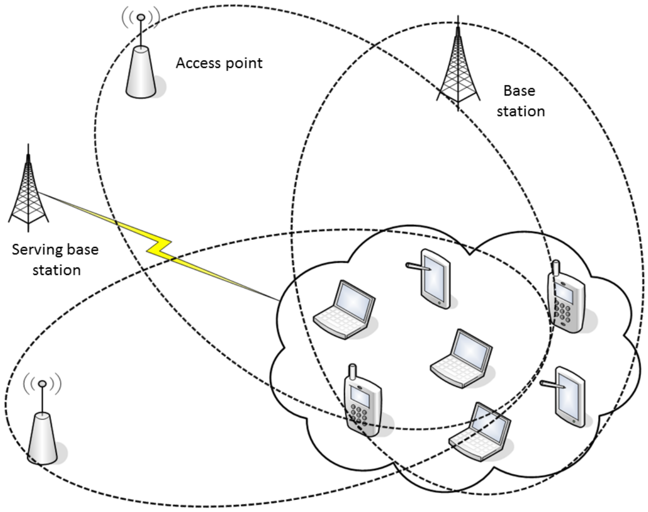


Figure 2. Network topology built upon multiple mobile networks with heterogeneous technologies to serve a group of clients.

Users located in the same cell of a mobile network can experience degradation in quality due to shortage of available bandwidth. Though admission control mechanisms are designed to ensure the quality of wireless connections and to prevent network congestion, there is still a possibility that a user is admitted to the network while requiring low bandwidth, e.g., for web browsing, will require more resources for video streaming shortly after. We consider also a situation when a base station may act in a proactive way and monitor the available resources in adjacent cells. The users that are going to move to a cell that, at the moment, is not able to admit new users can be notified to perform a vertical handoff to another available network. Since users may have different preferences and request different types of service their utility functions are built upon different criteria.

To provide better load balancing between the networks, and to avoid disturbing ping-pong effects, joint coordination, and information exchange between the users and the base stations is essential; both the clients and the networks can benefit from cooperative handoffs. However, due to strict bandwidth and power limitations of mobile networks, and also due to scalability issues, a complete information exchange between mobile users and networks is not feasible. Decentralized network selection is therefore essential.

B. Multicast Scenario

To illustrate the yet unsolved challenges for optimal network selection in multicast networks, we consider a multimedia streaming scenario for a group of mobile users that concurrently receive the same content from the Internet. We assume that a backbone proxy server (BPS) is placed at the network edge. The BPS is a member of a content distribution system (CDN). This scenario is an extension of a scenario that we previously have considered to illustrate an adaptive multimedia streaming architecture to mobile nodes [39].

The BPS streams content that either is hosted on a streaming server, or re-sends the streaming content as a part of an application layer multicast. The users of this network are located in an area with a substantial overlap in coverage of several mobile networks, and are connected to different networks. One of examples of such networking is Heterogeneous networks (HetNets) in LTE-Advanced [38].

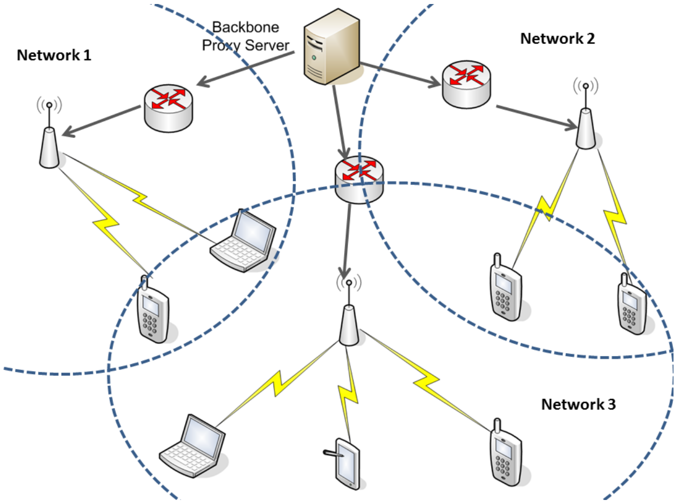


Figure 3. Multicast streaming scenario for a group of mobile clients served by several mobile networks before regrouping.

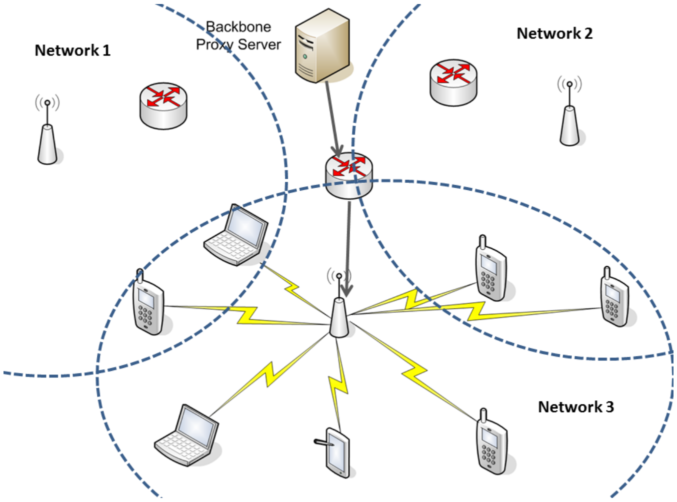


Figure 4. Multicast streaming scenario for a group of mobile clients switched to one mobile network after regrouping.

The base stations of the system have multicast capabilities, implementing, for example, Multimedia Broadcast Multicast Service [40]. A representative scenario of such networking is illustrated in Figure 3.

In our scenario, we assume that the mobile terminals are capable to connect to several access networks, and vertical handoffs between these networks are technically possible. Further, we assume that these terminals are equipped with GPS receivers, so that their location information can be transmitted to the BPS. The BPS can use this information to determine how users can be regrouped in multicast groups. Such regrouping is beneficial as it saves network resources. Hence, users that get the same content can exploit the same wireless link because the content can be broadcasted to them. The resources in the backhaul network are also better utilized because the content is now delivered only to one mobile network instead of being spread to several networks. An example of such regrouping is depicted in Figure 4.

Technically, to facilitate such a mechanism, the user terminals will have the possibility to switch to other mobile networks after receiving certain messages from the BPS. Since users may have different preferences depending on diverse criteria, for example, power consumption, security, or network cost of service, the interplay between the users' utilities and the networks' utilities is important to consider. Network selection support for multi-access networks can be implemented at any layer of the protocol stack. There are certain tradeoffs to consider at the design stage. Cross-layer signaling can potentially be added to allow the application level to control the process, hence, to prevent breakup of ongoing sessions.

IV. Problem Formulation for Load Balancing in Heterogeneous Wireless Networks

Decentralized network selection can be formulated as a team decision problem [41, 42] where several decision variables are involved. These decisions are made by different decision makers that have access to different information but participate in a common goal.

Team decision theory is concerned with determining the optimal decisions, given a set of information for each of several decision makers, that work together to achieve a payoff. In team decision problems, these sets of information are different though often correlated for different decision makers. These optimal decisions can be either person-by-person optimal or team optimal. In person-by-person optimal cases, each person makes the decisions that optimize the individual's payoff, but not necessarily the team payoff. These cases are optimal for a particular team member, given that the decision functions for other members are fixed. In team optimal cases, the group payoff is optimized. Team optimality is a stronger condition, and is thus harder to achieve. Taking into account that person-by-person optimal strategies may result in unfair distribution of the resources, we focus our research on team optimal strategies.

A. System Model

Taking into consideration our understanding about preferences of mobile nodes and the networks, we are now ready to formalize our observations into a system model. We consider a set of networks $N = 1, 2, ..., n$ and a set of mobile terminals $M = 1, 2, ..., m$. For each terminal m_j and network n_i the following is defined. Streaming bitrate requirements of mobile nodes are denoted by r_j; $rss_{i,j}$ is the received signal strength in network i for terminal m_j while power consumption and cost of service in network n_i for terminal m_j are denoted by $p_{i,j}$ and $c_{i,j}$, respectively. The total available bandwidth of network n_i is denoted by b_i. For each terminal m_j, we define a user preference profile that is described by a tuple containing $Th^p_{i,j}$, $Th^c_{i,j}$, and $Th^{rss}_{i,j}$. These denote thresholds, or user preferences, for respectively power consumption, cost of service and received signal strength. We define a time period $\tau_{i,j}$ during which terminal m_j is served by network n_i before performing a handoff and moving to the next cell of this network.

For each mobile terminal m_j and each network n_i we define the function $x_{i,j}$ which mimics the decision taken by a mobile terminal m_j to switch to or to stay in mobile network n_i.

$$x_{i,j} = \begin{cases} 1, & \text{if } m_j \text{ has roamed to or stays in } n_i \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

We define the common goal of the set of networks and users of these networks as maximization of consumed bandwidth over a period of time, minimization of the number of handoffs and reduction of signalling between the networks and terminals. To achieve this goal, the participating networks and the terminals need to cooperate while trying to maximize their own performance. To facilitate a decentralized approach, we define two components, the network component and the mobile node component. We formulate the problem and solve it separately for these two components.

1) Network Component: For each network n_i, we define a utility function U_i as a sum of consumed bandwidth of all users of this network over the time the user is connected to it, as defined in Eq. (2).

$$\forall\{i\} : U_i = \sum_j x_{i,j} \cdot r_j \cdot \tau_{i,j} \tag{2}$$

In this sense, the networks benefit if they select the users that not only request a higher bandwidth but also intend to stay in the network for longer periods of time, which also can eliminate the ping-pong effect when the user needs to change the network again recently after the handoff. Basing our decision on research done by other research groups [43–48], we assume that a mobile network is capable of predicting the residence time of a mobile node inside the network based on mobile node's velocity, movement patterns and the local area. We realize that this problem is an ongoing research work. For the purpose of this paper, we assume that the prediction can be performed with acceptable precision.

The common goal is the maximization of the expected value of the sum of network utility functions.

$$\max \sum_i \mathbb{E}[U_i] \tag{3}$$

The utility function is constrained by the network resources. As any network of the system has a limited knowledge about the resources and decisions of the rest of the system, $x_{i,j}$ is given as its expected value.

$$\forall\{i\} : \sum_j \mathbb{E}[x_{i,j}] \cdot r_j \leq b_i \tag{4}$$

```
At Mobile Node $m_j$ :
if Network Selection is triggered
search for available networks
for each available network $n_i$
  if ($p_{i,j} < Th^{p}_{i,j}$, $c_{i,j} < Th^{c}_{i,j}$, $rss_{i,j} < Th^{rss}_{i,j}$) then
    add the network to list of candidate networks
for each network $n_i$ in list of candidate networks
  send requests to candidate network $n_i$
wait for response from candidate networks
upon reception of response from candidate network $n_i$
if (admitted value == true) then
  add response to response list
if (response list == null)
  stay in the current network
else
  for all responses in response list
    choose the network with the highest $\tau_{i,j}$
```

Figure 5. Distributed Algorithm for Network Selection, Mobile Node Component

```
At Candidate Network $n_i$ :
wait for admission requests from mobile nodes
upon reception of admission requests from mobile node $m_j$
  using available knowledge solve Eq. (2) with constraints Eq. (4)
  return admission message containing
    admitted value = true/false and expected time
```

Figure 6. Distributed Algorithm for Mobile Node Admission, Network Component

2) Mobile Node Component: For each user we define a utility function f_j as a function of power consumption, cost of service, available bandwidth and received signal strength. We realize that the set of parameters that define user preferences can be larger than the one mentioned above and can also differ from user to user. We also realize that users might employ different utility functions but for this work we limit us to this definition.

$$\max f_j(p_{i,j},\ c_{i,j},\ r_j,\ rss_{i,j}) \tag{5}$$

Eq. (5) poses a multiparameter optimization problem that can be solved by introducing weights and normalization. Another solution is to relax the optimization problem by reducing it to a one variable optimization Eq. (6). Consequently, we introduce a set of constraints in Eq. (7), where some of parameters $z \in \{p, c, rss\}$), namely power consumption, cost of service, and received signal strength, are limited by their thresholds $Th^{z}_{i,j}$ defined by the user's preferences.

$$\max \sum_{i} x_{i,j} \cdot r_j \cdot \tau_{i,j} \tag{6}$$

$$p_{i,j} \le Th^{p}_{i,j}, \quad c_{i,j} \le Th^{c}_{i,j}, \quad rss_{i,j} \le Th^{rss}_{i,j} \tag{7}$$

B. Algorithm

The system model defined in Section IV-A is used in the decentralized algorithm for network selection outlined below. We build the algorithm based on the following *a*) maximization of the total consumed bandwidth by distributing the users between the networks taking into account the networks' available bandwidth, and *b*) minimization of the number of performed handoffs between the networks.

There are two events that may trigger the execution of the network selection algorithm: *1*) based on monitoring of available resources in its cells, and the prediction of users' location information, the network informs mobile nodes that are about to move to a congested cell to switch to another available network instead; *2*) mobile terminal m_j experiences degradation of network performance detected by increased packet loss or delay on the mobile terminal.

When the network selection is triggered, the effected terminal runs the selection algorithm as shown in Figure 5. As a consequence, the network receives calls from mobile nodes it runs the algorithm shown in Figure 6. To calculate the expected values of the utility function defined by Eq. (2), this algorithm takes as input the knowledge available for this network. Depending on how the knowledge of the system is shared among the networks, we differ between two versions of the algorithm: Algorithm A and Algorithm B.

1. **Algorithm A**: Each mobile node m_j, while sending a request to the network n_i, informs the network about the requests sent to other networks. Based on this information, each network calculates the probability of mobile node m_j to choose this network if accepted.
2. **Algorithm B**: Each network n_i shares its information with exactly one more network n_k. The network n_i does not accept a mobile node m_j if the node is accepted by the network n_k and $\tau_{i,j} < \tau_{k,j}$.

Also, we consider using a combination of these two versions, refered further as Algorithm AB. In this version, in addition to the information exchanged between the networks as in Algorithm B, the nodes specify in their requests the number of all requested networks, as in Algorithm A.

C. Algorithm Evaluation

To evaluate the algorithms, we define upper and lower bounds to their operation. The upper bound is achieved by applying a centralized solution with fully shared knowledge of the conditions in all evaluated networks and is further referred as global knowledge reference. This reference can also be viewed as a modification of the algorithms [26, 27, 31] discussed in Section II-B The algorithm [31] is now extended to a multi-user scenario. Its utility function is defined in Eq. (8). The utility function is constrained by resource limitations of networks as described in Eq. (4) and preferences of mobiles nodes described in Eq. (7). This problem belongs to the class of integer linear programming problems, which is known to be NP-complete, thus problematic for real-time tractable implementation and in most cases can be used only as a reference for evaluating algorithms.

$$U = \sum_{i} \sum_{j} x_{i,j} \cdot r_j \cdot \tau_{i,j} \tag{8}$$

The lower bound corresponds to a situation when all networks base their decisions only on local knowledge (Eq. (2)) and is further referred as local knowledge reference. The local knowledge reference can also be viewed as a modification of algorithm [28] discussed in Section II-B and applied to a multi-user scenario. In this sence, the algorithms are compared with the related work.

V. Simulations for Load Balancing in Heterogeneous Wireless Networks

The performance and functionality of the algorithms have been evaluated through multiple simulation runs. We have implemented both versions of our algorithm in the OMNet++ environment [49]. In Algorithm A, the network gets the information about how many other networks are requested by the same mobile node. This information is submitted to the network by the mobile node along with the request to join the network. As no other information is available, we assume that the probability of being assigned to any of the networks is equal for all participating networks. In Algorithm B, each network shares its information with exactly one more network. In our testing scenario, these networks do not overlap. We compare the algorithms with the global knowledge reference and the local knowledge reference.

A. Simulation Setup

For the sake of simplicity, we simulate a scenario with four wireless networks, which covers quite well the scope of the evaluation. In this scenario, a group of users from one network is about to move from one cell of the network to another cell of the same network that experiences a shortage of available bandwidth. In consequence, the cell that the users move to is not able to accommodate all these users.

For our evaluation, we run tests with 100, 200, and 300 users moving to this congested cell. Further, we divide the users into four categories in terms of requested bandwidth. To define these categories we use service class characteristics defined by Tragos et al. [31] as follows: *a*) at 64 kbps, for simple telephony and messaging *b*) at 512 kbps, for web browsing *c*) at 1024 kbps, for interactive media and *d*) at 2000 kbps, for video streaming, each category having approximately the same number of users.

Note that none of the networks have enough resources to accommodate all users alone. All four networks must be used in order to meet the requirements of all users. We run also tests when total bandwidth of all networks is not sufficient to accommodate all users. The tests are done for network conditions that result in 5%, 10%, 15%, 20%, 25%, 30% dropped calls if the global knowledge reference is applied. The time $\tau_{i,j}$ for the user j to stay in the network n_i before performing a horizontal handoff, or a cell residence time, is randomly distributed in the range $[1, 100]$ time units.

B. Simulation Results

We evaluate how mobile nodes are distributed among the networks after one iteration of the algorithm run. We calculate the number of decision errors as a number of users whose connection ends up in dropped calls due to wrong network allocation. These errors are the results of wrong assignments to networks that do not have sufficient bandwidth to accommodate the assigned users. For each group of users (100, 200, 300 users), we repeat the experiment 1000 times with different sets of $\tau_{i,j}$.

For all tests done, the top and bottom 5 % of the results are excluded from the evaluation. The results are averaged over these simulation runs and are depicted for minimum value results in Figure 7(a), for average value results in Figure 7(b), for maximum value results in Figure 7(c), for cumulative distribution function in Figure 8. The global knowledge reference is 0 for all experiments meaning that in the centralized solution, all users were assigned to the networks without any dropped calls. The results with dropped calls in the global knowledge reference are depicted in Figure 9. The figure shows the results for 200 mobile nodes. The results for 100 and 300 mobile nodes are very similar to the results for 200 nodes and, therefore, are not included in the paper.

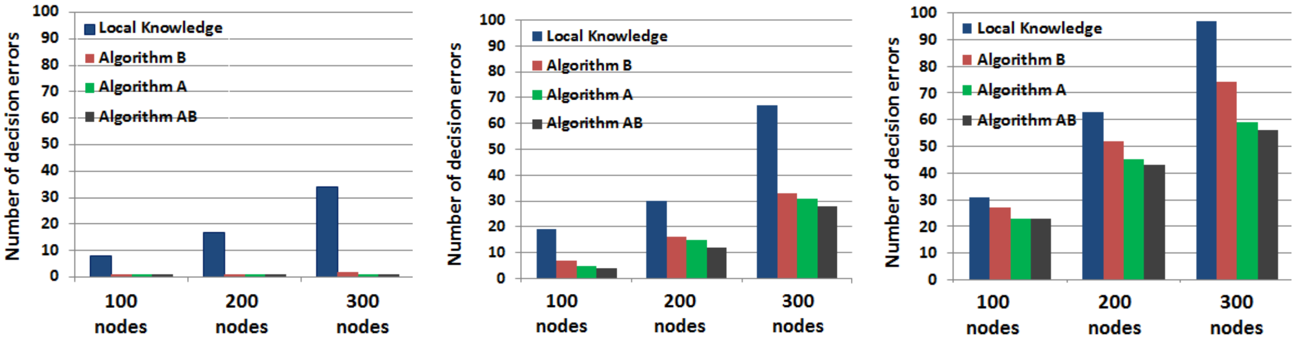
The tests show that all three proposed algorithms can distribute the users between the networks significantly better than the local knowledge reference. Algorithm AB performs better than Algorithm A and Algorithm A performs better than Algorithm B for all user groups through all tested values for dropped calls in the optimum solution.

It shows that sharing partial information about network status as in Algorithm B makes little use of extra information from just one network. It also shows that this information when used in Algorithm AB does not give any significant reduction of decision errors in comparison with Algorithm A. However, Algorithms B and AB require significantly more information to exchange between the networks than Algorithm A. It also requires more sophisticated mechanisms and protocols to be implemented in the networks, including security considerations and synchronization of the information flow. Though the information flow initiated by Algorithms AB and B is significantly less than the one initiated by the global knowledge reference, it still demands the exchange of network information across the mobile networks on a fast time scale and low-latency basis, making it quite challenging to implement the algorithms in practice for large scale networks, as the global knowledge reference.

We also evaluate the dynamic scenario. For these tests, the algorithms are run until all clients are assigned to the networks with sufficient bandwidth, also considering the arriving calls. The arrival rate of new calls is modeled with a Poisson stream. The graphs depicted in Figure 10 show the averaged results for 100, 200, and 300 users over 1000 test runs. The x-axis shows the number of iterations of the algorithm. The y-axis shows the percentage of decision errors. Clearly, Algorithms A, B and AB converge faster than the local knowledge reference. There is very little difference between Algorithms A, B and AB even though Algorithms AB and B rely on more information.

C. Signaling Overhead

We estimate signaling overhead S_o for the algorithms and the references. As signaling required to trigger network selection is the same for the references and the algorithms these



(a) Minimum value (b) Average values (c) Maximum value

Figure 7. Decision errors from the simulation using one iteration of the algorithm run after network selection is triggered. The results for 100, 200, 300 mobile nodes are based on 1000 simulation runs for each group of nodes.

messages are excluded from the estimation. For the global knowledge reference all n networks in consideration need to exchange the information about m users that get triggered network selection, and the overhead is estimated as follows (Eq. (9)).

$$S_o = n \cdot (n-1) \cdot m \tag{9}$$

To make estimations for Algorithms A, B and AB and the local reference defined respectively by Eq. (10), Eq. (11), Eq. (12), Eq. (13), we use the results of the dynamic scenario that are depicted in Figure 10.

$$S_o = 0.17 \cdot m \cdot (n-1) \tag{10}$$

$$S_o = 0.18 \cdot m \cdot (n-1) \tag{11}$$

$$S_o = 0.32 \cdot m \cdot (n-1) \tag{12}$$

$$S_o = 1.06 \cdot m \cdot (n-1) \tag{13}$$

Clearly, Algorithm A provides a significant reduction of the signaling overhead.

VI. Problem Formulation for Multicast Transmission in Heterogeneous Wireless Networks

In this section, the scenario discussed in Section III is formalized as a centralized system model, as illustrated in Figure 11. The system model for this scenario was previously presented [2]. For the sake of completeness, we revisit the model in this section. In addition, we implement some modifications to its prior definition.

A. System Model

We consider a set of networks $N = 1, 2, \ldots, n$, a set of mobile nodes $M = 1, 2, \ldots, m$ and a set of streaming contents $S = 1, 2, \ldots, s$. The contents are hosted in different BPSs. Each content s_k can be delivered to more than one mobile node m_j. Therefore, using multicast for data dissemination is beneficial. For each node m_j, content s_k and network n_i, the following is defined: available bandwidths of networks

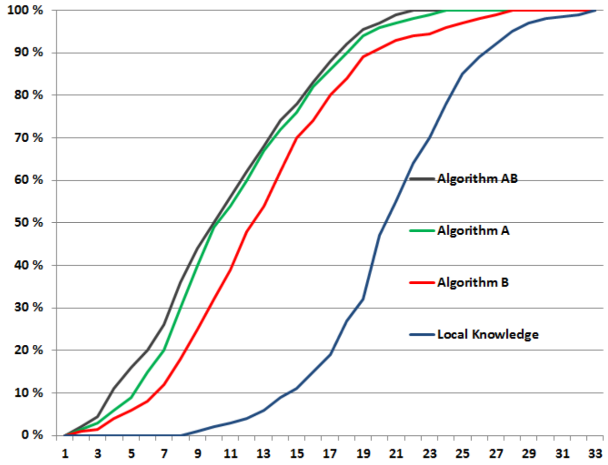


Figure 8. Cumulative distribution function for percentage of decision errors using one iteration of the algorithm run after network selection is triggered. The results are based on 1000 simulation runs for a system consisting of 200 mobile nodes.

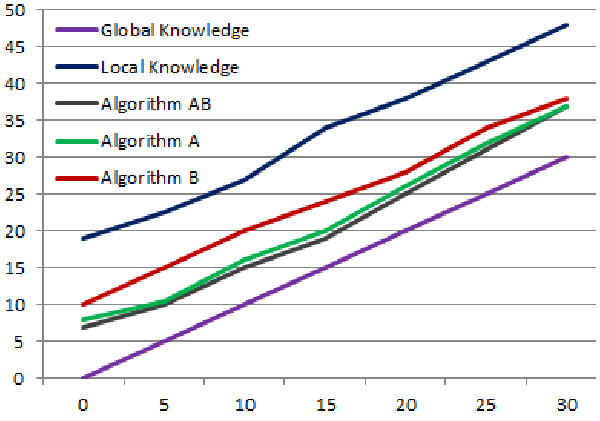


Figure 9. Dropped calls from the simulation using one algorithm run with total available bandwidth less than total required bandwidth.The results are based on 1000 simulation runs for a system consisting of 200 mobile nodes. The x-axis shows the percentage of dropped calls for the optimum (global knowledge reference). The y-axis shows the percentage of dropped calls.

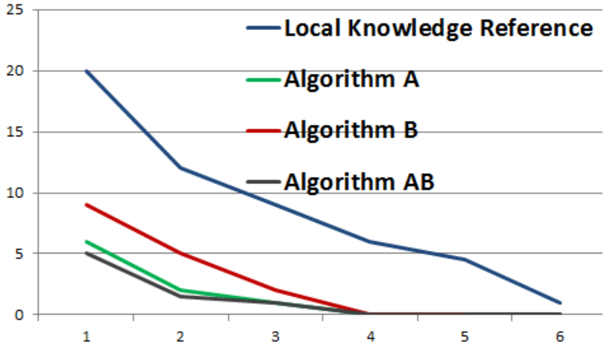


Figure 10. Percent of users received dropped calls, dynamic scenario.The x-axis shows the number of iterations of the algorithm. The y-axis shows the percentage of decision errors. The results are based on 1000 test runs.

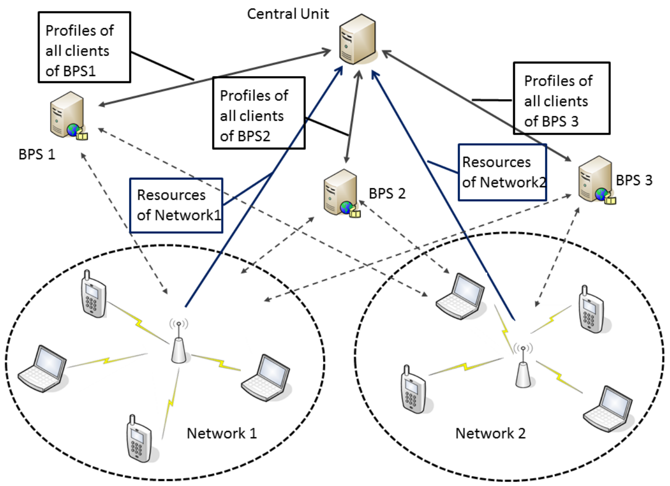

Central Unit
Profiles of all clients of BPS1
Profiles of all clients of BPS2
Profiles of all clients of BPS 3
BPS 1
BPS 2
BPS 3
Resources of Network1
Resources of Network2
Network 1
Network 2

Figure 11. Centralized Approach: Mobile nodes convey their data to their BPSs, the BPSs send the data to the central unit. The central unit collects information about resource availability from the networks, performs the network selection, sends results to the BPSs. The BPSs send the results to their mobile nodes.

are denoted by b_i; streaming bitrate requirements of mobile nodes that request content s_k are denoted by r_k; $rss_{i,j}$ is the received signal strength in network n_i for node m_j, while power consumption and the cost of service in network n_i for node m_j are denoted by $p_{i,j}$ and $c_{i,j}$, respectively.

For each node m_j, we define node preferences that are described by a tuple containing Th^p_j, Th^c_j, and Th^{rss}_j. These denote thresholds for, respectively, power consumption, cost of service and received signal strength. Without loss of generality, we consider these three parameters in our work, however, this list can include other conditions. The thresholds for these parameters are determined by each node according to its own optimization policies and used as an input that constrains optimization solved locally by each mobile node for available networks. The objective function is also defined based on the node's optimization policies. Its definition is beyond the scope of our work.

$$\forall\{i,j\}:\delta(i,j)\cdot p_{i,j}\leq Th^p_j \tag{14}$$

$$\forall\{i,j\}:\delta(i,j)\cdot c_{i,j}\leq Th^c_j \tag{15}$$

$$\forall\{i,j\}:\delta(i,j)\cdot rss_{i,j}\geq Th^r ss_j \tag{16}$$

The output of this optimization is a list of mobile networks that satisfy the user's requirements. It is further referred to as a node's *network profile* captured by a function δ_j. This function is used as input for computing an optimal allocation of mobile nodes to the available networks in the model and is defined as follows.

$$\delta_j(i)=\begin{cases}1, \text{if } n_i \text{ is selected by } m_j\\ 0, \text{otherwise}\end{cases} \tag{17}$$

We define a binary decision variable $x_{i,k}$ as follows:

$$x(i,k)=\begin{cases}1, \text{if } n_i \text{ is allocated for } s_k\\ 0, \text{otherwise}\end{cases} \tag{18}$$

To find the best possible allocation of the requested streams to the available networks in terms of minimization of consumed bandwidth, we minimize the following objective function:

$$\min\sum_{n_i\in N}\sum_{s_k\in S}x_{i,k}\cdot r_k \tag{19}$$

The objective function is subject to the set of constraints given below.

For each mobile node m_j, we need to guarantee that it can receive the requested content from at least one network belonging to nodes profile. We need to specify that user preferences defined in their profiles are satisfied.

$$\forall\{j\}:\sum_j\delta_j(i)\cdot x_{i,k}\geq 1 \tag{20}$$

For each network, the availability of its bandwidth is checked.

$$\forall\{i\}:\sum_k x_{i,k}\cdot r_k\leq b_i \tag{21}$$

After the results for $x_{i,k}$ are computed, these are send to the nodes. If there are several networks that receive the requested content, a node can narrow its selection criteria to choose among these alternatives.

The defined problem is a typical location allocation problem that belongs to a class of integer programming problems. To solve this problem, we have taken advantage of the GNU Linear Programming Kit (GLPK) version 4.49 [50]. This is an ANSI C package that is intended for solving large-scale linear programming and mixed integer programming problems. We tested the performance of the package for solving the aforementioned problem that consisted of respectively 500 and 1000 nodes, 5 mobile networks and 10 different streaming contents

(505 and 1005 constraints respectively and 50 variables). For this test, the constraint matrix and the coefficients of the objective function were randomly generated, as in the Monte Carlo simulation. For a 2.83GHz Intel processor, the average CPU time estimates based on 1000 algorithm runs are 710 ms and 230 ms for 1000 nodes and 500 nodes respectively. Though these estimates are computer configuration-specific, they show that the problem can be solved within reasonable time for a relatively large number of nodes. Here, we assume that all necessary information is locally available for the computation. For a component, the CPU time can be precomputed and used as a threshold for deciding whether or not the optimization can be applied to a particular problem scope. On the other hand, collecting this information from different network locations can become a bottleneck for the algorithm operation.

Please note, that a node can, in fact, exploit the multipath streaming scheme and receive data from several networks concurrently. The problem is then reduced to a class of linear programming problems, which is less complex to solve using, for example, the Simplex method. For this problem, the variable $x_{i,k}$ denotes the share of the content s_k delivered to the network n_i.

VII. System Components, their Functions and Feedback Exchange for Multicast Transmission in Heterogeneous Wireless Networks

In this section, we look at architectural aspects of a system that supports the scenario considered in Section III. For this, we consider several entities and decision making components that are responsible for decisions in the system and have access to different information necessary for optimal regrouping of the mobile users. Since the decision-making process requires access to various data that originate from different network components, our system needs to implement a signaling infrastructure for the exchange of such information.

A. BPS Component

The BPS component runs on a backbone server. It either hosts the content or acts as a proxy server that re-sends the content to the user. The component maintains multimedia sessions and controls multicast groups. It receives and processes feedback from its mobile clients and the access networks these clients are connected to. Based on results of processing the data, the component can trigger the network selection. The component can also send data to other components upon their requests.

B. Mobile Network Component

This component is located inside a mobile network. It monitors network's resources including available bandwidth. It also maintains various information about clients and multicast groups of the network. The component implements the following: *1*) processes this information; *2*) sends the information to other components for further processing; and *3*) initiate network selection for multicast groups.

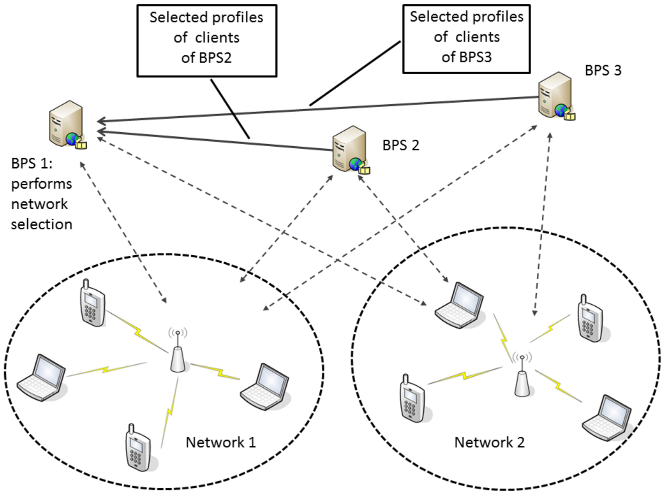

Selected profiles of clients of BPS2
Selected profiles of clients of BPS3
BPS 3
BPS 2
BPS 1: performs network selection
Network 1
Network 2

Figure 12. Decentralized Approach: BPSs convey some information to the BPS or other component that is elected to compute the selection. This component also collects information about resource availability from the networks and performs the network selection. The results are sent back to participated BPSs. The BPSs send the results to their mobile nodes.

C. Mobile Client Component

This component runs on mobile device and maintains its wireless channel state information monitoring the availability of mobile networks and received signal strength in these networks. It maintains its preferences towards these networks based on the following: *1*) power consumption in these networks; *2*) security issues; *3*) network cost of service; and *4*) channel state information.

D. Information Exchange

To achieve an optimal solution, all required information need to be exchanged between the decision components and to be communicated to some centralized unit that computes this optimal allocation, as shown in Figure 11. This centralized unit can be a predefined component of the system that other components are aware of or it can be elected on a vote-basis among the components from the mobile networks and the backbone network. The centralized approach demands the exchange of network data across the mobile networks on a fast time scale and low-latency basis. This makes it unrealistic to implement the algorithms in practice for large scale networks. It also requires the implementation of a centralized component that runs this operation and solving a computational problem of high complexity in real time.

The information overhead can be overcome by using distributed designs as illustrated in Figure 12. As an alternative to the centralized approach, we present a distributed solution in which the networks and the BPSes handle the problem completely independently from each other or in a cooperative and coordinated manner with some limited information exchange between the components. We consider the following problems:

1) How much information of any given BPS/network/user data needs to be exchanged?

2) From the architectural design, how to disseminate this information mostly efficiently?

In our work, we focus on the usage of the application specific message (APP) of the RTP/RTCP protocol suite [51] to convey client-related information to the respective mobile networks and BPS components. These protocols are designed for multicast architectures with multicast channels specified for data transmission from the sender to the receivers. In the considered model, only BPS and network components receive feedback from the clients, meaning that the client-to-client feedback exchange is not required. In addition, these feedback reports do not consume much bandwidth and are sent only when the client-related information is changed. Therefore, to deliver feedback reports, unicast transmission is used.

VIII. Decentralized Approach for Multicast Transmission in Heterogeneous Wireless Networks

In this section we consider two decentralized approaches that solve the previously defined network selection problem. Since not all knowledge for the network selection is available on these nodes the algorithms make their decisions based on the currently available knowledge. We consider the solutions (*1*) when all backbone proxies in the system perform the network selection independently from each other, further referred to as a *BPS solution*, and (*2*) when an access network performs the network selection for a set of multicast groups, further referred to as a *mobile network (MN) solution*.

Both approaches rely on the information acquired from mobile nodes regarding their network profiles. To maintain its network profile, the mobile node component periodically monitors the availability of mobile networks. As the node does not need to keep all its interfaces active all the time and the power consumption in idle mode is less than under receiving of data, this operation is not expected to drastically increase the battery use.

To disseminate the information, we make use of the application specific message (APP) of the RTP/RTCP protocol [51]. After a node's network profile changes, for example, a new network becomes available, this information is packed into the APP message. It requires that this protocol is implemented for communication and all needed modifications that allow interpreting of the message are made. This way, the information can also becomes available for the access network. If the RTP/RTCP protocol is not used, any application layer messaging can be implemented to convey feedback from nodes to other components for further processing.

A. BPS Solution

In the BPS solution, all backbone proxies in the system perform the network selection for their clients independently from each other, that is no cooperation or information exchange is performed between different proxies. We consider two versions of the solution: *a*) the network selection is run for all multicast groups of the BPS and *b*) the network selection involves only multicast groups that receive the same content.

```
At Backbone Proxy Server
read APP messages from RTP/RTCP stream
maintain nodes' network profiles
if network profile is changed
   if threshold $\leq$ number of nodes
      do selection for all nodes
   else
      do selection for one content
foreach node that changes network
   instruct node to change network: send switch message
```

(a) Backbone Proxy Server Component

```
At Mobile Node $m_j$
monitor available networks
if new network is available
   compute new network profile
   send APP message with network profile to BPS
wait for response from BPS
upon reception switch message from BPS
   switch to new mobile network
```

(b) Mobile Node Component

Figure 13. Network Selection Algorithm for BPS

These two versions perform the same operations and differ only in scale of involved nodes.

The network selection algorithm is depicted in Figure 13. It is initiated by the backbone proxy component when network profiles of nodes receiving any of the proxy's multicast streaming sessions change. The changes include: a new node joins the session, one of the nodes leaves the session, or a network profile of any node is updated, e.g., a new network becomes available. To check whether a reconfiguration of multicast groups is needed, the BPS solves the optimization problem defined in Section VI. We use the CPU threshold, also discussed in Section VI, to determine which of the two above versions of the algorithm to apply.

We solve the problem for all multicast groups of the BPS if the threshold allows for it. Otherwise, the problem is solved for one content only. We shall evaluate both possible algorithms separately in Section X. That way, the effect of reducing the problem's input data can be shown better.

B. Mobile Network Solution

In this solution, we consider an access network that initiates and performs the network selection for a set of multicast groups. The network maintains nodes' network profiles based on information extracted from the RTP/RTCP stream. The network selection operation is triggered by the network component when the network's available bandwidth goes below a predefined threshold. For each multicast group, the network defines a set of networks that is a conjunction of nodes' network profiles that comprise the group. The groups with high cardinality of such sets are selected. The network requests the BPSs that host the content received by the selected groups about network profiles of other nodes from other networks receiving the same content. The network solves the optimization problem defined in Section VI for the nodes detected by the above selection operation. The threshold is adjusted accordingly, if the result of the optimization exceeds the

```
At Mobile Network $n_i$
read APP messages from RTP/RTCP stream
maintain nodes' network profiles
monitor available bandwidth
if available bandwidth $\geq$ threshold
    compute conjunctions for each multicast group
    order groups by cardinality of conjunction sets
    select candidates for optimization
    request network profiles from involved BPSs
upon replies from servers
    do selection for defined scope of problem
foreach node $m_j$ that changes network
  if $m_j \in n_i$
      instruct node to change network: send switch message
  else
      send response to server
```

(a) Mobile Network Component

```
At BPS
upon request from mobile network
    select network profiles for requested contents
    send network profiles to mobile network
upon response from mobile network
  foreach node that changes network
      advise node to change network
```

(b) BPS Component

```
At Mobile Node $m_j$
monitor available networks
if new network is available
   compute new network profile
   send APP message with network profile to BPS
wait for response from BPS
upon reception switch message from BPS
   switch to new mobile network
wait for response from mobile network
upon reception switch message from mobile network
   switch to new mobile network
```

(c) Mobile Node Component

Figure 14. Network Selection Algorithm for Mobile Network

current threshold of the bandwidth. In this case, the network component periodically checks the bandwidth to detect if the threshold can be reduced to its previous value.

The selection algorithm depicted in Figure 14 requires the implementation of infrastructure that supports interactivity between the involved components and signaling mechanisms that invoke exchange of information about users' profiles and network conditions and initiate network selection.

IX. LTE-Advanced HetNet Approach

In this section, we consider a solution for a network system discussed in Section II-E that is depicted in Figure 1. Usually in such systems, small cells are connected to a macrocell via the wired backhaul infrastructure, thus, wireless resources are not used for the exchange of control messages and the exchange of the channel information and multicast decisions can be performed within reasonable time period. More importantly, these networks combine a high transmit power and long range base station with several lower power short range stations. Handoff of users moving with high velocity to short range small cells will require a new handoff soon when the user leaves this cell. Therefore, from the point of view of network capacity and avoidance of the ping-pong effect, it is important to consider the mobile terminal velocity for network selection. Here, we assume that a mobile network is capable to predict the residence time of a mobile node inside a cell of the network based on terminal velocity, the local area, movement patterns, and other statistical information. We base our decision making process on research done by other authors [43–45]. While the prediction of the residence time of mobile nodes is ongoing work, we consider this beyond our scope. For the purpose of this paper, we assume that the prediction can be performed with acceptable precision.

X. Simulations for Multicast Transmission in Heterogeneous Wireless Networks

In this section, we evaluate the algorithms described in Section VIII. We present the simulation setup, then the evaluation metrics, and finally discuss the simulation results.

A. Simulation Setup

Because full-scale field experiments for several wireless networks and several hundred users are problematic and expensive to carry out, we used simulations to evaluate the validity of the proposed solutions. The performance and functionality of the system is analyzed through multiple simulation runs. We evaluate two types of network systems: a general heterogeneous wireless network system and an LTE-Advanced HetNet network.

1) Setup for Users and Streaming Content: For the simulations, we consider a scenario with five backbone streaming servers each having five different streaming contents. Video streaming content is used for the evaluation. In terms of required bandwidth, we divide the requested content into five categories: 500 kbps, 800 kbps, 1200 kbps, 1800 kbps, and 2400 kbps. These rates are recommended bit rates for live streaming for the Adobe Media Server [52]. Further, we consider that mobile users are randomly assigned to the servers and their content. The users arrive at one user per time unit, and they stay in the system for 200, 300 or 400 time units. This time period is randomly selected for each user. Except for the initial stage of 200 time units, there are always at least 200 users in the system. The initial stage is excluded from the evaluation.

2) Modeling of Movements: We realise that usage of real world traces evaluates performance only for these particular scenarios. Therefore, we chose to use random generated sintetic data since these data allow more comprehensive performance evaluation by using a large number of variations. Several parameters for modeling of movements are important for our evaluation. For the simulation of movements of mobile nodes, we looked at different studies concerning mobility models for the wireless communications [53, 54]. In the random waypoint model, the location of mobile nodes, their velocity and direction of the movement are chosen randomly and independently of other nodes. We captured the randomness

```
**At Backbone Proxy Server**
*for each* request { $\delta_j(i)$, $s_k$ } received from mobile node $m_j$
    define set of mobile networks $N_k$ receiving $s_k$ and satisfying $\delta_j(i)$
    *if* $N_k \neq \emptyset$
        return $N_k$ to node $m_j$
    *else* return *NULL*
```

(a) Backbone Proxy Server Component

```
**At Mobile Node** $m_j$
compute $\delta_j(i)$
send request { $\delta_j(i)$, $s_k$ } to Backbone Proxy Server
*wait for* response from Backbone Proxy Server
upon *reply* from Backbone Proxy Server
    *if reply* $\neq$ *NULL*
        select network from *reply*
    *else* select network from network profile
```

(b) Mobile Node Component

Figure 15. Algorithm for Computing Lower Bound Reference

of these parameters by random time during which any mobile network in consideration is available to a mobile user.

In the simulation, we also have a number of users who do not move, e.g., people in public places like internet café or train stations. For these users, the subset of the available networks is the same during the whole simulation run. We also distinguish between single users moving alone and groups of users moving together using, e.g., public transport. For users belonging to the same group, the availability of the networks changes likewise while the streaming content can differ. Since we do not have any observations for realistic distribution of these different types of users, we model them as roughly equally distributed, varying from 25% to 40% of users in each group.

3) General Heterogeneous Wireless Network Setup: In this setup, we consider a scenario with four wireless networks. The networks only cover parts of the area in consideration. Therefore, only a selection of them is available for each user at a given time. We have implemented different scenarios for network availability. In all these scenarios, each user has access to at least one network continuously during the whole session of an experiment run. Some users have access to all networks during the whole session. We vary the percentage of these users in different scenarios. For the rest of the users and networks in each scenario, the users can access these networks during some period of the session randomly chosen from the session duration.

4) LTE Heterogeneous Network Scenario: We evaluate an LTE Advanced Heterogeneous Network scenario, the so called macro-pico scenario [38], with several picocells deployed inside a macrocell as illustrated in Figure 1. We consider four picocells deployed inside one macrocell. All users can access the macrocell and some of users have access to one of picocells. The access areas of the picocells do not overlap. The picocells that the users can access during the session are randomly chosen for each user. Accordingly, the periods when picocells are available for the users are randomly chosen from the session duration.

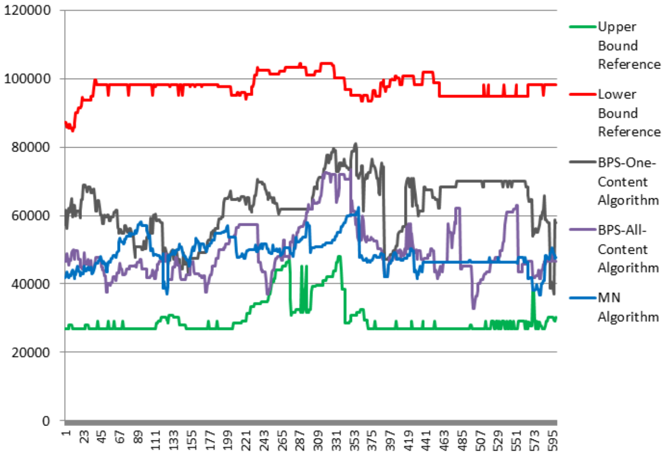


Figure 16. Total Bandwidth Consumption for simulations with background traffic changes applied. The x-axis shows time units. The y-axis shows consumed bandwidth in kbits. The results are an average of 500 simulation runs. The duration of one simulation run is 600 time units.

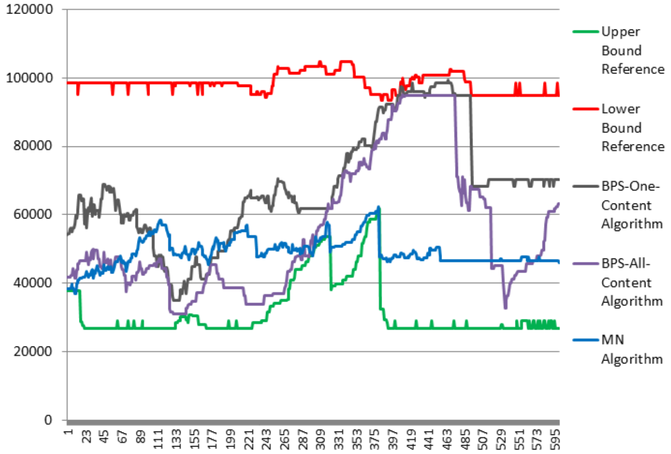


Figure 17. Total Bandwidth Consumption for simulations with applied background traffic changes and insertion of a flashcrowd. The x-axis shows time units. The y-axis shows consumed bandwidth in kbits. The results are an average of 500 simulation runs. The duration of one simulation run is 600 time units.

B. Evaluation Metrics

To evaluate the algorithms, we define upper and lower bounds to their operation. By the *upper bound*, we mean the theoretical best possible allocation of nodes to multicast groups that can be achieved in terms of resource utilization. It is established by applying a centralized solution with fully shared knowledge of the conditions in all evaluated networks and preference profiles of all nodes. It is defined in Section VI-A and is further referred to as an *upper bound reference*.

The lower bound corresponds to an algorithm depicted in Figure 15. Upon a request from a node, which specifies the requested content s_k and the node's network profile $\delta_j(i)$, the server assigns the node to a multicast group, if any that satisfies the node's network profile exists. If no such

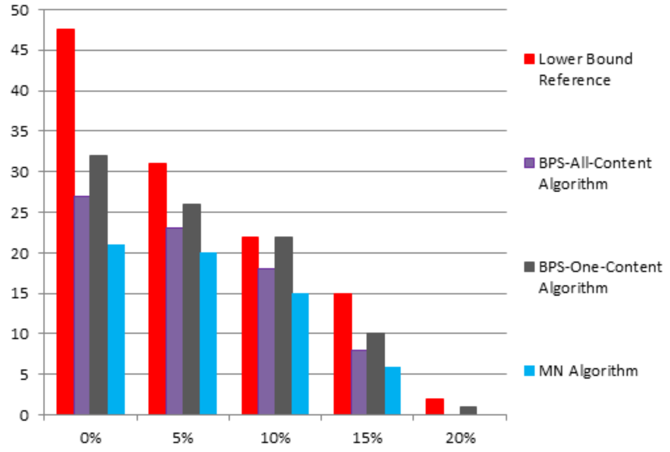


Figure 18. Dropped Connections. The x-axis shows the percentage of bandwidth surplus compared to the Upper Bound Reference. The y-axis shows the number of dropped connections for each algorithm. The number of dropped connections for the Upper Bound Reference is 0 for all algorithm runs. The results are an average of 2000 simulation runs.

group exists, the server opens streaming to a network that meets the user's preferences. This simple algorithm uses a low number of operations and input information to establish multicast groups. We decided to use this algorithm instead of a trivial unicast scenario to provide a fairer comparison of the proposed solutions. In our evaluation, it is referred to as a *lower bound reference*. The lower bound reference is also used as a initialization algorithm for the mobile network solution. Note that the MN algorithm is triggered only when the available resources in a network are dropped below a certain threshold. Thus, some initialization is required and the MN solution evaluated in this section is a hybrid approach.

We evaluate the performance of the algorithms using two performance metrics: total bandwidth consumption and a number of dropped calls, as follows.

Total bandwidth consumption is a direct measure of the bandwidth usage of all multicast groups in the system. We measure the total bandwidth consumption for two different bandwidth variation profiles for the available bandwidth of the access networks. These bandwidth variations model the background traffic for each network.

1) Changes in bandwidth are applied to all access networks, and their values are normally distributed in the range $[-0.1, 0.1]$ of the currently available bandwidth. The results for this test are depicted in Figure 16.
2) In addition to the bandwidth fluctuations formulated above, we simulate a flashcrowd scenario with an number of nodes arriving within short-time intervals. For all access networks, we inserted from 30 to 50 additional nodes once for each session. The simulation results are depicted in Figure 17.

Note that none of the networks have enough capacity to accommodate all multicast groups alone even if the nodes are optimally allocated to their networks. All four networks must be used in order to meet the requirements of all users.

The number of dropped connections is measured for all multicast groups in the system for the following five condition cases: *1*) The total available bandwidth of the system equals to the bandwidth utilized by the upper bound reference. *2*) The total available bandwidth exceeds the amount of bandwidth utilized by the upper bound reference by 5%; *3*) by 10% ; *4*) by 15%; *5*) by 20%, respectively. To compute the setup for this type of tests, we applied a relaxation to the upper bound reference. The optimization problem is relaxed by removing the network bandwidth constraint depicted in Eq. (21). The available bandwidths of the evaluated networks are then calculated from the optimization results. For these tests, bandwidth fluctuations have not been applied. Also, once assigned, the availability of the networks does not change within a test run, meaning that the algorithms have been evaluated statically. The results for this test are depicted in Figure 18.

C. Performance Results

The simulation results are drawn from the average of 500 simulation runs to evaluate the total bandwidth consumption, and of 2000 simulation runs to evaluate the number of dropped connections that are discussed in Section X-B. The performance metrics to evaluate the total bandwidth consumption are collected for 600 time units. We excluded results for which the optimum solution has not been found, i.e., when no optimum existed. For the evaluation of dropped connections, we also excluded the top and bottom 5% of the results of all performed tests.

When evaluated in terms of consumed bandwidth, the tests show that applying any of the proposed solutions can save up to 50% of available bandwidth if compared to the lower bound reference. As expected, the all-content version of the BPS algorithm gives better results than the one-content version though it requires longer processing time and, what is more important, more signaling and reconfigurations for mobile nodes. The trade-off between these two versions can be studied. The MN algorithm behaved very close to the BPS algorithm most of the time.

For the flashcrowd scenario, the MN algorithm was able to handle better the insertion of new nodes than the BPS algorithms. This can be explained by the fact that the MN algorithm is applied across several cutting planes of the total solution space while the BPS algorithm is applied across one cutting plane. In other words, the BPS algorithm relies only on the information from one server while the MN algorithm takes information from several servers as an input. Certainly, the MN algorithm requires exchange of significantly more information across the system. To disseminate this information, we need to implement an appropriate protocol and develop mechanisms that allow the components, namely the networks and the BPSs, to cooperate with each other and exchange information about their users and network conditions, which also involves certain security considerations. Contrary to the MN algorithm, the operation of the BPS algorithm can thoroughly rely on information received from the RTP/RTCP feedback messages.

Evaluations of dropped connections show the same tendency as the evaluation of consumed bandwidth. Both the MN solution and the BPS solution give a good reduction of dropped connections, up to 50%, if compared to the lower bound reference. Compared to the BPS solution, the MN algorithm gives roughly from 10% to 20% less dropped connections.

As expected, the BPS algorithm for all streams performs better than the algorithm for one stream, giving roughly 10% to 20% difference in evaluation. At the same time, the BPS algorithm for one stream still gives good reduction in dropped connections and consumed bandwidth compared to the lower bound reference. Therefore, this version can be applied if solving the all-streams version is not computationally reasonable.

XI. Conclusion

The paper studied the problem of load balancing and forming mobile multicast groups in heterogeneous network environments. An efficient decentralized network selection solution is important for future mobile networks, since it improves utilization of the network resources and QoS of users and reduces signalling overheads. We study how the solution results depend on the information available for the decision making. The problem is considered for a multi-stream multi-server scenario. The candidate networks are selected for multicast groups based on their mobile nodes' preferences and available resources of the networks.

For load balancing scenario, the solution provides a substantial improvement in reduction of decision errors and signalling overhead in comparison to the work specified in Section II. The simulation results of our algorithms show that blocked calls can be reduced with approximately 60-50 % compared to the local knowledge reference. The test results do not differ much for 100, 200 and 300 users, and we expect that these results can be extended to the general case.

All three evaluated algorithms deliver similar results in terms of number of blocked calls. The implementation of Algorithms AB and B requires development of mechanisms for synchronizing information about the network conditions and careful security considerations when information from one network is available to other networks. Operation of Algorithms AB and B requires significantly higher signalling between the networks and the users. We therefore conclude that Algorithm A is to be preferred over Algorithms AB and B.

For multicast scenario, we proposed two solutions that establish multicast groups and assign them to networks based on incomplete information of the whole system. The operation is also performed by different components of the system with limited cooperation between the components.

Compared to the work specified in Section II, our main achievement is decentralization of the network selection for multicast groups, consideration of the impact of several multicast groups and incompleteness of information. An efficient decentralized network selection solution for multicast is important for future mobile networks, since it improves utilization of the network resources and QoS of users and reduces signaling overheads.

We studied how the solution results depend on the information sets available for the decision making. Evaluating dropped connections shows that both algorithms provide a substantial improvement in reduction of dropped connections compared the the lower bound reference. According to our findings, the MN algorithm performs better than the BPS algorithm.

In terms of consumed bandwidth, both solutions deliver similar results for monotonous variations in available bandwidth and arrivals of nodes. For the tests with insertions of extra users, the MN solution performs better than the BPS solution. However, the operation of the MN solution requires complex signaling across several mobile networks and BPSs. In addition, it requires implementation and deployment of mechanisms and communication protocols that provide cooperation between the involved components. The disadvantage of using the BPS algorithm is the necessity of the network reconfiguration of mobile nodes each time the network profile of a node changes. Therefore, a mechanism that is similar to monitoring bandwidth threshold in the MN algorithm can be considered as a next improvement.

As a further step, we intend to investigate how the system can benefit from joint operation of these solutions and limited feedback signaling. We need to implement mechanisms that detect which of two solutions is preferable for certain events. Since, the BPS solution deliver good results for the most of the cases, the operation of the MN solution is going to be triggered only under predefined circumstances. Thus, we avoid unnecessary messaging between the components. We also intend to perform more expanded tests by extending the simulation scenarios by, for example, taking down one of the networks during the simulation.

Finally, we mention that the centralized approach still can be applied for some scenarios and network configurations like small cell networks deployed by the same provider. We intend to investigate better the conditions for applying this approach and consider also implementation of partially centralized solutions.

Acknowledgment

The work described in this paper has been conducted as a part of the ADIMUS (Adaptive Internet Multimedia Streaming) project, which is funded by the NORDUnet-3 programme.

References

[1] S. Boudko, W. Leister, and S. Gjessing, "Multicast group management for users of heterogeneous wireless networks," in CONTENT 2012: The Fourth International Conference on Creative Content Technologies. International Academy, Research and Industry Association (IARIA), 2012, pp. 24–27.

[2] ——, "Optimal network selection for mobile multicast groups," in ICSNC 2012 The Seventh International Conference on Systems and Networks Communications. In-

ternational Academy, Research and Industry Association (IARIA), 2012, pp. 224–227.

[3] ——, “Team decision approach for decentralized network selection of mobile clients,” in Proceedings of 2012 5th Joint IFIP Wireless and Mobile Networking Conference. IEEE Computer Society, 2012, pp. 88–94.

[4] S. Boudko and W. Leister, “Network selection for multicast groups in heterogeneous wireless environments,” in MoMM ’13: Proc. 11th Int’l Conf. on Advances in Mobile Computing and Multimedia. ACM, 2013, pp. 167–176.

[5] I. Gillott et al., “The potential for LTE broadcast/eMBMS,” iGR, white paper, January 2013.

[6] 3GPP, “Multimedia Broadcast/Multicast Service (MBMS); Stage 1,” 3rd Generation Partnership Project (3GPP), TS 22.146, Jun. 2008, [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/22146.htm, accessed September 6, 2013.

[7] ——, “LTE; evolved universal terrestrial radio access (E-UTRA); long term evolution (LTE) physical layer; general description,” ETSI, technical specification 3GPP TS 36.201, 2009, version 8.3.0 Release 8.

[8] Ericsson AB, Qualcomm Technologies, Inc., and Qualcomm Labs, Inc., “LTE broadcast,” white paper, February 2013, [Online]. Available: http://www.ericsson.com/res/docs/whitepapers/wp-lte-broadcast.pdf, accessed September 6, 2013.

[9] T. L. Paramvir, T. Liu, P. Bahl, S. Member, and I. Chlamtac, “Mobility modeling, location tracking, and trajectory prediction in wireless atm networks,” IEEE J. Sel. Areas Commun., vol. 16, 1998, pp. 922–936.

[10] I. F. Akyildiz and W. Wang, “The predictive user mobility profile framework for wireless multimedia networks,” IEEE/ACM Trans. Netw, vol. 12, 2004, pp. 1021–1035.

[11] C.-C. Tseng, L.-H. Yen, H.-H. Chang, and K.-C. Hsu, “Topology-aided cross-layer fast handoff designs for IEEE 802.11/mobile IP environments,” Communications Magazine, IEEE, vol. 43, no. 12, Dec. 2005, pp. 156–163.

[12] Y.-H. Choi, J. Park, Y.-U. Chung, and H. Lee, “Cross-layer handover optimization using linear regression model,” in ICOIN 2008. Int’l Conf. on Information Networking, Jan. 2008, pp. 1–4.

[13] S. Ray, K. Pawlikowski, and H. Sirisena, “Handover in mobile wimax networks: The state of art and research issues,” IEEE Communications Surveys Tutorials, vol. 12, no. 3, 2010, pp. 376–399.

[14] Q. Song and A. Jamalipour, “A network selection mechanism for next generation networks,” in IEEE Int’l Conf. on Communications, ICC, vol. 2, May 2005, pp. 1418–1422.

[15] A. Hasswa, N. Nasser, and H. Hassanein, “Tramcar: A context-aware cross-layer architecture for next generation heterogeneous wireless networks,” in IEEE Int’l Conf. on Communications, ICC, vol. 1, Jun. 2006, pp. 240–245.

[16] F. Zhu and J. McNair, “Optimizations for vertical handoff decision algorithms,” in IEEE Wireless Communications and Networking Conference, WCNC2004, vol. 2, Mar. 2004, pp. 867– 872.

[17] P. M. L. Chan, Y. F. Hu, and R. E. Sheriff, “Implementation of fuzzy multiple objective decision making algorithm in a heterogeneous mobile environment,” in IEEE Wireless Communications and Networking Conference, WCNC2002, vol. 1, Mar. 2002, pp. 332–336.

[18] L. Xia, L. Jiang, and C. He, “A novel fuzzy logic vertical handoff algorithm with aid of differential prediction and pre-decision method,” in IEEE Int’l Conf. on Communications, ICC, Jun. 2007, pp. 5665–5670.

[19] N. Nasser, S. Guizani, and E. Al-Masri, “Middleware vertical handoff manager: A neural network-based solution,” in IEEE Int’l Conf. on Communications, ICC, Jun. 2007, pp. 5671–5676.

[20] G. Karetsos, E. Tragos, and G. Tsiropoulos, “A holistic approach to minimizing handover latency in heterogeneous wireless networking environments,” Telecommunication Systems, 2011, pp. 1–14.

[21] A. H. Zahran, B. Liang, and A. Saleh, “Signal threshold adaptation for vertical handoff in heterogeneous wireless networks,” Mob. Netw. Appl., vol. 11, Aug. 2006, pp. 625–640.

[22] C. W. Lee, L. M. Chen, M. C. Chen, and Y. S. Sun, “A framework of handoffs in wireless overlay networks based on mobile IPv6,” IEEE J. Sel. Areas Commun., vol. 23, no. 11, Nov. 2005, pp. 2118–2128.

[23] K. Yang, I. Gondal, B. Qiu, and L. Dooley, “Combined SINR based vertical handoff algorithm for next generation heterogeneous wireless networks,” in IEEE Global Telecommunications Conference, GLOBECOM ’07, Nov. 2007, pp. 4483–4487.

[24] C. Chi, X. Cai, R. Hao, and F. Liu, “Modeling and analysis of handover algorithms,” in IEEE Global Telecommunications Conference, GLOBECOM ’07, Nov. 2007, pp. 4473–4477.

[25] N. Nasser, A. Hasswa, and H. Hassanein, “Handoffs in fourth generation heterogeneous networks,” IEEE Communications Magazine, vol. 44, no. 10, Oct. 2006, pp. 96–103.

[26] G. Zhang and F. Liu, “An auction approach to group handover with mobility prediction in heterogeneous vehicular networks,” in ITS Telecommunications (ITST), 2011 11th International Conference on, aug. 2011, pp. 584–589.

[27] L. Sun, H. Tian, and P. Zhang, “Decision-making models for group vertical handover in vehicular communications,” Telecommunication Systems, vol. 50, 2012, pp. 257–266.

[28] O. Ormond and J. Murphy, “Utility-based intelligent network selection,” in IEEE Int’l Conf. on Communications, ICC, 2006.

[29] A. Gluhak, K. Chew, K. Moessner, and R. Tafazolli, “Multicast bearer selection in heterogeneous wireless networks,” in IEEE Int’l Conf. on Communications, ICC,

vol. 2, May 2005, pp. 1372–1377.

[30] I.-S. Jang, W.-T. Kim, J.-M. Park, and Y.-J. Park, “Mobile multicast mechanism based mih for efficient network resource usage in heterogeneous networks,” in Proc. of the 12th Int’l Conf. on Advanced Communication Technology, ser. ICACT’10, 2010, pp. 850–854.

[31] E. Tragos, G. Tsiropoulos, G. Karetsos, and S. Kyriazakos, “Admission control for QoS support in heterogeneous 4G wireless networks,” Network, IEEE, vol. 22, no. 3, 2008, pp. 30–37.

[32] M. A. Khan, A. C. Toker, F. Sivrikaya, and S. Albayrak, “Cooperation-based resource allocation and call admission for wireless network operators,” Telecommunication Systems, vol. 51, 2012, pp. 29–41.

[33] D.-N. Yang and M.-S. Chen, “Efficient resource allocation for wireless multicast,” IEEE Transactions on Mobile Computing, vol. 7, no. 4, Apr. 2008, pp. 387–400.

[34] M. L. Fisher, “The lagrangian relaxation method for solving integer programming problems,” Manage. Sci., vol. 50, no. 12 Supplement, Dec. 2004, pp. 1861–1871.

[35] F. Hou, L. Cai, P.-H. Ho, X. Shen, and J. Zhang, “A cooperative multicast scheduling scheme for multimedia services in ieee 802.16 networks,” IEEE Transactions on Wireless Communications, vol. 8, no. 3, 2009, pp. 1508–1519.

[36] Multicast Mobility Working Group, “Charter for Working Group,” 2010, [Online]. Available: http://datatracker.ietf.org/wg/multimob/charter/, accessed July 30, 2013.

[37] 3GPP, “LTE; evolved universal terrestrial radio access (E-UTRA) and evolved universal terrestrial radio access network (E-UTRAN); overall description; stage 2,” ETSI, technical specification 3GPP TS 36.300, 2013, version 11.6.0 Release 11.

[38] ——, “Evolved universal terrestrial radio access (E-UTRA); mobility enhancements in heterogeneous networks,” ETSI, technical specification 3GPP TS 36.839, 2013, version 11.1.0 Release 11.

[39] W. Leister, T. Sutinen, S. Boudko, I. Marsh, C. Griwodz, and P. Halvorsen, “An architecture for adaptive multimedia streaming to mobile nodes,” in MoMM ’08: Proc. 6th Int’l Conf. on Advances in Mobile Computing and Multimedia. ACM, 2008, pp. 313–316.

[40] G. Xylomenos, V. Vogkas, and G. Thanos, “The multimedia broadcast/multicast service,” Wireless Communications and Mobile Computing, vol. 8, no. 2, 2008, pp. 255–265.

[41] R. Radner, “Team decision problems,” Ann. Math. Statist., vol. 33, no. 3, 1962.

[42] Y.-C. Ho, “Team decision theory and information structures,” Proceedings of the IEEE, vol. 68, no. 6, june 1980, pp. 644–654.

[43] B. Liang and Z. J. Haas, “Predictive distance-based mobility management for multidimensional PCS networks,” IEEE/ACM Trans. Netw., vol. 11, no. 5, Oct. 2003, pp. 718–732.

[44] I. Akyildiz, J. S. Ho, and Y.-B. Lin, “Movement-based location update and selective paging for PCS networks,” IEEE/ACM Trans. Netw, vol. 4, no. 4, 1996.

[45] G. Yavas, D. Katsaros, O. Ulusoy, and Y. Manolopoulos, “A data mining approach for location prediction in mobile environments,” Data Knowl. Eng., vol. 54, August 2005, pp. 121–146.

[46] B. Jabbari, Y. Zhou, and F. Hillier, “A decomposable random walk model for mobility in wireless communications,” Telecommunication Systems, vol. 16, 2001, pp. 523–537.

[47] M. Canales, J. Gállego, Á. Hernández, and A. Valdovinos, “An adaptive location management scheme for mobile broadband cellular systems,” Telecommunication Systems, 2011, pp. 1–17.

[48] A. Ulvan, R. Bestak, and M. Ulvan, “Handover procedure and decision strategy in lte-based femtocell network,” Telecommunication Systems, 2011, pp. 1–16.

[49] G. Pongor, “Omnet: Objective modular network testbed,” in MASCOTS ’93: Proc. Int’l Workshop on Modeling, Analysis, and Simulation on Computer and Telecommunication Systems. Society for Computer Simulation, 1993, pp. 323–326.

[50] A. Makhorin, “GLPK (GNU Linear Programming Kit),” Free Software Foundation, 2010–2012, [Online]. Available: http://www.gnu.org/software/glpk/, accessed September 6, 2013.

[51] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, “RTP: A Transport Protocol for Real-Time Applications,” RFC 3550 (Standard), Jul. 2003. [Online]. Available: http://www.ietf.org/rfc/rfc3550.txt [Accessed: 10. Dec 2013].

[52] A. Kapoor, “Recommended bit rates for live streaming,” January 12, 2009, [Online]. Available: http://www.adobe.com/devnet/adobe-media-server/articles/dynstream_on_demand.html, accessed July 25, 2013.

[53] C. Bettstetter, H. Hartenstein, and X. Pérez-Costa, “Stochastic properties of the random waypoint mobility model,” Wirel. Netw., vol. 10, no. 5, Sep. 2004, pp. 555–567.

[54] T. Camp, J. Boleng, and V. Davies, “A survey of mobility models for ad hoc network research,” Wireless Communications & Mobile Computing (WCMC): Special issue on Mobile ad hoc Networking: Research, Trends and Applications, vol. 2, 2002, pp. 483–502.