Pipeline
Technology for Service Providers.
www.pipelinepub.com Volume 15, Issue 3

Orchestration Imperative for 5G Slicing

By: Mark Cummings, Ph.D., Christos Kolias, Vinay Devadatta

Talk about 5G tends to center on the upsides—the new horizons, potential, and use cases that this next-generation network will make possible. While full of promise, that dialogue skirts the cold reality that 5G deployment will be very expensive. And that expense can only be justified if 5G fully meets its promise.

Key to realizing that promise is lowering the capital burdens CSPs must shoulder. This can best be accomplished by network slicing, which is sharing resources while improving products that allow for the construction of innovative end-to-end services from atomic units of resource that may transit multiple CSP domains. To successfully accomplish this slicing, CSPs must have distributed orchestrators that can negotiate with other orchestrators both inside and outside an individual CSP's domain.

Early studies show that in order to actually deliver these promises, a complete deployment of 5G will require massive capital expenditures that a single operator may not be able to afford with current business or service structures. Effective network slicing can allow CSPs to share the expense of 5G network components while at the same time creating the ability to offer innovative new services that increase revenues.

Before diving into the details, let's talk definitions. For the purposes of this article, network slicing is defined (consistent with 3GPP TR 21.905's) as the process of presenting a set of network functions and the resources (eg. Basestations, wireline, fiber, satellite, switches or routers, ePC, CPU, storage, and so forth) that can be arranged and configured to form a logical network and/or service. Each network slice is independent and isolated from other slices, but it still runs on the same, shared infrastructure. Each slice is implemented on an end-to-end basis and can be dynamically created and discontinued.

Making this model work requires cooperation between divisions inside a CSP, between CSPs, and between CSPs and other service providers. With the growth of SDN, NFV, and other software-based technologies, orchestrators controlling segments of infrastructure are beginning to appear. Slicing requires a different kind of orchestrator. These slicing orchestrators must be able to function in a distributed fashion and negotiate with other orchestrators that are owned and operated by different entities, as illustrated in Figure 1.

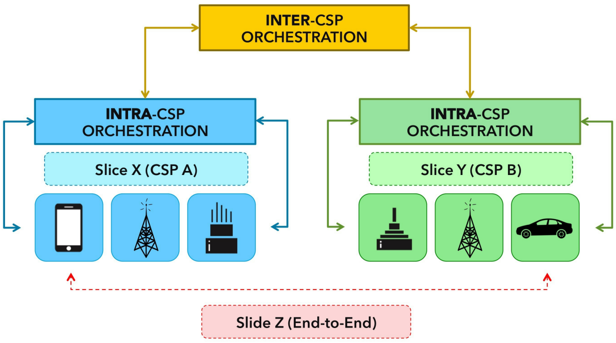

Figure 1: Generic Multi Administrative Unit Multi CSP Slice Example
(click to enlarge)

Entities may be different administrative units within a single CSP (intra-CSP), or may be in different

CSPs (inter-CSP) For example, Figure 1 shows a combination of both intra- and inter-CSP slicing.

End-to-End (E2E) connectivity for all possible human and Machine-to-Machine (M2M) requirements means that, in many cases, multiple CSPs will have to cooperate in forming slices across their domains and technologies. The E2E connectivity also means that communication not only span across technologies (satellite communications, traditional wireline, wireless, and more) but also CSP boundaries and geopolitical, regulatory and cultural boundaries. Because of its complexity, handling orchestration across dissimilar regulatory and political boundaries will be discussed in a subsequent article.

Let's look at a few use cases to illustrate the business and technical implications of 5G slicing. The low latency and high bandwidth of 5G have been targeted to, among other potential uses, the connected autonomous car. In one use case, two CSPs offering services in the same area face the expense of deploying 5G basestations to serve a 500-kilometer freeway covering flat rural land linking two metro areas. In 4G, basestations could be 50 kilometers apart, meaning there is a requirement of ten basestations to cover the road. But in 5G, because of the higher RF frequency and lower latency, basestations might have to be 1 kilometer apart, meaning 500 basestations to cover the area. This means not just more basestations, but more backhaul and more ePC. With fifty times as many basestations, the cost projections are daunting. So, the two CSPs decide to deploy one system and share the costs. There are many cost-sharing scenarios, from sharing just the basestations to sharing everything including backhaul and ePC. Although it is likely that each physical component will be operated by one entity, it is possible that different segments such as basestation and backhaul will be operated by different entities—for example, wireless and wireline groups in a single CSP, in different CSPs, and so forth.

Another use case example involves the connected autonomous car. A model of the Audi A8 has been introduced with a limited autonomous driving capability. Audi has stated that it will assume all financial liability for any accident that occurs while the car is in autonomous mode. This means that Audi will want to monitor the cars to make sure that there is no problem resulting from a bug, cybersecurity breach, or other outside interference. Because of the potential liability, Audi will want a very low latency connection to all equipped A8s in the world. For example, the German monitoring system will want low latency connectivity from Germany to an A8 on US highways. This scenario will likely involve a number of CSPs: at least one in Germany, another over the Atlantic and, because of coverage gaps, more than one in the US.

A third use case involves a merchant bank in New York that sees a short-term opportunity to do currency arbitrage transactions between New York and Tokyo. While there is a dedicated fiber link built just for this purpose, it is fully occupied by long-term leases. So, the bank purchases a temporary low latency and very expensive service for the transactions and a less expensive longer latency service for back office info. It makes this purchase a CSP, which we will call CSP A. CSP A has to quickly assemble both services. To get immediate connectivity to the bank, fixed 5G is used. For the transaction service, a transcontinental fiber line from CSP B is connected to a transpacific fiber link from CSP C. For the longer latency service at lower cost, a wireline connection to a satellite dish in northern New Jersey connecting to a dish outside Tokyo and then wireline to the Tokyo exchange is used.

In all three of these use cases, Service Level Agreements (SLAs) will be necessary. What is interesting here is how the SLAs are implemented and the implications for orchestration. In the past, we have thought of SLAs as static, long-term and single-targeted. In the world of 5G slicing, they will be dynamic, short-lived, and many-pronged. In the banking example, the two services may be created, used, and retired within a single day. There are also two different SLAs with the bank for the two purchased services, and those SLAs propagate between and within a number of different units in the customer-facing CSP, and between the customer-facing CSP and the other CSPs providing portions of the end-to-end service. A similar situation is apparent in the Audi use case.

At first glance, the highway use case may appear different. But there are likely to be more complex versions of it. For example, CSP A may deploy enough Owned and Operated (O&O) capacity to support users when communications traffic flow is moderate and only acquire shared slice capacity from CSP B during rush hours. This would be dynamic but within fixed time constraints. Another option might be CSP A acquiring an option to purchase slice capacity from CSP B whenever CSP A faced capacity or performance problems. Carrying this scenario a little further, in areas where there are 3 (China) or 4 (US) major wireless CSPs, CSP A might have options to acquire slices from CSPs B, C, and D on an “as available basis”—that is, if the other CSPs have available capacity. There could also be a sliding pricing scale: guaranteed-slice capacity at a high price, high-priority slice capacity at a medium price, and as-available slice capacity at a low price. In fact, all three of the use cases could have a range of SLA terms and pricing, and the range of variable parameters could be quite large.

Given these kinds of business arrangements a particular set of orchestration capabilities are required. In the simplest implementation, at each segment juncture (between administrative units, CSPs, etc.) there has to be an orchestrator at each side of the juncture that knows and understands its local resource situation and the range of business models it is authorized to offer. These distributed orchestrators must be able to negotiate based on their current resource situations and their available business models. Once an SLA contract has been established, orchestrators must have the ability to monitor ongoing operations to determine whether the SLA has been met. If it is not met, orchestrators must have an ability to alert, escalate, and resolve. Of course, inside each segment there has to be orchestrator(s) capable of insuring that the SLA the juncture orchestrator has contracted to will be met.

At each juncture, a CSP orchestrator may have a number of choices of orchestrators (and their associated resources) to satisfy its objectives. Thus, it has to be able to simultaneously inquire of each of the other orchestrators as to their current resource and business model status. Then, the orchestrator must choose and negotiate with them to create a contract with one.

It is possible that more complex implementations will emerge where SLAs are daisy-chained. In daisy-chaining, CSP A's orchestrator may pass on the entire customer-facing SLA minus the portion satisfied locally by itself. Then, the next segment orchestrator passes on the SLA minus the portion satisfied by it, and so forth. Other structures may also appear.

Because of the expense of 5G deployment, effective slicing will be critical to success—both to lower costs and to support the creation of new and innovative service revenues. We have seen how this requires orchestration across administrative and ownership boundaries that can only be achieved with distributed orchestrators that can negotiate across these boundaries. In addition to the business and technical challenges, there may be regulatory, political and cultural issues at some of the slicing junctures.

Please note: the views expressed in this article are those of the authors and do not necessarily reflect the views of the organizations they represent.