Load balancing and flow management under user mobility in heterogeneous wireless networks

Tom De Schepper, Steven Latré and Jeroen Famaey
University of Antwerp - imec, IDLab, Department of Mathematics and Computer Science, Belgium
firstname.lastname@uantwerpen.be

***Abstract*—The utilization and size of today's wireless networks is continuously increasing, as more and more wireless communication technologies and connected devices are being added. As the use of multiple communication technologies is supported by modern devices, efforts have been made to allow these devices to utilize simultaneously and handover between different technologies. However, existing management frameworks and standards lack the intelligence to provide fine-grained network-wide optimizations. This despite the potential of dramatically increasing overall network performance (e.g., throughput) and user experience. To this extent, we present a multi-technology load balancing approach that can manage devices and steer traffic across different wireless technologies, in order to maximize the global throughput. This dynamic approach can be deployed on top of existing solutions and takes into account the specific characteristics of wireless networks and the mobility of stations. We present a mathematical problem formulation of load balancing traffic and devices across different wireless technologies. We demonstrate its ability to significantly increase network-wide throughput and meet the demands of the users.**

***Index Terms*—heterogeneous wireless networks, traffic management, load balancing**

I. Introduction

Nowadays wireless networks are all around us and have become the standard for providing Internet connectivity. This is, among others, the case in Local Area Networks (LANs), Wide Area Networks (WANs), and backhaul networks (e.g., for vehicles). The utilization of these wireless networks is ever-increasing as a result of the rising number of connected devices and the expansion of the demands of users and applications (e.g., high quality live streaming) [1]. With the addition of new wireless technologies and devices, the heterogeneity and management burden of wireless networks is rapidly increasing. On one hand, modern multimedia services have stringent quality requirements and are very sensitive to network disruptions and degradations (e.g., high latency, congestion or link failures) On the other hand, current wireless networks are generally managed in a mostly static manner, unable to automatically react in a timely fashion to temporary disruptions that cause Quality of Service (QoS) degradations. This problem is expected to further increase as even more communication technologies, such as IEEE 802.11ay and 802.11ax, and application domains, like smart cities or autonomous vehicles, become available [2].

As both modern connected devices and wireless networks are equipped with multiple communication technologies, dynamic network and traffic management would allow for network optimizations. Examples of such optimizations are multipath routing, load balancing, and dynamic path reconfiguration. In order to enable these optimizations, it needs to be possible to seamlessly switch between or load balance traffic over different technologies. However, traditional approaches fail to offer the required dynamic management, as they typically delegate this to the application layer, or even worse, to the user. Making it thus impossible to automatically react in a timely fashion to dynamic network changes (e.g., disruptions or varying number of devices).

To this extent, dynamic multi-technology frameworks and standards have been proposed. The most important ones are Multipath Transmission Control Protocol (MPTCP) [3], LTE-Wireless Local Area Network Aggregation (LWA) [4], and ORCHESTRA [5]. MPTCP allows to split a Transmission Control Protocol (TCP) flow across different paths through the network, while LWA allows to offload traffic between an LTE base station and IEEE 802.11 (Wi-Fi) access points (APs) [3, 4]. ORCHESTRA offers a transparent management solution by introducing a Virtual MAC (VMAC), arching different technologies per device, and a centralized controller [5]. While these solutions introduce the features needed to enable dynamic flow rerouting and load balancing, they lack the intelligence to allow for network-wide optimizations.

Therefore, we present a multi-technology load balancing approach that can balance devices across different APs and steer traffic across different paths through the network. It can make use of the management functions offered by the above mentioned frameworks. This approach aims to find a global optimal scheduling configuration for all the traffic flows and stations in the network, in order to achieve maximum global throughput. In contrast to existing load balancing approaches, we do not assume full knowledge over the network and use real-time monitoring information as inputs. Furthermore, we also present a general approach that can be used without dependencies to specific technologies.

Our previous work introduced a load balancing solution for specific LANs consisting of only stationary devices [6]. The technologies under consideration were Ethernet and Wi-Fi, provided by a single AP. In this paper we extend this work in several ways. First, we shift the focus to the more challenging environment of wireless networks and take into account the presence of multiple APs (or base stations). Second, we specifically take into account the mobility of stations.

The contributions of this paper are twofold: first, we introduce a mathematical model of the load balancing prob-

978-3-903176-14-0 © 2018 IFIP

lem, for both devices and traffic flows, in heterogeneous wireless networks. The problem is formulated as a Mixed Integer Quadratic Program (MIQP), which can be solved using existing linear programming approaches. Second, we evaluate the resulting model and the heuristic in a variety of scenarios, using different network configurations, based on ns-3 simulations.

The remainder of this paper is structured as follows. We start by giving an overview of the current state of the art in Section II. Next, we introduce the mathematical problem formulation and load balancing algorithm in Section III. Finally, Section IV discusses the simulation results, while conclusions are provided in Section V.

II. Related work

In this section we discuss existing work on both the topic of multi-technology network management and load balancing in heterogeneous network environments.

A. Multi-technology standards and frameworks

MPTCP is a TCP extension that enables the transmission and reception of data concurrently over multiple network interfaces. Multiple regular TCP connections (denoted as subflows), are offered as one to the application layer, while under the hood each subflow can follow different paths through the network [3]. A scheduler can thus divide or duplicate application data across these sub-flows, based on the ever-changing network characteristics (e.g., increased RTT), to attain a higher throughput, or increased reliability [7]. MPTCP is actively being used, for instance, in consumer devices like smartphones (e.g., Siri) [8, 9]. While MPTCP aims at improving QoS and network resource utilization, it focuses only on the alternative paths between two hosts and not on a network-wide scale.

The ever-growing bandwidth and traffic speed demands have urged the 3GPP community to explore the wireless spectrum outside of the traditional licensed 3G/4G bands. In order to offload traffic, both the direct usage of LTE in the unlicensed spectrum (i.e., LTE-LAA/LTE-U) and the combined usage of LTE in the licensed and Wi-Fi technology in the unlicensed spectrum (i.e., LWA) have been proposed [10, 4]. While the first can cause severe performance degradations in coexisting Wi-Fi systems, the LWA approach clearly introduces less coexistence issues and no hardware changes are required on the infrastructure [11, 12]. From a user perspective, both LTE and Wi-Fi are used seamlessly as mobile traffic flows are tunneled over the Wi-Fi connection.

More recently, the ORCHESTRA framework has been proposed as the first solution that can be used transparently with all technologies and communication protocols [5]. The framework consists of a VMAC on the devices and a centralized or cloud-based controller. The VMAC unifies the underlying heterogeneous technologies per device, offering a single interface to the upper layers with a single IP address. Based on packet matching rules, the VMAC forwards packets to the designated underlying technologies. This allows for packet-level load balancing, vertical handovers, and duplication. The rules on the VMAC can be changed by the controller, based on the real-time monitoring information that is sent from the different VMACs to the controller.

Summarized, different solutions are proposed that allow for multi-technology management and features (e.g., handovers or duplication). In complement, there is a need for algorithms and intelligence (as the approach presented in this work) that use these frameworks and standards to optimize the network.

B. Load balancing in heterogeneous networks

Multi-technology load balancing has been mostly addressed in two different research areas, mainly LANs and WANs (4G/5G). Macone et al. propose a per-packet load balancing algorithm for LANs that runs centralized on the gateway and assumes full instantaneous knowledge of network resources and conditions [13]. Furthermore, a decentralized load balancing algorithm specifically for heterogeneous wireless access networks was proposed by Oddi et al. [14]. The algorithm is based on the Wardrop equilibrium and does not take into account the fact that users do not have dedicated network resources when using wireless technologies. In general, Olvera-Irigoyen et al. have shown that determining the actual available bandwidth on the links has a big impact on the results of distributing the flows [15]. Recent load balancing solutions for LANs focus also on energy optimization [16, 17]. However, this is done by assuming the energy consumption model is known in advance, and not by real-time measurements.

In WANs, most research proposes technology-specific solutions that are capable of load balancing across only two of these technologies (e.g., LTE and Wi-Fi or Wi-Fi and WiMAX) [18]. Load balancing policies are generally based on the number of connected devices to a base station, and different decision strategies have been proposed, using among others utility functions, multiple attributes decision making, Markov chains, and game theory [18, 19]. These strategies take only a limited number of parameters into account, with Received Signal Strength Indicator (RSSI) and Signal To Noise Ratio (SNR) being the most popular ones [20, 21]. Open issues include, for instance, the development of more generic solutions, better support for mobility, the use of multi-criteria decision functions, supporting different QoS classes and the increase of QoS during or after handovers [20].

Summarized, most existing work on load balancing in heterogeneous networks makes use of theoretical models that assume, unrealistically, full knowledge over the detailed state of the network. Furthermore, the specific nature of wireless networks is ignored and approaches are technology dependent. In contrast, the proposed approach is technology independent and focuses on wireless networks (taking into account the specifics), while using only real-time monitored information.

III. Multi-technology load balancing problem formulation

This section presents the proposed multi-technology load balancing model. We distinguish from our previous work as we created a novel formulation targeting heterogeneous wireless networks by taking into account specific elements such as station mobility and presence of multiple APs [6].

A. Network model

A heterogeneous wireless network is modeled as a multigraph defined as a tuple (S,T,B) where:

- S is the set of stations $\{s_1, s_2, ..., s_n\}$. These stations represent all kinds of connected devices, depending on the modeled network (e.g., smartphones, sensors, vehicles).
- T is the set of technologies $\{t_1, t_2, ..., t_n\}$. This can, for instance, be IEEE 802.11ac, IEEE 802.11ad, or LTE.
- B is the set of all Basic Service Sets (BSSs) $\{b_1, b_2, ..., b_n\}$. A BSS is defined as a set of stations $\{s_1, s_2, ..., s_n\}$ that are connected to an AP or base station using a certain technology. In other words, a BSS encapsulates all the stations that can interfere with each other since they share the capacity of a technology. We assume no interference between BSS that are in range of each other (i.e., use of different channels).

Furthermore, we define the following sets and elements:

- $\forall s \in S : T_s$: defines for each station the set of all technologies $t \in T$ that are supported by the station.
- $\forall b \in B : B_t$: is the set of BSS for a certain technology $t \in T$.
- $\forall s \in S : B_s$ set of BSSs to which $s \in S$ can belong. In other words these are all the BSS of which the AP are in range of the station (for a supported technology).
- Finally, we define $d_{s,b}$ and $b_{s,b}$ to be, respectively, the data rate (depending on the MCS) and bit error rate of the station $s \in S$ for a specific BSS $b \in B$. Note that these values depend on the mobility and position of stations and can change over time.

In addition to the network topology, traffic flows going through the network also need to be modeled. Therefore, we define F as the set of all flows. A flow $f \in F$ is a triple $< s_f, r_f^{in}, r_f^{out} >$ with $s_f \in S$ the station within the network that is the source or destination of the flow within the network, r_f^{in} the incoming desired rate of f $\in \mathbb{R}^+$ and r_f^{out} the outgoing desired rate of f $\in \mathbb{R}^+$. Note that we do assume that the gateway is always one of the two endpoints of the flow, while the other is denoted by s_f. Furthermore, we separate the desired rate of the flow between the incoming and outgoing rate. This allows us to more precisely schedule all flows across the different paths, as incoming and outgoing packets of a flow can be assigned a different route. To clarify, for a TCP flow originating from some web server, the incoming rate is the rate of the data traffic, while the outgoing rate is the one of the ACKs.

B. MIQP formulation

The load balancing problem considered in this paper is modeled as an MIQP, which consists of the necessary inputs, decision variables, an objective function, and a set of constraints. The inputs of the presented MIQP consist of the previously described network and flow model. Additionally, we need one more input: we define χ_b to be a linear function that approximates the capacity of the different BSSs, taking into account the number of stations and the particular technology [6]:

$$\chi_b(\alpha, \beta) = \alpha \cdot (\sum_{f \in F} \lambda_{f,b}^{in} + \lambda_{f,b}^{out}) + \beta$$

The parameters α and β are technology specific and account for the impact of contention and collisions under an increasing number of stations. They can be experimentally determined.

Next, we define the following decision variables:

- $\tau_f^{in} \in \left[0, r_f^{in}\right]$; this variable defines the total incoming rate assigned to a flow $f \in F$.
- $\tau_f^{out} \in \left[0, r_f^{out}\right]$; this variable defines the total outgoing rate assigned to a flow $f \in F$.
- $\lambda_{f,b}^{in} \in \{0, 1\}$; this variable represents the path for the incoming traffic of a flow. If the incoming traffic of flow $f \in F$ is scheduled over BSS $b \in B_{s_f}$ then $\lambda_{f,b}^{in} = 1$, otherwise it equals 0.
- $\lambda_{f,b}^{out} \in \{0, 1\}$; this variable represents the path for the outgoing traffic of a flow. If the outgoing traffic of flow $f \in F$ is scheduled over BSS $b \in B_{s_f}$ then $\lambda_{f,b}^{out} = 1$, otherwise it equals 0.
- $\gamma_{s,t,b} \in \{0, 1\}$; this variable represents the connection between a station and an AP. It is equal to 1 if a station $s \in S$ on technology $t \in S_t$ is part of the BSS $b \in B_s \cap B_t$, otherwise it equals 0. In other words, we assume that per technology a station can only be connected to one AP or base station.
- $\delta \in [0, 1]$: represents the maximal load over all BSS.

As an objective function, the model maximizes the total rate (bandwidth) of the network-wide traffic, both incoming and outgoing, while minimizing the relative maximal load over all BSS:

$$max(\omega \cdot (\sum_{f \in F} \tau_f^{in} + \tau_f^{out}) + (1 - \omega) \cdot (-\delta) \cdot (\sum_{b \in B} \chi_b))$$

This objective function consists of two weighted subfunctions that need to be optimized (with the weight denoted by ω). The first subfunction represents the total assigned rate over all flows (which needs to be maximized). The second part represents the division of load across all available BSSs. The idea is to minimize the maximal relative load, denoted by δ, across all BSSs [22]. As many mathematical solvers do not allow the usage of maximization or minimization functions within the objective function, δ is bounded by the final constraint. Note that the multiplication of δ with $\sum_{b \in B} \chi_b$ is only needed for normalization.

Finally, we define the following constraints: we first define a constraint that guarantees that the capacity of BSSs and their underlying technologies is not exceeded:

- $\forall b \in B : \sum_{f \in F} \lambda_{f,b}^{in} \cdot \tau_f^{in} + \lambda_{f,b}^{out} \cdot \tau_f^{out} \leqslant \chi_b$

Next, we define a constraint that limits the total rate over all traffic flows on a station, going over a certain BSS, by the maximal rate supported by the configuration of that station:

- $\forall s \in S, \forall b \in B_s : \sum_{f \in F_s} \lambda_{f,b}^{in} \cdot \tau_f^{in} + \lambda_{f,b}^{out} \cdot \tau_f^{out} \leqslant d_{s_f,b} \cdot b_{s_f,b}$

Furthermore, we define two constraints that guarantee the conservation of flows in the network (i.e., the right endpoints):

- $\forall f \in F : \sum_{b \in B_{s_f}} \lambda_{f,b}^{in} = 1$
- $\forall f \in F : \sum_{b \in B_{s_f}} \lambda_{f,b}^{out} = 1$

We also need to make sure that a device can be connected to only one BSS per technology:

- $\forall s \in S, \forall t \in T_s : \sum b \in B_s \cap B_t \gamma_{s,t,b} = 1$
- $\forall s \in S, \forall t \in T_s, \forall b \in B_s \cap B_t, \forall f \in F_s : \lambda_{f,b}^{in} \leqslant \gamma_{s,t,b}$
- $\forall s \in S, \forall t \in T_s, \forall b \in B_s \cap B_t, \forall f \in F_s : \lambda_{f,b}^{out} \leqslant \gamma_{s,t,b}$

Finally, we define the constraint that bounds the maximum value of δ for balancing the load across BSSs:

- $\forall b \in B : \sum_{f \in F} \lambda_{f,b}^{in} \cdot \tau_f^{in} + \lambda_{f,b}^{out} \cdot \tau_f^{out} \leqslant \delta \cdot \chi_b$

C. Network interaction and parameter estimation

In Section II-A we listed a number of existing multi-technology frameworks and standards that could be used to configure the network. While our load balancing approach can be deployed on all of these solutions, ORCHESTRA is the most suitable as it offers centralized control and monitoring, while also offering seamless handovers. These central control and monitoring features are important as real-time information on the network state is required as inputs of our load balancing approach and the calculated configuration needs to be rolled-out. Moreover, this monitoring information is also used to trigger the execution of the algorithm when dynamic changes to the network are detected (e.g., a variation in one of the flow rates of at least x %). While this repetitive execution allows to react to station mobility or changed traffic demands, this also requires a limited execution time of the algorithm.

Some of the gathered monitoring information, like station and traffic information, can be used directly without the need for further processing. Similarly, flow rates can be estimated by simply using the monitored rates [6]. However, this is not for all inputs the case: first, to avoid the use of complex theoretical models, we defined the approximation function χ_b to estimate the capacity of the wireless technologies [6]. The technology specific parameters α and β can be experimentally determined to capture the specifics of the wireless network under consideration. Second, a similar approach can be taken to determine the data rate (depending on the Modulation and Coding Scheme (MCS)) and bit error rate of the station $s \in S$ for a specific configuration, respectively denoted by $d_{s,b}$ and $b_{s,b}$. For the first parameter a mapping from measured RSSI values to MCS values (and theoretical data rate) can be constructed. A second linear function can map the measured RSSI values to packet loss, in order to correct the theoretical achievable data rate. Both functions can be experimentally determined by using the well-known fingerprinting approach to record MCS and packet loss values at different distances (and thus different RSSI values) in the network environment. This method can be applied to each heterogeneous environment to capture the specific characteristics and can be rapidly re-executed if needed.

TABLE I: Overview of the devices, and the supported flow rates, used in the scenarios

Device type	**Rate boundaries per flow type**		
(and mobility)	**Download**	**Video stream**	**Conference call**
Laptop (mobile)	10–30 Mbps	8–20 Mbps	4–10 Mbps
HD Television	5–25 Mbps	10–20 Mbps	5–10 Mbps
4K Television	5–25 Mbps	15–25 Mbps	7.5–12.5 Mbps
Tablet (mobile)	1–8 Mbps	2.4–9 Mbps	1.2–4.5 Mbps
Smartphone (mobile)	1–8 Mbps	2.4–9 Mbps	1.2–4.5 Mbps

IV. Results and discussion

In this section, we evaluate the proposed load balancing approach, using simulation results obtained from the ns-3 event-based network simulator [23]. First, the evaluation setup and scenario are discussed. Second, the performance of the approach, in terms of achieved throughput and execution time, is evaluated in a number of scenarios. For every scenario, we provide a comparison to a fully distributed baseline, where each device decides for itself to which AP to connect, based on the best RSSI values. In other words, the baseline corresponds to the current state of the art, where one of the discussed multi-technology management solutions (Section II-A) is in place, without the centralized intelligence, but with seamless handovers.

A. Evaluation setup

All simulations are conducted using the ns-3.27 network simulator, while the Gurobi Optimizer (7.5.2) is used to solve the MIQP formulation. The experiments take place using a single core of an Intel® Xeon® E5-2680 Processor running at 2.8 GHz and with 8 GB RAM. In the ns-3.27 simulator, we implemented the entire ORCHESTRA framework [5] and the MIQP approach. Furthermore, we also extended the basic ns-3.27 implementation to allow for multi-channel Wi-Fi networks. During all of our experiments, we assume two technologies present: IEEE 802.11n and IEEE 802.11ac (respectively, 2.4 GHz and 5 GHz Wi-Fi). Every scenario has at least two APs that support both technologies. Dynamic rate adaptation for all devices is made possible through the Minstrel rate adaptation algorithm.

In order to generate representative network topologies and conditions, several types of devices are defined, each with different mobility and traffic rates. This information is depicted in Table I. The exact number of devices, the assigned flow type, and the rate of the flow are chosen uniformly at random between an upper and lower bound, based on the involved device and depending on the scenario. Each mobile device (all except for the televisions) moves around according to

TABLE II: Setup for static scenarios

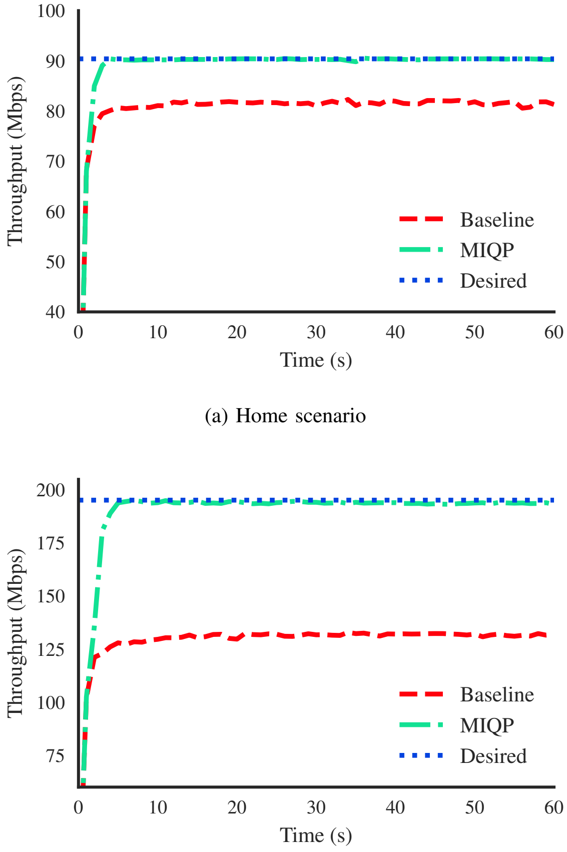
Device	**Home** (20x10 m)	**SME** (25x10 m)	**Flows**
APs	2	3	N/A
Laptop	2	9	Download/Conf. call
HD TV	0	1	Video stream
4k TV	1	0	Video stream
Tablet	2	1	All types of flows
Smartphone	3	5	All types of flows
Total	10	19	

the Random Waypoint Model within a certain area, with a random start position and a uniformly random chosen speed between 0.3-0.7 $\frac{m}{s}$. The size of the area and the wait times at the waypoints are depending on the scenario. Moreover, in the static scenarios the flow rates do not change over time, while in the other scenarios the download flows will consume as much bandwidth as possible (reflecting their actual behavior). Assuming a static flow rate for the first part of the evaluations allows us to better estimate the impact of only the mobility aspect. The size of the download is uniformly at random chosen between 10 MB and 10 GB. We assign one flow per device and as such do not assume the concurrent usage of both Wi-Fi interfaces, as this is generally not supported by current hardware. Note, that the flow rates were selected based on representative figures from literature of existing applications and that we use only TCP traffic flows, as current Internet traffic is dominated by TCP [24].

For every described scenario, results are averaged over 20 different randomly generated flow and topology configurations. For the fully distributed baseline, we assume that when the RSSI of the current connection drops below a threshold of -75, a better connection is selected (if present). The execution of the algorithm is triggered by the real-time monitoring component when dynamic changes to the network have been detected (e.g., a variation in one of the flow rates of at least 25 %) or when it has been 10 s since the last execution. To avoid oscillations in the decision making, there should be at least 2 s between two consecutive executions. Furthermore, a time limit of 900 s is set for solving the MIQP. This time limit ensures the continuation of the experiments, while still being a magnitude larger then required for reactive real-time optimizations. Finally, the following parameters were experimentally determined: for the function χ_b, α and β are respectively, for 2.4 GHz Wi-Fi -1.74 and 57.58, and for 5 GHz Wi-Fi -3.21 and 112.99. These values are experimentally determined using the method described in our previous work [6]. For the objective function a weight of 0.91 is used. Different values were compared but the value of 0.91 showed the best average result across a number of experiments.

B. Static scenarios

In order to get a first impression of the performance of the different approaches we created two basic scenarios with varying topologies. As depicted in Table II, these scenarios slightly differ in size and density. The results for the two scenarios are shown in Figure 1. The graphs compare the baseline and MIQP formulation to the sum of the desired flow rates (known as we use fixed flow rates here). Across both graphs we clearly see a significant improvement by our approach in comparison to the distributed baseline.



(a) Home scenario

(b) SME scenario

Fig. 1: Throughput as a function of time for different scenarios, comparing the MIQP formulation and the baseline

For the Home scenario, we can report the following rates (± the standard error), respectively for the baseline and MIQP: 81.61 Mbps (±2.62) and 90.15 Mbps (±2.36). Their is thus an improvement of, respectively 10.46 % compared to the baseline. As the total desired rate is 90.40 Mbps (±2.35), it is clear that our approach succeeds in providing the optimal network configuration. Similarly for the SME scenario, the following average rates are achieved: 131.46 Mbps (±3.73) and 193.90 Mbps (±3.76) for respectively, the baseline and MIQP. The increases towards the baseline is larger than for the Home scenario: 47.50 %. The same can be said for meeting the requirements of the flows as the total desired rate is 195.21 Mbps (±3.46).

Furthermore, finding the optimal solution took, on average, 16.38 s (±4.28) and 736.58 s (±39.71), respectively for the home ans SME scenario. This is execution times are rather high and are significantly above the minimal interval (of 2 s) between two consecutive runs of the algorithm. We will

TABLE III: Impact of mobility on throughput

	Wait times	Baseline	MIQP
Home	0-10 s	83.16 Mbps (±3.31)	89.67 Mbps (±2.35)
	5-15 s	81.61 Mbps (±2.62)	90.15 Mbps (±2.36)
	10-20 s	80.32 Mbps (±2.88)	90.24 Mbps (±2.29)
SME	0-10 s	157.19 Mbps (±4.70)	189.03 Mbps (±4.80)
	5-15 s	131.46 Mbps (±3.73)	193.90 Mbps (±3.76)
	10-20 s	135.46 Mbps (±3.98)	194.32 Mbps (±3.35)

TABLE IV: The execution time for the MIQP under increasing network load

Load	Flows	Exec. time MIQP
10	6	8.17 s (± 1.08)
15	8	12.14 s (± 2.69)
20	10	29.75 s (± 6.84)
25	12	87.52 s (± 9.39)
30	14	478.36 s (± 36.39)

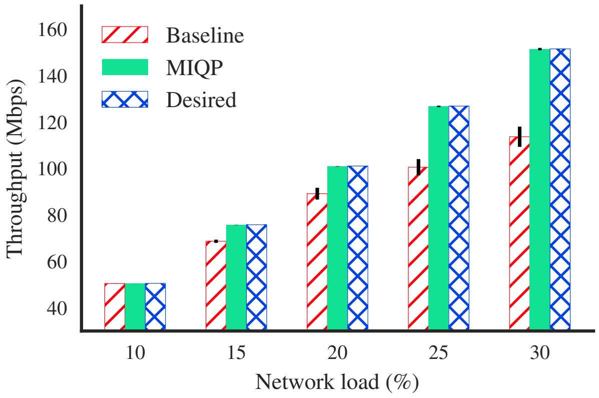


Fig. 2: Throughput as a function of network load, error bars depict the standard error

discuss the scalability the MIQP in more detail in the next section.

Finally, we considered the impact of mobility on the overall throughput. Therefore, we varied the waypoint wait times for both scenarios by additional experiments for times between 0-10 s and 10-20 s. The results, listed in Table III, show that the algorithm always significantly outperforms the baseline. However, for the case with the highest mobility (and lowest wait times) the baseline performs significantly better, than in the other cases. We believe this to be due the higher number of handovers, triggered by the mobility.

C. Impact of network load

To investigate the scalability of the algorithm in terms of traffic and execution time, the following scenario was created: a set of devices was randomly generated, each with a uniform randomly assigned flow with a randomly chosen type and rate. The total desired rate of all flows equals a certain percentage of total theoretical network capacity. Experiments were performed for loads of 10, 15, 20, 25, and 30 % of the theoretical network capacity. Moreover, the presence of 3 APs was assumed in a space of 20 by 15 m with a waypoint wait time of 5-15 s.

From Figure 2 it is clear that our load balancing approach offers a significant improvement towards the baseline. This improvement grows when the percentage of network traffic increases. For instance, for a load of 30 % there is an increase from 113.68 Mbps (±4.36) for the baseline to 151.42 Mbps (±0.51) for the MIQP. This is an increase of 33.20 %. More importantly, we see that the MIQP allows to satisfy the traffic demands of all flows. For instance, at a load of 30 % there is only a negligible difference of 0.10 Mbps or 0.06 % between the desired rates and the achieved throughput.

Furthermore, we measured the time it takes to calculate the optimal solution. Table IV shows the averages of the measured values across the different network loads. It is clear that the computation time for the MIQP scales exponentially. For instance, for only 14 flows (i.e., load of 30 %) it takes already 478.36 s (±36.39) to compute the configuration. Furthermore it showed to be infeasible to calculate a solution for higher loads within the time limit of 900 s. This clearly indicates that the MIQP solution can not be used in very dynamic real-life wireless networks. Finding a scalable near-optimal solution will be our prime focus in future work.

V. Conclusions

This article addresses the need for intelligent management of heterogeneous wireless networks. We introduce a multi-technology load balancing approach that can balance devices across different APs and steer traffic across different paths through the network, on top of existing management frameworks and standards (like MPTCP). Our approach focuses on the dynamic and challenging environment of wireless networks and takes into account specific parameters such as mobility of users and coexistence of multiple APs. This allows us to optimize the performance of the network in terms of network-wide throughput. We present a mathematical problem formulation, through a MIQP, that can calculate the optimal network configuration. In our evaluation we show that the presented approach offers a significant improvement in terms of throughput. The scalability of this approach will be improved in future research.

References

[1] Cisco, “Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021,” *CISCO white paper*, 2016.

[2] M. S. Afaqui, E. Garcia-Villegas, and E. Lopez-Aguilera, “IEEE 802.11ax: Challenges and Requirements for Future High Efficiency WiFi,” *IEEE Wireless Communications*, vol. 24, no. 3, pp. 130–137, 2016.

[3] A. Ford, C. Raiciu, M. Handley, and O. Bonaventure, “TCP extensions for multipath operation with multiple addresses,” Internet Requests for Comments, RFC Editor, RFC 6824, January 2013. [Online]. Available: http://www.rfc-editor.org/rfc/rfc6824.txt

[4] C. Hoymann, D. Astely, M. Stattin, G. Wikström, J. F. T. Cheng, A. Höglund, M. Frenne, R. Blasco, J. Huschke,

and F. Gunnarsson, “LTE release 14 outlook,” *IEEE Communications Magazine*, vol. 54, no. 6, pp. 44–49, 2016.

[5] T. De Schepper, P. Bosch, E. Zeljkovi, J. Haxhibeqiri, J. Hoebeke, J. Famaey, and S. Latre, “ORCHESTRA: Enabling Inter-Technology Network Management in Heterogeneous Wireless Networks,” *IEEE Transactions on Network and Service Management*, pp. 1–1.

[6] T. De Schepper, S. Latré, and J. Famaey, “Flow Management and Load Balancing in Dynamic Heterogeneous LANs,” *IEEE Transactions on Network and Service Management*, vol. 15, no. 2, pp. 693–706, 2018.

[7] C. Paasch, S. Ferlin, O. Alay, and O. Bonaventure, “Experimental evaluation of multipath TCP schedulers,” in *Proceedings of the 2014 ACM SIGCOMM workshop on Capacity sharing workshop - CSWS ’14*, 2014, pp. 27–32.

[8] F. Rebecchi, M. D. De Amorim, V. Conan, A. Passarella, R. Bruno, and M. Conti, “Data offloading techniques in cellular networks: A survey,” *IEEE Communications Surveys and Tutorials*, vol. 17, no. 2, pp. 580–603, 2015.

[9] Q. De Coninck, M. Baerts, B. Hesmans, and O. Bonaventure, “A first analysis of multipath TCP on smartphones,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9631, no. September 2015, pp. 57–69, 2016.

[10] A. Mukherjee, J. F. Cheng, S. Falahati, H. Koorapaty, D. H. Kang, R. Karaki, L. Falconetti, and D. Larsson, “Licensed-Assisted Access LTE: Coexistence with IEEE 802.11 and the evolution toward 5G,” *IEEE Communications Magazine*, vol. 54, no. 6, pp. 50–57, 2016.

[11] F. M. Abinader, E. P. Almeida, F. S. Chaves, A. M. Cavalcante, R. D. Vieira, R. C. Paiva, A. M. Sobrinho, S. Choudhury, E. Tuomaala, K. Doppler, and V. A. Sousa, “Enabling the coexistence of LTE and Wi-Fi in unlicensed bands,” *IEEE Communications Magazine*, vol. 52, no. 11, pp. 54–61, 2014.

[12] N. Zhang, S. Zhang, S. Wu, J. Ren, J. W. Mark, and X. Shen, “Beyond coexistence: Traffic steering in LTE networks with unlicensed bands,” *IEEE Wireless Communications*, vol. 23, no. 6, pp. 40–46, 2016.

[13] D. Macone, G. Oddi, A. Palo, and V. Suraci, “A dynamic load balancing algorithm for quality of service and mobility management in next generation home networks,” *Telecommunication Systems*, vol. 53, no. 3, pp. 265–283, 2013.

[14] G. Oddi, A. Pietrabissa, F. D. Priscoli, and V. Suraci, “A decentralized load balancing algorithm for heterogeneous wireless access networks,” in *World Telecommunications Congress*, 2014, pp. 1–6.

[15] O. Olvera-Irigoyen, A. Kortebi, and L. Toutain, “Available bandwidth probing for path selection in heterogeneous home networks,” in *IEEE Globecom Workshops (GC Wkshps)*, 2012, pp. 492–497.

[16] O. Bouchet, A. Kortebi, and M. Boucher, “Inter-MAC green path selection for heterogeneous networks,” in *IEEE Globecom Workshops (GC Wkshps)*, 2012, pp. 487–491.

[17] A. Kortebi and O. Bouchet, “Performance evaluation of inter-mac green path selection protocol,” in *12th Annual IEEE Mediterranean Ad Hoc Networking Workshop (MED-HOC-NET)*, 2013, pp. 42–48.

[18] M. Zekri, B. Jouaber, and D. Zeghlache, “A review on mobility management and vertical handover solutions over heterogeneous wireless networks,” *Computer Communications*, vol. 35, no. 17, pp. 2055–2068, 2012.

[19] G. Gódor, Z. Jakó, Á. Knapp, and S. Imre, “A survey of handover management in lte-based multi-tier femtocell networks: Requirements, challenges and solutions,” *Computer Networks*, vol. 76, pp. 17–41, 2015.

[20] J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. S. Dhillon, “An overview of load balancing in hetnets: Old myths and open problems,” *IEEE Wireless Communications*, vol. 21, no. 2, pp. 18–25, 2014.

[21] B. Ng, A. Deng, Y. Qu, and W. K. Seah, “Changeover prediction model for improving handover support in campus area wlan,” in *Network Operations and Management Symposium (NOMS), 2016 IEEE/IFIP*. IEEE, 2016, pp. 265–272.

[22] Y. Donoso and R. Fabregat, *Multi-objective optimization in computer networks using metaheuristics*. CRC Press, 2016.

[23] G. F. Riley and T. R. Henderson, “The ns-3 network simulator,” in *Modeling and tools for network simulation*. Springer, 2010, pp. 15–34.

[24] D. J. Lee, B. E. Carpenter, and N. Brownlee, “Media Streaming Observations: Trends in UDP to TCP Ratio,” *International Journal on Advances in Systems and Measurements*, vol. 3, no. 3, pp. 147–162, 2010.