Network Slicing as an Ad-Hoc Service: Opportunities and Challenges in Enabling User-Driven Resource Management in 5G

Madhumitha Harishankar, Patrick Tague, Carlee Joe-Wong

Abstract— **Creation of virtualized network instances, aka network slicing, has been one of the fundamental architectural paradigms that allow 5G to meet the widely diverse service requirements of IoT devices, autonomous vehicles, and mobile apps alike. Recent work has shown the benefits in dynamic rather than static allocation of physical resources to these virtual slices as well as the feasibility of creating and enforcing slices dynamically even in the radio access network. Encouraged by this, we make a case for offering network slicing as a real-time service to end users. This is a significant shift from the predominant business model wherein content providers own these slices and end-users are subject to the resource allocation policy of content providers. By instead enabling users to customize and acquire resources for their desired session performance spontaneously, real-time network virtualization as a service allows diverse applications to drive their network allocation. This is particularly useful for real-time applications with immediate and potentially time-varying resource needs such as cyber-physical systems and edge computing. The operator's centralized control of resources is reduced and thereby also the burden of meeting sufficient traffic demand in a slice under sparse availability (esp. in the edge). Instead, this burden is shifted to the design of incentive schemes that allocate these ad-hoc dynamic virtual slices of limited physical resources to users/devices/applications that value them the most. While enabling diverse service requirements in a decentralized fashion, our model brings up new challenges and opportunities in designing mechanisms that capture the spontaneous wireless end device's session needs and valuations.**

I. Network Slicing - Current Model and Concerns

One of 5G's key architectural innovations, network slicing, results from the aggressive network function virtualization that enables complete programmability of network components [1]. Even radio resource allocation and scheduling policies, previously coupled with the physical hardware in the base station, are virtualized and hence susceptible to real-time programmability. As a result, the network may be divided into virtual slices that potentially span resources all the way from the edge to the core of the network and are dedicated to satisfying demands of a specific service level. This enables 5G's vision of supporting the highly diversified network needs of existing and emerging applications like cyber-physical systems (CPS). Machine to machine scenarios (e.g., tactile internet [2] or telepresence [3]) that require low latency, high bandwidth and high reliability simultaneously are integral CPS use-cases. So are, for example, smart-city scenarios [4] that require periodic transmission of IoT data to the cloud and low-latency computing resources at the edge for making actuation decisions in real-time. Simultaneously, the end users' diversity in network requirements also grows, as newer apps such as Pokemon Go become popular.

Virtual network slices catering to different service level agreements (SLAs) emerge as the solution. The network enforces these SLAs by allocating sufficient physical resources to a slice to appropriately satisfy the traffic demand[5]. While slicing techniques have been explored extensively (in the mobile core [6], [7], [8] and to some level in the radio access network(RAN) [9], [10]), the *usage models* have not. Simply put, how should end users/devices/applications (collectively referred to as *edge entities* going forward) attempt to acquire guaranteed service for themselves? A B2B (business-to-business) model is predominantly assumed today, wherein content providers work with the operators to define and reserve slices specific to the resource needs of their content, and thereby influence edge entities' quality of experience (QoE) for their applications[11], [12]. However, this model has limiting consequences for *slice usability and utility for edge entities* as well as *slicing efficiency and operating costs*.

Usability concerns: Offering slices as long-term contracts (whether spanning hours or months) to internet content providers or other types of service owners retains the model of centralized resource control and renders a large part of the diverse requirements and use cases of edge entities unfulfilled. Previously at the mercy of the resource allocation policy of a centralized network operator, now edge entities and their network performance rely on the resource allocation policy of the corresponding content-provider/service-owner. This can be severely limiting. First, the edge entity cannot acquire any SLAs for services it values if the service owners have not acquired a dedicated slice from the network, potentially also harboring net-neutrality issues. For example, a user uploading a business-critical document to a server within a tight deadline foreseeably does not have a specific content provider that can allocate a guaranteed slice to complete the task in time. Similarly, in emerging deployments of dense sensors in smart buildings and offices, the sensing systems might have different services to send their periodic data to[4] and may not be able to fully anticipate their future network needs that depend on these services.

Even in the few slicing architectures that account for a B2C (business-to-customer) model alongside B2B, static contracts between the customer and the operator for slices are assumed[11]. Alternatively, content-provider sponsored slices are assumed to exist in which case the architecture tackles the problem of slice discovery and association[13]. This system is hence of *limited use* to a large bulk of the heterogeneous edge entities to whom guaranteed SLAs would be expected add significant value. Due to reliance on service owners to procure and offer these, or the infeasibility of accurately forecasting their changing resource needs that may be ad-hoc and temporal[14], they realize limited utility from slicing. The failure of these relatively static B2C models to leverage the full power of the network's dynamic virtualization capability is alluded to by Zhang et al.[15].

Second, a central content provider or network operator does not know how to *prioritize among its users*. For instance, a provider like Skype contracts a slice that delivers low latency and high bandwidth. This enables high-quality Skype calls for users admitted to the slice but physical resources within the slice continue to be limited. Presumably, when multiple users make Skype calls at congested times, they compete for admission into the slice and have no way to influence the outcome, just as is the case today. For example, a user cannot demonstrate to the slice allocation algorithm their higher call *utility value* for a job interview over a recreational call from another user. **Hence, while admission into a slice largely guarantees slice-specified SLA, admittance itself is entirely controlled by the centralized operator or the content provider in any case.**

Efficiency concerns: Apart from the usability limitations in the centralized slice-ownership model that do not cater to diverse CPS use cases, severe implications on resource utilization have been recently studied [16]. By way of this slicing model, *traffic multiplexing capacity is significantly diminished in the network*. Resources of a slice may be multiplexed only between the traffic demand for that slice rather than between all traffic demand over all resources, as possible today under ad-hoc resource provisioning [5]. Given the inherently sparse nature of radio and other resources in the edge, these utilization losses are highly undesirable.

The lowered utilization efficiency in slicing is empirically analyzed by Marquez et al. [16]. The imposition of a *guaranteed time fraction* in advance as part of slice specifications, i.e., a guarantee that at least a certain percentage of all traffic demand for that slice will be served satisfactorily over fixed time windows, has a steep cost for the operator. Due to the spiky nature of traffic demand, especially closer to the edge, the provider must provision for peak demand within these time windows to ensure the guaranteed time fraction, leading to efficiency loss as high as 80%.

Relaxation of this time window or time fraction does not help beyond a certain extent[16]; efficiency loss may only further be improved if slices acquire *frequent reconfigurability*. For example, if physical resource allocation to slices can be reconfigured every 30 minutes, then efficiency loss is reduced to a best-case scenario of about 20%. Beyond this, the unavoidable effect of multiplexing loss inherent in slicing dominates with no substantial further improvements. As the authors note, given the expenses of operating the network infrastructure and operationalizing such virtualized capabilities, such high resource utilization losses may prohibit monetary feasibility of realizing this model.

These usability and economic viability concerns are implicitly addressed in the model we propose. In this work, we introduce the concept of offering **slicing as an ad-hoc service to edge entities** who, consequently, drive their resource allocation and thereby the QoE for their network sessions. As we detail in the next section, creating network slices in response to edge entities when they express specific SLA needs alleviates the multiplexing loss from up-ahead resource dedication to slices. It creates new opportunities for monetization of value-added services to the network while empowering edge entities and adding value to them.

II. Slicing as a Service to Edge Entities

As Andrews et al. point out [17], network virtualization in 5G revives a radical concept that first emerged in the 1990s: "the provision of user-controlled management in network elements". By offering network slicing as a service in real time to edge entities instead of exposing it via periodic contracts to content providers, 1) current usability and efficiency issues are addressed and 2) new incentive challenges are introduced.

We first note that this proposed shift in the service model of network slicing has significant impact on utilization efficiency. In their empirical evaluation, Marquez et al. [16] show that efficiency loss with slicing becomes negligible only when there are a minimal number of slices (i.e., one dedicated to high-volume, SLA-driven traffic and one mainly serving low-volume, SLA-free traffic) with frequent reconfiguration of assigned physical resources to the slices.

In such a scenario, statistical multiplexing gains are largely re-captured, since they are higher when done over a larger portion of the physical resources and traffic. This is best done by maintaining a limited number (if any) of larger slices (perhaps two as above), and instead largely deciding the SLA feasiblily of a flow and its requisite physical resources in real time. A slice is created ad-hoc if the flow with its customized SLA is deemed feasible and sufficient physical resources dedicated to ensure performance isolation. In-fact, recent work [9], [10] has shown that slices can be created and their isolation enforced dynamically even in the highly contended RAN layer while simultaneously maintaining high radio efficiency, making our proposed model feasible.

Secondly, such an offering allows entities with diverse needs (and uncertainty or variation in their future resource requirements, as in several CPS scenarios) to use the system. Further, virtualization shields the edge entity from complex radio-layer details and predictions in computing the resources they need. Edge entities may simply relay requisite network service in terms of application-layer needs (such as bitrate and latency) and have operator compute the mapping to physical resources. Hence any application may procure its session needs, whatever they may be.

In the earlier Skype example, however, we see that the resource needs of a user are not simply the application or device's inherent network requirements; users also hold subjective preferences and relative utilities for network sessions. These subjective utilities are especially important to capture when edge resources are limited and not every slice request can be met. To enable users to drive their resource allocation, it is not only important to offer suitable slices (or equivalents) in response to their diverse slice specifications, but also allow them to influence its successful allocation by expressing their utility for it. The burden is then placed on the operator to design an effective incentive scheme that aligns the entity's stated utility with its true valuation. The operator may now monetize its virtualization capability (not pre-existing slices but the ability to create them) by offering it as a real-time value-added service to edge entities, while retaining much of its desirable statistical multiplexing capability. Hence, we realize **the fundamental shift of control from the operator/content-provider to the edge entity, which now drives its network resource allocation in real time as aligned with its incentives.**

III. The Role of Incentives

Historically, edge entities have had limited scope to express the utility they derive from network resources to operators. Prevalent mobile data plans are month-long contracts that do not capture finer-grained information about user preferences and utilities. The current trend towards high diversification of applications and the network services they require, combined with empirical observations that heavy users of cellular internet exhibit non-periodic, sporadic usage [18], indicates that this lack of fine-grained information likely induces a considerable loss in value for both the operator and edge entities. We posit that **offering network slices as a service in real time re-captures this source of value** by allowing operators to offer, and users to pay for, services customized to spontaneous user needs.

Capturing end-user preferences in the form of their utility or valuation and using this to drive resource allocation models is the aim of incentive design mechanisms. The seminal work on Paris Metro Pricing[19], for example, allows users to state their value for the supported QoS levels by explicitly choosing their tier and paying accordingly, assuming that the network can guarantee these tiered performances to all users who pay the price. Since then, several incentive mechanisms for network usage have been proposed with varying goals[20]. However, there has not been significant attention on the problem of capturing user (or generic entity, with the rise of CPS) valuations in real time for application-oriented session preferences. Until now, it has been difficult to realize dynamic QoS policies in today's network architecture with limited flexibility. However, 5G's virtualization features makes such a paradigm entirely feasible. We turn our attention to several open challenges in designing these incentive mechanisms and, more broadly, in realizing a fully functional offering of slicing as a service to edge entities.

IV. Research Challenges

Incentive design: The operator must benefit monetarily by offerring virtualization capabilities to edge entities. Since slice specifications and their resource costs are hereby known only in real time, the operator must price slices in real time as a function of the associated SLA. On the other hand, dynamic pricing schemes have been known to have usability limitations [21] since typical end users are budget-constrained and make economic choices over longer time spans. *Building user-friendly mechanisms for dynamic pricing that provably incentivize users to state their true valuations for customized slice specifications is an open problem.* In fact, the most user-friendly method might be to develop agents that act as a proxy for the edge entity in engagements with the network. Once the entity's preferences are captured by the agent, such as budget, applications for which service guarantees are desired, and preferred resolution rates, the agent may transparently engage with the network to acquire a guaranteed slice. The design of such agents and the various learning tasks they may have to perform (see below) poses questions.

Interaction model and slice parameterization: The network operator must provide a framework for edge entities or the corresponding agents to state their desired SLAs, also parameterized by duration of consumption, entity location during the session etc. Rather than the entity explicitly engaging with the network to specify such parameters for each network session (thereby degrading usability, given the volume of mobile data activity in a day [22]), agents may instead learn from the entity's preferences and usage patterns to estimate these parameters. Depending on the incentive scheme being employed by the network, the agent may also need to estimate the entity's valuation for the slice and corresponding utility-optimizing slice specifications in real time.

Cost of being dynamic: Since CPS traffic is expected to be largely machine driven [2], establishing slices in real time would likely involve complex and frequent communication of the requirements between the various edge entities and the network. While recent work [9], [10] has established the feasibility of dynamic RAN slicing without significant efficiency loss, further work is required in studying the signaling overhead/stress caused by slice requests. Further, the turnaround times for resolution of dynamic slice requests must be minimal to facilitate ad-hoc sessions with SLA needs. This requires further study of the delays incurred by the incentive mechanism in determining allocations.

Slice Policies: The operator may wish to enforce generic slice policies to allow opportunities for other edge entities to acquire slices. For example, by enforcing a slice occupation duration of no more than thirty minutes, the limited edge resources are guaranteed to free up periodically for occupancy by other entities. This may also influence the operator's revenue and incentive mechanism, as the market rates for resources presumably change under congested times. However, such policies involve tradeoffs. For example, periodically terminating slices may expose the operator to the risk of decreased market rates during the subsequent time period as demand drops off.

V. Our Preliminary and Ongoing Work

We currently explore incentives for dynamic slice offerings in the context of real-time applications. Sessions of real-time applications are especially hard to provision for due to their immediate resource needs that do not lend to buffer-based adaptations. We enable such applications to acquire guaranteed QoE for their sessions by negotiating with the operator for a slice of their desired SLA.

We first introduce the slice model considered by Marquez et al. [16], where a slice is fully (pre-)specified as $z = (f,w)$, where f and w are such that the operator satisfies at least a fraction f of the slice demand, averaged over discrete time windows w. The operator provisions for peak demand to satisfy f, thereby necessitating additional deployment of physical resources and considerable economic strain [16].

In our entity-driven slice model, a requested slice or SLA may be fully characterized as $z_t = (s(t,d),c)$, where $s(t,d)$ is the dynamic slice specification of the edge entity for consumption duration d at time t and c is the cost of the slice. A slice s is entirely feasible only if it complies with operator's policies and resource availability. The slice cost c may or may not be conveyed by the network (for example, prices may not be explicitly set in an auction bidding scenario). As seen in this model, the operator only needs to decide if enforcing s is feasible, without the burden of guaranteeing the minimum traffic fulfillment f. Instead the burden of fulfillment is offloaded to the incentive scheme. If the entity's valuation is sufficiently high, its traffic is served by the requested slice, otherwise it is turned away.

Based on this, we provide a combinatorial auction mechanism for edge entities to compete for slice allocation in periodic real-time auctions that maximize social welfare. By exploiting the nature of real-time applications, we achieve auctions with winner determinations that are simultaneously fast and incentive compatible, properties not readily achieved in this setting. An agent, on behalf of the edge entity, submits a bid b for z_t while adhering to the entity's daily budget, thereby addressing usability concerns of dynamic pricing models. We also explore learning mechanisms for the agent to learn the edge entities' QoE preferences and place bids proportionally, making the slice procurement process more transparent and user friendly.

VI. Conclusion

We address the diversified service requirements of IoT and CPS application scenarios in 5G networks and the resource scarcity that characterizes the network edge through consideration of user-driven mechanisms for real-time resource management. We propose to offer these limited resources as virtualized network slices tailored for the network needs and aligned with the incentives of edge entities. By allowing the edge entity to explicitly influence its resource allocation, we fundamentally reduce the centralized control of the operator/content provider on the entity's network experience. We highlight the promising and feasible directions for such dynamic and ad-hoc slicing, and the efficiency and usability gains that can be achieved using a user-driven approach. We outline the research challenges in realizing slicing offerings, with a focus on incentive design and usability. Finally, we provide a brief overview of our ongoing work in this space.

References

[1] M. Chiosi, D. Clarke, P. Willis, A. Reid, J. Feger, M. Bugenhagen, W. Khan, M. Fargano, C. Cui, H. Deng, *et al.*, "Network functions virtualisation: An introduction, benefits, enablers, challenges and call for action," in *SDN and OpenFlow World Congress*, vol. 48, sn, 2012.

[2] G. P. Fettweis, "The tactile internet: Applications and challenges," *IEEE Vehicular Technology Magazine*, vol. 9, no. 1, pp. 64–70, 2014.

[3] B. Kang, I. Hwang, J. Lee, S. Lee, T. Lee, Y. Chang, and M. K. Lee, "My being to your place, your being to my place: Co-present robotic avatars create illusion of living together," in *Proc. 16th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 54–67, ACM, 2018.

[4] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of things for smart cities," *IEEE Internet of Things journal*, vol. 1, no. 1, pp. 22–32, 2014.

[5] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network slicing in 5g: Survey and challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 94–100, 2017.

[6] X. Foukas, N. Nikaein, M. M. Kassem, M. K. Marina, and K. Kontovasilis, "Flexran: A flexible and programmable platform for software-defined radio access networks," in *Proc. 12th International Conference on emerging Networking EXperiments and Technologies*, pp. 427–441, ACM, 2016.

[7] A. Banerjee, R. Mahindra, K. Sundaresan, S. Kasera, K. Van der Merwe, and S. Rangarajan, "Scaling the lte control-plane for future mobile access," in *Proc. 11th ACM Conference on Emerging Networking Experiments and Technologies*, p. 19, ACM, 2015.

[8] X. Costa-Pérez, J. Swetina, T. Guo, R. Mahindra, and S. Rangarajan, "Radio access network virtualization for future mobile carrier networks," *IEEE Communications Magazine*, vol. 51, no. 7, pp. 27–35, 2013.

[9] X. Foukas, M. K. Marina, and K. Kontovasilis, "Orion: Ran slicing for a flexible and cost-effective multi-service mobile network architecture," in *Proc. 23rd Annual International Conference on Mobile Computing and Networking*, pp. 127–140, ACM, 2017.

[10] A. Ksentini and N. Nikaein, "Toward enforcing network slicing on ran: Flexibility and resources abstraction," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 102–108, 2017.

[11] X. Zhou, R. Li, T. Chen, and H. Zhang, "Network slicing as a service: enabling enterprises' own software-defined cellular networks," *IEEE Communications Magazine*, vol. 54, no. 7, pp. 146–153, 2016.

[12] K. Samdanis, X. Costa-Perez, and V. Sciancalepore, "From network sharing to multi-tenancy: The 5g network slice broker," *IEEE Communications Magazine*, vol. 54, no. 7, pp. 32–39, 2016.

[13] X. An, C. Zhou, R. Trivisonno, R. Guerzoni, A. Kaloxylos, D. Soldani, and A. Hecker, "On end to end network slicing for 5g communication systems," *Transactions on Emerging Telecommunications Technologies*, vol. 28, no. 4, p. e3058, 2017.

[14] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, C. Ziemlicki, and Z. Smoreda, "Not all apps are created equal: Analysis of spatiotemporal heterogeneity in nationwide mobile service usage," in *Proc. 13th International Conference on emerging Networking EXperiments and Technologies*, pp. 180–186, ACM, 2017.

[15] H. Zhang, N. Liu, X. Chu, K. Long, A.-H. Aghvami, and V. C. Leung, "Network slicing based 5g and future mobile networks: mobility, resource management, and challenges," *IEEE Communications Magazine*, vol. 55, no. 8, pp. 138–145, 2017.

[16] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, "How should i slice my network? a multi-service empirical evaluation of resource sharing efficiency," 2018.

[17] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5g be?," *IEEE Journal on selected areas in communications*, vol. 32, no. 6, pp. 1065–1082, 2014.

[18] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in *IEEE INFOCOM*, pp. 882–890, IEEE, 2011.

[19] A. Odlyzko, "Paris metro pricing for the internet," in *Proc. 1st ACM conference on Electronic commerce*, pp. 140–147, ACM, 1999.

[20] S. Sen, C. Joe-Wong, S. Ha, and M. Chiang, "Pricing data: A look at past proposals, current plans, and future trends," *CoRR, abs/1201.4197*, 2012.

[21] S. Ha, S. Sen, C. Joe-Wong, Y. Im, and M. Chiang, "Tube: time-dependent pricing for mobile data," *ACM SIGCOMM Computer Communication Review*, vol. 42, no. 4, pp. 247–258, 2012.

[22] H. Wang, F. Xu, Y. Li, P. Zhang, and D. Jin, "Understanding mobile traffic patterns of large scale cellular towers in urban environment," in *Proc. Internet Measurement Conference*, pp. 225–238, ACM, 2015.