Received December 16, 2019, accepted January 28, 2020, date of publication February 6, 2020, date of current version February 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2972105

End-to-End Network Slicing in Radio Access Network, Transport Network and Core Network Domains

XU LI, RUI NI, JUN CHEN, YIBO LYU, ZHICHAO RONG, AND RUI DU
Shenzhen Research Development Center, Huawei Technologies Company, Ltd., Shenzhen 518129, China

Corresponding author: Rui Du (ray.du@huawei.com)

ABSTRACT Network slicing (NS) has been well discussed in the transport network (TN) and core network (CN) domains. This paper extends it to the radio access network (RAN) domain, and the NS in RAN, TN and CN domains is defined as end-to-end (E2E) NS system. The advantages of using NS in the RAN domain with two-level resource allocation scheme are studied and shown by numerical simulations. Then the E2E NS system architecture and components are proposed and demonstrated with hardware and software. The demonstration shows the capability with very fine spectral granularity, and the slice creation, delete and adjustment schemes in sub-minute time, which could be used in the operator's network.

INDEX TERMS End-to-end network slicing, radio access network, spectrum allocation, demonstration.

I. INTRODUCTION

With the development of wireless communication system, one network fitting for all is technically infeasible for diverse scenarios, including enhanced mobile broadband (eMBB), ultra-reliable and low-latency communication (uRLLC) and massive machine type communication (mMTC), hence network slicing (NS) is adopted as a major character in 5G [1]. In 2016, NS is formally presented by next generation mobile networks (NGMN), which consists of service instance layer, slice instance layer and resource layer [2]. The service instance layer refers to an instance of an end-user service or a business service. The slice instance layer is a set of network functions and resources to run these functions. The resource layer are physical computation, storage and radio access resources.

In industry, many companies have designed their system structure for NS. For example, Ericsson proposes an automated service-oriented lifecycle management platform to solve the complex challenge of managing services with different requirements throughout their lifecycles [3]. Nokia shows a network slice demonstration to practice the orchestration of network functions and mapping from service instance layer to network slice instance layer [4]. AT&T proposes its enhanced control, orchestration, management & policy (ECOMP) architecture, which is a top-down, service-lifecycle focused, metadata-driven software platform [5]. In the system structure, with specific requirement, the orchestrator generates and sends commands to network function virtualization (NFV) [7] and software-defined networking (SDN) [6] controllers. Then virtual functions created by NFV, are connected by the networks created by SDN, which is regarded as a network slice instance. However, all the above works study NS in the domains of core network (CN) and transport network (TN). An important question to consider is whether NS could be used in the domain of radio access network (RAN).

In the domain of traditional RAN, each user has three signaling radio bears (SRBs) and eight data radio bears (DRBs) [8]. The SRBs are used to carry control signal, while DRBs are used to carry traffic data. Different DRBs are designed for diverse traffics, which is similar to NS. However, all the DRBs for different users are viewed and scheduled by a shared medium access control (MAC), which greatly threatens the security [9]. For isolation requirement, NS is also a promising method in RAN in the future. After isolation, according to the specific requirements of each slice instance, the packet data convergence protocol (PDCP), radio link control (RLC) and MAC protocols could be further optimized and standardized to improve the system performance.

Meanwhile, the open-air spectrum is also scheduled by the MAC functions in RAN. The spectrum numerology in LTE is fixed to 15kHz subcarrier spacing with 1ms subframe length [10], which is upgraded in 5G to support 15, 30, 60, 120 and 240kHz subcarrier spacing with 1,

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wang.

0.5, 0.25, 0.125 and 0.0625ms sub-frame length [11]. Since the sub-frame length is the minimum schedule granularity, the upgrade enables the service with less than 1ms delay requirement. Without NS, multiple kinds of numerology are jointly scheduled by MAC, which increases the complexity and degrades the robustness of the system. With NS, there could be one master MAC to optimize the amount of resource among slice instances, and multiple self-designed slave MACs schedule the resource in slice instances. The two-level scheme is simple and easy to be implemented.

This paper proposes the end-to-end (E2E) NS, where NS is adopted not only in the TN and CN domains, but also in the RAN domain. First, the two-level scheme with one master and multiple slave MACs is studied. Numerical simulations are provided to evaluate the system performance and show the advantages with NS in RAN domain. Then a Huawei's E2E NS system is demonstrated with real hardware. The hardware infrastructure is scheduled by orchestrators assisted by NFV and SDN controllers, where the system architecture along with the runtime command and real-time response are described. The demonstration achieves a balance among real-time response, compatibility and scalability.

The remainder of this paper is organized as follows. Section II describes the related works and contributions of this paper. Section III introduces the E2E NS architecture. Section IV proposes and analyzes the two-stage spectrum allocation scheme and discusses the benefits with NS in the RAN domain. Section V presents the E2E NS system, including the system components, impacting to RAN protocols and discussions. Section VI shows the E2E NS demonstration followed by conclusions in VII.

II. RELATED WORKS AND CONTRIBUTIONS

State of the art of the two-level resource allocation scheme in the RAN domain can be traced back to the works with multiple mobile virtual network operators (MVNOs) [12]–[15]. The resources of base stations owned by the infrastructure provider (InP) are shared to multiple MVNOs. A two-level scheme named network virtualization substrate (NVS) could be used to allocate resources to different MVNOs and their mobile users [15]–[17]. The first stage allocates resources to MVNOs and maximizes the InP revenue based on the feedback from the MVNOs, such as bandwidth requirement, resources utilization and biding price. The second stage allocates resources to users within each MVNO following traditional MAC methods. The two-level scheme in this paper considers different subcarrier spacing and sub-frame length using 5G frame structure, which could be directly used in the real system. Meanwhile, comprehensive comparison of modified largest weighted work first (M-LWWF) and modified largest weighted delay first (M-LWDF) methods at the second stage are discussed.

After verifying the system performance benefits, it is straightforward to extend the E2E NS to the RAN domain. Indeed, [20] identifies requirements for NS in the RAN domain, including efficiency, protection, differentiation and slice awareness. Reference [21] discusses different concepts, challenges and management for NS in the RAN domain, but without detailed system architecture. Meanwhile, remote radio unit (RRU) and baseband unit (BBU) mapping [22], [23] and functions split and placement [24] are also studied for NS in the RAN domain. Similar to the system structure proposed by many companies for NS in the TN and CN domains [4]–[7], this paper proposes the system structure for E2E NS including RAN domain. We consider virtual and physical resources in RAN, TN and CN domains for different slice instances. The virtual resources are the functions and protocols, and the physical resources include router, CPU, memory, antenna, time, and open-air spectrum especially. Besides, the structure is also demonstrated with real hardware and software to show the very fine spectral granularity and slice creation, delete and network breathing in sub-minute time.

The main contributions of this paper are summarized as:

- *(Section IV: NS in the RAN domain)* The benefits with NS in the RAN domain are studied, besides discussion, a two-level resource allocation scheme is proposed and numerically analyzed with sliced open-air spectrum. The scheme takes multiple kinds of subcarrier spacing and sub-frame length of 5G frame structure into consideration, and comprehensively compares with the cases without NS in the RAN domain.
- *(Section III & V: E2E NS architecture and system)* After verifying the benefits with NS in the RAN domain, the E2E NS including RAN, TN and CN domains is proposed, whose detailed components and impacting to RAN protocols are described and discussed.
- *(Section VI: E2E NS demonstration)* The E2E NS is demonstrated with real hardware and software to show its capability to support very fine spectral granularity and operating mechanism in sub-minute time. The very fine spectral granularity shows the feasibility with sliced open-air spectrum. The operating mechanism in sub-minute time enables the system to be deployed in the operator's network.

We note that extending E2E NS system to the RAN domain not only benefits the system performance, but also the ecosystem. Once the open-air spectrum is sliced, over the top (OTT) service providers could sense the spectrum to guarantee that their slice instances are truly dedicated for themselves, which could highly increase their willingness to pay for the NS system. Then the mobile network operators and telecommunications equipment manufacturers are motivated to upgrade their system and equipment to provide better but differentiated communication services.

III. ARCHITECTURE

As shown in Fig.1, our architecture of 5G E2E NS is consisted of RAN, TN and CN in east-west direction, and infrastructure, control framework and management plane in south-north direction.

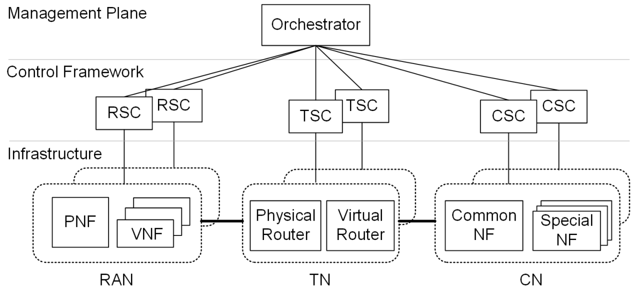


FIGURE 1. Architecture of the E2E NS system.

In the following, the architecture is introduced from the south to north direction. The infrastructure layer consists of physical and virtual resources. In the RAN domain, there are physical network functions (PNFs) and virtual network functions (VNFs), where PNFs include the hardware resources and the VNFs contain functions and protocols. In the TN domain, there are physical and virtual routers. In the CN domain, there are VNFs classified as common network functions and special network functions. The common network functions are shared by different slice instances, and the special network functions are dedicated for the related slice instances.

The control framework layer is comprised of radio slice controllers (RSCs), transport slice controllers (TSCs) and core slice controllers (CSCs), which are in RAN, TN and CN domains respectively. The reason to place controllers in different domains is that, for security or policy requirement, the equipment in the real telecommunication system is always supplied by different telecom equipment manufactures. Because of competition, it seems to be more reasonable for the telecom equipment manufactures to produce their own equipment along with the self-designed controllers, unless the open white box telecom equipment is universally deployed in the future. As a result, there are multiple RSCs, CSCs and TSCs, which may be developed by different telecom equipment manufactures.

The management plane has the orchestrator, which receives the business order from the north and status reports from the south, and sends feedback to the north and commands to the south control framework layer. The business order could be the requirement to create, delete or reserve a slice instance. The status reports could be the real-time key performance indicator (KPI) of the service, slice instance status, and physical resource usage status, etc. The feedback is either a success or fail response to the business order. The commands could be create, delete, reserve or adjust the physical and virtual resources for slice instances.

The major difference of our E2E NS compared with other NS systems is that, besides the slicing in the TN and CN, slicing is also adopted in the RAN domain especially the open-air spectrum. In the following, we first study the benefits with slicing in the RAN domain, then introduce the details of the proposed E2E NS system and the demonstration with real hardware.

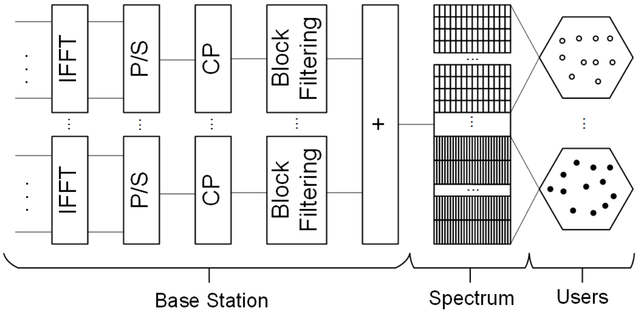


FIGURE 2. RAN slicing model.

IV. NS IN THE RAN DOMAIN

With NS in the RAN domain, there is one master MAC allocates resource between slice instances, and multiple self-designed slave MACs schedule the resource within slice instances.

A. RAN SLICING MODEL

The RAN slicing model is shown in Fig. 2, where users are classified into groups and communicate with a base station. We assume M groups and m_i users in the i-*th* group. The groups are indexed by $i \in \{0, 1, \ldots, M-1\}$ and users are indexed by $j \in \{0, 1, \ldots, m_i - 1\}$.

Each group is served by one slice instance and uses one type of numerology, whose slot length is assumed to be prefixed as $\Gamma_i \in \{1\text{ms}, 0.5\text{ms}, 0.25\text{ms}, 0.125\text{ms}, 0.0625\text{ms}\}$. Let Γ_0 be the maximum one, and others satisfy:

$$\Gamma_i = \Gamma_0 / 2^{n_i} \tag{1}$$

where $n_i \in \mathbb{N}$ and $n_0 = 0$.

With pre-fixed slot length, the subcarrier spacing of each type of numerology is also fixed as $B_i \in \{15\text{kHz}, 30\text{kHz}, 60\text{kHz}, 120\text{kHz}, 240\text{kHz}\}$. The minimum granularity of spectrum resource is physical resource block (PRB), which has 12 sub-carriers [11]. Let K_i be the number of PRBs allocated to the i-*th* group, which satisfy:

$$\sum_{i=1}^{M} 12 K_i B_i \leq W. \tag{2}$$

where W is the total bandwidth. Moreover, let $k_{i,j}$ be the number of PRBs allocated to the j-*th* user in the i-*th* group, which satisfy:

$$\sum_{j=1}^{m_i} k_{i,j} \leq K_i. \tag{3}$$

We study the down-link channel and the up-link channel follows the same way. The position of the base station is assumed to be $(0, 0)$. The transmit power spectrum density for and the position of the j-*th* user in the i-*th* group are defined as $p_{i,j}$ and $(x_{i,j}, y_{i,j})$, respectively.

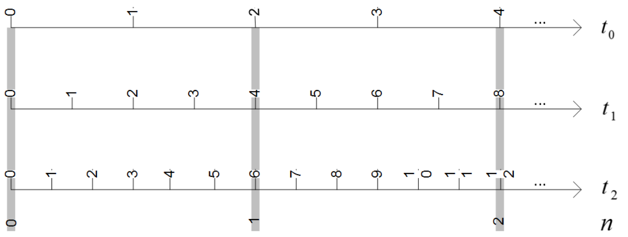


FIGURE 3. One schedule time example with $M = 3, n_0 = 0, n_1 = 2, n_2 = 3$ and $N = 2$. The shadowed parts are the schedule time of master MAC, and others are the schedule time of slave MAC.

Rayleigh fading model is used for links between the base station and users [26], [27]. The path loss of a link is:

$$h_{i,j} = \alpha l_{i,j}^{-\beta} \tag{4}$$

where $l_{i,j} = \sqrt{x_{i,j}^2 + y_{i,j}^2}$, α is a constant value that considers shadowing and antenna gain, β is the path loss exponent.

At the user, in the absence of interference, the signal-to-noise ratio (SNR) is:

$$\mathrm{SNR}_{i,j} = \frac{p_{i,j}\alpha l_{i,j}^{-\beta} g_{i,j}}{\sigma^2}. \tag{5}$$

where $p_{i,j}$ is the transmit power spectral density, $g_{i,j}$ is the power gain of the Rayleigh fading channel following the exponential distribution with unit mean, and σ^2 is the power spectral density of white Gaussian noise.

The data rate for each user is determined by the SNR and allocated spectrum, while the data arriving rate depends on the real traffic. When the instance data rate is lower than the arriving rate, the remaining data will be cached in the buffer. However, the buffer size of the user is limited and may overflow. Hence for delay insensitive traffic, the traffic will be dropped once the buffer overflows. For delay sensitive traffic, the traffic will be dropped once the buffer overflows or its delay exceeds a pre-defined threshold.

B. SPECTRUM RESOURCE ALLOCATION BY MASTER MAC

The master MAC allocates resource between slice instances. According to 5G, the radio frame is indexed by an increasing positive integer, which are divided into multiple slots. We index the slot as an increasing non-negative integer from 0 to infinity.

As shown in Fig. 3, the slot length varies among different types of numerology. This paper defines the shadowed parts as the schedule time of master MAC, and others as the schedule time of slave MAC. At the schedule time of master MAC, $(K_0, K_1, \ldots K_{M-1})$ are optimized to allocate spectrum resources to different slice instances. At the schedule time of slave MAC, $k_{i,j}, i \in \{0, 1, \ldots, M-1\}, j \in \{0, 1, \ldots, m_i - 1\}$ are optimized to allocate PRBs to users within each slice instance. The schedule time interval of the master MAC is:

$$\Delta T = N\Gamma_0 = N2^{n_i}\Gamma_i \tag{6}$$

where $N \in \mathbb{N}^+$ determines the interval length. We remind that Γ_0 and Γ_i are the slot length defined in equation (1).

The master MAC takes the traffic drop rate at its previous schedule time interval into consideration. The traffic drop rate of each slice instance is:

$$P_i(l) = \frac{\sum_{j=1}^{m_i} \mathcal{D}_{i,j}\langle (l-1)\Delta T \to l\Delta T\rangle}{\sum_{j=1}^{m_i} \left(\mathcal{D}_{i,j} + \mathcal{T}_{i,j}\right)\langle (l-1)\Delta T \to l\Delta T\rangle} \tag{7}$$

where $\mathcal{D}_{i,j}\langle T_1 \to T_2\rangle$ is the amount of traffic dropped from T_1 to T_2, $\left(\mathcal{D}_{i,j} + \mathcal{T}_{i,j}\right)\langle T_1 \to T_2\rangle$ is the amount of traffic dropped and transmitted from T_1 to T_2, and l is the index of master schedule time interval. In the following, we use P_i instead of $P_i(l)$ for ease of expression.

The spectrum resource are allocated to achieve fairness between slice instances, whose weighted traffic drop rate intends to be identical in the current schedule time interval of master MAC as [28], [29]:

$$w_1 P_1 = \cdots w_i P_i = \cdots w_M P_M \tag{8}$$

where w_i is the weight. The slice instances with more critical drop rate requirement has larger weight value, and vice versa.

Let $(K_0^*, K_1^*, \ldots, K_{M-1}^*)$ be the spectrum resource allocated to slice instances at the previous schedule time interval by master MAC. In the current master schedule time interval, the slice instances with lower traffic drop rate may release the allocated resources and the ones with higher traffic drop rate prefer to gain more resources. All the slice instances share the spectrum resources with another weight η_i as:

$$\eta_i = 12K_i^* B_i \exp\left(w_i P_i\right). \tag{9}$$

Then using equality (2), the spectrum resource allocation intends to be:

$$12K_i' B_i = \frac{\eta_i}{\sum_{j=1}^{M} \eta_j} W. \tag{10}$$

where $(K_0', K_1', \ldots, K_{M-1}')$ are the intended spectrum resource allocation result at the current schedule time interval. We note that the equation holds with floating value of K_i'.

The intended spectrum resource allocation highly depends on the traffic drop rate. But in the real system, the traffic drop rate may fluctuate dramatically due to the traffic burst and transient interference, where substantially revising the spectrum allocation may be un-necessary. Meanwhile, frequently revising the spectrum allocation increases the system complexity and threats the system stability. In this paper, we adopt smoothing average method [30] and regard K_i' as the current value with $\xi \in [0, 1]$ as the smoothing factor. The current spectrum allocation $(\hat{K}_0, \hat{K}_1, \ldots, \hat{K}_{M-1})$ are revised as:

$$\hat{K}_i = (1 - \xi) K_i^* + \xi K_i'. \tag{11}$$

Moreover, the number of PRBs allocated to each slice instance should be an integer. Let K_i be the nearest integer of $\hat{K}_i$. We also need to guarantee that almost all the spectrum resources are allocated. If $\sum_{i=0}^{M-1} 12K_i B_i > W$, the amount of PRBs minus one for the slice instance with the maximum $\frac{K_i - \hat{K}_i}{K_i}$. The slice instance with more PRBs and more gaining part caused by the rounding operation has higher priority to minus one PRB. If $\sum_{i=0}^{M-1} 12K_iB_i < W$, the amount of PRBs plus one for the slice instance with the maximum $\frac{\hat{K}_i - K_i}{K_i}$. The slice instance with less PRBs and more missing part caused by the rounding operation has higher priority to gain one PRB.

C. SPECTRUM RESOURCE ALLOCATION BY SLAVE MACS

After the spectrum resource allocated to slice instances, we allocate PRBs to users within each slice instance during the current master schedule time interval. In the following, the schedule time of slave MAC is every slot as the non-shadowed part shown in Fig.3.

The slave MAC takes the capacity of each PRB into consideration as [31]:

$$c_{i,j} = 12B_i \log_2 \left(1 + \frac{p_{i,j}\alpha l_{i,j}^{-\beta} g_{i,j}}{\sigma^2}\right) \tag{12}$$

which varies for different user and PRB. The index of the PRB is omitted to simplify the expression without misunderstanding.

1) DELAY INSENSITIVE TRAFFIC

For the delay insensitive traffic, M-LWWF method is used to allocate PRBs to users [18], [19]. The PRBs are allocated one by one. In each step, one PRB is allocated to the user with the largest weighted amount of cached traffic times capacity whose index is:

$$I_1(i,j) = \arg\max\left(\zeta_{i,j}c_{i,j}b_{i,j}\right) \tag{13}$$

where $\zeta_{i,j}$ is the weight and $b_{i,j}$ is the amount of cached traffic in the buffer. We note that each user may obtain multiple PRBs. Once a PRB is allocated, the amount of cached traffic reduces to $b_{i,j} - c_{i,j}$ to compete for others.

2) DELAY SENSITIVE TRAFFIC

For the delay sensitive traffic, M-LWWF method is suitable only if the traffic arrives smoothly, where more cached traffic implies larger delay of the head-of-line traffic.

In general, M-LWDF method is more suitable to allocate PRBs to users with delay sensitive traffic [18], [19]. Similarly, the PRBs are allocated one by one. In each step, one PRB is allocated to the user with the largest weighted delay of the head-of-line traffic times capacity whose index is:

$$I_2(i,j) = \arg\max\left(\zeta_{i,j}c_{i,j}d_{i,j}\right). \tag{14}$$

where $d_{i,j}$ is the delay of the head-of-line traffic. Once a PRB is allocated, the delay of the head-of-line traffic is reduced to compete others.

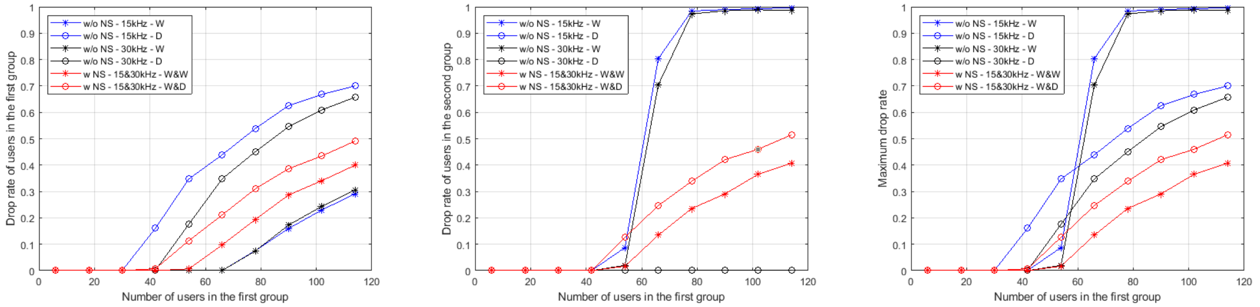
The two-level spectrum resource allocation scheme with NS is summarized in *Algorithm* 1, where step 12 means more resource is allocated or the remaining resource is not less than the minimum PRB, and step 30 guarantees that one PRB is allocated to no more than one user.

Algorithm 1 Spectrum Resource Allocation With NS

```
1:  **Input:** $\alpha$, $\beta$, $\sigma^2$ $W$, $\xi$, $\Delta T$, $B_i$, $\Gamma_i$, $w_i$, $p_{i,j}$, $(x_{i,j}, y_{i,j})$, $\zeta_{i,j}$, $\forall i \in \{0, 1, \ldots, M-1\}$, $\forall j \in \{0, 1, \ldots, m_i - 1\}$.
2:  **Output:** The spectrum resource allocation to slice instances $K_i$ and to users $k_{i,j}$ within each slice instances.
3:  Initialize the schedule index of master MAC $l \leftarrow 0$.
4:  while True do
5:      Stage1: Spectrum resource allocation by master MAC
6:      if $l == 0$ then
7:          $\hat{K}_i \leftarrow \frac{m_iW}{\sum_{j=0}^{M-1} m_j}/12B_i, \forall i \in \{0, 1, \ldots, M-1\}$.
8:      else
9:          Compute $(\hat{K}_0, \hat{K}_1, \ldots, \hat{K}_M)$ using equality (11).
10:     end if
11:     $K_i \leftarrow round(\hat{K}_i), \forall i \in \{0, 1, \ldots, M-1\}$.
12:     while $W - \sum_{i=0}^{M-1} 12K_iB_i < 0 || W - \sum_{i=0}^{M-1} 12K_iB_i \geq \min\{12B_i\}$ do
13:         if $\sum_{i=0}^{M-1} 12K_iB_i > W$ then
14:             $K_{i'} \leftarrow K_{i'} - 1$ where $i' = \arg_i \max\{\frac{K_i - \hat{K}_i}{K_i}\}$.
15:         else
16:             $K_{i'} \leftarrow K_{i'} + 1$ where $i' = \arg_i \max\{\frac{\hat{K}_i - K_i}{K_i}\}$ and $12B_{i'} \leq W - \sum_{i=0}^{M-1} 12K_iB_i$.
17:         end if
18:     end while
19:     Stage2: Spectrum resource allocation by slave MACs
20:     for $i \in \{0, 1, \ldots, M-1\}$ do
21:         for $t_i \in [l\Delta T, (l+1)\Delta T - 1]$ do
22:             Let $\#_i \leftarrow 0$ be the number of PRBs allocated.
23:             while $\#_i < K_i$ and $\sum_{j=0}^{m_i-1} b_{i,j} > 0$ do
24:                 if Delay insensitive traffic then
25:                     Allocate the $k'_i$-th PRB to the $j'$-th user using equality (13).
26:                 else
27:                     Allocate the $k'_i$-th PRB to the $j'$-th user using equality (13) or (14) and update its delay of the head-of-line traffic.
28:                 end if
29:                 $b_{i,j'} \leftarrow b_{i,j'} - \min\{c_{i,j'}, b_{i,j'}\}$, where $c_{i,j'}$ is the capacity at the $k'_i$-th PRB for the $j'$-th user.
30:                 $c_{i,j} \leftarrow 0$ at the $k'_i$-th PRB for all users.
31:             end while
32:         end for
33:     end for
34:     $l \leftarrow l + 1$.
35: end while
```

D. NUMERICAL RESULTS

In this section, numerical simulations are used to evaluate the performance of the proposed algorithm. The main parameters chosen according to 3GPP standards [25] are listed in TABLE 1, where Exp(•) and Poisson(•) are the exponential and Poisson distributions with mean •, respectively.



(a)

(b)

(c)

FIGURE 4. Traffic drop rate with the increase of number of users with and without NS in the RAN domain, where the weight $w_1 = w_2 = 1$ and MAC in the two groups could use M-LWWF or M-LWDF method. (a): drop rate of users in the first group; (b): drop rate of users in the second group; (c): maximum drop rate.

TABLE 1. Main simulation parameters.

Parameter	Value	
Bandwidth: W	100 MHz	
Transmit power of the base station: $p_{i,j}W$	30 dBm	
Coverage of the base station where users distributed: $l_{i,j}$	circle with radius 750 m	
Path loss: $h_{i,j}$	$15.3 + 37.6\log_{10}(l)$ dB, l in meters	
Rayleigh fading: $g_{i,j}$	Exp(1)	
Power spectral density of Gaussian noise: σ^2	-174 dBm/Hz	
Smoothing factor: ξ	0.5	
Number of groups: M	2	
Group schedule time interval: ΔT	2 ms	
	1st group	2nd group
Number of users: m_i	$2m$	m
Position of users: $(x_{i,j}, y_{i,j})$	uniform distribution	uniform distribution
Subcarrier spacing: B_i	$\{15, 30\}$ kHz	$\{15, 30\}$ kHz
Slot length: Γ_i	$\{1, 0.5\}$ ms	$\{1, 0.5\}$ ms
Traffic weight: w_i	1	{1,2,3,4}
Buffer size	10000	10000
Traffic type	delay in-sensitive	delay sensitive
Delay threshold	–	2 ms
Number of arrived packets per 1 ms	Poisson(6)	Poisson(6)
Size of arrived packets	Poisson(300)	Poisson(30)

The algorithms with and without NS in the RAN domain are compared, which are defined as:

- w/o NS-15kHz-W: Without NS in the RAN domain, where the SCS is 15kHz and all users use M-LWWF method as equation (13) in two groups.
- w/o NS-15kHz-D: Without NS in the RAN domain, where the SCS is 15kHz and all users use M-LWDF method as equation (14) in two groups.
- w/o NS-30kHz-W: Without NS in the RAN domain, where the SCS is 30kHz and all users use M-LWWF method in two groups.
- w/o NS-30kHz-D: Without NS in the RAN domain, where the SCS is 30kHz and all users use M-LWDF method in two groups.
- w NS-15&30kHz-W&W: With NS in the RAN domain, where the SCS is 15kHz and 30kHz and users use M-LWWF and M-LWWF method in the first and second group respectively.
- w NS-15&30kHz-W&D: With NS in the RAN domain, where the SCS is 15kHz and 30kHz and users use M-LWWF and M-LWDF method in the first and second group respectively.

The user in the first group has delay in-sensitive traffic, which will be dropped once its buffer overflows. The user in the second group has delay sensitive traffic, which will be dropped once its buffer overflows or the traffic delay exceeds 2ms. Similar to equation (7), the traffic drop rate $\hat{P}_1$ and $\hat{P}_2$ are defined as the amount of traffic dropped divided by the amount of traffic dropped and transmitted in the first and second groups of all the simulation time. With weight w_1 and w_2 in equation (8), the weighted traffic drop rate are $w_1\hat{P}_1$ and $w_2\hat{P}_2$.

Using Monte Carlo simulation with three times initialization of the users' location and ten thousands slots, the traffic drop rate of the two groups with weight $w_1 = w_2 = 1$ with different algorithms are shown in Fig.4. It could be seen in Fig.4a and Fig.4b that without NS in the RAN domain, the M-LWWF method prefers to allocate resources to users in the first group, which have more cached traffic volume with more users and larger size of arrived packets as shown in TABLE 1. We note that with more than 80 users, almost all the resources are allocated to the first group, which highly increases the drop rate of the second group. The M-LWDF method takes the delay of the head-of-line traffic into consideration, which almost transmits all the data of the second group with smaller size of arrived packets, but increases the drop rate of users in the first group. We note that with smaller size of arrived packets, the second group hasn't exhaust the

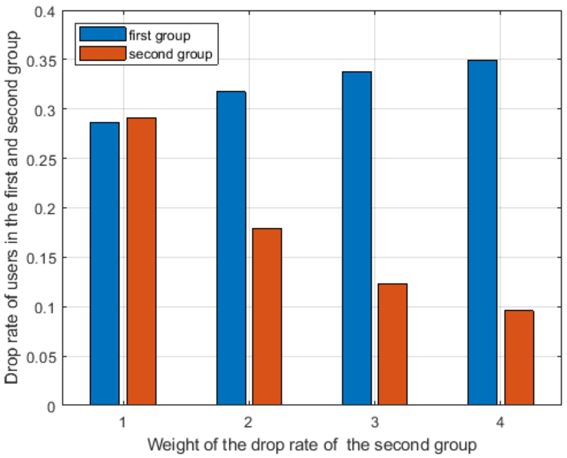


FIGURE 5. Using w NS-15&30kHz-W&W algorithm, the traffic drop rate of users in different groups with the weight $w_1 = 1$ and the increase of weight w_2.

resource, and the first group could have resource to transmit data. Thus, the two groups influence each other and the methods will benefit one at the sacrifice of the other. With NS in the RAN domain, either M-LWWF or M-LWDF achieves a balance between the two groups and highly decreases the maximum drop rate as shown in Fig.4c. Meanwhile after slicing, the slave MAC prefers M-LWWF method in the two groups and w NS-15&30kHz-W&W is the best algorithm. This is because of that M-LWDF may allocate resource (the minimum size is one PRB) to users with only a little bit of data to transmit. Then part of the resource will be unloaded and wasted. Hence in the real system with NS in the RAN domain, M-LWWF method is suitable for both delay insensitive and delay sensitive slices instances.

Moreover, the weight w_1 and w_2 are used to ensure fairness between slice instances, where the slice instances with more critical drop rate requirement has larger weight value, and vice versa. Using w NS-15&30kHz-W&W algorithm, the traffic drop rate of users in two groups with the weight $w_1 = 1$ and weight $w_2 = \{1, 2, 3, 4\}$ are shown in Fig.5. It could be seen that, with w_2 increases, more resources will be allocated to the second group and the traffic drop rate of the first and second group increases and decreases respectively. The slight difference of the drop rate of the two groups with $w_1 = w_2 = 1$ is because of that, the algorithm achieves roughly but not accurately equal weighted drop rate of each master schedule time interval. Meanwhile, increasing the simulation time and number of initialization of the users' locations could reduce the difference.

Therefore, the system performance could be improved with NS in the RAN domain, whose improvement could also be studied from different perspectives, such as security and robustness. In the following, we propose our E2E NS system along with the demonstration.

V. E2E NS SYSTEM

The proposed E2E NS system including RAN, TN and CN domains, is consisted of infrastructure, control framework and management plane layers. This section describes the system components, impacting to RAN protocols, and the difference between eMBB, uRLLC and mMTC slices.

A. SYSTEM COMPONENTS

As discussed in Section II, the E2E NS system considers RAN, TN and CN domains, whose details are illustrated in Fig. 6. The RAN domain is the edge data center close to the users. The conventional RAN places radio frequency (RF), inter-media frequency (IF) and baseband functions on the outdoor tower. The distributed RAN places parts of the functions in the indoor as BBU and others on the outdoor tower as RRU [32]–[34]. The CN domain has a regional data center and a remote data center. The regional data center is placed close to the RAN to decrease latency. The TN domain is the connection between RAN and CN domains, which is transport fabric composed of routers.

Meanwhile, the E2E NS system has three layers. The infrastructure layer has PNFs and VNFs in the RAN domain, physical and virtual routers in the TN domain, common VNFs and special VNFs in the CN domain. For the RAN domain, the PNFs include spectrum, time, antenna, RRU pool, BBU pool, and CPU pool, etc, while VNFs contain digital filter, FFT/IFFT module, modulation/demodulation module, coding/decoding module, hybrid automatic repeat request (HARQ) function, MAC protocol, RLC protocol, and PDCP protocol, etc. [35], [36]. For the TN domain, the physical routers are the real hardware such as Huawei's router OTN9800, while the virtual routers are the ones generated by open source codes such as Open vSwitch (OVS) and Mininet [37], [38]. For the CN domain, the VNFs could be mobility management entity (MME), home subscriber server (HSS), and policy control and charging rules function (PCRF), etc. in LTE [39], [40], and authentication server function (AUSF), core access and mobility management function (AMF), structured data storage network function (SDSF), unstructured data storage network function (UDSF), network exposure function (NEF), NF repository function (NRF), policy control function (PCF), session management function (SMF), unified data management (UDM), user plane Function (UPF), and application function (AF), etc. in 5G [41], [42].

The control framework layer has multiple RSCs, TSCs and CSCs in the RAN, TN and CN domains. The major components in RSC are resource allocation and protocol custom modules. The resources allocation module distributes the hardware resources for slice instances, and reports the RAN resource status to the orchestrator. The protocol custom module specifies the VNFs for different slice instances, and reports the RAN VNFs status to the orchestrator. For example, HARQ and ARQ functions in acknowledged mode (AM) RLC protocol are critical for error-free transmission at the cost of delay [43]. For the eMBB slice instance with no error requirement, HARQ and ARQ functions may be its basic functions. But for the uRLLC slice instance with less than 1ms delay requirement, re-transmission greatly increases the

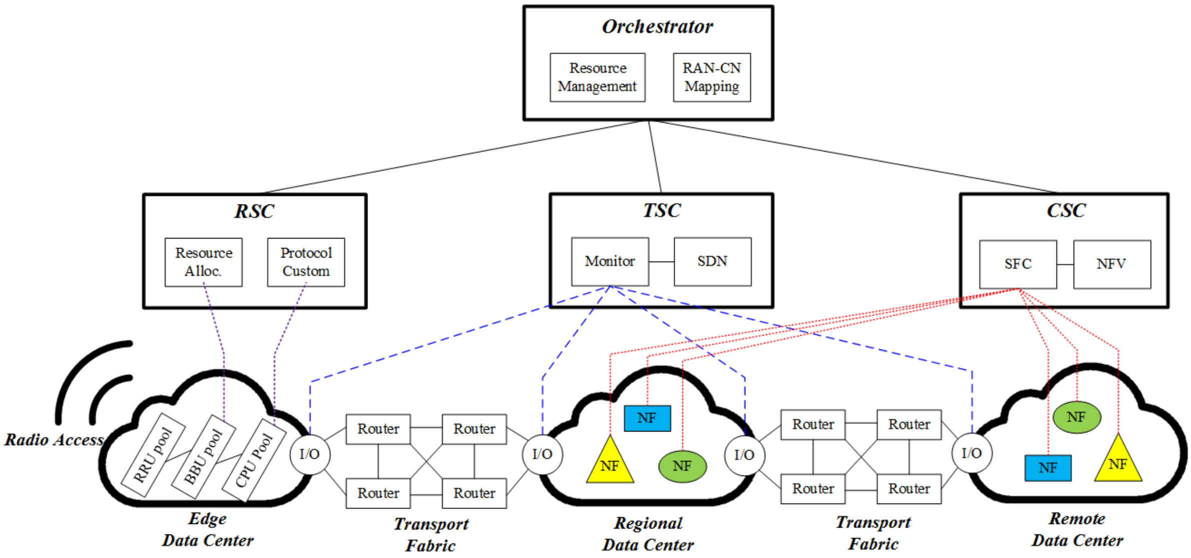


FIGURE 6. **Components of control framework.**

delay, and HARQ and ARQ functions will not be needed. The major components in TSC are monitor and SDN modules. The monitor module keeps on tracking traffic loads in the I/O ports at the edge data center and regional data center. The traffic status are also reported to the orchestrator for traffic balancing among routers. The SDN module creates and deletes routing paths among PNFs and VNFs. The major components in CSC are slice function chain (SFC) and NFV modules. The SFC module receives the function status reports from each VNF instance, and forwards the reports to the orchestrator. The NFV module creates and deletes VNF instances.

The management plane layer has the orchestrator, whose major components are resource management and RAN-CN mapping modules. The orchestrator receives business order from the north and reports from the south control framework layer. The resource management module records the resource usage status, and generates the commands for functions creation, delete and adjustment. The RAN-CN mapping module records the routing paths and status, and generates the commands for path creation, remove and adjustment. Then commands are sent to the control framework layer for execution. We note that, the commands could specify what to be done such as creating an unacknowledged mode (UM) RLC module on a specified BBU, or only specify the requirements such as creating an uRLLC slice instance with less than 0.4ms, 0.2ms and 0.4ms delay at RAN, TN and CN domains. With the requirements, the control framework layer decides where and how to create the slice instance.

In the existing 5G system architecture [41], [44], network slice instance (NSI) management plane has communication service management function (CSMF) and network slice management function (NSMF). The CSMF receives business order and translates to network slice requirements, such as capacity, throughput, delay, and etc.. The NSMF manages resources and generates commands for functions creation, delete and adjustment. The management plane in this paper involves features of the NSI management plane in 5G, and also extends the management to RAN and TN domains. Meanwhile, additional control framework layer has RSC, TSC and CSC modules in RAN, TN and CN domains. The control framework layer executes the commands from and collects resource and slice instances status to the management layer. We note that for security consideration, the carrier operator's system may contain RAN, TN and CN hardware from different telecommunications equipment manufacturers. However, it is difficult for the control framework layer from one manufacturer to control and collect status and data from the hardware developed by other manufacturers. Hence this paper prefers to have dedicated control modules for RAN, TN and CN domains, and their interfaces are left to be standardized in future 6G network to achieve the E2E slicing feature.

B. IMPACTING TO RAN PROTOCOLS

For traditional RAN, there are PDCP, RLC and MAC protocols. The major functions in PDCP are header compression, ciphering and integrity protection. In RLC, the data is transferred in one of three modes named AM, UM and transparent mode (TM). The AM and UM have concatenation, segmentation, reassembly, reordering, duplicate detection, data discard functions. Besides, the AM has ARQ and re-segmentation functions. The TM transfers data transparently. The major function in MAC is resource allocation, which schedules and allocates resource to multiple logical channels [9], [43], [45].

With NS in RAN, there could be a specific bandwidth as anchor for synchronization. Then users monitor/scan the bandwidth to receive the synchronization sequence for synchronization. Meanwhile, the major control information could be transmitted in a common control slice, which includes the slice configuration information. The common control slice could utilize the synchronization resource or other dedicated resource.

After synchronization and configuration, the slice instances could be created. For eMBB slice, PDCP, RLC and MAC protocols are similar to the traditional ones, and HARQ and AM RLC could be used to guarantee the successful delivery of the packets. For uRLLC slice, with long transmission distance such as remote surgery, dedicated and redundant resources in the RAN, CN and TN domains could be allocated to satisfy the delay and jitter requirements. With short transmission distance such as traffic control, we recommend to place the application along with RAN functions to reduce the latency. Moreover, short sub-frame length already adopted in 5G could be used to reduce the latency [11]. If the delay requirement is extremely strict, HARQ, AM RLC and even PDCP layer are not suitable. Directly connecting the application to the MAC interface may be the most time-saving scheme. Meanwhile, a timestamp could be transmitted along with the data for nodes in the path or destination to assess the timeliness of the packets and drop the time-out ones. For random access (RA), there could be a pre-established DRB pool. Once there is a slice requirement, the DRB could be immediately allocated to save time. For mMTC slice, the major requirement of RAN is the collection of the short packets. In PDCP, header compression is necessary and could be enhanced with packet aggregation to aggregate packets from/to the same source/destination, which could avoid plenty of short transmission control protocol (TCP) or user datagram protocol (UDP) packets and save resource in CN and TN domains. Ciphering and integrity protection could also be removed, where the OTT applications design their own protocols. Moreover, grant free scheme instead of traditional random access scheme is more suitable. The later one uses too much resource to set up a dedicated DRB, but transmit a few short packets. The former one could transmit the short packets, user id along with preamble in random access channel (RACH) to save resource.

C. EXAMPLES AND DISCUSSIONS

1) STADIUM EMBB SLICE

Time tidal effect of bandwidth is significant in the stadium. During a hot sporting event, there are tens of thousands of people cheering and communicating together. While it is almost empty in the free time. A temporary eMBB slice is a suitable method for stadium compared to fix deployed ultra-dense Wi-Fi or small cellular methods.

Orchestrator gets a business order with information: geographical position of stadium, minimum open-air resource request and the duration of sporting events. Firstly, the resource management module in orchestrator selects some base stations to cover this stadium and allocates enough open-air resource during the sporting event. Secondly, the RSC configures the open-air resource parameters in physical layer, and deploys enough baseband functions and protocol programs on the selected base stations. Thirdly, the CSC deploys empty atom network functions in the regional data center for the sporting event. Then, these atoms are configured for authentication, mobility management, and policy control functions, etc. After receiving acknowledges from RSC and CSC, orchestrator triggers TSC to active transport link between RAN and CN. Finally, the RSC, TSC and CSC keep on tracking the status of functions and routing links, and report to the orchestrator periodically to trigger network breathing once needed.

2) REMOTE SURGERY URLLC SLICE

Ultra-low latency and predictable jitter are the basic requirements for remote surgery. The incompressible propagation delay is the distance divided by the speed of light. Here, ultra-low latency points to the margin of requested latency minus incompressible propagation delay. In the practice, we find that main latency is caused by TCP/UDP layer in the wireline link of TN, and serious jitter is accompanied by TCP congestion.

Orchestrator gets a business order with information: geographical position of source and destination, latency and jitter requirements, the link type of last hop (wireless or wireline). Firstly, orchestrator notices user terminals to run multipath TCP (MPTCP) protocol for the scenario of remote surgery. Some middle boxes as the anchor between two MPTCP-enabled user terminals are deployed in edge data centers nearby the users. Then, TSC calculates and creates routing path for each path of MPTCP to meet the latency and jitter requirements. In our solution, the private buffers in all physical or virtual routers along these paths are allocated for the uRLLC traffic. If the link type of last hop is wireless, RSC will creates a private *radio spectrum slicing* for the scenario of remote surgery. Short sub-frame length supported by 5G is suitable for low latency request and is verified on our demonstration. Without traffic going through the CN domain, the CSC is almost useless for remote surgery.

3) METER READING MMTC SLICE

There are a large number of sensors in downtown, which are on idle state in most of time. The data of meter reading is delay insensitive and low data rate. However, the LTE/LTE-A system needs to do the whole authentication process in CN for each sensor. And the router should look up the routing table in TN for each small packet. Simplify authentication process and efficient small packet aggregation are the main improvement in our mMTC slice depending on control framework.

Orchestrator has the priori information about coverage area and number of sensors. Firstly, a narrow bandwidth *radio spectrum slicing* is created by the resource allocator module in RSC. In our solution, protocol customer module in RSC configures no authentication model in RAN and SFC module in CSC configures no session model in CN. For the scenario of meter reading, the security of data is supplied by the application layer instead of link layer. Then, the CSC deploys a lot of packet aggregation middle boxes in the edge and regional data centers nearby coverage area. Finally, the spectrum is released and middle boxes are removed after the data acquisition from all sensors. And the orchestrator refreshes its available resource table for next request of slice instance. Without specific routing path, the TSC is almost useless for meter reading.

Besides the above, the control elements of framework (RSC, TSC and CSC) could also be flexibly combined to satisfy varieties of use cases by the following methods in our solution:

- *Blueprint* - The spectrum allocation strategy for typical use cases have been calculated in advance and saved as blueprint by RSC. When new request of slice instance coming, RSC chooses a closest blueprint to configure the physical layer as soon as possible, and then uses a close-loop adaptive algorithm to adjust spectrum resource with feedback of practice KPI.
- *Decouple* - Each control element uses multiple processes to run algorithms for different slice instances. The identity of slice instance and context of its state machine are decoupled in our solution. Accordingly, some programming has been done to support this decouple effect of some communication protocols. The management of control elements, e.g. identity and deployment, is the duty of orchestrator.
- *Private* - We build up a private subnet to connect all the control elements and orchestrator, which only transmits control signal of NS. This subnet is consisted of private virtual routers, and is isolated with the data slices of eMBB/uRLLC/mMTC. Control elements and subnet are collectively called *common control slice* in our solution.

VI. E2E NS DEMONSTRATION

A demonstration has been developed to verify the architecture of the proposed E2E NS system, where NS is adopted not only in the TN and CN domains, but also in RAN domain.

4) HARDWARE AND SOFTWARE

The hardware is shown in Fig.7, and the software of orchestrator, RSC, CSC, TSC and NFs are developed using C++ and Python languages. In UE and the RAN domain, the physical layer baseband functions are implemented on FPGA and DSP following the 5G standards. Besides, NS is supported with configurable characters and additional filter OFDM techniques [46], [47]. The MAC, RLC and PDCP protocols are based on the open-source OpenAirInterface using Huawei FusionServer 2488, and modified to be seamlessly compatible with the self-designed baseband functions. In the TN domain, the physical router is Huawei's OTN9800, and the virtual routers are generated by open source Open vSwitch (OVS) and Mininet. In the CN domain, one kind of vEPC software is adopted. Furthermore, lots of enhanced network functions are self-developed to meet the special KPI of eMBB/uRLLC/mMTC. Third party testing instruments are used to verify the results, such as Rohde&Schwarz spectrum analyzer for radio spectrum analysis and Spirent TestCenter for latency monitoring.

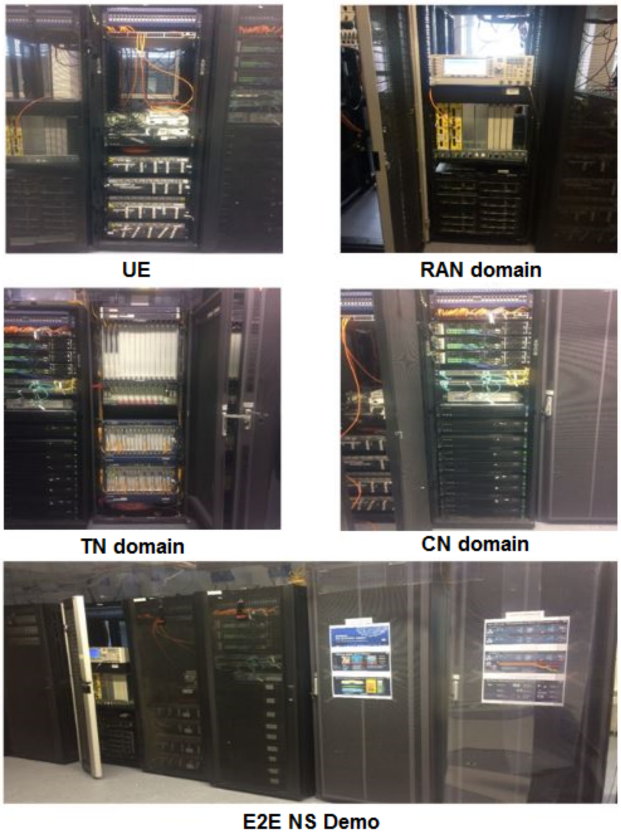


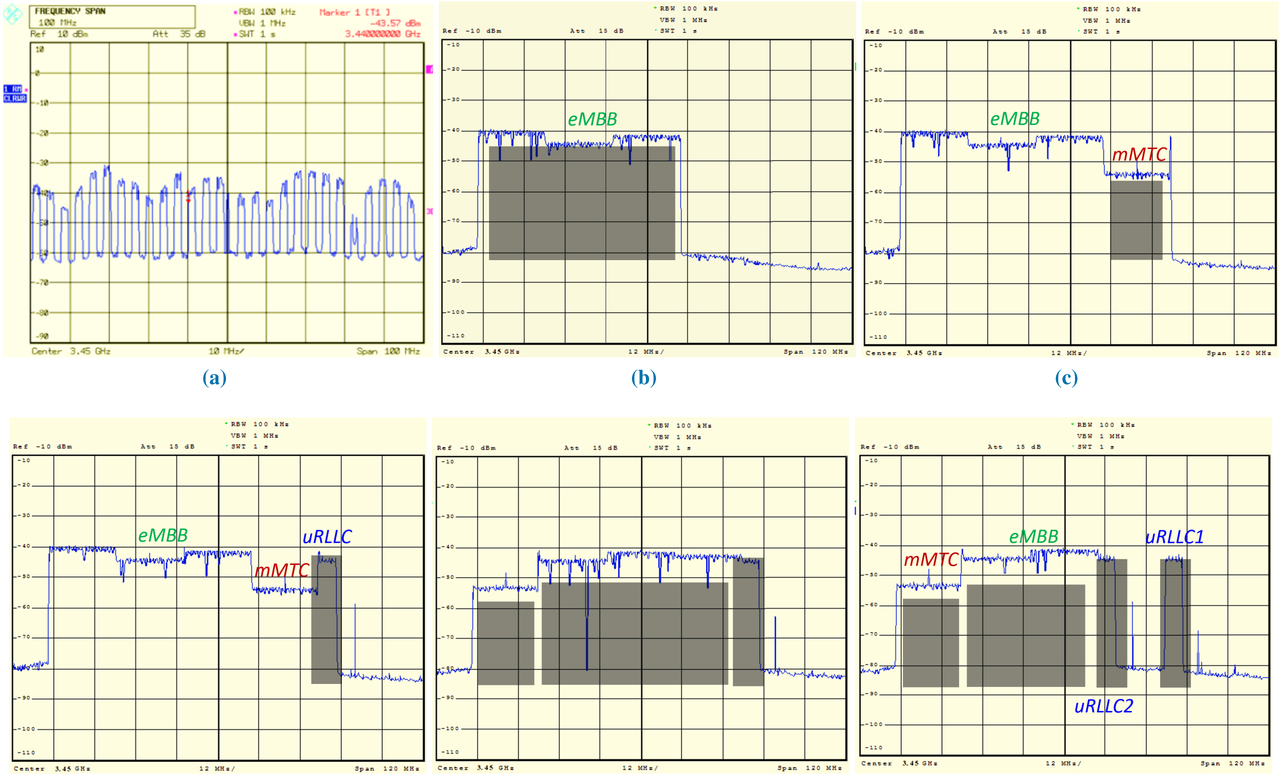
FIGURE 7. E2E NS demonstration.

5) FINE SPECTRAL GRANULARITY

Since NS is adopted in RAN domain, the spectrum are allocated to and isolated between different slice instances. The capability of the spectrum slicing is shown in Fig.8a, where 27 comb narrow bandwidth are realized on the total 100MHz spectrum at 3.45GHz central frequency. Our demonstration could support ultra-flexible spectrum pattern, with arbitrary number of the fine spectral granularity narrow bandwidth within the 100MHz spectrum. The fluctuation of signal amplitude of different isolated narrow bandwidth is caused by the raised cosine filter. Each narrow bandwidth could be demodulated and decoded independently, and the mapping between narrow bandwidth and slice instance could be flexibly configured by RSC.

6) SLICE CREATION AND DELETE

The slice creation are shown in Fig. 8b- d, where eMBB, mMTC and uRLLC slice instances are created sequentially. It could be seen that the eMBB slice instance uses the left 60MHz spectrum, and mMTC slice instance uses the middle 20MHz spectrum, and uRLLC slice uses the right 20MHz. The mMTC slice has lower transmit power than eMBB slice,



(a)

(b)

(c)

(d)

(e)

(f)

FIGURE 8. Captured spectrum from the Rohde&Schwarz spectrum analyzer equipment. (a): flexible spectrum capability with 27 isolated narrow bandwidth on the total 100MHz spectrum; (b): eMBB slice instance creation; (c): mMTC slice instance creation; (d): uRLLC slice instance; creation; (e) spectrum switch between eMBB and mMTC slice instances; (f): spectrum re-allocated from eMBB to the new uRLLC slice instance. *The x axis is the frequency with label as center 3.45GHz, 10MHz per division and span 100MHz for (a), and center 3.45GHz, 12MHz per division and span 120MHz for (b-f). The y axis is the power with label as 10dBm per division from -90dBm to 10dBm for (a), and 10dBm per division from -110dBm to 10dBm for (b-f). The legend has RBW 100kHz, VBW 1MHz, SWT 1s for (a-f), Ref 10dBm and Att 35 dB for (a) and Ref -10dBm and Att 15 dB for (b-f).*

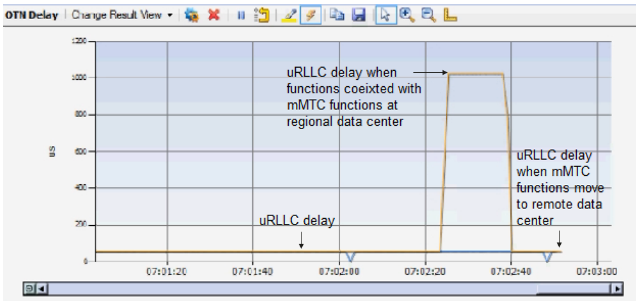

OTN Delay
Change Result View
1200
1000
800
600
400
200
0
us
07:01:20
07:01:40
07:02:00
07:02:20
07:02:40
07:03:00
uRLLC delay when functions coeixted with mMTC functions at regional data center
uRLLC delay when mMTC functions move to remote data center
uRLLC delay

FIGURE 9. Captured delay of uRLLC slice, which varies with the resources in regional data center changes.

and the captured uRLLC spectrum is its control channel. Meanwhile, NFs and routing paths for each slice instance are also created in the RAN, TN and CN domains. The slice delete follows the same way and omitted here, and the time for slice creation and delete in sub-minute time.

7) NETWORK BREATHING

Network breathing adjusts resource among slices to deal with traffic variation, unpredictable open-air interference and equipment damage, etc. A simple demonstration is shown in Fig. 8d and 8e, where the spectrum of eMBB and mMTC is switched. Another complicated demonstration is described as following. A user sends a request to establish a new uRLLC slice instance to orchestrator using the existed uplink wireless channel. The orchestrator verifies the authority and accepts the request. However, there is not sufficient spectrum resource, hence RSC re-allocate 20MHz spectrum from eMBB to the new uRLLC slice instance as shown in Fig. 8f in the RAN domain. Meanwhile, CSC deploys dedicated VNFs in the regional data center in the CN domain, and monitors its KPI. As captured by Spirent TestCenter in Fig. 9, when resources are enough in the regional data center, the uRLLC delay is low, but increases once the resources is not sufficient. The CSC detects the status and reports to orchestrator. The orchestrator recalculates and relocates mMTC VNFs, from regional to remote data center, and the uRLLC delay decreases. The reason to migrate mMTC VNFs is that, the mMTC slice has lower data rate and latency requirement. The time for network breathing is also in sub-minute time.

VII. CONCLUSION

In this paper, the E2E NS is proposed where NS is adopted in the RAN, TN and CN domains. The benefits with NS in the RAN domain, besides discussion, are numerically studied with a two-level resource allocation scheme. With sliced open-air spectrum, the isolation between slice instances is achieved, which dramatically decreases the maximum drop rate. Then the architecture, system components, RAN protocols, and demonstration of the E2E NS system are discussed. The E2E NS system has management plane layer, control framework layer and infrastructure layer, where each layer has orchestrator, RSC and TSC and CSC, functions and resources in the three domains, respectively. The demonstration with plenty of hardware and software shows the capability to support sliced open-air spectrum with very fine spectral granularity, and the slice creation, delete and breathing in sub-minute time, which could be used in the operator's network. For the sliced open-air spectrum, its benefits are numerically studied and feasibility is demonstrated with very fine spectral granularity. Meanwhile, deploying the proposed resource allocation scheme with sliced open-air spectrum in the real-time system is left for future work, where protocol standardization and hardware acceleration are crucial to reduce the response time to microseconds.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests regarding the publication of this paper.

REFERENCES

[1] *5G System Architecture for the 5G System*, document TS 23.501, Version 16.0.2, Release 16, 3GPP, Apr. 2019.

[2] *Description of Network Slicing Concept*, NGMN Alliance, Frankfurt, Germany, Jan. 2016.

[3] R. Inam, A. Karapantelakis, K. Vandikas, L. Mokrushin, A. V. Feljan, and E. Fersman, "Towards automated service-oriented lifecycle management for 5G networks," in *Proc. IEEE 20th Conf. Emerg. Technol. Factory Automat. (ETFA)*, Sep. 2015, pp. 1–8.

[4] R. Pries, H.-J. Morper, N. Galambosi, and M. Jarschel, "Network as S service—A demo on 5G network slicing," in *Proc. 28th Int. Teletraffic Congr.*, 2016, pp. 209–211.

[5] *ECOMP Architecture White Paper*, AT&T Inc., Dallas, TX, USA, 2016.

[6] E. Haleplidis, K. Pentikousis, S. Denazis, H. Salim, D. Meyer, and O. Koufopavlou, *Sofeware-Defined Networking (SDN): Layers and Architecture Terminology*, document RFC 7426, IETF, Jan. 2015.

[7] *Network Functions Virtualisation—Introductory White Paper*, ETSI, Darmstadt, Germany, Oct. 2012.

[8] *LTE E-UTRA Radio Resource Control (RRC) Protocol Specification*, document TS 36.331, Version 15.4.0, Release 16, 3GPP, Feb. 2018.

[9] *LTE E-UTRA Medium Access Control (MAC) Protocol Specification*, document TS 36.321, Version 15.6.0, Release 15, 3GPP, Jun. 2019.

[10] *LTE E-UTRA Physical Channel and Modulation*, document TS 36.211, Version 11.5.0, Release 11, 3GPP, Jan. 2014.

[11] *5G-NR Physical Channel and Modulation*, document TS 38.211, Version 15.5.0, Release 15, 3GPP, Apr. 2019.

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

[13] S. Khatibi, L. Caeiro, L. S. Ferreira, L. M. Correia1, and N. Nikaein, "Modelling and implementation of virtual radio resources management for 5G cloud RAN," *EURASIP J. Wireless Commun. Netw.*, vol. 1, no. 1, pp. 128–143, 2017.

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

[15] Y. K. Tun, N. H. Tran, D. T. Ngo, S. R. Pandey, Z. Han, and C. S. Hong, "Wireless network slicing: Generalized kelly mechanism based resource allocation," 2019, *arXiv:1907.02182v2*, [Online]. Available: https://arxiv.org/abs/1907.02182v2

[16] R. Kokku, R. Mahindra, H. Zhang, and S. Rangarajan, "NVS: A substrate for virtualizing wireless resources in cellular networks," *IEEE/ACM Trans. Netw.*, vol. 20, no. 5, pp. 1333–1346, Oct. 2012.

[17] S. Gendy and Y. Gadallah, "LTE-based network virtualization schemes adaptation for M2M deployments," in *Proc. IEEE Int. Black Sea Conf. Commun. Netw. (BlackSeaCom)*, Sochi, Russia, Jun. 2019, pp. 1–3.

[18] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar, and P. Whiting, "CDMA data QoS scheduling on the forward link with variable channel conditions," *Bell Labs Tech. Memorandum*, vol. 4, pp. 1–45, Jan. 2000.

[19] J. Y. Lee, S. Sorour, S. Valaee, and W. Park, "Dynamic parameter adaptation for M-LWDF/M-LWWF scheduling," *IEEE Trans. Wireless Commun.*, vol. 11, no. 3, pp. 927–937, Mar. 2012.

[20] I. da Silva, "Impact of network slicing on 5G radio access networks," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Athens, Greece, Jun. 2016, pp. 153–157.

[21] S. E. Elayoubi, S. B. Jemaa, Z. Altman, and A. Galindo-Serrano, "5G RAN slicing for verticals: Enablers and challenges," *IEEE Commun. Mag.*, vol. 57, no. 1, pp. 28–34, Jan. 2019.

[22] S. D'Oro, F. Restuccia, T. Melodia, and S. Palazzo, "Low-complexity distributed radio access network slicing: Algorithms and experimental results," *IEEE/ACM Trans. Netw.*, vol. 26, no. 6, pp. 2815–2828, Dec. 2018.

[23] N. Salhab, S. E. Falou, R. Rahim, S. E. E. Ayoubi, and R. Langar, "Optimization of the implementation of network slicing in 5G RAN," in *Proc. IEEE Middle East North Afr. Commun. Conf. (MENACOMM)*, Jounieh, Lebanon, Apr. 2018, pp. 1–6.

[24] Y. Tsukamoto, R. K. Saha, S. Nanba, and K. Nishimura, "Experimental evaluation of RAN slicing architecture with flexibly located functional components of base station according to diverse 5G services," *IEEE Access*, vol. 7, pp. 76470–76479, 2019.

[25] *LTE E-UTRA Radio Frequency (RF) Requirements for LTE Pico Node B*, document TR 36.931, Version 11.0.0, Release 11, 3GPP, Sep. 2012.

[26] X. Li, D. Guo, J. Grosspietsch, H. Yin, and G. Wei, "Maximizing mobile coverage via optimal deployment of base stations and relays," *IEEE Trans. Veh. Technol.*, vol. 65, no. 7, pp. 5060–5072, Jul. 2016.

[27] X. Chen, X. Li, D. Guo, and J. Grosspietsch, "Resource allocation in public safety broadband networks with rapid-deployment access points," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1660–1671, Feb. 2018.

[28] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *Proc. IEEE Veh. Technol. Conf.*, Tokyo, Japan, May 2000, pp. 1854–1858.

[29] F. Afroz, K. Sandrasegaran, and P. Ghosal, "Performance analysis of PF, M-LWDF and EXP/PF packet scheduling algorithms in 3GPP LTE downlink," in *Proc. Australas. Telecommun. Netw. Appl. Conf.*, South Wharf, Australia, Nov. 2014, pp. 87–92.

[30] Y.-L. Chou, *Statistical Analysis*. Salem, OR, USA: Holt International, 1975, sec. 17.9.

[31] D. N. C. Tse and P. Viswanath, *Fundamentals of Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[32] I. A. Alimi, A. L. Teixeira, and P. P. Monteiro, "Toward an efficient C-RAN optical fronthaul for the future networks: A tutorial on technologies, requirements, challenges, and solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 708–769, Nov. 2018.

[33] L. M. P. Larsen, A. Checko, and H. L. Christiansen, "A survey of the functional splits proposed for 5G mobile crosshaul networks," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 146–172, Oct. 2019.

[34] S. Sarmiento, J. A. Altabas, S. Spadaro, and J. A. Lazaro, "Experimental assessment of 10 Gbps 5G multicarrier waveforms for high-layer split U-DWDM-PON-based fronthaul," *J. Lightw. Technol.*, vol. 37, no. 10, pp. 2344–2351, May 15, 2019.

[35] H. da Silva, L. M. Correia, and P. Costa, "Design of C-RAN fronthaul for existing LTE networks," M.S. thesis, Sci. Degree Electr. Comput. Eng., Tecnico Lisboa, Lisboa, Portugal, pp. 1–112, 2016.

[36] *RAN Architecture Components ĺC Final Report*, document H2020-ICT-2014-2 5G NORMA/D4.2, 2017.

[37] R. Pang, H. Li, G. Wang, Y. Ji, X. Man, and S. Shen, "An end-to-end IP and optical collaborative solution for 5G transport network," in *Proc. Asia Commun. Photon. Conf.*, Hangzhou, China, Oct. 2018, pp. 1–3.

[38] J. S. Wey and J. Zhang, "Passive optical networks for 5G transport: Technology and standards," *J. Lightw. Technol.*, vol. 37, no. 12, pp. 2830–2837, Jun. 15, 2019.

[39] A. Basta, W. Kellerer, M. Hoffmann, H. J. Morper, and K. Hoffmann, "Applying NFV and SDN to LTE mobile core gateways the functions placement problem," in *Proc. 4th Workshop All Things Cellular*, 2014, pp. 33–38.

[40] R. Mijumbi, J. Serrat, J.-L. Gorricho, N. Bouten, F. De Turck, and R. Boutaba, "Network function virtualization: State-of-the-art and research challenges," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 236–262, 2016.

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

[42] D. Sattar and A. Matrawy, "Optimal slice allocation in 5G core networks," *IEEE Netw. Lett.*, vol. 1, no. 2, pp. 48–51, Jun. 2019.

[43] *LTE E-URTA Radio Link Control (RLC) Protocol Specification*, document TS 36.322, Version 15.2.0, Release 15, 3GPP, Jun. 2019.

[44] A. Kaloxylos, "A survey and an analysis of network slicing in 5G networks," *IEEE Commun. Stand. Mag.*, vol. 2, no. 1, pp. 60–65, Mar. 2018.

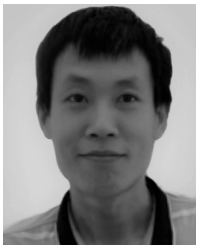
[45] *LTE E-UTRA Packet Data Convergence Protocol (PDCP) Specification*, document TS 36.323, Version 15.4.0, Release 15, 3GPP, Jun. 2019.

[46] J. Abdoli, M. Jia, and J. Ma, "Filtered OFDM: A new waveform for future wireless systems," in *Proc. IEEE 16th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Stockholm, Sweden, Jun./Jul. 2015, pp. 66–70.

[47] P. Guan, D. Wu, T. Tian, J. Zhou, X. Zhang, L. Gu, A. Benjebbour, M. Iwabuchi, and Y. Kishiyama, "5G field trials: OFDM-based waveforms and mixed numerologies," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1234–1243, Jun. 2017.

XU LI received the B.Sc. and Ph.D. degrees in electrical and electronics engineering from the University of Science and Technology of China (USTC), in 2010 and 2015, respectively. From 2013 to 2014, he was a visiting Ph.D. student with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA. Since 2015, he has been a Wireless Senior Engineer with the Central Research Institute, 2012 Laboratory of Huawei Tech. Company Ltd. His research involves various architectures of radio access networks in 5G ear, especially end-to-end network slicing. He has a rich wireless research experience, including ultrawide band (UWB) chips, interference alignment, relay networks, stochastic geometry, and public safety wireless broadband networks. His current interests include microwave wireless communication, radio over fiber communication, and nonlinear compensation methods.

RUI NI received the B.Sc. and Ph.D. degrees in electrical and electronics engineering from the University of Science and Technology of China (USTC), in 2006 and 2011, respectively. Since 2011, he has been a Wireless Senior Engineer with the Central Research Institute, 2012 Laboratory of Huawei Tech. Company Ltd. His research involves various architectures of radio access networks and core networks from 2G to 5G, especially network slicing in 5G era. He has a rich wireless network related experience in engineering practice. His current interests include orbital angular momentum, microwave quantum, and time reversal signal processing.



JUN CHEN received the M.Sc. and Ph.D. degrees in electrical engineering from the South China University of Technology, in 2010 and 2014, respectively. From November 2014 to June 2019, he has been with the Department of Central Research Institute of Huawei Technology at Shenzhen China as a Senior Engineer Researcher on ultradense networks (UDN), networking slicing of 5G, satellite communication, and interference cancellation in Bluetooth. Since June 2019, he was with the algorithm and technology development in global technology service at Dongguan Huawei, China. His research interests include multiple antenna systems, convex optimization in communications, and statistical signal processing.

YIBO LYU received the M.Sc. degree in information and signal processing from the Chongqing University of Posts and Telecommunications, in 2011, and the Ph.D. degree in circuits and systems from Xiamen University, in 2016. Since 2017, he has been a Wireless Engineer with the Central Research Institute, 2012 Laboratory of Huawei Tech. Company Ltd. His research interests include channel coding, chaotic communications, radio over fiber communication, and nonlinear compensation methods.

ZHICHAO RONG received the B.Sc. and M.Sc. degrees in electrical and electronic engineering from the University of Liverpool, U.K., in 2011 and 2012, respectively, and the Ph.D. degree in engineering from the University of Warwick, U.K., in 2018. In 2013, he was a System Engineer with Seagate Technology, Suzhou, China. He is currently a Senior Engineer with the Central Research Institute, 2012 Laboratory of Huawei Tech. Company Ltd. His research interests include wireless communications, terahertz communications, and nanoscale networks.

RUI DU received the M.S. and Ph.D. degrees in information and communication engineering from Northwestern Polytechnical University, in 2014 and 2018, respectively. From 2015 to 2017, he was a visiting Ph.D. student with the Microwave Integrated Systems Laboratory, University of Birmingham. He joined Wireless Technology Laboratory, Huawei, in 2018, as Senior Research Engineer. His research interests include wireless communication, the Internet of Things, radar target identification and classification, and array signal processing.