Part I

Physical Layer for 5G Radio Interface Technologies

1

Emerging Technologies in Software, Hardware, and Management Aspects Toward the 5G Era: Trends and Challenges

Ioannis-Prodromos Belikaidis,[1] Andreas Georgakopoulos,[1] Evangelos Kosmatos,[1] Stavroula Vassaki,[1] Orestis-Andreas Liakopoulos,[1] Vassilis Foteinos,[1] Panagiotis Vlacheas,[1] and Panagiotis Demestichas[1,2]

[1] *R&D Department, WINGS ICT Solutions, Athens, Greece*
[2] *Department of Digital Systems, University of Piraeus, Piraeus, Greece*

1.1 Introduction

As the number of smartphones and demand for higher data-rate connections keep explosively growing, the technology has to pursue this trend in order to be able to provide the suitable communication schemes. By 2020, it is calculated that the total global mobile wireless user devices will be over 10 billion, with the mobile data traffic growing more than 200 times over compared to 2010 numbers. It is foreseeable that 4G mobile communication system could no longer meet the need of users service requirements. Thus, the new 5G communication systems are promising a complete network structure with unlimited access to information, providing the requested service demands to users far beyond what the current 4G offers by supporting innovative new wireless technologies and network architecture to meet the extremely high-performance requirement (Table 1.1).

These features include support for new types and massive number of devices, for very high mobile traffic volumes, universal access for users, very high frequency reuse and spectrum reuse in wireless technologies, automated provisioning, configuration and management of a wide range of new network services, ultrareliable, ultralow latency, ultradensification, and even more. 5G networks would be a heterogeneous networks (HetNets), meaning that different networks will be integrated all together to a unified system, enabling

5G Networks: Fundamental Requirements, Enabling Technologies, and Operations Management, First Edition.
Anwer Al-Dulaimi, Xianbin Wang, and Chih-Lin I.

Table 1.1 Main comparison between 5G and previous generations.

5G	4G and Earlier
New waveforms based on filter banks and other novel technologies	Waveforms based mainly on OFDM and variations
Gbps performance	Up to hundreds Mbps
End-to-end latency of some milliseconds	End-to-end latency of hundreds of milliseconds
Support of massive MIMO	SISO and limited MIMO technologies
Support of mm-wave bands up to hundreds GHz	Operation mainly below 6 GHz bands
Efficient support of massive number of devices in ultradense environments	Support of limited number of devices in dense/congested areas

aggregation of multiple existing radio access technologies (RATs) such as LTE-A, Wi-Fi, D2D, and even lightly licensed. Delivering all 5G requirements in order to support the new features and services, a substantial change on the network architecture is inevitable.

Normally, in the past such a process would need deployment of specialized devices build for a specific application and with fixed functionalities. Thus, any development and transformation to follow the constantly increasing and heterogeneous market requirements demands a huge investment to change/deploy hardware. Nowadays, various technologies and architectures have been utilized in order to solve this problem and provide a faster introduction and adaptation of new technologies to the communications systems. One of these elements that offers reprogrammability of the network elements in order to solve new problems or to establish new more suitable functions is software defined networking (SDN). This architecture provides dynamic, manageable, cost-effective, and adaptable solutions, making it ideal for the high-bandwidth, dynamic nature of today's applications. SDN decouples the dependence of implementing instructions that are provided by multiple, vendor-specific devices and protocols, making it bound to an exact hardware. Such aspects are also related to the flexible deployment of functionality to hardware, software, or mixed. As Section 1.3 proposes, hardware has better execution performance, but software offers greater flexibility.

Networking, computing, and storage resources would be integrated into one programmable, unified, and flexible infrastructure that will be more cost-effective and with higher scalability at minimum cost. This unification will allow for an optimized and more dynamic utilization of various distributed resources, and the pooling of fixed, mobile, and broadcast services.

Also, 5G will be designed to be a viable, robust, and scalable technology. Another positive result with the introduction of 5G communication systems will be the drastic energy consumption reduction and energy harvesting that will help the industry to have an astounding usage growth. Since network services will rely progressively on software, the creation and growth will be further encouraged. In addition, the 5G infrastructures will provide network solutions and involve vertical markets such as automotive, energy, food and agriculture, city and buildings management, government, health care, manufacturing, and public transportation.

The rest of the chapter is structured as follows: Section 1.2 elaborates on the main 5G requirements, while Section 1.3 presents the status and challenges in hardware and software development. Section 1.4 elaborates on the status and challenges in 5G wireless communications by focusing on physical layer, MAC and RRM. Finally, Section 1.5 investigates the benefits of machine learning in 5G network management.

1.2 5G Requirements and Technology Trends

In the digital era, users and devices are becoming more dependent on various applications and services that involve the creation, access/ communication, processing, and storage of digital content. These developments have been tremendously accelerated by wireless/mobile technologies, which have offered unparalleled access/ communication opportunities to users. 5G is expected to be dominated mainly by the following application classes, including massive machine-type communications (mMTC), enhanced mobile broadband (eMBB), and ultrareliable and low latency communications (URLLC). These main classes will facilitate scenarios related to critical and demanding applications for the realization of smart cities, as well as the realization of applications for Industry 4.0 and automation aspects. Also, self-driving aspects are expected to pose strict requirements especially on the latency in order to ensure reliable and secure service with very high rate of availability. Figure 1.1 illustrates the main application areas of 5G, as mentioned before.

Also, 5G has to support tight quality-related requirements in order to provide enhanced user experience in more demanding environments compared to previous generations. As mentioned also in Reference [1], such requirements include the following:

- Support of very high bitrates of more than 1 Gbps in heterogeneous environments.
- Support of very high number of devices (massive machine communications) especially in ultradense environments.
- Support of very high mobility (e.g., up to 500km/h).
- Support of aircraft communications.

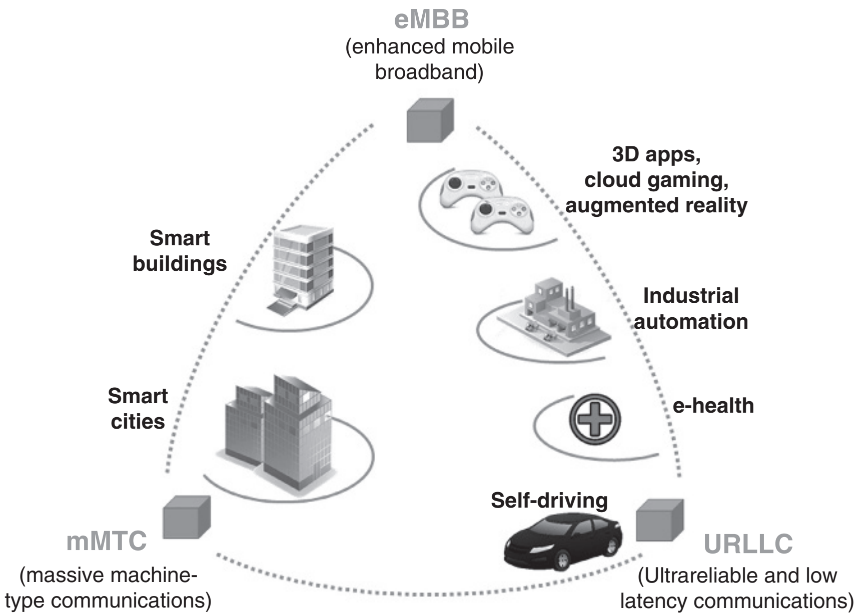


Figure 1.1 5G main application areas

- Support of mission critical communications with very low latency (down to 1ms).

For the successful realization of the aforementioned application areas, there are certain technological and networking trends that can lead toward the 5G direction. In terms of technology, trends are as follows:

- *Network Slicing:* Dynamic network slicing in 5G enables the design, deployment, customization, and optimization of different network slices running on a common network infrastructure. It leverages innovations in cloud mobile access and core [2].
- *Cloud and Fog/Mobile Edge Computing:* Cloud aspects can involve the allocation of resources to physical components in content servers, activation of the appropriate volume/type of functional components, and the determination of the interconnections and links between the physical elements. In this respect, they will manage in an aggregate manner all types of resources, namely, communication, computing, and storage, by taking advantage of significant processing powers/storage facilities associated with cloud platforms and virtualization through abstractions of resources and service components (which are pooled and universally accessible/sharable), in an on-demand, elastic, and scalable manner.

- *Separation of Control and User Plane:* Separation on control and user plane can potentially lead to more efficient usage of resources and energy efficiency as well. For example, the Green Touch initiative proposed that user plane (data) can be served mainly by small cells while a limited number of macro cells can serve as a signaling umbrella in order to handle control plane aspects.
- *Virtualization of Networking Functions:* Virtualization may include solutions based on software-defined networking (SDN) and network function virtualization (NFV) principles, which may be used for accelerating the application/service deployment times and flexibility (in general) in the network. For instance, through the standardized interfaces of the SDN model there can be instructions on how to handle new applications/services. Likewise, through NFV there can be an easier implementation of applications/services and networking intelligence (activation in cloud and instructions toward forwarding elements). The overall challenge is to evaluate the potential of these concepts, in terms of impact in the application/service deployment times, QoS/QoE.

In terms of wireless network trends there are the following (Figure 1.2):

- *Massive MIMO and Utilization of mmWave:* An important direction is related to the usage of massive multiple input–multiple output (MIMO) that can significantly enhance the spectral and energy efficiency of the wireless network. Moreover, the extra capacity that is needed by 5G networks for facilitating massive IoT, mobile broadband communications, and so on. can be provided through the utilization of extra spectrum by exploiting bands above 6 GHz. Mm-wave frequencies can be used for outdoor point-to-point backhaul links or for supporting indoor high-speed wireless applications (e.g., high-resolution multimedia streaming) [3]. Moreover, as the millimeter waves have a short wavelength, it becomes possible to pack a large number of antenna elements into a small area, which consequently helps realize massive MIMO at both the base stations and user devices [3].
- *Novel Multiple Access Schemes:* Novel multiple access schemes, such as nonorthogonal multiple access (NOMA), is one of the techniques being considered that uses cancellation techniques in order to remove the more powerful signal. Of course, well-known techniques such as orthogonal frequency division multiple access (OFDMA) can be exploited as well.
- *Ultradense Infrastructures:* Another direction that will continue to progress is the constant decrease in the cell size, at the expense of a corresponding constant increase in the number of cells that will be deployed. Cells of different sizes, characterized as macrocells, microcells, picocells, or femtocells will continue to be deployed. Specifically, hundreds of small cells are deployed per macrocell. The challenge with ultradense networks is to deploy and operate the appropriate set of cells, so as to carry data traffic, without severely increasing the signaling traffic (increases with the number of cells),

by minimizing the impact of mobility and radio conditions, and by achieving cost and energy efficiency.
- *New Waveform and Advanced Coding:* New candidate waveforms for 5G would be needed in order to serve specific service requirements, for example, higher or lower bandwidths, sporadic traffic, ultralow latency, and higher data rates compared to legacy technologies. Also, robustness toward distortion effects such as interference, RF impairments, and so on is important to be supported by new waveforms.

1.3 Status and Challenges in Hardware and Software Development

The increasing network demands require new network services, higher performance, increased bandwidth, lower energy consumption, and increased resilience. These demands imply a higher number of network devices and stations, increasing the cost and the energy consumption. The network centralization and functions virtualization become more significant as they enable better distribution of the available resources, less hardware utilization, and an easier to upgrade network as current devices and architectures are meeting their limits. New implementation techniques must be introduced in order to further increase reusability, flexibility along with performance and energy consumption at the same time. According to this approach, the network functions can be moved to software as much as possible, without affecting networks latency. The functions virtualization can be achieved at any level, using a partitioning technique between software and hardware functions, which takes into account the available resources. The current network systems introduce a static and customized functions virtualization. This manual and static partitioning may lead to high performance but it is not reusable and not reconfigurable, thus limiting

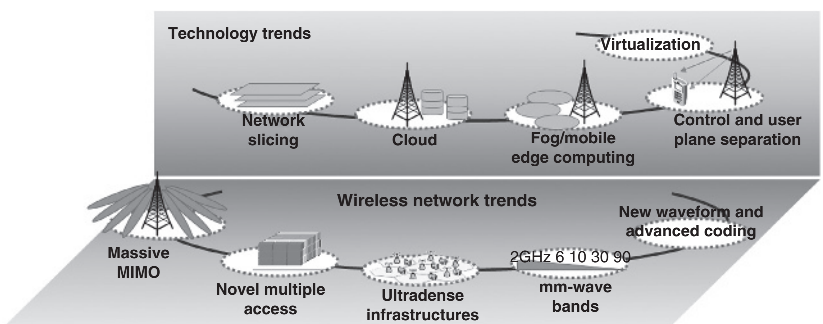


Figure 1.2 5G technology and wireless network trends.

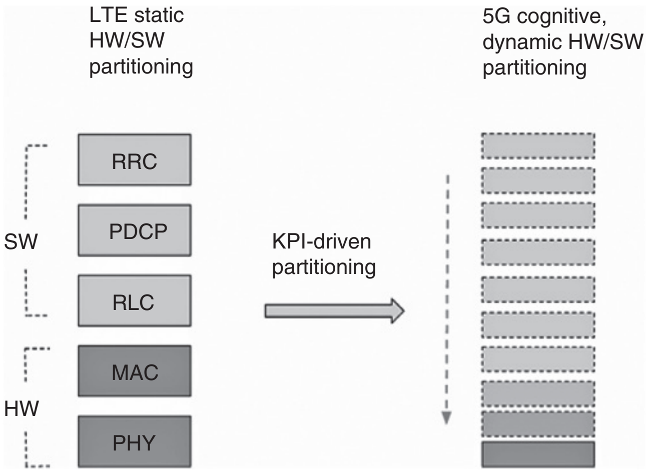


Figure 1.3 Cognitive, dynamic HW/SW partitioning in network stack layers.

network upgrades and resource allocation. Also, the processing power cannot be shared among nodes offering limited efficiency and spectrum capacity. Full virtualization is not always available as the devices of the underlying network might not be able of such a task, or virtualization might be limited according to available physical and computational resources. A cognitive and dynamic HW/SW partitioning can provide reconfigurable and flexible HW/SW partitioning to both device and network element architectures in 5G technologies, considering high performance and energy consumption reduction, according to the specified performance scenarios. The HW/SW partitioning is applicable for either inside a network stack layer and/or between multiple network stack layers, as shown in Figure 1.3.

1.3.1 Problem Statement

The cognitive, dynamic partitioning takes into account a set of given network functions, the KPIs that have to be optimized and the KPI constraints relative to the available resources. The technique result provides the HW or SW implementation decision for each given function, according to the given policies. These policies consist of the KPIs and their constraints. The result of the implementation has to change according to the policies alteration. The cognitive dynamic HW/SW partitioning task is to provide the best HW/SW partitioning of the 5G network stack functions, considering the given KPIs per scenario and the available HW/SW resources. The partitioning algorithm's result can be parsed to the management programs for further decision-making on the implementation part. The partitioning solution has to communicate with other programs to be aware of the available resources. Moreover, the partitioning

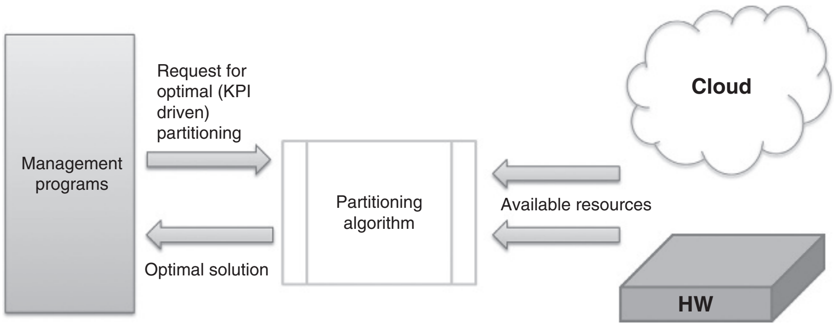


Figure 1.4 Partitioning algorithm functionality and communication.

must be aware of the performance scenario and the KPI constraints; thus the partitioning would interact with management programs and monitoring agents, as seen in Figure 1.4.

1.3.2 Solution

The cognitive dynamic HW/SW partitioning algorithm considers a multiple and diverse set of objectives, relative to 5G requirements. To this purpose, the current solution contains an evolutionary multiobjective algorithm, aiming to solve a specified optimization problem. This implies that the optimization problem formulation is one of the most significant parts of the solution, as it has to accurately address the requirements, the KPIs, and the objectives of a 5G network. The multiobjective algorithm receives as input the functional graph of the defined functions, along with the systems and functions KPIs/Constraints and optimization goals. Then the number of possible solutions is populated according to the optimization problem formulation. After having the set of solutions, a multiobjective algorithm will search for the best solution, according to the given KPIs, constraints, and optimization goals.

1.3.2.1 Functions Definition (LTE, 3GPP-Based PHY Functions)

The cognitive dynamic HW/SW partitioning algorithm is able to decide on device and network element functions, either inside a network stack layer and/or between multiple layers. Most of the current research and implementations are using functions from the LTE MAC layer or its reconfiguration while trying to introduce softwarization and reconfiguration of the LTE PHY layer functions. Some implementations are also able to simulate full LTE networks. The most notable among them are the LENA ns-3 [4] LTE simulator from CTTC, which provides full software implementation of virtual LTE networks, the OpenAirInterface [5], OpenLTE [6], and srsLTE [7]. In order to provide a first realization of the HW/SW partitioning challenges, the selected subset of functions used for the evaluation of the HW/SW partitioning is derived from the OpenLTE physical layer software implementation in Octave source code. The functions that compose the Octave code are identified, manually implemented in Verilog code, and characterized according to the KPIs. Specifically, the functions used for the evaluation of the algorithm untill now are the following:

- eNodeB LTE PHY [8] layer, downlink, OpenLTE, Octave (SW functions). The Octave source code that describes an LTE-based frequency domain downlink transmitter implementation in SW. This code contains the functions that are specified from the existing LTE, 3GPP standards.
- eNodeB LTE PHY layer function, broadcast channel:
  - Cyclic redundancy check (CRC), Verilog code (HW function)
  - Convolutional encoding, Verilog code (HW function)
  - Rate matching, Verilog code (HW function)
- eNodeB LTE PHY layer function, physical downlink control channel:
  - Cyclic redundancy check (CRC), Verilog code (HW function)
  - Convolutional encoding, Verilog code (HW function)
  - Rate matching, Verilog code (HW function)
  - Pseudorandom sequence generation, Verilog code (HW function).

The seven latter functions were manually implemented in Verilog code and evaluated in order to be bit accurate in conjunction to the corresponding SW functions. These functions are the first to be implemented in HW as they introduce a significant amount of delay in the SW execution time. Furthermore, while the implementation of broadcast channel and physical control channel, in Verilog, is similar, their HW mapping results in different power consumption that makes them perfect candidates to exercise the capabilities of the multiobjective algorithm.

1.3.2.2 Parameters (KPIs)/ Constraints Definition

The partitioning algorithmic solution considers the most critical KPIs for the 5G networks regarding partitioning and virtualization, so far. The considered KPIs are described in the following lines.

- Execution time
  - The measured execution time of the LTE network functions when implemented in SW.
  - The measured execution time of the LTE network functions when implemented in HW.
- Energy consumption of Verilog modules derived from power analysis using appropriate FPGA IDE.
- Measured SW memory utilization (e.g., RAM) of LTE functions when implemented in SW.
- Communication time, considering measured time of data transferring with respect to send and receive communication functions, between HW and SW implemented LTE functions.

- Reusability referring to the available HW resources, inversely relative to HW functions utilization.

1.3.2.3 Functional Graph (Dataflow Graph) Provision

The communication scheme between the utilized LTE functions forms a dataflow graph, including nodes that represent the functions implemented in HW or SW and edges that represent the link/interconnection between two functions. The communication overhead is also applied as weight to each corresponding link. The GUI accompanying the algorithmic solution provides the user with the ability to add or remove nodes/functions and edges. The partitioning algorithm receives this information and forms an array structure that includes the mapping of the interconnected components. The array items provide the multiobjective algorithm with the ability to find the optimal decision that also considers the communication overhead. An example is provided in Figure 1.5.

Figure 1.5 is generated by the user interface that controls the partitioning algorithm. The solution includes a predefined set of functions and interconnections, referring to the already implemented functions that form the graph in the figure. The user interface provides the user with the ability to add or

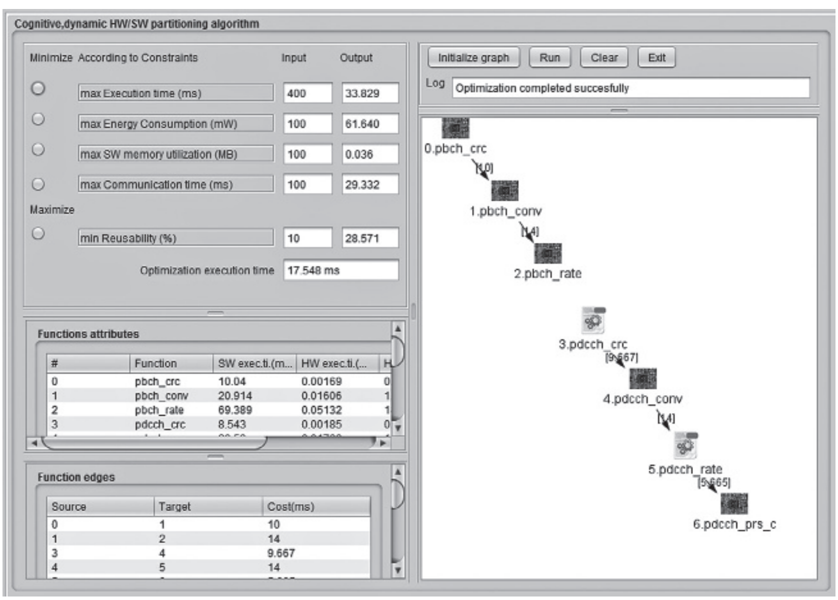


Figure 1.5 Graphical User Interface with generated data flow graph. The cyclic redundancy check function pdcch_crc and the rate matching encoding pdcch_rate functions of the physical downlink control channel are implemented in SW.

remove functions, interconnections, and their KPIs in order to create a graph that better suits his/her needs.

1.3.3 Optimization Problem Formulation

The core process of the optimization procedure is the optimization problem formulation as it guides the whole optimization process toward the goals of the problem. The main steps of the procedure are the following:

Input: This process receives as input the implemented or user provided functions along with their KPIs.

- N is the number of considered functions.
- M is the number of considered KPIs.
- An initial set of decision variables vector is created using random processes
  - $x = [x_1, \cdots x_N]$ is the decision variables vector, with x_n, $n = 1 \cdots N$, representing HW implementation ($x_n = 1$) or SW implementation ($x_n = 0$) of function n.

Process: When a KPI value is minimized, another KPI value is increased. In order to specify this relationship, the algorithm includes equations for each KPI with factors representing the relative difference. The objectives to be optimized are similar or even exact predefined KPIs, implemented by objective functions as described in the following paragraph– KPI-driven binary optimization:

- $f(x) = [f_1(\mathrm{x}) ,...f_M(x)]$ is the vector of objective functions to be minimized, each objective function f_m, $m = 1...M$, representing a corresponding KPI optimization.
- $f_m(x) = a_1mx_1 + \cdots + a_Nmx_N, where, a_nm$ is the difference of the KPI m value between HW implementation and SW implementation of function n. The factor a_nm is used to make the algorithm aware of the inverse relationship between the KPIs.
- Each objective function f_m, $m = 1\cdots M$, can have a corresponding constraint, for example, if m is the full model execution time-latency and it should be less than 80 s, this mean $f_m < = 80$.

A multiobjective algorithm will create a set of possible solutions (solution: $x = [x_1,\cdots,x_N]$) by measuring the objective function values for every solution. Then the algorithm evaluates the available solutions and sorts them considering the optimization target (specific KPI e.g., power).

Output: Finally, the algorithm selects the optimal solution ($x = [x_1,\cdots,x_N]$) representing HW implementation ($x_n = 1$) or SW implementation ($x_n = 0$) of function n) as described earlier.

The next section provides a description of the multiobjective algorithmic solution that is utilized toward this direction.

1.3.4 Evolutionary Multiobjective Algorithmic Solution

This section provides a description of the multiobjective algorithmic solution that is utilizing the optimization problem formulation. The multiobjective algorithm derives a subset of solutions from the set of all possible solutions according to the provided multiple objective functions, their KPIs, their constraints, and the optimization goal. The initial set of all possible solutions is provided from the objective functions that are considered in the optimization problem formulation. The functions and KPIs/Constraints are provided by the user utilizing the solutions user interface. The selected multiobjective algorithm for the HW/SW partitioning is the nondominated sorting genetic algorithm NSGA II [9]. The algorithm begins its search with an initial population of individual's solutions usually created at random within a specified lower and upper bound on each variable. Once the population in initialized, the population is sorted based on nondomination into each front. Once the nondominated sort is complete, the crowding distance (the Euclidian distance between each individual's solution) is assigned. After the population members are evaluated, the selection operator chooses better solutions with a larger probability to fill an intermediate mating pool. For this purpose, the tournament selection procedure takes place in which two solutions can be picked at random from the evaluated population and the better of the two can be picked. The crossover operator is to pick two or more solutions (parents) randomly from the mating pool and create one or more solutions by exchanging information among the parent solutions. Each child solution, created by the crossover operator, is then mutated with a mutation probability so that on an average one variable gets mutated per solution. In the context of real-parameter optimization, a simple Gaussian probability distribution with a predefined variance can be used with its mean at the child variable value. This operator allows an EO to search locally around a solution and is independent of the location of other solutions in the population. The elitism operator combines the old population with the newly created population and chooses to keep better solutions from the combined population. Such an operation makes sure that an algorithm has a monotonically nondegrading performance. Finally, the user of an EO needs to choose termination criteria. Often, a predetermined number of generations is used as a termination criterion. In most cases, this algorithm is able to find much better spread of solutions and better convergence near the true Pareto-optimal front compared to other multiobjective algorithms, on a diverse set of difficult test problems.

1.3.5 Testbed Setup

The current implementation of the cognitive dynamic HW/SW partitioning algorithm is tested considering the dynamic hot spot use case, resulting in the LTE eNodeB PHY layer reconfiguration, according to prespecified mea-

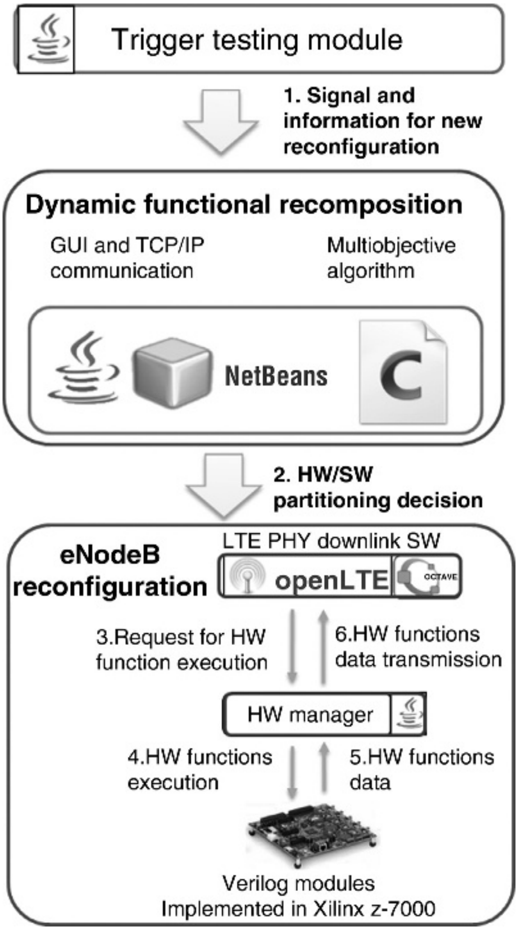


Figure 1.6 Cognitive dynamic HW/SW partitioning algorithm test-bed.

sured KPIs and optimization goals mostly referring to execution time or energy consumption reduction. The testbed of the demonstration includes the SW environments that will host the SW modules and a HW development environment along with an FPGA board, regarding the HW modules, as provided in Figure 1.6:

- The trigger module sends a signal that indicates a hot spot or not a hot spot condition. It is implemented in Java and it is connected to the dynamic functional recomposition module via TCP protocol.
- The dynamic functional recomposition module is the main module under test and contains the cognitive dynamic HW/SW partitioning algorithm

along with its GUI, TCP send/receive functions to communicate with the trigger module, and the SW implementation of the LTE eNodeB PHY layer. The algorithm provides the optimal decision vector and sends the decision to the SW LTE eNodeB (OpenLTE) PHY implementation via TCP.

- The LTE eNodeB PHY downlink SW function is provided by OpenLTE and it is implemented in Octave representing the SW functions of the network stack. It is facilitated by java TCP functions, in order to be able to receive the HW/SW partitioning decision from the DFR and to send/receive data to/from Vivado SDK control mechanism of the HW functions. The Octave code has been modified accordingly in order to exploit the timing results of the functions execution time, along with the communication time between HW and SW functions.
- The HW manager module is a java implemented module that receives, via TCP connection, the name of the function that must be implemented in HW along with the data that it has to measure and activates the HW function by sending appropriate synchronization messages to Vivado SDK environment that controls the HW functions inside an FPGA. The HW manager reads the results of the HW function execution and sends them back to the Octave OpenLTE module.
- The Vivado SDK environment controls and executes the aforementioned HW functions implemented in Verilog inside a Xilinx zc702 FPGA board and provides the corresponding data back to the HW manager. The energy consumption of the two aforementioned HW simulated functions is derived from the Xilinx Vivado power analysis tool.

1.3.6 Preliminary Test Results

The partitioning algorithmic solution provides cognition in terms of optimal decision based on information about functions KPIs; derived from management programs and monitoring agents, dynamicity referring to the ability to dynamically move functions from HW to SW implementation and vice versa according to the specified policies, considering also the communication overhead. Current results show improved overall performance (execution time, latency/communication overhead) by 70% and power consumption reduction by 50% with respect to the digital baseband processing of an LTE eNB PHY layer of Tx. Figure 1.7 illustrates the improvements in overall execution time of the addressed and tested physical layer functions, when HW/SW partitioning is performed targeting high-performance (a) and normal performance scenarios (b). In a high-performance scenario, more functions will be executed in HW while in normal performance more functions will move to SW in order to reduce power consumption.

The first set of results imply that the partitioning algorithm can achieve high performance with respect to algorithmic solution execution time, below thresh-

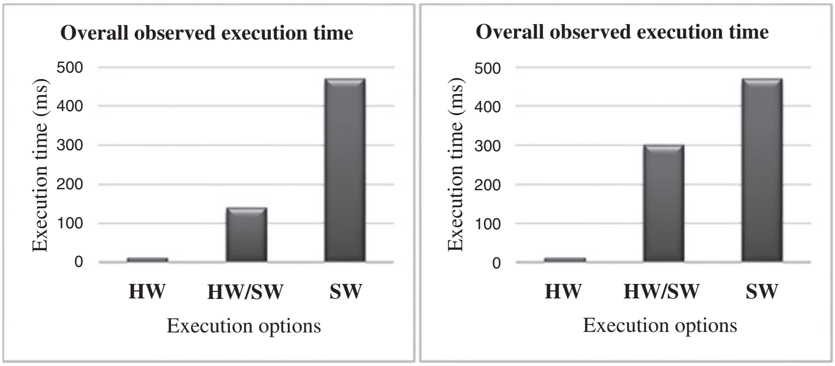


Figure 1.7 i) SW implementation, ii) HW/SW optimal partitioning, iii) HW implementation results in execution time (ms) for high-performance and normal performance scenarios using Octave

olds for LTE current handover time and user experience maintenance and algorithmic solution complexity. In addition, the cognitive, dynamic HW/SW partitioning algorithm provides reduction of the power consumption, limited memory overhead, enabling the interaction between algorithmic solution with management programs and monitoring agents. The proposed HW/SW partitioning will be further refined and upgraded considering improvements on performance, memory overhead, and portability of the current algorithmic solution. Furthermore, there is ongoing investigation for integration with the CTTCs LENA ns-3 simulator in order to have a full stack LTE network implementation to better clarify the benefits and challenges of HW/SW partitioning. Moreover, since the algorithmic solutions design enables KPIs and functions extensions future implementations and KPIs will be investigated in order to meet a diverse set of objectives such as user data rate and even capacity.

1.3.7 Status and Challenges in 5G Wireless Communications

This section provides useful insights on the advancements in physical layer, MAC, and RRM. Details on the aforementioned advancements are provided in the sections that follow.

1.3.7.1 Novel Physical Layer Aspects

One of the key aspects that will evolve with 5G has to do with the novel design of a unified, flexible air interface, its components, and procedures, so as to effectively deal with the issue of 5G requirements through such an adaptation. Therefore, it is important to develop a new spectrum agnostic 5G air interface for carrier frequencies below 6 GHz. This is motivated by the fact that todays licensed bands for cellular usage are all below 6 GHz, and the World Radio

Conference (WRC) in 2015 also focused on below 6 GHz spectrum among other aspects. Furthermore, even if higher frequency spectrum bands are made available for 5G operation in the future, having effective means for utilizing 5G below 6 GHz is still of relevance due to the more favorable radio propagation properties. A unified, flexible 5G air interface would have the following key characteristics:

- Flexibility to support the broad class of services with their associated KPIs
- Scalability to support the high number of devices
- Versatility to support the diverse device types and traffic/transmission characteristics
- Efficiency to support the requirements on energy consumption and resource utilization
- Future-proofness to support easy integration of new features.

The new air interface will meet the requirements on the 5G main KPIs (e.g., for increased throughput, reduced latency, etc.) with increased flexibility, reliability, future-proofness, and cost and energy efficiency. Notably, devices will be designed to support one common air interface for all services and more devices will be produced applying similar/common chip sets; thus achieving economy of scale for vendors. Also, the wireless system is expected to be more scalable and thus better suited to follow load variations between the services– both temporal and in different locations.

1.3.7.2 Novel Frame Design Based on Service Requirements

As previously mentioned, among the main objectives of a new, unified 5G air interface should be its flexibility to be adapted to the diverse requirements imposed nowadays by the heterogeneous service demands. This diversity of requirements creates a very challenging environment in terms of service-specific KPIs and channel characteristics as it should be flexible enough to satisfy these needs while in parallel optimize the resource utilization and minimize the overhead introduced by the multiservice support functionalities/mechanisms. The 5G services are foreseen to include mobile broadband (MBB) services, supporting high data rates and high coverage, massive machine communication (MMC) services, supporting small packet sizes and infrequent transmissions, mission critical communication (MCC) services with strict delay bounds and reliability factors, and vehicle-to-anything communication (V2X) services supporting both bs-to-device and device-to-device transmissions.

The current status of a stiff frame structure with, for example, a fixed transmission time interval (TTI) value either cannot satisfy the extremely strict requirements of specific services (e.g., delay requirements of MCC services) or results in underutilization and waste of resources due to inefficient resource management. Therefore, a flexible frame structure supporting the coexistence of different transmission time intervals (TTIs) is more than necessary in order to accomplish these diverse service requirements. The introduction of a flexible TTI supporting different TTI durations will both accomplish ultralow latency capabilities (e.g., in case of MCC services) by facilitating short TTI durations and high spectral efficiency gains (e.g., in case of MBB services) by utilizing long TTI durations. The selection of TTI scaling, which is the set of available TTI values, is of high importance, because it directly affects the effectiveness of the proposed flexible frame structure solution.

We proposed two methods for TTI scaling:

- 2^N *Scaling:* Definition of a set of TTI durations with double duration in each one.
- *Scaling Based On Service Classification:* Definition of a minimum set of TTI durations (mapped to a set of services)

In case of 2^N scaling, the minimum and maximum TTI length is defined and then a set of TTI lengths is generated based on the 2^N approach. For the 5G services mentioned already, reasonable values of TTI duration can be between 0.125 and 4 ms, therefore, the generated TTI values belong to the set (0.125, 0.25, 0.5, 1, 2 and 4ms). This approach is graphically depicted in Figure 1.8.

According to the second method, a set of predefined services (e.g., MBB, MMC, MCC, V2X) are analyzed based on their requirements in order for each service to estimate the TTI value that best reflects its characteristics, satisfy its requirements in terms of KPI values (e.g., average delay, max delay, throughput), and minimize the system resource overhead. Then a set of TTI lengths are generated that are mapped to the aforementioned services or group

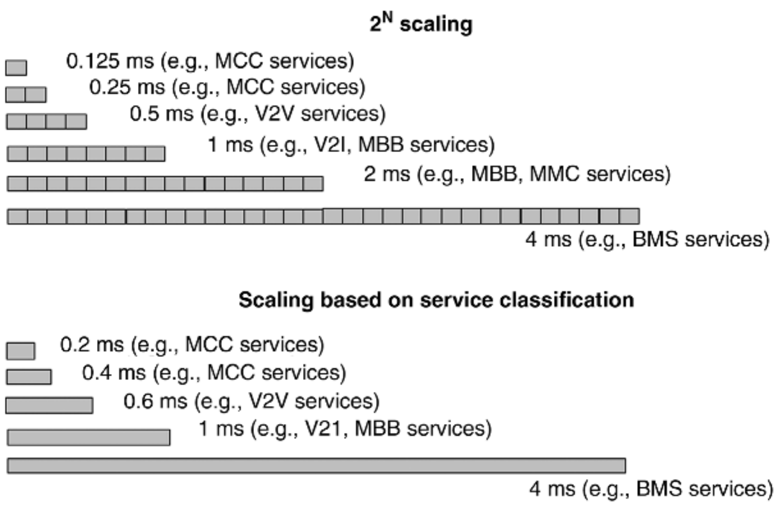


Figure 1.8 Flexible TTI - TTI Scaling

Table 1.2 Numerologies used in simulation scenarios.

RB Type	TTI (ms)	Subcarrier Spacing (kHz)	Number of Symbols per TTI
RB1 (LTE like)	1	15	14
RB2	0.5	15	7
RB3	0.25	30	7

of services. In Figure 1.8, an indicative set of TTI lengths is depicted based on an initial analysis and service classification. Each TTI length is mapped to one or several services. Regarding the comparison between the two scaling methods, the second method is better fitted to the service requirements as it is based on analysis of the special characteristics and optimum TTI selection for each service type. However, in case of absence of resource partitions in the spectrum, because of the different multiply factors between TTI values, this method cannot succeed high multiplexing gain between different services in the frequency domain as resource gaps may emerge for a set of selected TTI values. The first method may not be optimally, fitted to the services, although it eliminates the gaps of the spectrum (in case of no recourse partitions), while it occupies the minimum length in the packet header (for eight different TTI values, 3 bits are required).

1.3.7.3 Support of Different Numerologies

The following parameters are made available for configuration in order to create different numerologies: TTI length, subcarrier spacing, number of subcarriers, and tiling type (horizontal or vertical related to time/frequency axes). Table 1.2 illustrates the set of numerologies used during the preliminary system level evaluation.

A certain simulation set can be used by a proprietary system level simulation tool. The set evaluates a selected one-stage access protocol [10] in combination with a different numerology in order to estimate the combined gains from the use of both technical components. The results are evaluated against the LTE-A environment (ARP protocol and LTE-A frame structure). The general parameters of simulations are based on the parameters defined mainly in 3GPP case 1 reference scenario. The use case specific parameters for the set are presented in Table 1.3.

In the aforementioned simulation set, the performance of the combination of one-stage protocol with numerologies with smaller TTI sizes are evaluated. Figure 1.9 depicts the performance in terms of latency (primary KPI) for four alternative configurations. The first configuration is the baseline LTE-A scenario that includes the ARP protocols and the LTE-A frame structure with TTI length of 1 ms. In the other three configurations, the one-stage protocol is

Table 1.3 Simulation scenario parameters.

Parameters	Value
Reference scenario	3GPP case 1
Network topology	19 3-sectorized base stations (57 cells)
Intersite distance	500 m
Bandwidth	FDD 40MHz - 20MHz (uplink)
Request generation	Poisson
UE data traffic size	100 bytes / data report
PRACH allocation	1 PRACH allocation / TTI
Backoff window	Discrete uniform distribution (2,20) TTIs
Max connection attempts	4
Preambles	64

adopted in combinations with three different numerologies with TTI lengths of 1, 0.5, and 0.25 ms.

Figure 1.9 shows that the LTE-A (ARP, TTI = 1ms, PRACH per 1ms) has latency values between 5 and 7.5 ms, which is a performance far from the KPI target of 1 ms. The other configurations show improved latency values highly affected by the request rate. In detail, the numerology of TTI = 1 ms have relatively low-latency values only for low request volumes, while the numerologies of TTI = 0.5 and 0.25ms have relatively low-latency values for all

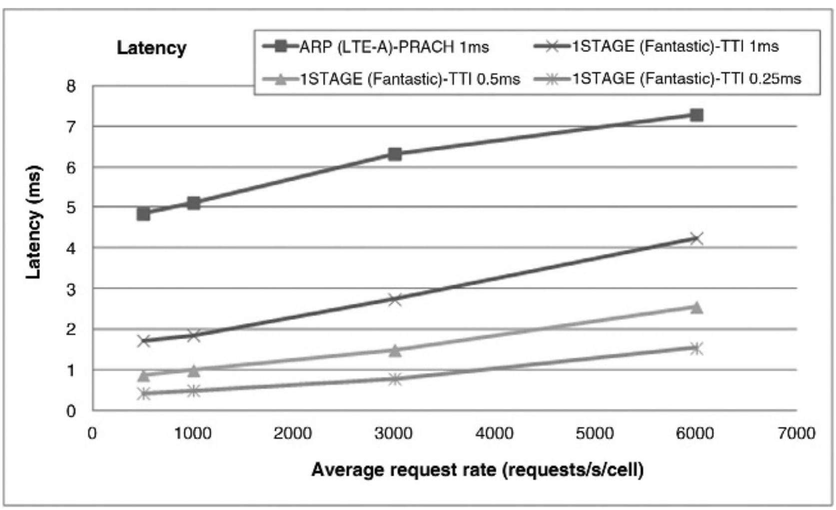


Figure 1.9 KPI 2 latency (primary KPI) - second set of simulation results.

the examined request volumes. Latency values below the KPI target are observed only for the numerologies of TTI = 0.5 and 0.25 ms and for low request rates.

The aforementioned latency results depict the latency from the UE to the BS (uplink), while no processing delay was taken into consideration.

1.3.8 Enhanced Radio Resource Management (RRM) and MAC Adaptation for 5G

In 5G networks it is important to proceed to the investigation and development of technologies that address the well-known challenges of predicted growth in mobile connections and traffic volume by successfully addressing the lack of dynamic control across wireless network resources that is leading to unbalanced spectrum loads and a perceived capacity bottleneck. Resource management with three degrees of freedom can be taken into consideration: (i) densification, (ii) rationalized traffic allocation over heterogeneous wireless technologies, and (iii) better load balancing across available spectrum bands in licensed, lightly licensed, and unlicensed spectrum portions. Moreover, the MAC has to be adapted in order to be able to support the deployment of various 5G services that call for increased reliability, reduced latency, and higher throughput in licensed, lightly licensed, and unlicensed bands.

Traffic steering provides operators with the necessary functionality in order to let them optimize resource utilization, QoS/QoE, and power consumption of cells and UEs by directing the traffic to the RAT or layer that is the most appropriate/suitable for a certain type of service. Steering of certain traffic flows depending on their type and availability of resources will enable devices to obtain guidance on how to optimally access content in indoor and outdoor environments. Various factors could be taken into account, including signal strength, interference levels, availability of RATs and channels, requirements of certain services (e.g., mission critical, massive access, etc.).

Figure 1.10 shows the actions that are considered for dynamic steering by taking into account certain prioritization of traffic flows (for the selection of RATs), load balancing (for the selection of cells), and channel assignment (for the selection of most appropriate/ suitable channels). Specifically, as depicted in the figure, reception of low quality from UEs triggers the initialization of traffic steering. RAT and channel with high load and low quality is identified in order to seek for solutions.

Furthermore, a prioritization algorithm receives reports on QoS and QoE from all UEs. In the case of triggering a low QoE/QoS, then the RRM mechanism must begin some procedures to satisfy the UEs requirements. Taking into account the requested conditions and demands from each UE can then steer a UE to its more appropriate RAT. Also, a load balancing algorithm focuses explicitly on achieving a good load balance between cells of the same RAT.

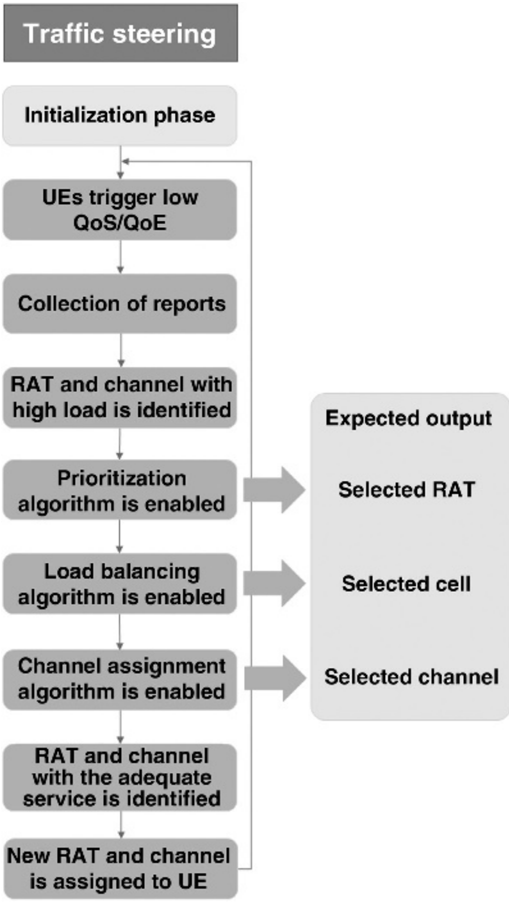


Figure 1.10 Traffic steering

Thus, based on sensing mechanism, the system is monitoring the loads and collects, the measurements from all cells of a particular RAT; an overloaded cell is identified and then under loaded cells that are in the vicinity are identified. An active and eligible UE can then be moved from the overloaded cell to the under loaded adjacent cell that meets its requirements in order to gradually arrive at the preferred load balance. In addition to the network load, the user experience (QoE) is also monitored throughout the sensing mechanism, and if needed an inter-RAT handover procedure is invoked to move the UE to a different RAT. Finally, at the time of connection establishment, the decision mechanism with the utilization of these algorithms and mechanisms is pursued to select the right RAT, cell, and channel to be used by the user throughout the connection.

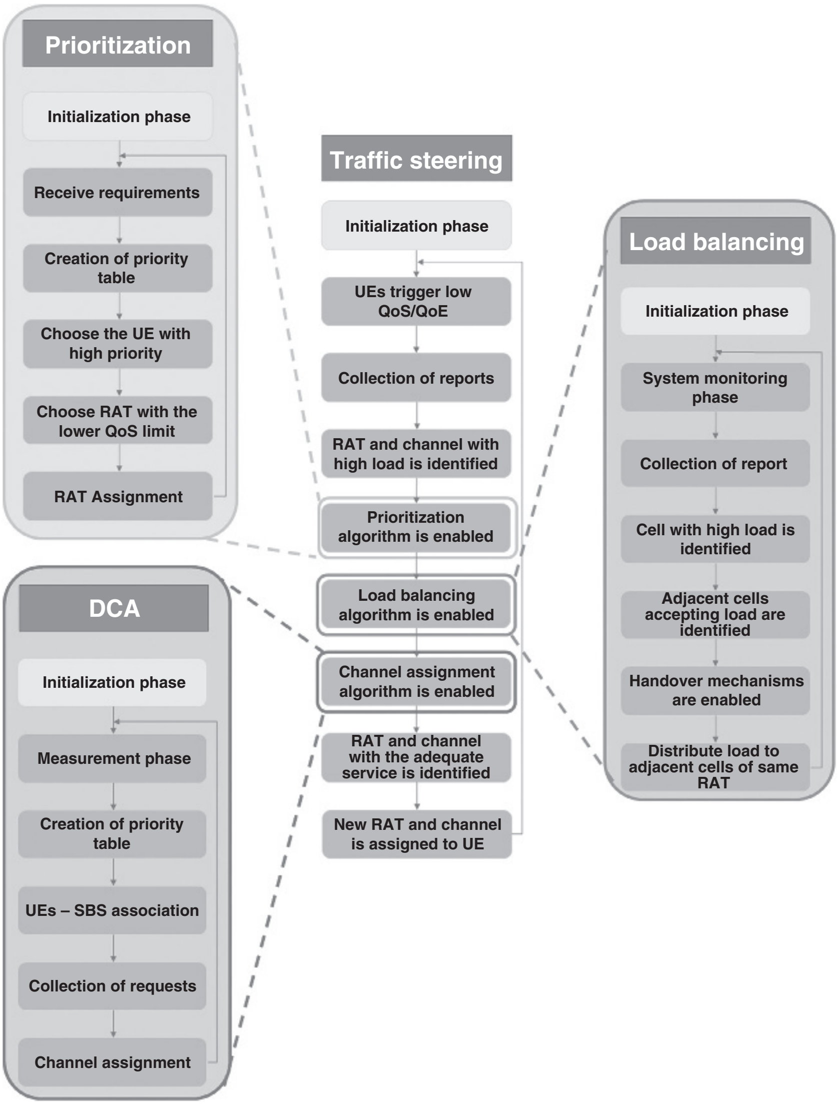


Figure 1.11 Dynamic Channel Assignment (DCA), prioritization, and load balancing as integral parts of traffic steering.

Figure 1.11 illustrates the integral parts of the Dynamic Channel Assignment (DCA); prioritization; load balancing, which have been previously discussed. Such concepts can be evaluated through system level simulations. Therefore, it is critically important to develop a system level simulator in a way to make a flexible, accurate, and efficient simulation of this complex heterogeneous net-

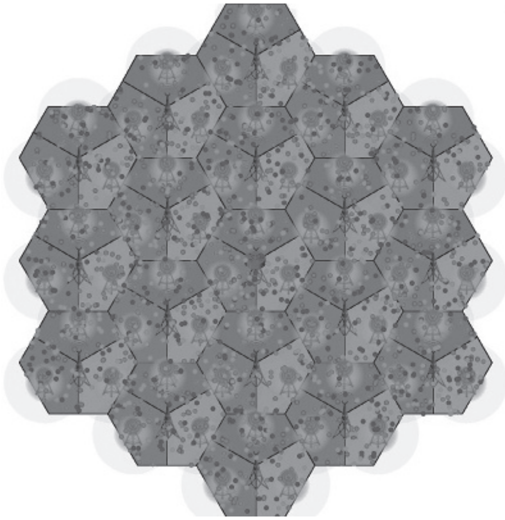


Figure 1.12 5G main application areas

works with multiple RAT, cells, and channels for 5G networks. Specifically, it should be possible to support heterogeneous networks consisting of LTE macrocells (licensed bands), small cells that utilize lightly licensed bands, and Wi-Fi with unlicensed bands, with BSs and APs densely deployed to provide users with seamless connectivity and demanded services, and finally, the radio resource management mechanism that assigns the most appropriate RAT, cell, and channel based on the requirements of each UE and at the same time keeping the system as load balanced as much as allowed. The particular simulator is a proprietary system-level simulation tool that is fully developed in Java with various capabilities and has been calibrated according to the 3GPP specifications. It takes into account various parameters such as traffic level, available infrastructure elements, available channels, and evaluates the various test cases. The calibration state of the proprietary simulator has been evaluated against the reference results of the 3GPP LTE calibration campaign [36.814] [11]. As a result, the cumulative distribution function (CDF) of coupling loss and downlink SINR have been examined in order to calibrate the tool with leading operators and vendors such as Nokia, Ericsson, DoCoMo, Huawei, and Telecom Italia.

The simulators playground is illustrated in Figure 1.12, where multiple macrocells and small cells were deployed in order to simulate a cellular environment as an example. The simulator is capable of creating various scenarios for sparse to dense and even ultradense deployment of various cells and technologies.

1.4 5G Network Management Aspects Enhanced with Machine Learning

1.4.1 Machine Learning for Service Classification in 5G Networks

A challenge for future wireless communication networks is the satisfaction of the diverse requirements coming from heterogeneous services. In 5G networks, the coexistence of different services like Mobile Broadband (MBB), massive machine-type communications (MMC), and mission critical communications (MCC) having various requirements in terms of both capacity and QoS will constitute a key prerequisite. Hence, one of the main issues that should be addressed by the 5G management system is the simultaneous provisioning of these services satisfying the corresponding requirements so as to optimize the network in order to be resource and energy efficient. A first step toward this direction is to be able to identify each service type in order to prioritize the services and be able to allocate efficiently the network resources.

Knowledge of QoS requirements per service flow could be provided by the higher layers as, for example, assumed in the HSPA and LTE, where sets of QoS parameters are available for RRM functionalities such as admission control and packet scheduling decisions. As an example from LTE, each data flow (bearer) is associated with a QoS profile consisting of the following downlink related parameters:

- Allocation retention priority (ARP)
- Guaranteed bit rate (GBR)
- QoS class identifier (QCI)

In particular, the QCI includes parameters like the layer-2 packet delay budget and packet loss rate. However, for the cases where detailed QoS parameters are not made available from the higher layers, the use of novel service classification techniques should be considered, in which the base stations monitor the traffic flows to extract more detailed service classification information and identify the service type providing this information input to packet scheduling algorithms and other RRM functionalities.

The support of fast and reliable traffic characterization is a necessary step in order to understand the network resource-usage and to provide differentiated and high QoS/QoE through prioritization targeting in increasing resource-usage and energy efficiency. In addition, the service classification process can interact with new services and procedures provided in 5G networks to support flexibility and adaptability to traffic variability.

In this section, the use of various machine learning mechanisms for the service classification problem is described and the performance of different algorithms is investigated. The considered classification methods reside in the area of statistical-based classification techniques and they are realized by

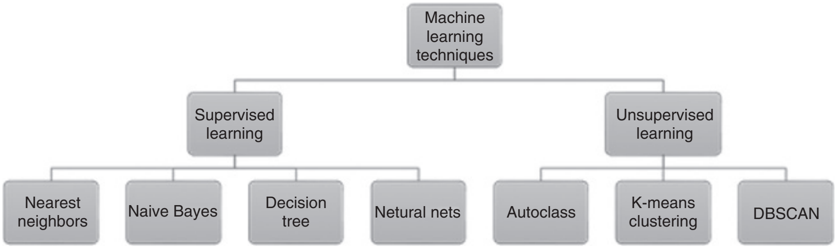


Figure 1.13 Overview of machine learning techniques.

exploiting several flow-level measurements (e.g., traffic volume, packet length, interpacket arrival time, and so forth) to characterize the traffic of different services. Then, to perform the actual classification, supervised machine learning techniques are applied to these measurements. It should be noted that in contrary to other methods of traffic classification, like payload-based classification, which need to analyze the packet payload or need to use deep packet inspection technologies, statistical-based classification techniques are usually very lightweight, as they do not access packet payload and can also leverage information from flow-level monitors.

1.4.2 State-of-the-Art Machine Learning Mechanisms for Traffic Classification

In the literature, there are a lot of studies that focus on application and service discrimination based on traffic classification learning techniques as presented in detail [12], [13], [14]. Various machine learning mechanisms are usually employed belonging to either unsupervised or supervised machine learning, as illustrated in Figure 1.13. It should be noted that recently semisupervised machine learning, mechanisms have also been proposed. However, considering that the majority of the existing works considers the other two categories, we focus only on supervised and unsupervised machine learning [15]. In the first case, clustering algorithms like K-Means, DBSCAN, and Autoclass [16] are investigated. The objective of these mechanisms is to group flows that have similar patterns into a set of disjoint clusters. The major advantage of these schemes is that they do not require a training phase like the supervised ones but they automatically discover the classes via the identification of specific patterns in the data set. However, the resulting clusters do not certainly map 1:1 to services as usually the number of clusters is greater than the number of service types and even in the case of 1:1: mapping, the clusters still need to be labeled in order to be mapped to the corresponding services.

Regarding the supervised machine learning techniques, which is also the approach that is analyzed in this section, there are various classification schemes that have been proposed for the traffic classification problem like Nave Bayes, Decision trees, Random forests, and others [17]. Authors in Reference [18] present the Bayesian classification techniques that use the naive Bayes approach. During the training phase, flow parameters are used to train the classifier and create a group of services. Then, when new flows arrive, they are subjected to probabilistic class assignment by calculating their probabilities of class membership and assigned to that class to which maximum probability is attained.

In addition, statistical fingerprint-based classification techniques as presented in Reference [19] classify traffic based on a set of preselected parameters (e.g., packet size, interarrival time). During the training phase, a data set of flows from each service are used in order to analyze the data set and create the service fingerprint. This fingerprint is usually a PDF (probabilistic density function) vector used to identify the service. During the classification phase, the algorithm checks the behavior of a flow against the available set of PDF vectors. Also, support vector machine (SVM) techniques, first proposed in Reference [20], are binary supervised classification algorithms that transform a nonlinear classification problem in a linear one, by means of what is called a kernel trick.

Furthermore, artificial neural networks (NNs) consist of a collection of processing elements that are highly interconnected and transform a set of inputs to a set of desired outputs that is inspired by the way biological nervous systems works. In Reference [21], authors proposed a NN, in which a multilayer perception classification network is used for assigning probabilities to flows. A set of flow features are used as input to the first layer of network, while the output classifies flow into a set of traffic classes by calculating the probability density function of class membership. Also, decision tree algorithms, which are mentioned in Reference [22], represent a completely orthogonal approach to the classification problem, using a tree structure to map the observation input to a classification outcome. In these supervised classification algorithms, the data set is learnt and modeled, therefore, whenever a new data item is given for classification, it will be classified according to the previous data set. The problem of ML data set validation is discussed in Reference [23] that highlights three training issues that should be considered in ML classification. These issues refer to the algorithm's impact when training and testing data sets are collected from same/different network, when the real online traffic classes of the training data set are not presented, and finally the impact of the geographic place where the network traffic is captured. Real Internet traffic data sets collected from a campus network are used to study the traffic features and classification accuracy for each validation training issue, demonstrating the impact of each issue.

1.4.3 Classification Approach and Evaluation Metrics

In this section, the problem of service classification is investigated employing a set of different classification mechanisms that belong to the supervised ma-

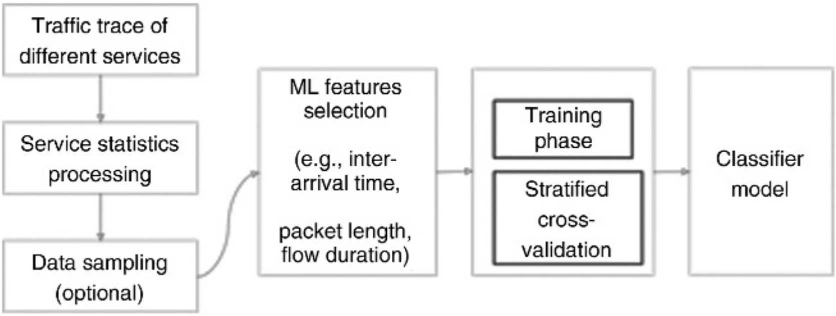


Figure 1.14 Proposed mechanism for the service classification process.

chine learning category. Before presenting the performance evaluation of each mechanism, the algorithmic procedure that has been followed is described and each step is explained in detail.

Figure 1.14 presents the algorithmic procedure that is followed for the classification mechanism. As can be seen, the first step refers to the collection of a number of traces from different services. For the considered simulation scenario, three service types are considered referring to MCC, MMC, and MBB communication while other services (like broadcast/multicast services) will also be considered in the future. The different traces of each service have been generated using specific traffic models. More specifically, the generation of different types of MCC/MMC traffic was following the traffic models presented in 802.16p [24], while for the generation of MBB traffic, video streaming traffic (YouTube) that follows the traffic models presented in Ref. [25] is assumed.

The second step refers to the statistical processing of these traces in order to separate them in flows. In particular, a flow is considered as a series of packets transmissions that have the same source and destination and for which the interarrival time is below a specific threshold. After this processing, a number of features for each flow is generated, including interarrival time statistics (mean value, standard deviation), packet size statistics (total value, mean/max/min value, standard deviation), and other flow characteristics like total number of packets, source, destination, and flow direction. Subsequently, some feature engineering tasks are performed before proceeding to the main classification mechanism. These tasks include the selection of the most representative features, the transformation of categorical features into numerical values, the normalization of features values and other tasks that guarantee a high data quality (e.g., replace missing values). Then, the implementation of machine learning mechanism follows including two main phases: the training phase and the cross-validation phase. It should be noted that stratification is applied in this case in order to randomly sample the flow data set in such a way that each service type is properly represented in both training and testing data sets.

Service \ Classification result	MMC service	Other services
MMC service	TP	FN
Other services	FP	TN

Figure 1.15 Confusion matrix of the service classification problem.

For the simulation scenario, a splitting of 70–30% for training and testing sets has been considered. Obviously, for the training set, the label service type of each flow is considered as known whereas for the testing set, this label is considered as unknown and each flow is labeled using the classifier model. The outcome of the proposed mechanism is a classifier model that can be employed in unknown flows in order to recognize them and label them in an accurate way.

To evaluate the performance of the classification mechanisms, various metrics have been defined and can be used in the train/test sets to select the most adequate mechanism for the specific problem. To illustrate the relationship between the different evaluation metrics, a very useful tool that provides a holistic view of each algorithm's performance is the confusion matrix. The confusion matrix is actually a two-dimensional matrix, in which the horizontal axis represents the predicted class (outcome of the algorithm) whereas the vertical axis represents the true class. In Figure 1.15, the confusion matrix for the considered classification problem is presented, where FP, TP, FN, and TN stand for false positives, true positives, false negatives, and true negatives, respectively, and they are defined as follows:

- FP: The percentage of other services flow that are incorrectly classified as MMC service.
- TP: The percentage of MMC services flow that are correctly classified as MMC service.
- FN: The percentage of MMC services flow that are incorrectly classified as other services.
- TN: The percentage of other services flow that are correctly classified as other services.

Some of the most common evaluation metrics used for classification problems are the accuracy metric, the precision, the recall, and the F1 score. More specifically, these are defined as follows:

- Accuracy is defined as the percentage of correct predictions to the total number of predictions and is given by

$$\frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN})}.$$

- Precision is defined as the percentage of the instances that were correctly predicted as belonging in a class among all the instances that were classified as belonging in this class and is given by

$$\frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FP})}.$$

- Recall is defined as the percentage of the instances of a specific class that were correctly classified as belonging to this class and is given by

$$\frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FP})}.$$

- F1 score is defined as the harmonic mean of the precision and recall and is given by

$$\frac{2\times \mathrm{Precision}\times \mathrm{Recall}}{(\mathrm{Precision}+\mathrm{Recall})}.$$

- To be able to choose the best mechanism for a classification problem, the investigation of a single metric, like accuracy, is not always enough as the misclassification of a specific class instances may be more important than the correct classification of others. For this reason, other evaluation metrics have also to be applied to make the most appropriate choice depending on the problem's characteristics.

1.4.4 Evaluation Performance of Classification Mechanisms

In the considered simulation scenario, the performance of a set of different machine learning mechanisms has been investigated, including base classifiers such as naive Bayes classifier, support vector machines, tree classifier, k-nearest neighbor classifier, logistic regression as well as ensemble-based classifiers like random forest classifier. The goal of ensemble methods is to combine the predictions of several base estimators built with a given learning algorithm in order to improve generalizability and robustness over a single classifier. Usually, two families of ensemble methods are distinguished: the averaging methods (e.g., random forests [26]) in which several classifiers are developed independently and then the average of their predictions is used, and the boosting methods (e.g., AdaBoost– Adaptive Boosting) where base classifiers are built sequentially and one tries to reduce the bias of the combined estimator. To be able to compare the various machine learning mechanisms, in Table 1.4, the accuracy metric of each algorithm is presented, where a Dump classifier that classifies all the flows as type 0 (MMC service) is also considered resulting in 0.512 accuracy. From this table, it can be seen that Decision tree and the

Table 1.4 Accuracy score for each classification mechanism.

Classification Mechanism	Accuracy
Naive Bayers	0.808
Support vector machine	0.662
Decision tree	0.976
k-nearest neighbor classifier	0.952
Logistic regression	0.685
Random forest classifier	0.988

Random Forest algorithms lead to the highest accuracy values, outperforming the other machine learning algorithms.

However, to provide a more complete view of each classifiers performance, the corresponding confusion matrices are illustrated in Figure 1.16. The horizontal axis of this matrix represents the predicted class whereas the vertical axis represents the true class. It should be noted that Class 0, Class 1, and Class 2 refer to MMC, MCC, and MBB service types, respectively. In the considered scenario, considering that it is desired to eliminate the possibility that a MCC service is misclassified as another service type, the optimal model should have high values of recall whereas high accuracy values for the case of MMC and MBB services are required. The results of confusion matrix show that the Decision Tree and the Random Forest algorithms result in extremely good results as they misclassify only a few flows, also resulting in high values of recall and precision, as can be seen in Figure 1.17. Therefore, these two classification mechanisms can be selected for further consideration for the problem of service classification.

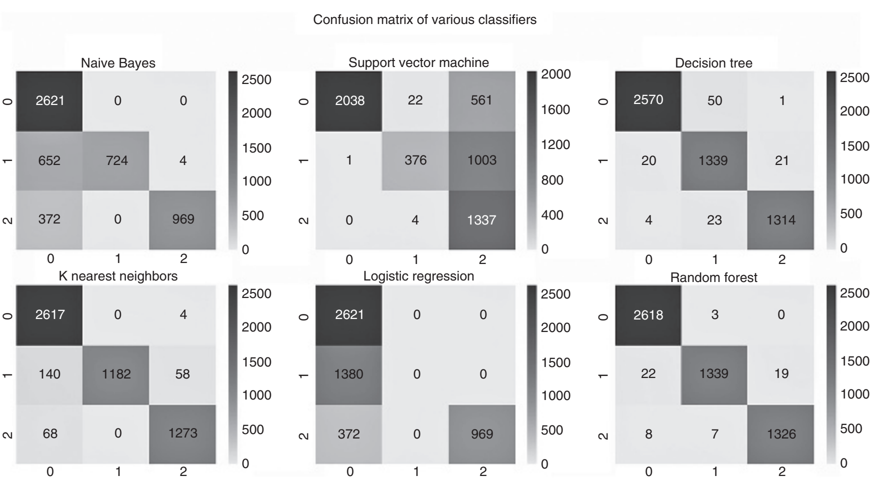


Figure 1.16 Confusion matrices of different classifiers.

Class\ metrics	Precision	Recall	F1 score	Class\ metrics	Precision	Recall	F1 score
MMC	0.99	1.00	0.99	MMC	0.99	0.97	0.98
MCC	0.99	0.97	0.98	MCC	0.94	0.97	0.96
MBB	0.99	0.99	0.99	MBB	0.98	0.99	0.99
Avg/total	0.99	0.99	0.99	Avg/total	0.98	0.98	0.98

Figure 1.17 Evaluation metrics for selected classification mechanisms.

1.5 Conclusion

5G is the next frontier of innovation for entire mobile industry. Consequently, the three major objectives for 5G are support of massive capacity and massive connectivity; support for an increasingly diverse set of services, applications, and users; and in addition flexible and efficient use of all available noncontiguous spectrum for widely different network deployment scenarios. Framed in this context, this chapter elaborated on the status and challenges in hardware/software development and in 5G wireless communications by focusing on physical layer, MAC, and RRM. Also, the benefits of machine learning in 5G network management were discussed. By taking into account the diversity of infrastructure, radio resources, and services that will be available in 5G, an adaptive network solution framework will become a necessity. Breakthrough developments in several RAN technologies will be required for realizing novel, 5G solutions. Such technologies include, among others, multiple access and advanced waveform technologies combined with coding and modulation algorithms, massive access protocols, massive MIMO, and virtualized and cloud-based radio access infrastructure.

Acknowledgment

Part of this work has been performed in the framework of the Horizon 2020 projects, SPEED-5G (ICT-671705), FANTASTIC-5G (ICT-671660), and Flex5Gware (ICT-671563) by receiving funds from the European Union.

References

1 NGMN, 5G White Paper. Available at www.ngmn.org (accessed April 2017).

2 J. Elliott and S. Sharma, Dynamic end-to-end network slicing unlocks 5G possibilities. Available at https://insight.nokia.com/dynamic-end-end-network-slicing-unlocks-5g-possibilities (accessed Nov. 2016).

3 T. E. Bogale and L. B. Le, “Massive MIMO and mmWave for 5G wireless HetNet: potential benefits and challenges,” *IEEE Veh. Technol. Mag.*, vol. 11, no. 1, pp. 64–75, 2016.

4 LENA, LTE-EPC Network Simulator. Available at http://iptechwiki.cttc.es/LTE-EPC.

5 N. Nikaein, R. Knopp, F. Kaltenberger, L. Gauthier, C. Bonnet, D. Nussbaum, and R. Ghaddab, "OpenAirInterface 4G: an open LTE network in a PC, MOBICOM 2014," *20th Annual International Conference on Mobile Computing and Networking*, 2014, Maui, Hawaii, Sept. 7–11, 2014.

6 Q. Zheng, H. Du, J. Li, W. Zhang, and Q. Li, "Open-LTE: an open LTE simulator for mobile video streaming," in *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, July 2014.

7 I. Gomez-Miguelez, A. Garcia-Saavedra, P. D. Sutton, P. Serrano, C. Cano, and D. J. Leith, *srsLTE: An Open-Source Platform for LTE Evolution and Experimentation*, Feb. 2016.

8 ETSI, LTE, Evolved Universal Terrestrial Radio Access (E-UTRA), Multiplexing and channel coding (3GPP TS 36.212 version 10.1.0 Release 10), 2011.

9 K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, April 2002.

10 S. Saur, A. Weber, and G. Schreiber, "Radio access protocols and preamble design for machine-type communications in 5G," in *2015 49th Asilomar Conference on Signals, Systems and Computers*, Nov 2015, pp. 3–7.

11 3GPP TR 36.814, Further advancements for E-UTRA physical layer aspects, March 2010.

12 N. Namdev, S. Agrawal, and S. Silkari, "recent advancement in machine learning based Internet traffic classification," *Procedia Comput. Sci.*, vol. 60, pp. 784–791, 2015.

13 S. Valenti et al. Reviewing traffic classification, in *Data Traffic Monitoring and Analysis*, Springer, 2013, pp.123–147.

14 P. Amaral, J. Dinis, P. Pinto, L. Bernardo, J. Tavares, and H. S. Mamede, "machine learning in software defined networks: data collection and traffic classification," *2016 IEEE 24th International Conference on Network Protocols (ICNP)*, Singapore, 2016, pp. 1–5.

15 T. Glennan, L. Christopher, and S. M. Erfani. "Improved classification of known and unknown network traffic flows using semi-supervised machine learning," *Australasian Conference on Information Security and Privacy.* Springer International Publishing, 2016.

16 T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," *IEEE Commun. Surv. Tutor.*, vol 10, no. 4, pp. 56-76, 2008.

17 J.S. Aafa et al. "A survey on network traffic classification techniques," *Int. J. Eng. Res. Technol.*, vol. 3, no. 3, 2014.

18 A. Moore and D. Zuev, "Internet traffic classification using Bayesian analysis techniques," in *ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS) 2005*, Banff, Alberta, Canada, June 2005.

19 M. Crotti, M. Dusi, F. Gringoli, and L. Salgarelli, "Traffic classification through simple statistical fingerprinting," *Comput. Commun. Rev.*, vol. 37, no. 1, pp. 516, 2007.

20 X. Peng, L. Qiong, and L. Sen, "Internet traffic classification using support vector machine [j]," *J. Comput. Res. Dev.*, vol. 3, p. 010, 2009.

21 T. Auld, A. W. Moore, and S. F. Gull, "Bayesian neural networks for Internet traffic classification," *IEEE Trans. Neural Netw.*, vol. 18, no. 1, pp. 223–239, 2007.

22 N. Williams, S. Zander, and G. Armitage, "A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification," *Comput. Commun. Rev.* vol. 36, no. 5, pp. 5–16, 2006.

23 H. A. H. Ibrahim, O. R. Aqeel Al Zuobi, M. A. Al-Namari, G. MohamedAli, and A. A. A. Abdalla, "Internet traffic classification using machine learning approach: datasets validation issues," *2016 Conference of Basic Sciences and Engineering Studies (SGCAC)*, Khartoum, 2016, pp. 158–166.

24 IEEE 802.16, Machine to Machine (M2M) Evaluation Methodology Document (EMD), 2011.

25 P. Ameigeiras et al. "Analysis and modelling of YouTube traffic," *Trans. Emerg. Telecommun. Technol.*, vol. 23, no. 4, p. 360–377, 2012.

26 G. Biau, L. Devroye, and G. Lugosi, "Consistency of random forests and other averaging classifiers," *J. Mach. Learn. Res.*, vol. 9, pp. 2015–2033, 2008.

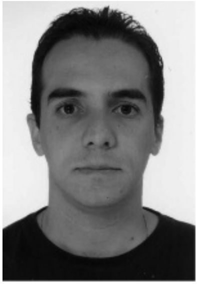
Mr. Ioannis-Prodromos Belikaidis is an electrical and computer engineer. He has been involved in routing and security of VANET (Vehicular Ad-hoc NETworks) projects and also in WiMAX propagation research. He also has extensive experience through 5G and energy-related projects that were conducted in the scope of his under- and post-graduate studies and internships in companies of the energy sector in Greece. He has knowledge on algorithms and procedures for complex networks, mobile, and pervasive, distributed computing, and currently he is active in 5G R&D.

Dr. Andreas Georgakopoulos is a solution architect and he has been involved in projects and R&D initiatives for almost a decade. Currently, he is active in 5G and Energy R&D activities and intelligent management of dynamic access networking paradigms such as opportunistic networking. He has also participated in the GreenTouch Initiative for energy-efficient networks (led by Bell Labs USA) and Wireless World Research Forum. He has also acted as work package and task leader in recent H2020 projects related to ICT and Energy.

Dr. Evangelos Kosmatos received his Dipl-Ing. and Ph.D. degrees from the school of Electrical and Computer Engineering (ICCS) of National Technical University of Athens (NTUA), Greece, in 2002 and 2008, respectively. He has participated in several EU projects (IST AQUILA, ENAMORADO, SMS, INCASA, STRONGEST, IDEALIST, CONFES, FANTASTIC-5G) as well as national projects (CONFES, WisePON). His research interests include 4G, 5G networks, optical networks (EPON, GPON, WDM-PON), Radio-over-Fiber (RoF) networks, protocols and algorithms supporting Quality of Service in heterogeneous networks, sensors/actuators networks (Internet of Things), middleware, and distributed technologies.

Dr. Stavroula Vassaki received the diploma degree in electrical and computer engineering (ECE) from the National Technical University of Athens (NTUA), Greece. Also, she received the M.Sc. degree in economics and administration of telecommunication networks from the National and Kapodistrian University of Athens, Greece and the Ph.D. degree from the NTUA focusing on the area of radio resource management in future telecommunication networks. Since 2006, she has been an associate researcher/ project engineer with the Mobile Radio Communications Laboratory, ECE, NTUA, participating in various research-oriented projects. From August 2009 to January 2011, she was an RF Engineer with the Non-Ionizing Radiation Office, Greek Atomic Energy Commission. She has been involved in various international and national research projects. Her research interests focus on wireless and satellite communications networks and include resource management mechanisms, optimization theory, game theory, and machine learning.

Mr. Orestis-Andreas Liakopoulos received the bachelor's degree in Physics from National and Kapodistrian University of Athens, Department of Electronics, Computers, Telecommunications and Control in 2012 and master's degree in electronic automation from national and Kapodistrian University of Athens, Faculties: Physics, Informatics and Telecommunications in Electronic Automation, Programming, Computer Science in 2014. In 2014, he participated in MIMO System Physical Layer Design and Implementation, in NKUOA cooperation with ANTCOR S.A. (ICT4 Growth, GSRT). Currently, he is a solution engineer at WINGS ICT Solutions. At the European level, he has been actively involved in the following international research and development programs: H2020-PHANTOM and H2020-Flex5Gware. His research interests are in the areas of structured programming, communication networks, parallel and distributed systems, object oriented programming, design and Use of database systems, and parallel computer systems technology.



Dr. Vassilis Foteinos graduated from the Department of Informatics & Telecommunications of the National and Kapodestrian University of Athens in 2008, received his M.Sc. diploma from the Department of Digital Systems of the University of Piraeus in 2011, and Ph.D. from the same department in 2015. He has conducted lab lessons for undergraduate students on the topic Computer Networks and Network Management in the Department of Digital Systems, University of Piraeus. Currently, he is participating in the H2020-Flex5Gware project. His main interests include the (co-)design of Hardware–Software systems and the development of innovative mechanisms for reconfigurable platforms in future 5G networks.

Dr. Panagiotis Vlacheas received the diploma and Ph.D. degree in electrical and computer engineering from the National Technical University of Athens (NTUA), School of Electrical and Computer Engineering, Athens, Greece, in 2003 and 2010, respectively. In 2004, he received the Ericssons awards of excellence in Telecommunications for his diploma thesis. Currently, he is a senior solution architect at WINGS ICT Solutions. At the European level, he has been actively involved in a number of international research and development programs, among others H2020-PHANTOM acting as Technical Manager, H2020-PROTEUS, H2020-Flex5Gware, H2020-iKaaS and MigraineNet project. His research interests are in the areas of cloud-IoT-Big Data integration, cognitive IoT, mobile edge computing, embedded systems, M2M communications, and reconfigurable and cognitive networks.

Prof. Panagiotis Demestichas is an experienced, senior electrical engineer with 25 years of experience in R&D and a professor in the University of Piraeus, Department of Digital Systems. He is active in the areas of 5G systems (services, environment conditions, ultradense infrastructures, spectrum aspects, virtualization, and softwarization), technoeconomic management and optimization, sustainability, simulation platforms, proof of concept, and validation by means of experiments involving environment condition generation and advanced analytics. He was also the project coordinator of ICT-FP7 OneFIT project and the technical manager of ICT-FP7 E3 project.