Integration and virtualization of new computing architectures for next generation distributed mobile networks

Future generation of mobile network will require to engineer new cutting edge and fine-tuned computing infrastructures to achieve at the same time the need for increased performance in term of data-throughput as well as reduced latency while becoming as efficient as possible in their usage of computing resources and energy. Such computing architecture should leverage the potential of the more advanced computation devices such as DPUs, high performance RISC processors (ARM & RISC-V), or purpose-built accelerators based on FPGAs or even ASICs. At the same time, future generations of mobile networks should feature enhanced flexibility and multi-tenancy as well as disaggregation. All those requirements will make the operation of the underlying computing architecture complex. Heterogeneous computing resources would be scattered asymmetrically across multiple sites but should still be shared among multiple network tenants and adapt to the workload. The state-of-the-art method to operate shared computing infrastructure relies on virtualization. Virtualization technologies such as the one used in cloud infrastructures can abstract the infrastructure and its shared nature to the users while operating the infrastructure in an efficient manner. But state-of-the-art virtualization technologies may not be yet able to operate the infrastructures of future mobile network technologies. There are still some missing features that need to be integrated and tested. As of today, some new kinds of computing devices such as DPUs or purpose-built accelerators are not commonly supported within virtualization technologies. On the other hand, such devices do not easily adapt to the current capability of virtualization technologies for implementing mobile network functions. Virtualization usually relies on containers to ship and run a workload to a device. This is fine if the device is a general-purpose machine with eventually a few embedded accelerators. But this is not easy to map to a standalone accelerator on a network like a DPU. There is a real need to figure out how to properly use these accelerators within a cloud-type infrastructure. Then there is also a need to know if these accelerators can be shared among multiple tasks and users and how. The ability of orchestrators to operate on complex and distributed architectures may also need to be studied and improved. Cloud infrastructures are often large homogeneous pools of resources centralized in data centers. This uniformity and localization of resources makes the efficient usage of resources easier to achieve than on mobile networks. An efficient implementation of a mobile network - even if it is shared between tenants – would more likely consist in a constellation of cellsites, edge sites and data centers with heterogeneous resources and backhaul connection statically assigned by the network operators avoiding overprovisioning. The orchestrator should be able to satisfy the performance requirements set by the user with this difficult architecture. This is a complex issue within which it is possible to look at the problem of managing heterogeneous resources efficiently. Having a pool of resources with many different types of computing devices and accelerators including general-purpose and purpose-built ones means that there are many ways to achieve the same process. This brings the question of the

optimal way to execute all the tasks that come to the infrastructure. This thesis will focus on the integration of most recent computing devices in distributed multi-tenant mobile networks. It will figure out how to integrate such devices in cloud-type infrastructure. It will also investigate ways to orchestrate a variety of devices while respecting optimal implementation schemes. The developed methods will then be benchmarked regarding the state-of-the-art technology at the time of release. The purpose of this work is to deliver a sample alternative virtualization technology that could address some of the many challenges in operating a distributed multi-tenant mobile network. It could then be reused as a previous experience for future generations of mobile networks that will have to address this challenge.