Enhanced Connectivity in Wireless Mobile Programmable Networks

PhD Thesis
Author: Luca Cominardi
Advisor: Carlos J. Bernardos

Technical summary

Packet-based mobile networks have experienced a huge success in the last years, with the number of subscribers and traffic volume constantly growing. Several reports show that the mobile traffic growth will not decelerate, but increase 10-fold instead from 2015 by the end of 2020. The envisioned scenario will not only assume a large increase in data volume, but also a profound diversification of traffic and service demands, leading to a new environment for the telecommunication industry which has been identified as the 5$^{\text{th}}$ generation of mobile networks. On the one hand, 5G aims at improving the network infrastructure to meet users' demands while reducing the associated deployment and operational costs for network operators. On the other hand, 5G enables a plethora of new services and business opportunities for solutions providers as documented by the 3rd Generation Partnership Project (3GPP)'s New Services and Markets Technology Enablers (SMARTER). As a result, the 5G wireless networks are expected to support the needs of an hyper-connected society which is continuously demanding very high data rate access, requiring a wider coverage, and offering an increasing number of almost permanently connected devices. In order to cope with this ever-increasing traffic demands, nowadays network technologies are hence experiencing a shift towards *softwarisation*. The key idea is to bring the flexibility and reduced cost of software development to network deployment.

This idea is materialised by the Software Defined Networking (SDN) paradigm, which moves the intelligence residing in the network elements to a central controller, which implements the network functionality through software. In traditional approaches, the network's control plane is distributed throughout all network devices, while SDN logically centralises the control plane. This removes the need of complex and costly changes in equipment or firmware updates in order to introduce new characteristics in the network, as only a change in the software running at the controller is required. The communication between the centralised controller entity and the merely traffic forwarding devices could be supported by what is called southbound interface protocols. One of the candidate protocols for it is OpenFlow, which is being standardised by the Open Networking Foundation (ONF). The main advantage of this approach is that operators can benefit from an increased *flexibility* to manage their networks and implement new services. Following this new technology (SDN), the 3GPP's architectural study for next generation mobile systems focuses on enhancing the mobile network's core part, considering the evolution of the network towards a distributed and softwarised ecosystem.

One service of paramount importance in mobile networks, which intrinsically differentiates them from fixed networks, is the mobility management of the users. Indeed, mobile users are characterised by the fact that they move within the network and may change point of attachment while moving (e.g., antenna they are connected to). The control and data planes of these users are converged at the same Packet Data Network Gateway (PGW) in today's 4G systems, making the 4G architecture highly centralised and hierarchical. The PGW hence is in charge of terminating the mobility signalling of the users as well as forwarding their traffic to and from the mobile network enforcing any policy functions. By doing so, the gateway acts as mobility anchor following the user movements by simply re-routing the packets over tunnels created with the access router where the user terminal is currently connected. But this simplicity comes with some penalties: the mobility anchor represents a single point of failure, it poses scalability issues overloading the network core, and, in general, it leads to sub-optimal paths between the mobile users, also known as Mobile Node (MN), and their communication peers, also known as Correspondent Node (CN). Flattening the network architecture is regarded as one of the most promising approaches to designing the architecture of the next generation system, and the Distributed Mobility Management (DMM) paradigm goes precisely in such direction. Both 3GPP and Internet Engineering Task Force (IETF), which are the main standardisation bodies in this area, have looked and are still looking at DMM-alike solutions.

One of the main contributions of this thesis is the analytic and experimental evaluation of two key DMM protocol families: IP mobility and SDN-based. While the Proxy Mobile IPv6 (PMIPv6)-based solution is available in literature, this thesis thoroughly designed, modelled, and implemented the SDN-based solution. Additionally, this thesis walked the path of decomposing the functions that a DMM solution should have and identify how these can be implemented in an DMM-based solution. Moreover, existing state-of-the-art solutions are not generally studied both analytically and experimentally as it is done in this thesis, thus providing solid insights on how to apply DMM concepts in future mobile networks. By implementing the proposed SDN architecture and testing it on a medium size test-bed, this thesis demonstrated how easy and quick would be for an operator to create and put into operation new services, like the proposed SDN-based DMM. The results obtained from analysis and experiments show that the performance of the analysed solutions depends on the scenario being considered, but also indicate that SDN approaches have a big potential: (*i*) achievable performance is good and even better than the one of the PMIPv6-based solution, (*ii*) the solution can be easily implemented, and (*iii*) provides additional flexibility in regards of how it behaves and provides service differentiation.

In order to provide the desired flexibility, the 5G transport network needs to undergo a profound transformation, evolving from a static connectivity substrate towards a service-oriented infrastructure capable of accommodating the various 5G services, including Ultra-Reliable and Low Latency Communications (URLLC). The network infrastructure for mobile networks traditionally envisages a backhaul segment for assuring connectivity between the core network and access network. Such segment is typically composed of several sub-systems including heterogeneous wired/wireless forwarding networks and fibre-based aggregation/routing networks, thus including several packet nodes, such as switches, routers, aggregators, etc. This heterogeneity leads to the use of various transport protocols for transporting packets between these nodes such as Carrier-grade Ethernet, Optical Transport Network (OTN), Synchronous Digital Hierarchy (SDH), Multiprotocol Label Switching (MPLS), IP, etc. The packets transmitted on the backhaul segment are also referred to as backhaul traffic. Recently, a new network segment called fronthaul has emerged, as the result of more centralised radio access network (C-RAN) architectures where the Evolved Node B (eNB) is split into two elements, a Remote Radio Head (RRH) and a Baseband processing Unit (BBU). The RRH simply keeps the RF functions necessary for the signal radiation at the cell site while the BBU takes all the baseband heavy computational functions to a separate location. The simplicity of the antenna sites would allow a more cost effective and flexible deployments of the 5G Radio Access Network (RAN), which is also expected to be denser so as to increase the spectrum re-utilisation and capillarity. To enable this functional split, new protocol interfaces have been designed, such as Common Public Radio Interface (CPRI), enhanced CPRI (eCPRI), Open Radio equipment Interface (ORI), and Next Generation Fronthaul Interface (NGFI). The data generated by these interfaces and then transported on the fronthaul segment is referred to as fronthaul traffic. While the location of the RRH is determined by the physical environment and deployments, the physical location of the BBU is variable (e.g., fully at the edge, in a local cloud, or fully central cloud). This can create situations where fronthaul and backhaul traffic may eventually share the same physical segment of the transport network. Different functional splits, and hence interfaces, impose different requirements (e.g., bandwidth, latency, jitter, BER) on the fronthaul interface that must be guaranteed within the transport network, which comprises heterogeneous transmission technologies.

A second major contribution of this thesis is therefore the analysis and design of a SDN-based unified data plane architecture for 5G, namely crosshaul, based on two main components: (*i*) the Crosshaul Forwarding Element (XFE) and (*ii*) the Crosshaul Common Frame (XCF). The XFE is a multi-layer switch based on packet – Crosshaul Packet Forwarding Element (XPFE) – and circuit – Crosshaul Circuit Switching Element (XCSE) – switching elements. While backhaul traffic is usually transmitted over the packet switch network, CPRI and diverse fronthaul traffic with stringent timing constraints are transmitted over the circuit switch network due the tight bandwidth and latency requirements the interface imposes to the network, which makes this interface quite rigid and costly. Aligned with new radio functional splits under study, NGFI and eCPRI relax the requirements of today's fronthaul in order to reach a more scalable interface so cheaper transport technologies can be used. At this purpose, packet switching enables statistical multiplexing when the peak to average radio access traffic load in 5G is high enough. Unified forwarding is enabled by the XCF format that is common across the various types of traffic and the various link technologies in the network. As a consequence, the unified data plane enables a common management of the integrated network in a SDN fashion. Therefore, traffic requirements, and hence services, could be easily enforced onto the network by leveraging the integrated and harmonised view provided by the unified data plane. As a result, the network operational costs can be significantly reduced.

In addition to the backhaul and fronthaul traffic, 5G transport networks will hence need to accommodate different kind of services with very distinct requirements on top of the same physical infrastructure. Specifically, these services span across a large variety of new use cases which go beyond the natural evolution of voice and data delivery in 4G mobile networks, such as multi-access network integration, even across operators and less trusted networks, Internet of Things (IoT), localised real-time control, vehicular communication, etc. All of this poses significant challenges to the 4G monolithic and centralised network architecture, both in terms of flexibility and scalability, making new services hard to introduce and scale. In addition, sharing the physical network assets through multi-tenancy is seen as a viable path for reducing the ever-increasing costs involved in the deployment and management of future networks. The above is key to understand the new Quality of Service (QoS) capabilities that are required to be supported in 5G transport networks so as to simultaneously fulfil the disparate traffic and tenant requirements associated to the

5G services.

The 3GPP identifies three main categories for the 5G services, namely enhanced Mobile Broadband (eMBB), URLLC, and Massive Internet of Things (MIoT). Each of them present different inherent characteristics spanning from ultra-low latency to high bandwidth and high reliability. According to Next Generation Mobile Networks (NGMN), these multiple services may be provided by customised network slices, which provide the necessary traffic treatment over the same physical substrate. Traffic differentiation is enforced at the access border of the network in order to ensure a proper forwarding of the traffic according to its class through the backbone, where it is more feasible to have high capacity. While existing networks can be considered as a continuum from the access to the interconnections points forming an end-to-end path, network slicing breaks this situation since now the end-to-end path becomes a composition of segmented paths within different slices that could even pertain to distinct administrative organisations or providers. This means that the end-to-end path traverses now many edges where the traffic should be enforced, discriminated and ensured, according to the service and tenants needs. Thus, transport networks move from a single-edge continuum towards a multiple-edges structure in 5G. Apart from the technical complexity added, cost implications can be expected, since the specialised and more expensive hardware today used only in the border for implementing fine-grained QoS should be generalised throughout the network.

This thesis therefore presents a characterisation of a 5G transport network and the expected traffic mixture of network slices. Several simulations have been performed to understand the role of queueing disciplines in different scenarios, such as urban, industrial, and rural. This characterisation is key for properly engineering operator's networks to support next 5G services and satisfy the very stringent and diverse needs intrinsic to each of them. It indeed provides powerful insight on the candidate nodes in the network where a given service should be provided in order to fulfil its traffic requirements. The results have been compared with the constraints of the traffic flows defined in 3GPP and criticality has been identified for the motion control traffic part of the URLLC slice. Jitter requirements for such flow are only satisfied when the traffic is terminated in the access ring and a strict priority with preemption queueing discipline is used. Regarding the other flows and slices, traffic requirements are fulfilled in a failure-free scenario where the protection ring in the access and aggregation is not activated.

Furthermore, this thesis has identified a gap between current SDN solutions and carrier grade network requirements under Operations, Administration and Maintenance (OAM) point of view. An analysis of widely-deployed OAM and SDN technologies has been hence performed showing that the stateless nature of OpenFlow poses significant scalability and accuracy problems in monitoring and managing the network. To overcome these issues, this thesis proposes an Adaptive Telemetry System (ATS) to enable locally on the switches active measurements (e.g., delay, bandwidth, etc.) and their reporting (e.g., alarms). The design approach chosen for ATS showed to provide compatibility with standard OpenFlow switches and controllers. An Application Programming Interface (API) has been defined for enabling the remote configuration of telemetry procedures, which adopt a Finite State Machine (FSM) implementation. This enables the switches to locally execute the stateful procedures required for active monitoring. Finally, an experimental evaluation has been presented, showing the benefits of ATS compared to legacy-SDN solutions. Particularly, ATS proved to bring significant benefits in terms of offloading the control plane, and the Network Controller (NC), as well as higher accuracy in the performed measurements, which comply with the performance requirements defined by 3GPP for 5G networks. To that end, the delay and bandwidth measurements obtained with ATS have proven to match the ones obtained with reference non-SDN tools, while providing higher flexibility in the type of measurements that could be performed. Moreover, ATS proved to be able to manage the periodical generation of messages over a large number of ports (up to 256) while running on a single CPU core. Finally, this thesis provided some implementation insights on ATS and some deployment considerations regarding the clock distribution in the network.

In addition to the separation of control and data planes (i.e., SDN), the telecommunication industry is embracing virtualisation technologies for evolving towards a cloud-based infrastructure. This trend has led to the creation of the European Telecommunications Standards Institute (ETSI) Network Function Virtualisation (NFV) Industry Specification Group (ISG) who pioneered the idea of bringing virtualisation capabilities into mobile operator networks to increase flexibility in service offerings and network management. By decoupling the network functions from the underlying hardware platform, NFV allows operators to dynamically deploy services in response to the needs of the traffic. In addition to NFV, ETSI Mobile Edge Computing (MEC) ISG brings computing capabilities close to the end users to cope with the ever-increasing amount of data (e.g., generated by IoT) and the low latency required by some use cases (e.g., vehicular communication). NFV and MEC jointly represent a paradigm shift for mobile operator networks, which evolve from a centralised architecture based on monolithic and hardware-integrated functions to a software-based distributed architecture. Such evolution enables a common hosting environment, namely edge computing, at the network edge characterised by low latency and high bandwidth as well as real-time access to radio network information. Network functions and software applications can be hence deployed close to the end users, thus alleviating congestion at the mobile network core and serving efficiently local purposes, such as data aggregation for IoT, localised real-time control, and single aggregation point for multi-access connectivity.

As ultimate final contribution, this thesis identifies the edge and fog as key pillars of future networks where intelligence and innovations will be increasingly applied. There is however not yet a common unified platform that integrates and federates these two pillars together. Whilst the edge is more infrastructure-oriented and hence easier to integrate, the fog tends to be more volatile with resources appearing and disappearing on the go, and belonging to different owners. The opportunities for such unified framework are clearly acknowledged, but there remains to be several challenges that need to be addressed first before such a common framework could emerge. These include: (*i*) the dynamic discovery of volatile and non-volatile resources, (*ii*) the federation of these resources when they belong to different domains and owners, (*iii*) the support of multi-tenancy in particular for the volatile fog resources, (*iv*) the customisation and interworking of different virtualisation technologies suitable to each type of resources (i.e., edge and fog), (*v*) the dynamic placement of functions and applications across the continuum of fog and edge, (*vi*) the automation and dynamic allocation and management of the resources, and finally (*vii*) the security, trust and privacy considerations. The overcoming of these challenges would hence enable a convergent 5G multi-Radio Access Technology (RAT) access through the integrated virtualised edge and fog solution, which envisages two main components: the Edge and Fog computing System (EFS) and the Orchestration and Control System (OCS). The former provides a low latency integrated virtualised environment distributed across the fog and edge to support multi-RAT convergence. The latter instead leverages on extended SDN, NFV, and MEC tools to build and maintain the EFS, by enabling the automatic integration and federation of EFS resources into a unified hosting environment, despite their heterogeneity, ownership, and volatility.