Demystifying Network Slicing: From Theory to Practice

Thomas Soenen*, Ratul Banerjee*, Wouter Tavernier*, Didier Colle* and Mario Pickavet*
*UGent - iMinds: {thomas.soenen, ratul.banerjee, wouter.tavernier, didier.colle, mario.pickavet}@intec.ugent.be

***Abstract*—Network slicing is the emerging paradigm in which operators use their resources to provide multiple logical networks and associated resources, with varying configurations and at the same time. More and more vertical industries need their machines and devices connected in networks with specific requirements. In order to provide networks fitted to these usecases, and not require that they adapt to the one-size-fits-all network as is currently the case with the mobile Internet, the telecom community vowed to include network slicing functionality within its next generation of mobile networking, 5G, as an end-to-end network solution. As the concept is new and still not fully grasped, we develop and refine the concept of a network slice both from a business and a technological point of view. We investigate how network slicing in the context of a vehicular network could be implemented and how it advances the state of the art. This involves a detailed study of the involved technologies across a range of infrastructures and network segments, as well as the resulting gaps in the existing technology landscape. Based on the lessons learned in this concrete usecase, network slicing is considered in a broader 5G landscape. We capture the main challenges and potential directions in order to make network slicing a true enabler of 5G-driven vertical industries.**

I. Introduction

The current generation of the mobile network doesn't offer much flexibility to vertical (i.e. independent) industries that are in need for a network to connect their devices. Most of the time, these industries adapt their connectivity requirements to fit the one-size-fits-all mobile network, often making them unviable. Network slicing is a new paradigm that aims to terminate this issue by providing logical networks tailored to the requirements of each usecase. Over the last couple of years, new concepts have changed the field of telecommunication networks. New virtualization techniques like network function virtualization (NFV), where network functions are implemented in software and deployed on-demand on general purpose hardware, have emerged. Software defined networking (SDN), in which SDN controllers program the routing tables of SDN switches, and radio access network (RAN) spectrum sharing have also been introduced. Building further on these new ideas, operators target to gain management and orchestration control over every part of the infrastructure. This control allows them to modify the behavior of the network and adapt it to satisfy specific requirements.

As it is a fairly new concept, we start with explaining network slicing from both a business and a technological point of view in section II. In the available literature (for example in [1]), multiple usecases as to why network slicing is needed can be found. As we feel that most of them lack detail, we use section III to describe a low-level usecase that cannot be realized with the current mobile network. We also design a network slice that could support the usecase in a 5G, the next generation of mobile Internet, slicing context. To provide slicing capabilities, an operator will need a variety of tools to control and manage its infrastructure. In section IV, we consider the enablers that will provide management and orchestration control over the different parts of the infrastructure, and the tools and technologies that are being developed in that aspect. To finish this article, section V concludes this work.

II. What is network slicing?

The term network slicing was introduced to describe the sharing of a telecom operator's resources between multiple logical networks. These logical networks are using the same physical resources, but serve different business usecases and therefore have to meet different requirements. Suppose that a client of a telecom operator is in need of a low-latency connection between its mobile user equipment that is located in geographical different places. An operator could provide this client with a spectrum segment of the RANs that the user equipment is connected to, and a segment of the bandwidth of the links in the core network that connect these RANs, all reserved for traffic originating from the client. This reservation of resources constitutes a slice of the operator's network. Operators can host multiple of those slices on their infrastructure. [1] states that network slices should be end-to-end, indicating that user equipment should be able to connect to them.

The slicing concept can be described by a 3-layered model [1]: an infrastructure, a control and a business layer. This model can be seen in figure 1. The infrastructure layer (IL) contains all the parts of the physical infrastructure of an operator, as slices should be end-to-end. Slices can be required to span multiple domains, so the IL contains the infrastructure of multiple operators. Among the infrastructure you can find RANs, computing and storage resources (i.e. datacenters), wired links connecting all the parts, etc.

The control layer (CL) resides on top of the IL. The CL can be seen as a library of modularized virtual network functions (VNF), and resources to run them on. These VNFs are used to manage and control the infrastructure to form a logical network that meets the network characteristics required by a slice request. Routing traffic along a path that is different from the path defined by the standard routing protocols requires functionality in the network that can control the routing tables. To reserve a segment of a RAN's spectrum, a VNF is needed that prioritizes the traffic near the basestation, so that the reservation requirements are met. The CL contains all the tools at an operator's disposal to control the behavior of the network. The CL also contains frameworks to manage and orchestrate these VNFs. These frameworks deploy the VNFs, they monitor the performance of the different segments of the network slice, they use this information to adapt the slice if performance requirements are not met by replacing VNFs or by changing traffic priorities, etc. These frameworks allow the maintenance

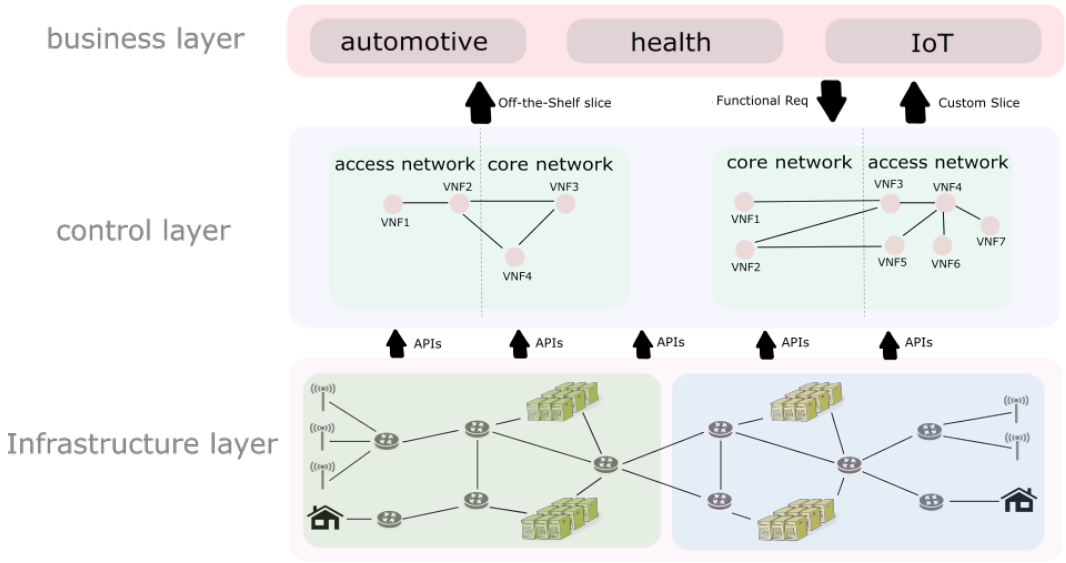

Fig. 1: 3-layered model for network slicing with infrastructure, control and business layer.

of slices to be automated, which is an important aspect for the viability of network slicing. The VNFs are stand-alone modules, meaning that 3rd parties are able to design their own VNFs to customize control mechanisms. Some frameworks, such as SONATA[1], provide a software development kit that contains tools for 3rd parties to develop VNFs. Through the framework, 3rd parties can request the deployment of these VNFs. Such mechanisms give slice customers, next to the functional requirements, additional tools to define their slice.

The business layer (BL) contains the services and usecases for which the network slices are created. They are provided by the operator or requested by 3rd parties. Figure 1 shows usecases in the BL for the automotive, health and the IoT sector. The BL contains catalogs with templates of slices that can be deployed by the operator, off-the-shelf. These are network slices for which the instructions for the CL are known and are ready to be deployed. The BL also needs a tool that translates the functional requirements of 3rd party slice requests into instructions for the control mechanisms. These instructions indicate which VNFs to deploy, which basestations to connect, what to do when monitored metrics reach thresholds, and which user equipment to allow connection to the slice. These tools ensure that slices can be deployed automatically.

The current mobile network contains few control tools and is not flexible. Therefore, multiple usecases cannot be deployed, as the network is not functioning in a way that supports them [1]. For this reason, telecom operators vowed to include network slicing capabilities in 5G. This incorporation should provide the much needed network flexibility that supports all possible usecases, including those that haven't been thought of yet. Network slices should be developed to provide a network specifically optimized for a usecase, instead of requiring usecases to adapt to networks. The performance of a network slice should be isolated, so it doesn't suffer from activity on other slices. This ensures that network slices appear to the connected devices as isolated physical networks.

[1]https://www.sonata-nfv.eu

The usecases for network slicing that have been described so far can be divided into three categories based on their requirements for the network. This division is shown in figure 2. The first category are enhanced mobile broadband (eMBB) usecases which require very high data rates and a high mobility, security and coverage. Typical examples are ultra high-definition video streaming to smart phones or tablets, tactile Internet and augmented reality. The second category of usecases is defined as massive machine-type communications (mMTC), in which a large number of devices transmits low volumes of data that is not sensitive to delay and usually does not involve mobile devices. Among their requirements is preventing that this traffic overloads the core network, which can be done by processing it before it reaches that core. This category contains the Internet of Things (IoT), which includes sensor networks that enable smart buildings and cities. The third category contains critical services which are characterized by ultra-low latency, high reliability and availability. Examples include remote surgery and communication networks for emergency services in case of disasters.

Different requirements need different new technologies. eMBB usecases require new radio access technologies such as Massive MIMO and mmWave to improve the data rates. They also need context awareness of the user equipment, such as battery information and location, for a personalized quality of service. mMTC would greatly benefit from device to device communication, allowing them to communicate with each other without sending traffic over the network and enabling them to send out grouped messages. Reducing the range of the basestation cells also serves mMTC by lowering the traffic per basestation. Critical services need wireless backhaul between the RAN and the core network, as well as meshed networks in the core, for reliable communication. The technical requirements for each category of usecases are shown in figure 2.

III. A LOW-LEVEL USECASE

Multiple work have described usecases for network slicing, but in our opinion, they mostly lack detail in their description

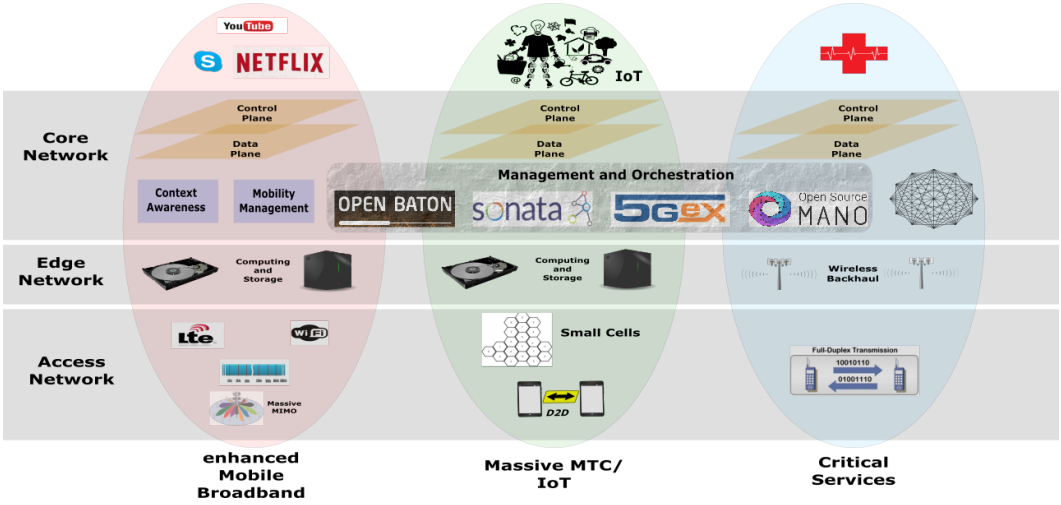

Fig. 2: usecase and technology model for network slicing.

and how they can be implemented in a 5G network which supports network slicing. Therefore, we use this work to describe a low-level usecase, point out why it is not feasible in the current mobile context and formulate what the slice should look like to implement this usecase in 5G.

The usecase finds its origin in the vehicular network domain. Inter-car communication can greatly increase the performance of self-driving cars. It enables them to distribute unexpected road conditions (e.g. an oil spot on the road, or light from the setting sun influencing camera images) towards upcoming cars. These cars can then make a more informed decision when passing that specific location. Due to their range limitations, ad hoc vehicular networks (VANET) are not the best fit for this usecase. VANETs are networks in which cars are directly connected with each other, without the help of any external infrastructure (e.g. basestations or routers). If a car notices unexpected conditions, but no other car is within VANET range, this information cannot be used by other cars. Since cellular networks have near unlimited range, they allow us to define a range that marks a territory around a car in which all other cars can be informed of the noticed road conditions.

We propose the following simple work flow. When a car notices a road condition that can be useful for other cars, it sends this information together with the GPS coordinates towards all the cars within a certain range R. Cars that are currently within this range receive this information and add it as a new entry in a list L. Cars manage this list by removing entries when the distance between the location of the entry and the location of their car exceeds R. When a car is approaching the location of one of the entries, the entry is forwarded to the auto pilot, which then anticipates to it. When a car notices that the road condition of an entry in L has changed, it informs the cars within range with an update of this entry.

In the current mobile Internet, a car that wants to distribute information to other cars needs to know their IP addresses. As the car is unaware of which cars are in its range, there is need for an external process P, a location analyzer, that keeps track of distances between the cars that are joined in the network. P sends the car the IPs of all the cars within its range, which can be seen on the left panel of figure 3. Once the car received these IPs, it can unicast the newly noticed road condition to them. None of this communication is constraint to the access network. Interaction with P is done through the core, as P is running in a datacenter or on private servers of the car manufacturer. The messages with road information are also passing through the core, even if the receiving cars are connected through the same basestation as the sending car. The evolved packet core (EPC), the mechanism that controls the routing in mobile access networks, checks the destination of uplink traffic in the PDN-gateway (PGW), which is an application that runs in the core. If the receiving car is using the same PGW as the sending car, this PGW forwards the message towards the basestation that the receiving car is connected to. If it is using a different PGW, it sends the message through the core to this PGW, to go down to the basestation from there. These processes are shown on the right panel of figure 3.

As every receiving car is contacted with an individual message, the bandwidth associated with this usecase scales quadratic with an increasing number of cars connected in the network. As all this traffic goes through the core network, the associated traffic load for the core network becomes too high if the number of cars increases too much. This makes the usecase unfeasible for a higher number of cars.

The unfeasibility of the usecase is merely a consequence of the fixed usage of the infrastructure. Since receiving cars are located in each others vicinity, they are connected to the network through the same or neighbouring basestations. Therefore, messages intended for each individual car travel almost identical paths. If we were able to broadcast a message from a basestation on a frequency that is known by the cars, it becomes possible to reach all the cars connected through the same basestation with one packet: an instruction to that basestation to broadcast a message on a specified frequency. Since the range R can span the territories of multiple basestations, each of them must be targeted with a message to reach all the cars that should receive the new road condition. To route the data packets towards the basestations, we make use

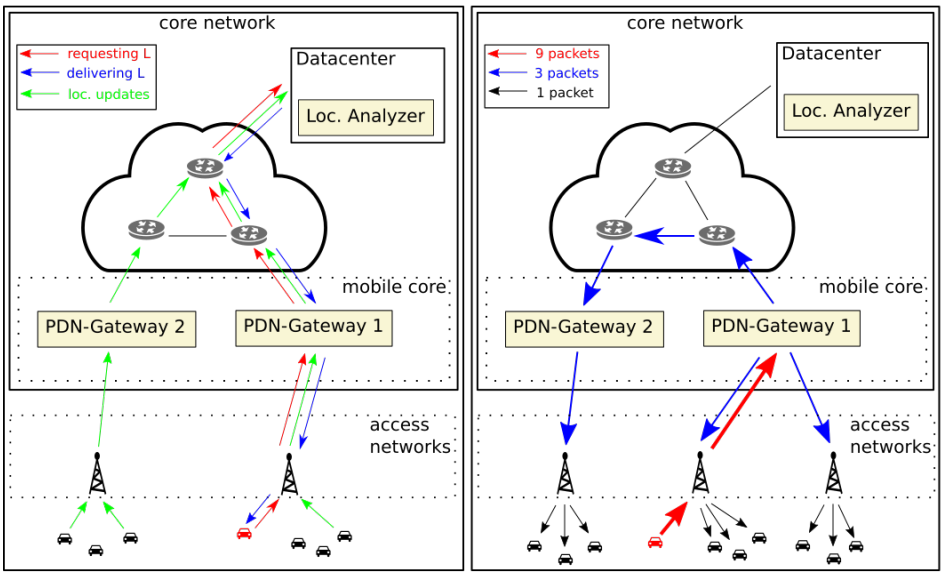

Fig. 3: Vehicular usecase without slicing. The left panel shows periodic location updates and a car requesting and receiving the IP addresses of the cars in range. The right panel shows the different unicasts of the sending car.

of SDN. A SDN controller is used to program the routing tables of SDN switches. By configuring this controller, one gains control over the paths that traffic going through the SDN switches is following.

By using SDN, the traffic reaches the basestations in the following way. The car that is sending out the information, sends out one packet. A SDN switch along the upstream path from the basestation picks up this packet. The car manufacturer has instructed the SDN controller to configure the routing tables of the SDN switches in such a way, so that when the SDN switch receives a packet from a sending car, the SDN switch forwards this packet towards the basestation of this car (to reach cars in the same basestation as the sending car) and towards all its neighbouring basestations. Once arrived near a basestation, this packet is decoded as a broadcast instruction, and the basestation broadcasts the information on the defined frequency. Figure 4 depicts this process. The uplink traffic is shown in red, the traffic after the message reaches the SDN switch is shown in blue. The control interfaces are shown in green: the car manufacturer can instruct the SDN controller to adapt the range and the SDN controller maps this on the routing table of the switch. On the right, the figure shows what network slicing provides to enable the usecase.

Each car has an IP address for all sorts of communication, but as this inter-car communication is not expecting a response to an outgoing message, it is not required to put the IP address in the outgoing packet. This freedom in setting the source IP address of the packets can be used to reduce the load of the SDN controller. If a predefined value is used, the SDN controller doesn't need to update the routing tables of the switches every time new cars enter the territory that communicates through this switch, as they just once add this predefined value that is equal for all the cars.

Although the traffic that goes through the SDN switches scales linearly with the amount of cars in the network and the load for the core is much lower than in the 4G scenario, it is possible to reduce this load even further. With the deployment of 5G, it is expected that the infrastructure of the operators will be extended with computing and storage resources in the edge of the network, to enable mobile edge computing (MEC). MEC has as purpose to keep traffic that is intended for access networks close to the originating access network out of the core, to further reduce latency and to lower the load in the core. Traffic that needs to be processed by an application can be processed on these resources, instead of sending it to the core. Our usecase is a good example for MEC. The traffic is intended for neighbouring basestations, but needs to be rerouted by an application (SDN switch) first. By embedding the SDN switches on these edge resources, all traffic remains in the access networks (except the communication with the SDN controller, which is running in the core).

After the introduction of the edge resources, the infrastructure to realize this usecase is available. What is missing are tools and the permission to control it. In 5G, we need tools to reserve bandwidth in the spectrum of basestations for broadcasting and tools that allow the deployment of applications on the computing resources. Once available, the slice that supports this usecase contains: i) a reserved part of the spectrum of the basestations, ii) reserved computing power on edge resources that is running SDN switches, iii) some SDN controllers that are running in datacenters in the core that are programming the SDN switches and iv) control tools that allow to manage the different parts of the slice (e.g. changing the range by instructing the SDN controller to update the forwarding tables, or changing the broadcast frequency).

A similar solution for this usecase was proposed by the Third Generation Partnership Project (3GPP), a collaboration between telecom associations and responsible for the description of the EPC. In [2], a study is made on the changes needed in LTE, a 3GPP radio access technology, to be able to reach all receiving cars with a broadcast message on the same and neighbouring basestations. They propose to place a

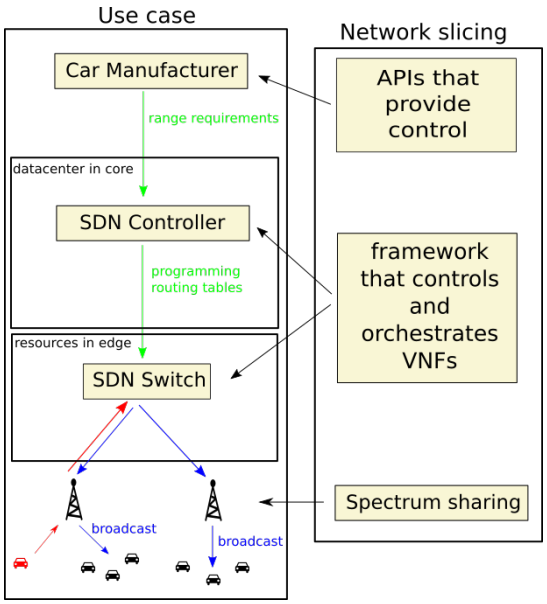

Fig. 4: Vehicular usecase in a network slicing context.

server near the basestation that is targeted by the sending car. This server sends the messages as a broadcast instruction to the basestations that are in range. The study concludes that with changes to the LTE infrastructure, the usecase can be supported. Although this solution is similar to our solution, there are some major differences. Our solution is radio access technology independent, as long as the technology supports broadcasting, where as the 3GPP solution is LTE specific. Where the 3GPP server is dedicated for this usecase, the SDN switches in our usecase can be shared among multiple slices.

IV. Slice enabling mechanisms

since a network slice is an end-to-end service, each part of the infrastructure should support it. We divide the infrastructure in user equipment, access networks and a core network. So far, no technology has emerged that allows to control the entire infrastructure. This section describes mechanisms en techniques that enable slicing in the different parts of the infrastructure. Table I gives an overview of challenges and the progress that is made for slicing support by the infrastructure.

A. User Equipment

There is need for a mechanism that provides user equipment with information on the available slices and that allows them to request a connection with one. In [3], the authors propose a device triggered network controlled (DTNC) slice selection mechanism for 5G user equipment. In this mechanism, access points of RANs broadcast system information (SI) containing information on all the slices that are accessible through this access point. 5G devices, when triggered, scan for this broadcast and interpret the SI to determine which slice is most fitting for its purpose. Once a slice has been chosen, the device makes a request to the access point to connect to this slice, to which the access point will respond based on the device's credentials, resource availability or a different criterion. This mechanism can be used by the usecase proposed in section III. When a car is started, it can connect to the slice by processing the SI that is send out by the basestation and request access to it.

B. Core Network

The core network is the central part of the infrastructure that connects all the access networks together and contains the datacenters. These datacenters contain commodity hardware that provides computing, storage and memory resources. Using these resources for network slicing can be done by leveraging NFV. An NFV management and orchestration (MANO) framework, as described in [12], is able to deploy VNFs on these resources upon request. The framework can manage the life cycles of these VNFs based on general or specific policies, such as monitoring inputs. Examples of frameworks that are being developed are Open Baton[2], SONATA and OSM[3].

A second technology that can be leveraged in close collaboration with NFV is SDN, which was already introduced in section III. SDN offers flexibility in how traffic is routed through a network. SDN can be used to steer traffic between VNFs. Motives to customize the traffic pattern are reducing the load in the network, prioritizing certain streams, etc.

As mentioned before in section III, the concept of MEC [13] raised the question of providing computing, storage and memory resources in the edge of the network, closer to the user equipment. These resources can unburden the traffic load of the core network, which is expected to increase with a factor of 8 between 2014 and 2020 [14], and they can reduce the latency. MEC enables network slicing by providing additional locations to host VNFs. Recently, the European Telecommunication Standards Institute (ETSI) released requirements and characteristics of a MEC facility [13]. Nokia Networks introduced a MEC platform [4] and an NFV/SDN based MEC platform, called WiCloud, has also been developed [4].

For improved slicing support, the control architecture should be modular and flexible so that it can support a wide variety of usecases. To achieve this, larger VNFs should be decomposed into atomic VNFs. This modularity improves the innovation process and gives operators more options in choosing which atomic VNFs to use, in order to construct the overall service. The network slice is then created by chaining these functions together, so that the service requirements are satisfied [10].

C. Access Network

When it comes to the access networks, the biggest challenge for network slicing is the sharing of RAN resources. Traditional RAN sharing methods were all static, meaning that resources are being reserved whether they are used or not. 3GPP has defined three categories of RAN sharing [6]: i) Only the basestation equipment (i.e. antennas) is shared, ii) both the equipment and the spectrum are shared, and iii) next to the RAN resources, also core network elements like the PGW are shared.

To further improve the efficiency of resource usage, RAN sharing should become dynamic, so no resources are blocked

[2]https://openbaton.github.io
[3]https://osm.etsi.org

TABLE I: Challenges, accomplishments and future work for network slicing.

Challenge for network slicing	Work achieved or under development	Future work
Slice selection and attachment procedure for user equipment	DTNC slice selection protocol [3]	User equipment should be able to be connected to multiple slices at the same time
On-demand slice creation	MANO frameworks that automatically deploy NFV services are being developed	A platform that combines NFV MANO with aspects like spectrum reservation to build the entire slice
Latency and QoS improvements	Nokia Networks MEC platform and WiCloud [4]	Mapping MEC resources on the ETSI standards
Integrating multiple radio access technologies such as Massive MIMO	Light Radio from Alcatel Lucent, Antenna-integrated Radio from Ericsson, FluidNet from NEC and CloudIQ Framework [5]	Defining a unified interface for various radio access technologies, integrating their processing resources, joint resource allocation, ...
Resource sharing in RAN	1) Static sharing based on fixed contractual sharing agreements [6]	1) Enable feedback on spectrum usage from basestations and user equipment towards control layer
	2) Dynamic sharing [7]: software defined RANs like SoftRAN [8] and CloudRAN [5]	2) Ensuring that the delay caused by this feedback and the response to it is not significant
Slice isolation and security	NGMN recommendations [9] on security tackle the main issues by adding several levels of isolation between slices	Translate these recommendations into design patterns for slices in a multi-slice environment
Architecture modularization and decomposition	Work on function decomposition of the core network at DocomoLab [10], service function chaining in NFV	Continued process of identifying new VNFs to support new services
Sharing context information between slices	Research performed into context aware resource allocation in RANs [11]	Identify usable info from user equipment, RAN, core network and applications to create a richer context

by a reservation while not being used. For this, SDN and NFV can be leveraged to introduce software defined RANs. Examples are SoftRAN [8] and CloudRAN [5]. They allow control of spectrums through flexible allocation and scheduling. As these architectures will be key to enabling proper radio resource management within slices, 3GPP started in 2016 with 5G RAN slicing research projects [15].

The vehicular usecase can benefit from shared RAN resources by requesting a part of the spectrum of the basestations for its broadcasting.

D. Other Factors

Slicing introduces new security concerns, as sharing physical infrastructure adds new vulnerabilities. This led to the consensus that network slices should be isolated [1]. The isolation refers to three aspects [9]: i) slice A should not be influenced by slice B when slice B is exhausting its provisioned resources, ii) direct communication between slices should not be allowed to prevent eavesdropping and iii) mechanisms should be in place to prevent 'hacking through the walls' of network slices by malicious intent. [9] contains a list of possible security risks and recommendations to deal with them.

A 5G system should adapt to context information, such as battery information of user equipment, its location, the state of the network (traffic load and congestion) and application usage patterns of users. As knowledge of this information can hugely benefit different slices, a framework that allows slices to exchange this info in a secure way would have great value.

V. Conclusion

Vertical industries are in need for a unique combination of network and compute resources across different network segments to support their usecases, a requirement that operators currently are unable to provide. The concept of network slicing, which is clarified in this article from both business and technological points of view, offers new perspectives for this problem, drawing huge interest of operators who are currently trying to prepare their infrastructure for it. As we have demonstrated, many network slicing enablers exist, and the technologies to allow operators to fully exploit these enablers are in the process of being developed. Once all parts of the infrastructure can be managed by one platform, we can expect the roll-out of network slicing, and we expect it to greatly impact the telecommunication community.

Acknowledgement

This work was performed in the framework of the SONATA and the 5G CHAMPION project, both funded by the EC in the scope of the Horizon 2020 and 5G-PPP programs.

References

[1] NGMN Alliance, "NGMN 5G White Paper," Tech. Rep., Feb. 2015.
[2] 3GPP TR 36.885, "Study on LTE-based V2X Services," Tech. Rep. v14.0.0, June 2016.
[3] X. An *et al.*, "On End to End Network Slicing for 5G Communication Systems," *Transactions on ETT*, June 2016.
[4] H. Li *et al.*, "Mobile Edge Computing: Progress and Challenges," in *Proc. 4th IEEE MobileCloud*, Apr. 2016, pp. 83–84.
[5] R. Wang, H. Hu, and X. Yang, "Potentials and Challenges of C-RAN Supporting Multi-RATs Toward 5G Mobile Networks," *IEEE Access*, pp. 1187–1195, Oct. 2014.
[6] 3GPP TS 23.251, "Network Sharing; Architecture and Functional Description," Tech. Rep. v11.3.0, Sept. 2012.
[7] X. Costa-Pérez *et al.*, "Radio access network virtualization for future mobile carrier networks," *IEEE Commun. Mag.*, July 2013.
[8] A. Gudipati *et al.*, "SoftRAN: Software defined radio access network," in *2nd ACM SIGCOMM workshop on Hot topics in SDN*, Aug. 2013.
[9] NGMN Alliance, "5G security recommendations Package 2: Network Slicing," Apr. 2016.
[10] M. R. Sama *et al.*, "Reshaping the Mobile core network via function decomposition and network slicing for the 5G era," in *Proc. IEEE WCNCW*, Apr. 2016, pp. 90–96.
[11] H. Ghouma and M. Jaseemuddin, "Context aware resource allocation and scheduling for mobile cloud," in *IEEE CloudNet*, Oct 2015, pp. 67–70.
[12] ETSI, "Network Functions Virtualisation - White Paper #3," Oct. 2014.
[13] ETSI, "Mobile Edge Computing: A Key Technology Towards 5G - White Paper," Nov. 2015.
[14] P. Cerwall *et al.*, "Ericsson Mobility Report - White Paper," June 2012.
[15] I. da Silva *et al.*, "Impact of Network Slicing on 5G Radio Access Networks," in *Proc. IEEE EUCNC*, June 2016.