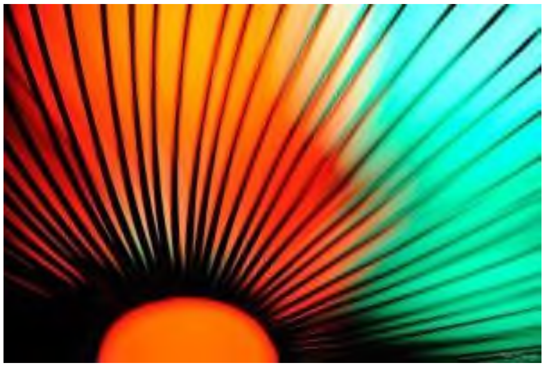
Special Report: The In’s and Out’s of Network Slicing

By Ernest Worthman

October 4, 2017

With the anticipated ramp up of mobile data, new approaches to bandwidth efficiency are needed. Network slicing promises to help ease that congestion



Network slicing is a theory designed to offer network services across a wide and varied application base. It is a form of virtual network architecture utilizing similar principles as are found in software-defined networking (SDN) and network function virtualization (NFV), in fixed networks.

Because the primary objective 5G is to be an umbrella network under which any number of dissimilar wireless platforms and technologies coexist and assimilate, the theory of creating virtual environments for these platforms, where they can coexist as unique systems, is quintessential. 5G is being touted as the great enabler for a completely seamless and interoperable global network regardless of technology or platform.

A rather tall order. We are quite a long way from realizing its full potential, but there are innovations and evolutions coming down the channel that promise to get some of that vision going. One of those is network slicing.

The technology is definitely available. So is the hardware and software. The computer industry has been using virtualization for quite some time. There is no reason the same principles cannot be applied to wireless networks.

Network Slicing 101

A very high-level explanation of network slicing is that it is similar to server farms in that you have the core hardware upon which virtual elements are carved out. With the cloud, for example, there is the hardware (server farms) within which any number of virtualized systems can be deployed, and customized to the demand. In addition, they need not have common denominators.

Some can be based on Windows, others UNIX- LINUX, or some other operating system. They all run concurrently on the base hardware as instantized machines.

What makes this so attractive is that server, or network resources, can be allocated dynamically. So as the user makeup shifts, so do the instantized machines.

The same concept is used in network slicing (also called network scaling). In this case, the servers (or chips, as the case may be) are used to virtualize parallel network functions for wireless networks. This platform can host multiple carriers, multiple bands and any protocol on LTE networks, for example. There are other use cases – unlicensed is one, so is the Internet of Everything/Everyone (IoX), which, provides two prime examples; massive machine-type communications (mMTC) and critical machine-type communications (cMTC). The IoX has other use cases, as well, but many scenarios can fall under these two criteria.

In these cases, the former is large-scale systems that demand low throughput and can tolerate latency. The latter is just the opposite; it requires low latency, high throughput and ultra-high reliability. Obviously, it would be difficult to have both cases in the same band.

A good pictograph of how network slicing would look in a 5G world is shown in Figure 1. These use cases cover a number of dissimilar technologies that would kill a single network with common bandwidth. And there are more use cases as well, millimeter wave, incumbent users, public safety, location services, satellites, enhanced mobile broadband (eMBB), and the list goes on.

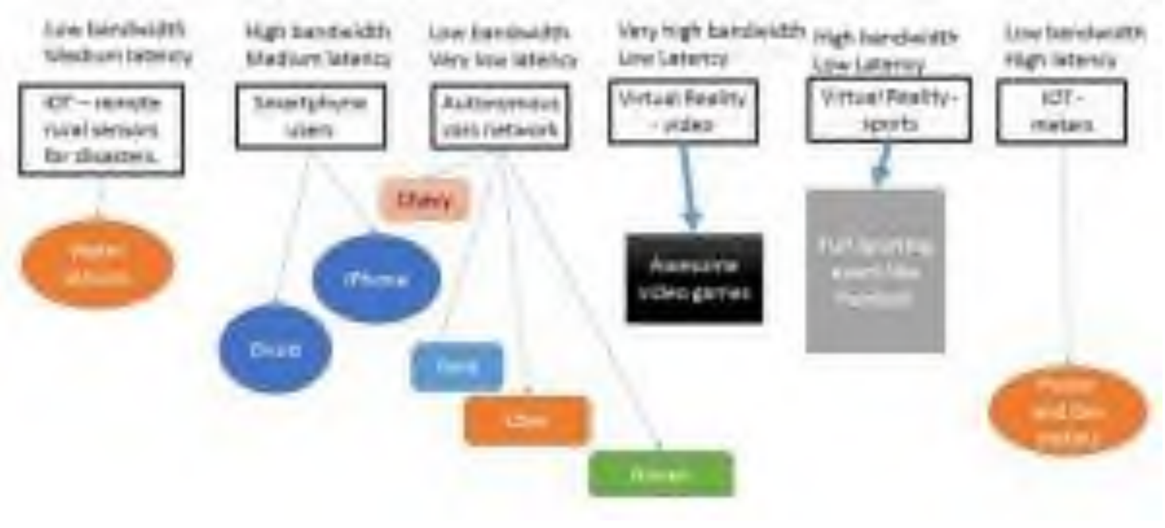


Figure 1. Network slicing use cases. Source: wade4wireless.com

Network slicing 201

Network slices are deployed Concurrently, as logical, independent, and self-contained, multiple partitioned network functions and resources on a common physical infrastructure. They are specifically configured to support the requirements of a particular use case. They are created in software and use common resources such as memory, storage and processors.

However, network slicing introduces an additional layer of abstraction by the creation of logically or physically isolated groups of network resources and network/virtual network function configurations separating its behavior from the underlying physical networks. It will require autonomous and autonomic systems (AI, anyone?) to organically grow or contract resources as the demand fluctuates.

How this is accomplished is by separating the control plane from the user plane. This facilitates moving servers and RANs to the edge (where there exists a centralized pool of virtual networks), while the core maintains control over certain functions such as virtual resource management, mobility management, interference management and other global functions. It also reduces control plane signaling and improves data throughput. [1]

The edge now is responsible for local data management, store-and-forward and certain control plane functions (baseband processing, for example). User plane functions — such as mobility management entity, policy charging rules function and packet/service-gateway — are also shifted to the edge. This reduces the load on backhaul. Once this is configured, network slicing can be implemented.

What makes this so elegant is that each use can have a set of optimized resources, specific to the demands of the application. It will provide the resources and network topology for the specific service and traffic that will use the slice.

Typically, functions that will be optimized for a network include speed, capacity, connectivity and coverage. Such functions can also be optimized for time, such as rush hours or sporting events.

Simply put, virtualization technology is kind of like a house with no physical components. Walls are dynamically placed as are room contents depending upon what the function of the room is and who lives in the room. This determines how big the room needs to be. As room capacities vary, so do the contents. And each room will be constrained so that no one in another room can interfere with the functionality (in this case wireless traffic) in another room.

Use cases

One of the hot topics of 5G is eMBB, which requires a large slice of the network to ensure the high-data rates for apps such as HD video and augmented reality. Because of the high bandwidth and high buffering requirements, this particular slice needs a larger share of resources such as caching.

Another example is autonomous vehicles. This scenario demands ultra-reliable, low-latency communication and security. By instantizing these functions in the network slice at the edge, the criterion can be met.

Finally, a third, and very typical, example is the IoX, which is a classic example of a mixed-use demand environment. There will be countless applications, with many varied bandwidth demand slices. That isn't to say each case will require a unique slice. Rather, many apps will fall into one type of bandwidth slice or another. By grouping similar demands into a network slice, each demand group can be optimized for its particular requirements (static or dynamic machines, sensors, security monitoring, smart homes/appliances, etc.).

Network Management

For the network to function efficiently there must be reliable communication between the control parts, which is accomplished through controllers or some other type of interface. There is a network function manager that is responsible for mapping and allocating the physical network resources to the virtual machines. This is done by the software-defined networks (SDN) controller in coordination with the virtualized network function manager (VNFM). By interconnecting the data layer and the vertical application layer via pre-defined interface protocols, it runs the complete virtual network.

This is called virtualized infrastructure management (VIM) and, essentially, allocates resources to the VMs and monitors resource utilization. VIM is an integral function of the slicing management because it is responsible for dynamic slice management, that is creating, activating or deleting network slices according to the service requirements.

A Drill Down on Mobility Management

eMBB was touched on briefly earlier. However, one of the most demanding use cases for network slicing will be mobility management. Radio access technology (RAT) will be radically different in 5G networks. This is why network slicing is so critical to 5G communications.

The practical way of doing that will be to separate the control and user planes in the core network, which has the effect of reducing the control signaling, one of the high overhead functions. However, it is not yet clear how all this will shake out, due to the uber-densification and ultra-mobility of 5G networks. Currently, research is being conducted on new techniques for mobility management.

To peel back some of the issues, take, for example, two separate scenarios on rather opposite ends, one being railway mobility. By comparison, there will be far fewer instances of railway communications than there will be for the IoX. Communications for moving trains are much easier to model than the diversity of devices on the IoX (latency/bandwidth/platform/application requirement, for example). Therefore, railway communication network slices are much less complex that those for the Internet of Everything/Everyone (IoX), where multiple requirements exist and are all over the map.

However, in all cases, one of the biggest challenges in mobility management is handovers. Traditionally, handovers are event-triggered and controlled by the base station. This will not work in 5G scenarios. The reason for that is beyond the scope of this article due to the complexities of the demands in the different slices.

A quick flyover of one possible scenario is to move or create an SDN in the RAN itself. This allows for the creation of software-defined wireless networks (SDWNs). This approach calls for a hierarchical control plane to be deployed close to the edge. This can enable cooperation between controllers and RATs. This is possible because of the implementation of virtual servers using IP protocols rather than traditional. A much more complete dissertation on this can be found in the reference cited below.

There are many more challenges that will have to be considered to make true, unrestricted and reliable mobility management a reality. Many are working on that; however, it will take some time. Do not expect this to be operational, at least not beyond fundamental functionality for several years, and even longer for mmWaves.

Conclusion

Network slicing is a very practical way to address the diversity of dissimilar network requirements that the 5G wireless world will bring. There is a lot of buzz around it and for good reasons. However, it is still in the development stages and there is no real consensus as to how to implement it.

There are lots of good theories, however, and several organizations, such as the GSMA, TM Forum, NGMN and MEF are all moving to try to find stasis on network slicing.

As well, there are still many challenges. One of the most pressing is getting ubiquitous edge networks. Without them, the real-time analytics, bandwidth and latency issues for emerging applications will gridlock the current wireless infrastructure.

Still, once the bugs are ironed out, platforms are standardized and everybody is on board with the technology, it promises to be one of the great enablers of 5G.

---

[1] Haijun Zhang, Member, IEEE, Na Liu, Xiaoli Chu, Senior Member, IEEE, Keping Long, Senior Member, IEEE, Abdol-Hamid Aghvami, Fellow, IEEE, and Victor C.M. Leung, Fellow, IEEE Network Slicing Based 5G and Future Mobile Networks: Mobility, Resource Management, and Challenges