SLA Decomposition for Sustainable End-to-End Multi-Domain Multi-Technology Network Slicing

Haitham H. Esmat and Beatriz Lorenzo

Abstract

The sixth generation (6G) networks will integrate many autonomous networks operating in different administrative domains and wireless technologies. Hence, service provisioning involves negotiation between multiple stakeholders (e.g., users, operators, manufacturers) to reach a service level agreement (SLA). To ensure its sustainability under dynamic network conditions, the SLA must be dynamically decomposed into portions each domain can support. Failure to do so will result in SLA violations and user dissatisfaction. In addition, management and orchestration schemes for SLA decomposition are needed to facilitate mapping service requirements to the most suitable domains, providers, and technologies, and perform SLA lifetime monitoring. However, SLA decomposition is challenging with management decisions at different levels and interdependent performance implications. Despite the relevance of this topic, it remains unexplored. This article aims to fill this gap and presents an SLA decomposition management and orchestration architecture for multi-domain, multi-technology networks that supports centralized, semi-distributed, and fully distributed implementation. The proposed architecture is well aligned with the guidelines provided by standardization bodies. We evaluate its performance for typical 6G use cases and show that our approach achieves up to six times higher reward with half the cost of existing schemes. Finally, remaining challenges and promising future research directions are outlined.

Introduction

6G networks are envisioned to be highly heterogeneous regarding services, operation environments, spectrum bands, and providers, making service management more complex [1]. Expected 6G applications [2, 3], like multisensory extended reality, teleoperated driving, and remote monitoring, have different requirements in terms of data rates, latency, reliability, computing, and storage. Network slicing [4] is a promising approach to meet diverse application requirements by creating logically isolated virtual networks, that is, slices, on top of the physical network. Network slicing decisions must select the most suitable wireless technology (e.g., WiFi, millimeter wave [mmWave]), spectrum band (e.g., sub-6 GHz, THz band), and computing resources (e.g., fog, edge, or cloud) together with energy and other cost-efficient considerations. To this end, it is crucial to specify the application requirements and service options in the SLA. The SLA is a contract between the user and a service provider (SP) that states the service guarantees, possible failures, and corresponding indemnification.

As networks become more heterogeneous and dynamic, SLA decomposition is crucial to achieving sustainable network slicing, and user satisfaction. Multiple providers may collaborate across administrative domains to enable the necessary communication, computing, and caching resources. Therefore, the SLA must be decomposed into portions each domain, provider, and technology can support. De Vleeschauwer *et al.* [5] study SLA decomposition in a static network under the assumption of independent acceptance probabilities of SLA portions per domain. However, in practice, domains may collaborate to achieve the end-to-end (e2e) SLA, and the SLA decomposition must be dynamically managed to allocate the necessary resources.

Several works have studied multi-domain network slicing (NS) but ignored SLA management and decomposition, as summarized in Table 1, which is the focus of this article. In [1], a modular architecture with in-slice embedded management and orchestration is developed to simplify the composition of multi-domain slices independently of the available technology. Li *et al.* [6] consider scaling for applications, services, and resources to meet a given SLA. Abbas *et al.* [9] present an intent-based approach to automate the management and orchestration of multi-domain network slices with no collaboration between domains. These works lack the flexibility to use network resources efficiently across domains, technologies, and operators.

A few works study SLA decomposition in single-domain networks [10, 11]. Chen *et al.* [10] study SLA management in cloud computing and translate high-level service-level objectives to low-level system-level thresholds to determine the resources needed based on system performance bounds. Kapassa *et al.* [11] propose an SLA management framework in single-domain slicing based on Artificial Neural Networks (ANNs) to estimate the resources needed to meet the SLA. Qureshi *et al.* [12] describe the challenges of assuring the

H. H. Esmat and B. Lorenzo (corresponding author) are with the University of Massachusetts Amherst, USA.

Digital Object Identifier: 10.1109/MWC.008.2300157

1536-1284/24/$25.00 © 2024 IEEE

	NS	Multi-domain networks	6G services	e2e SLA	Resources-aware SLA decomposition	Cost-aware offloading decisions	Intra-domain routing	Inter-domain routing	Radio, computing, storage resource allocation
[1]	✓	✓	✓	×	×	×	×	×	✓
[3]	✓	×	✓	✓	×	✓	×	×	✓
[5]	✓	✓	×	✓	×	×	×	×	×
[6]	✓	✓	×	✓	×	×	×	×	✓
[7]	✓	✓	×	×	×	×	✓	✓	×
[8]	✓	✓	×	×	×	×	×	×	✓
Proposed work	✓	✓	✓	✓	✓	✓	✓	✓	✓

TABLE 1. Literature review.

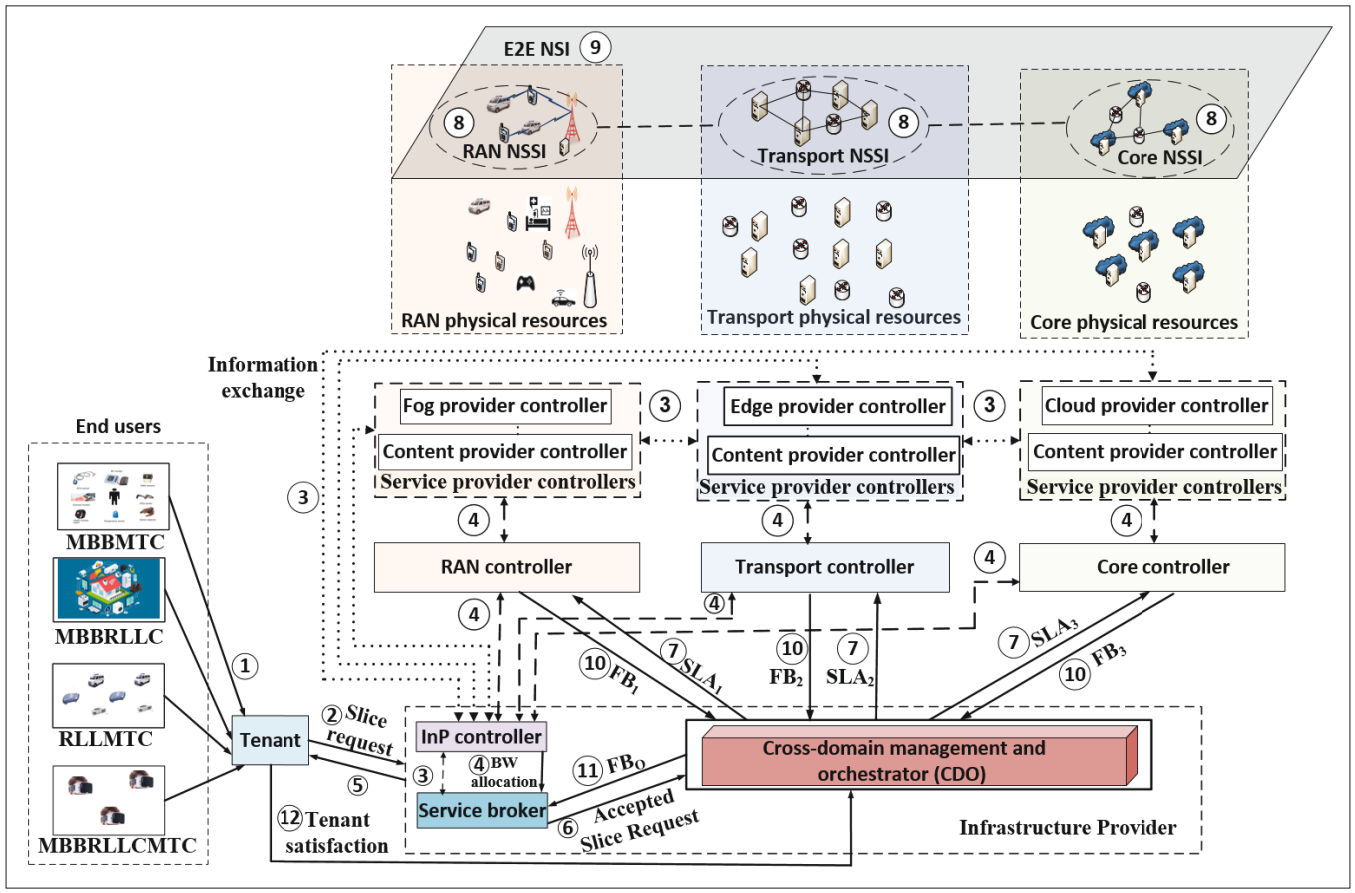

FIGURE 1. Multi-domain Multi-technology Management and Orchestration scheme.

SLAs in healthcare and the risks associated with service degradation, delay, and disruption. Indeed, the design of multi-domain, multi-technology network slicing schemes is challenging due to the coordination of multiple entities for resource sharing, pricing, management, control, and evaluation of the SLA.

This article pioneers SLA decomposition management and orchestration across multiple domains, technologies, and providers toward sustainable network slicing. Our proposed scheme admits centralized, semi-distributed, and fully distributed implementation. We present possible solutions and a case study using typical 6G use cases. Finally, the main challenges to achieving sustainable network slicing and promising future research directions are discussed.

Network Slicing in Beyond 5G and 6G Networks

Multi-Domain Multi-Technology Network Slicing

Network slicing is a virtualization technique where the infrastructure is shared by multiple tenants (operators) to serve different traffic classes simultaneously with an agreed-upon SLA, [2, 3] as shown in Fig. 1. It is supported by software-defined networks (SDN) that facilitate centralized network management and network function virtualization (NFV) techniques to virtualize network resources. Network slicing has played a key role in 5G networks [6, 9] to support diversified services for many vertical industries (e.g., industrial automa-

tion, healthcare, smart cities), and will continue to do so beyond 5G and 6G networks [2, 3]. In these networks, the infrastructure includes multiple cell sizes (e.g., Femto, Pico, Micro, Macro) [13], multiple technologies (e.g., cellular, WiFi, mmWave), spectrum bands (e.g., sub-6 GHz, THz), and administrative domains managed by different providers (e.g., communication/computing/content providers) to serve new applications.

The main principle of network slicing is constructing logically isolated network slices on top of the physical infrastructure. In multi-domain network slicing, each logical network slice instance (NSI) is an independent virtual network created through multiple network slice subnet instances (NSSIs), each belonging to a different domain, such as Radio Access Network (RAN), transport, and core. An NSI contains the necessary services and resources to meet the requirements for each traffic class, as defined in the service level agreement (SLA). Therefore, the e2e SLA associated with a slice, needs to be decomposed into portions attributed to each of these domains.

SLA decomposition has been acknowledged as one of the main challenges in resource allocation for network slicing [3, 5] since it involves management decisions at different levels with interdependent performance implications. A multi-domain, multi-technology network slicing scheme is presented in Fig. 1, which consists of the following entities.

Infrastructure Provider (InP): owns the physical infrastructure (e.g., base stations, access points, routers) and provides communication resources to serve the demand from tenants. Different domains may have different InPs that collaborate to create the NSIs on the shared physical network based on the e2e SLA requirements.

Computing and Content Providers: facilitate the computing resources and data needed in different domains. For instance, an intelligent traffic light system application in which users' data is collected in real-time for accident prevention requires fog computing resources for a quick response. However, data collection for traffic light management across a smart city needs edge computing resources to execute near real-time data. In contrast, the evaluation and improvement of the overall traffic light system require cloud computing for the collection and execution of large amounts of data delay-tolerant.

Tenants: request slices to serve their users' requirements. We consider the following 6G traffic classes [2, 3]:

- Mobile Broadband Machine Type Communication (MBBMTC), for example, remote pervasive monitoring, supports high broadband data rates and massive connectivity
- Mobile Broadband Reliable Low Latency Communication (MBBRLLC), for example, smart city applications, offers high broadband data rates and reliable and low latency communication
- Reliable Low Latency Machine Type Communication (RLLMTC), for example, teleoperated driving, supports massive connectivity, reliability, and low latency
- Mobile Broadband and Reliable Low Latency Machine Type Communication (MBBRLLMTC), for example, immersive VR video transmission, supports high data rates, reliability, low latency, and massive connectivity.

End Users: run their applications on the slices their operator (tenant) provided. If the requirements are dynamic, the SLA must specify that the number of resources may vary. Cloud providers can scale up or down the service to meet the resource demand. For instance, an e-health application may require low-latency communication when performing telesurgery. Once the surgery terminates, the demand for resources will change, as specified in the SLA.

Service Broker: interacts with the tenants and potential providers able to facilitate the resources and services for the slice request and negotiates with them the price for the slice.

Cross-Domain Management and Orchestrator (CDO): decomposes the SLA and coordinates with the domain controllers to achieve their part of the SLA. The CDO leverages domain knowledge to find a sustainable SLA decomposition.

Domain Controllers: every domain must collect the data to implement the necessary functions. The domain controllers interact with SPs' controllers (e.g., computing provider controller, content provider controller) and map the SLA requirements to physical resources facilitated by different SPs. In addition, they must determine the resource allocation and scheduling policies that ensure slice isolation between slices using the same or different technologies. This is crucial to meet throughput, delay, and reliability requirements and avoid wasting resources.

Given the heterogeneity of services and applications, network slicing schemes in beyond 5G and 6G networks [2, 3] must include admission control policies per-slice SLA and rules to decompose and manage the SLA.

Service Level Agreement

The SLA is a contract between the tenant and the SP(s) that the latter has to guarantee to avoid user dissatisfaction with the service and affecting the business model. In legacy networks, SLAs were defined based on the fixed quality of service requirements, and one provider facilitated the service. However, in 6G networks [1–3], service is provided by a collaboration between multiple stakeholders that span multiple domains and network technologies, and service requirements may be dynamic. These new characteristics render former SLA specifications obsolete. Therefore, SLAs must include the following specifications:

- Service requirements (e.g., required bandwidth, mean time to service recovery, ...)
- Service aspects (e.g., multiple band selection, multiple access point selection, edge computing)
- Design aspects (e.g., availability, reconfigurability, security, robustness)
- Legal aspects to identify which party is responsible for the service degradation and the corresponding compensation. In addition, to guarantee the SLA requirements in a sustainable way throughout the slice lifecycle, the following phases are needed:

Definition and Negotiation: The process of SLA definition and negotiation consists of the next steps, as illustrated in Fig. 1.

Step 1: The tenant receives their users' traffic requests (e.g., MBBMTC, MBBRLLC, RLLMTC, MBBRLLMTC) with different performance requirements. Possible requirements for these traffic classes are given in [3] Table 2.

The main principle of network slicing is constructing logically isolated network slices on top of the physical infrastructure. In multi-domain network slicing, each logical network slice instance is an independent virtual network created through multiple network slice subnet instances, each belonging to a different domain, such as Radio Access Network, transport, and core

Class	Throughput (Importance)	Delay (Importance)	Reliability (%)	No. users/ slice	Fog data size (KB)	Edge data size (KB)	Cloud data size (KB)	Fog computing resources (cycles/bit)	Edge computing resources (cycles/bit)	Cloud computing resources (cycles/bit)	Technology
MBBMTC	43 Mb/s (0.8)	250 msec (0.2)	99	11	0.15	0.25	0.35	20	30	40	mmWave
MBBRLLC	18 Mb/s (0.5)	50 msec (0.5)	99.99	8	0.8	1	1.2	80	100	120	mmWave
RLLMTC	640 Kb/s (0.2)	5 msec (0.8)	99.999	14	0.4	0.536	0.6	170	200	230	Microwave
MBBRLLMTC	36 Mb/s (0.4)	30 msec (0.6)	99.9	5	1	1.25	1.5	120	150	180	THz

TABLE 2. Requirements for each traffic class.

Step 2: The tenant requests the slice to the InP and negotiates with the service broker the SLA and the price for the service. The service broker is responsible for network slice admission control and pricing strategy.

Step 3: The service broker interacts with the InP to identify other SPs (e.g., specialized in computing, storage) to serve the slice.

Step 4: The SPs collaborate to provide the necessary communication, computing, and storage resources per domain to meet the SLA. Each domain controller forwards the information regarding the number of resources needed to the service broker.

Step 5: The service broker sets up the price for the SLA based on the required resources and services in negotiation with the tenant.

Step 6: The service broker forwards the accepted slice request with its e2e SLA requirements to the CDO for its decomposition.

The slice admission control problem is solved in [4] by using deep reinforcement learning (DRL) to maximize the long-term reward of the InP given the uncertain availability of network resources. However, further work is needed on collaboration schemes between multiple providers to share time-varying resources and study their impact on pricing.

Decomposition: The e2e SLA should be decomposed efficiently and dynamically among the domains according to their available resources and services provided. The decomposition is performed in the following steps, as illustrated in Fig. 1:

Step 7: The CDO decomposes the e2e SLA into partial SLAs and assigns them to each domain controller to meet their part of the SLA. Partial SLAs are obtained using a *decomposition rule* that indicates how the e2e SLA is decomposed per domain, technology, and provider. For instance, an e2e delay requirement can be decomposed as the sum of the delays per domain, while the e2e throughput can be obtained as the minimum throughput in all domains. On the other hand, the e2e reliability can be decomposed as the product of the reliability per domain. Similarly, the decomposition parameters related to network design, such as availability, can be decomposed through different technologies and operators to guarantee link availability with a high probability. The importance of each metric differs depending on the traffic class and should also be specified in the SLA. For instance, in MBBMTC, achieving minimum throughput is more critical than achieving low latency. In MBBRLLC and MBBRLLMTC, the importance of throughput is the same as latency, while in RLLMTC, throughput is less important than latency.

Step 8: Every domain controller allocates radio, computing, and storage resources to meet the partial SLA requirement and creates an NSSI. If a domain lacks resources, it can collaborate with others and offload the computing task to a resource-richer domain. Multi-provider collaboration may involve multiple network technologies and intra- and inter-domain resource sharing, as described below.

Step 9: After each domain has created its NSSI, the InP instantiates the e2e NSI.

A risk model is presented in [5] to find an SLA decomposition with a higher probability of meeting the e2e SLA in a static network. To adapt the resources to the service needs, and resource availability, the decomposition rule should be scaled to the network condition [6]. If a failure occurs in a domain, and there are not enough resources in other domains to serve the task, the provider will compensate the tenant for the service degradation, as detailed in the SLA legal aspects specifications.

Resource Provisioning and Management: After the slice is created, the domain controllers send feedback to the CDO regarding their achieved partial SLAs (FB_1, FB_2, FB_3) in *step 10*. The CDO uses this feedback to revise the decomposition rule if similar requests are received again. Similarly, the domain controllers interact with the SPs' controllers to inform them of the achieved performance. SPs use this information to revise their resource provisioning strategies.

Monitoring: The SLA is monitored during the slice lifecycle. In this process, the collaboration between providers (*step 3*) and the price (*step 5*) will be revised according to the feedback (FB_0) received by the service broker in *step 11*. Likewise, the resource provisioning (*step 4*), and the decomposition (*step 7*) will be revised based on the feedback received in *step 10*.

Evaluation: Finally, the tenant assesses if the slice has met the service requirements and sends the tenant satisfaction (TS) to the CDO in step 12. This feedback is also taken into account by the CDO to adapt the decomposition rule.

Standardization

The proposed architecture aligns well with the guidelines provided by main standardization bodies and advances network slicing in multi-domain, multi-technology networks to fulfill service requirements by SLA decomposition. Regarding the standardization process for multi-domain network slicing, ETSI NFV MANO framework is working on transitioning from core network slicing toward peer-to-peer cross-domain orchestration and management [1]. Existing shortcomings and recommendations related to multi-domain slicing are also discussed in [1].

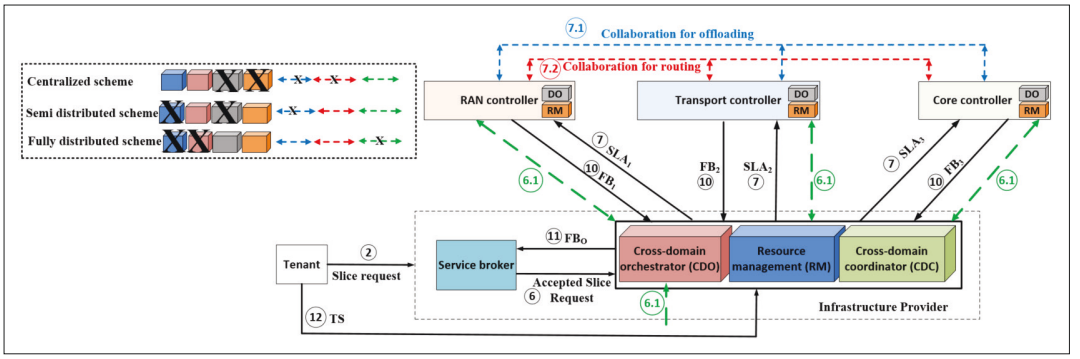

FIGURE 2. SLA decomposition management and orchestration schemes.

SLA Decomposition Management and Orchestration Architecture

SLA decomposition in multi-domain, multi-technology networks is a multi-dimensional problem in which multiple and interdependent aspects must be addressed. In particular, solving the SLA decomposition involves the following challenges: a) heterogeneous requirements that translate into different SLAs decomposition rules, resource mapping, and pricing; b) coordination among different domains for resource sharing, task offloading, and routing; c) negotiation between multiple providers with interdependent performance implications; and d) dynamic management and orchestration of the SLA decomposition to guarantee the e2e SLA throughout the slice lifetime.

In the sequel, we discuss three approaches to solve the SLA decomposition based on centralized, semi-distributed, and fully distributed implementation, as shown in Fig. 2. These approaches rely on our proposed management and orchestration scheme shown in Fig. 1.

Centralized Scheme

The centralized implementation assumes that the CDO knows the state of the resources in each domain and thus acts as a central entity to coordinate, decompose the SLA, and manage the overall resources. The centralized implementation consists of the following steps.

SLA Decomposition Rule: The CDO adjusts the decomposition rule to request partial SLAs from each domain based on the state of their resources (*step 6.1*), the price, and the feedback on previous decompositions (*steps 10 and 12*).

Offloading Decision Policy: If there is collaboration between domains for task offloading, the CDO will offload the computing task to a richer-resource domain to improve the performance. Offloading may happen when a domain has insufficient resources to satisfy the assigned partial SLA, or when a service has changed its requirements and demands more resources temporarily. In that case, the domain controllers share with the CDO the new state of their resources (*step 6.1*). Then, the CDO chooses the appropriate offloading policy which includes selection of the offloading domain (i.e., RAN, transport, and core), amount of computing task to be offloaded, and offloading price.

Resource Management (RM): The cross-domain coordinator (CDC) indicates to the domain controllers the resource allocation policy and performs intra-domain and inter-domain routing. The latter refers to routing the traffic between adjacent domains while the former refers to routing within a domain.

The CDO can solve the centralized scheme as the joint optimization of the resource allocation and inter- and intra-domain routing constraint by the e2e SLA requirements. However, solving this problem dynamically as the network evolves is complex. DRL algorithms such as the actor-critic framework can be adopted to predict the available resources in the domains [3]. Existing MANO solutions by 3GPP are mainly centralized [1]. However, centralized solutions lack scalability, and providers may not be willing to share information regarding available resources due to privacy and bargaining concerns. Besides, it has high overhead due to the exchange of messages between CDO and controllers to update the state of the resources, and the CDO is a single point of failure. In the sequel, we assume that the CDO progressively delegates management and orchestration to the domains toward a fully distributed implementation.

Semi-Distributed Scheme

The semi-distributed scheme assumes that the CDO is aware of the state of the resources in the domains, but each domain performs resource management and allocation.

SLA Decomposition and Offloading Decision Policies: the CDO decomposes the SLA based on the state of the resources, pricing, and previous feedback from domains, as in the centralized scheme.

Resource Management: each domain performs resource management and allocation, and intra-domain routing. Domains collaborate to solve the inter-domain routing that achieves the e2e SLA requirement as in step 7.2.

For a given SLA decomposition, the domain controllers can solve the resource allocation and routing distributively. Therefore, the complexity is lower than the previous scheme and will depend on the algorithm adopted to solve it. For instance, multi-agent DRL can be adopted in which domain controllers collaborate to decide the amount of communication, computing, and storage resources allocated per domain. The actions can be taken sequentially. The RAN controller determines the resource allocation and the intra-domain routing and collaborates with the transport controller to find the most convenient ingress node for the inter-domain route and bandwidth allocation. Similarly, the transport and core controllers collaborate to solve the inter-domain routing and find ingress

For a given SLA decomposition, the domain controllers can solve the resource allocation and routing distributively. Therefore, the complexity is lower than the previous scheme and will depend on the algorithm adopted to solve it.

and egress nodes. In [8], an inter-domain routing scheme is presented that considers optimal traffic volumes for different paths as a result of an iterative auction game.

Fully Distributed Scheme

The fully distributed scheme assumes that the CDO has no knowledge of the available resources per domain and each domain performs resource management and orchestration.

SLA Decomposition Rule and Offloading: The CDO decomposes the e2e SLA requirement of the slice request based on the price and previous feedback from the domain controllers but delegates the offloading decisions to the domain orchestrators (DOs), as shown in *step 7.1*. Accordingly, each DO decides whether to execute its computing task or offload it to another domain based on the availability of resources and cost-efficient considerations.

Resource Management: As in the semi-distributed scheme, the domain controllers share information to manage the resource allocation and find the best intra-domain and inter-domain paths distributively in *step 7.2*.

The complexity and overhead of this scheme is the lowest, and it is the most reliable since there is no single point of failure. To solve the resource allocation, task offloading, and routing distributively, we can explore federated learning (FL). Each domain controller trains their DRL model and send it for aggregation. This way, the domain controllers can keep data private since only trained models are shared.

SLA Decomposition and Offloading Decision: Case Study

We have conducted extensive simulations to illustrate the performance of our scheme under different SLA decomposition policies. The goal is to maximize the reward, which includes throughput, delay, and cost of serving the users. We consider four traffic classes (MBBMTC (c = 1), MBBRLLC (c = 2), RLLMTC (c = 3), and MBBRLLMTC (c = 4)) with different SLA requirements, as summarized in Table 2, and the physical layer models are described in [3]. The SLA is given in terms of the minimum throughput (and its weight), maximum latency (and its weight), reliability, number of users, and the amount of bandwidth, computing, and storage resources needed per domain. We evaluate the performance of four scenarios with different distributions of initial computing resources per domain such that:

- 57% of users were served in scenario 1 due to a lack of fog computing resources.
- 70% users were served in scenario 2 given the low edge computing resources.
- 43% users were served in scenario 3 due to limited cloud computing resources
- 20% users were served in scenario 4 given the limited fog and cloud computing resources.

Since the scenarios are different based on the available computing resources, we will refer to the domains as fog, edge, and cloud domains. The price per unit of resources is selected to incentivize every domain to serve its own tasks if it has enough resources. The fog, edge, and cloud node cache speeds are 80 Mb/s, 200 Mb/s, and 550 Mb/s, respectively. The network topology has 15 fog nodes, 9 edge nodes, and 6 cloud nodes.

We consider four policies for SLA decomposition and offloading decisions:

- *SLA-Decomp-Off:* joint SLA decomposition and resource sharing (i.e., offloading tasks from a resource-lacking domain to a resource-rich domain). This policy results in optimum performance.
- *SLA-Decomp-no-Off:* SLA decomposition and no offloading.
- *SLA-Decomp-sub-Off:* SLA decomposition and offloading to the other domain non-selected in the SLA-Decomp-Off policy.
- *Fixed-SLA-Decomp-Off:* offloading for a given SLA.

Reward, Cost, and Tenants' Satisfaction

In Fig. 3, we show the reward and cost for RLLMTC traffic. We have observed that the other traffic classes achieve the same trend and thus are not included for space limitations. We can see that the SLA-Decomp-Off policy that jointly optimizes the SLA decomposition and offloading achieves the highest reward in all scenarios compared with the other policies. In particular, the SLA-Decomp-Off policy achieves a reward of up to 2.5, 3, and 6 times higher than the SLA-Decomp-sub-Off, Fixed-SLA-Decomp-Off, and SLA-Decomp-no-Off, respectively. We have seen that the SLA-Decomp-Off policy serves all users, resulting in 100% tenant satisfaction. Moreover, the SLA-Decomp-Off policy has the lowest cost of creating the slice compared to SLA-Decomp-sub-Off and Fixed-SLA-Decomp-Off when serving the same number of users. This is because the SLA-Decomp-Off scheme jointly considers SLA decomposition, resource allocation, and pricing in their offloading decisions. On the other hand, SLA-Decomp-sub-Off offloads to the other domain non-selected in the SLA-Decomp-Off policy which is sub-optimal and thus it has a higher cost and lower reward. Similarly, the Fixed-SLA-Decomp-Off has a predefined SLA decomposition and optimizes the selection of the domain which has inferior performance than SLA-Decomp-Off that jointly optimizes both. In SLA-Decomp-no-Off policy each domain serves its tasks without offloading. Thus, if a domain has not enough resources available (as described in each scenario), it will serve only a fraction of the users, which results in a lower cost but also less reward. In scenario 4, since two domains lack resources (i.e., fog and cloud), the SLA-Decomp-sub-Off policy that selects a different domain than the edge (optimal one) results in a zero reward.

SLA Decomposition and Offloading

The final SLA decompositions are shown in Table 3 for all traffic classes and scenarios by using the different policies. We present the decomposition in terms of the fraction of the e2e delay assigned to each domain. The minimum e2e throughput requirement is achieved per domain. The SLA decompositions are different per scenario and traffic class since the initial resources per domain and the traffic requirements are different. This translates into the different amount of resources (radio, computing, and storage) required per domain. First, we analyze the SLA-Decomp-Off policy. In scenario 1, the available fog computing resources at the fog domain (F) are not enough to serve the require-

ments for any class (c), and thus the tasks are offloaded to the cloud domain (C) which results in a larger portion of the delay (SLA_C) assigned to this domain. Similarly, in scenario 2, the edge domain (E) lacks resources and the optimum offloading decision is also to offload to the cloud domain. In scenario 3 when the cloud domain lacks resources, the optimum performance is obtained by offloading to the fog domain. Consequently, SLA_F is the largest. Finally, in scenario 4 the fog and cloud offload to the edge which results in a larger fraction of the delay (SLA_E) assigned to the edge domain. In SLA-Decomp-no-Off policy, the SLA is optimized without offloading. Since we have assumed that cloud resources are cheaper than the edge and fog resources (i.e., cost of cloud resources < edge resources < fog resources), a higher amount of resources is allocated to the computing tasks in the cloud which results in a lower fraction of the delay in the cloud domain (SLA_C). The offloading decision of the SLA-Decomp-sub-Off policy is sub-optimal, and thus it offloads the computing task of the resource-lacking domain to the other domain non-selected in the SLA-Decomp-Off policy. Finally, in the Fixed-SLA-Decomp-Off policy, the SLA is given, and the offloading decision is the same as in SLA-Decomp-Off since it is the optimum one. The reason for this is that the resource-lacking domain offloads to the lowest-cost resource domain as long as it has enough resources to serve the offloaded computing tasks.

Open Challenges and Future Research

To achieve sustainable network slicing in 6G networks, the following research challenges must be addressed.

SLA-Based Network Slice Management and Orchestration

The heterogeneous traffic requirements and service characteristics result in SLAs with different structures, business models, and performance metrics. Therefore, the translation of service requirements to SLAs is challenging. Network slice management and orchestration schemes should be designed having this heterogeneity in mind. Accordingly, slice pricing, admission policies, and collaboration schemes for resource sharing should be defined per slice-SLA. To achieve sustainable network slicing it is crucial to guarantee the SLA throughout the slice lifecycle and thus mechanisms to monitor and detect SLA violations are to be explored. In this regard, service scaling has been successful in adapting the resource allocation per slice to traffic priorities [6]. However, further understanding of the dynamics of per-slice SLA management is needed to scale the resources proactively.

Multi-Domain Multi-Technology Network Slicing

End-to-end network slicing in multi-domain and multi-technology networks is challenging due to varying logical architectures and performance requirements. As more entities participate in providing the service, management and orchestration schemes need distributed solutions that scale with the number of domains and providers. This includes solutions for distributed SLA price negotiation, decomposition, monitoring, and collaboration policies between providers. Moreover, resource allocation policies should be aware of the performance characteristics of each technology and varying traffic requirements and ensure slice isolation. A network slicing scheme that incorporates mmWave and THz frequency bands for fog computing is presented in [4]. However, further research is needed to decompose the SLA through different technologies based on the application requirements.

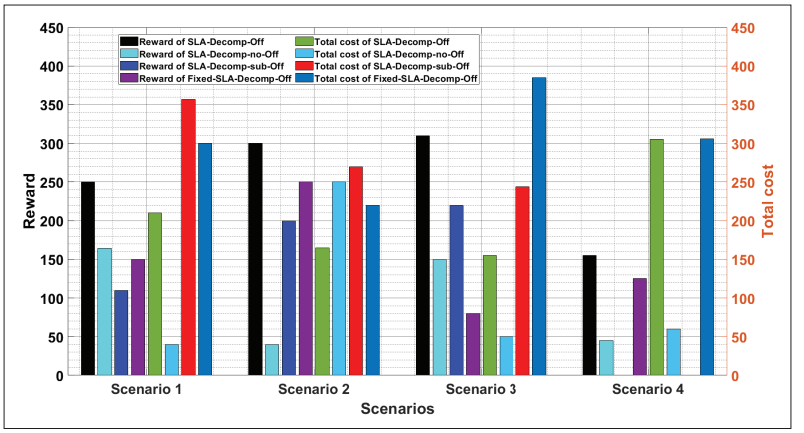

FIGURE 3. Reward and total cost for RLLMTC vs scenarios.

Secure SLA

Distributed network slicing brings security challenges due to the diversified security vulnerabilities introduced by multiple stakeholders. Weerasinghe *et al.* [14] introduce Secure SLA (SSLA) to ensure that SPs meet the security and privacy requirements defined in the SLA in terms of integrity, availability, confidentiality, and non-repudiation. The violation of these requirements could impact the performance of many slices that share the same infrastructure and thus could affect the reputation and business model. To this end, legal aspects should be defined in the SLA to identify which entity is responsible for reporting service failures or paying fees. Moreover, further work is needed to preserve SPs' private information when sharing resources between domains, such as resource availability and intra-domain topology. In [15], privacy-preserving end-to-end network slice orchestration is presented based on blockchain and a trusted execution environment. Further research is needed to perform a risk analysis of achieving the SLA requirements related to security threats.

Conclusion

This article presents a multi-domain multi-technology network slicing scheme and SLA decomposition management and orchestration policies. A comprehensive discussion is included on the challenges of decomposing the SLA in multi-domain multi-technology networks to achieve sustainable network slicing. We elaborate on the different steps to define, decompose, manage, monitor, and evaluate SLAs in these heterogeneous networks. Then, we discuss centralized, semi-distributed, and fully distributed implementations of our proposed scheme and present a case study to illustrate the performance using typical 6G use cases. Our results set a benchmark for future research in this area. Finally, future research directions and open challenges are highlighted.

		SLA-Decomp-Off				SLA-Decomp-no-Off			SLA-Decomp-sub-Off				Fixed-SLA-Decomp-Off			
Scen.	**c**	**Case**	SLA_F	SLA_E	SLA_C	SLA_F	SLA_E	SLA_C	**Case**	SLA_F	SLA_E	SLA_C	**Case**	SLA_F	SLA_E	SLA_C
1	1	F → C	0.01	0.28	0.66	0.4	0.4	0.2	F → E	0.01	0.77	0.18	F → C	0.01	0.32	0.63
	2	F → C	0.01	0.19	0.77	0.5	0.3	0.2	F → E	0.01	0.77	0.18	F → C	0.01	0.32	0.64
	3	F → C	0.02	0.18	0.74	0.4	0.4	0.2	F → E	0.02	0.75	0.17	F → C	0.02	0.31	0.61
	4	F → C	0.01	0.19	0.77	0.5	0.3	0.2	F → E	0.01	0.77	0.18	F → C	0.01	0.32	0.64
2	1	E → C	0.29	0.02	0.65	0.4	0.4	0.2	E → F	0.76	0.02	0.18	E → C	0.32	0.02	0.62
	2	E → C	0.3	0.01	0.67	0.5	0.3	0.2	E → F	0.78	0.01	0.18	E → C	0.33	0.01	0.63
	3	E → C	0.29	0.039	0.63	0.4	0.4	0.2	E → F	0.75	0.03	0.17	E → C	0.32	0.03	0.6
	4	E → C	0.29	0.01	0.67	0.5	0.3	0.2	E → F	0.77	0.01	0.18	E → C	0.32	0.01	0.63
3	1	C → F	0.55	0.38	0.03	0.3	0.5	0.2	C → E	0.19	0.75	0.03	C → F	0.63	0.23	0.03
	2	C → F	0.57	0.39	0.02	0.3	0.5	0.2	C → E	0.2	0.76	0.02	C → F	0.63	0.24	0.02
	3	C → F	0.54	0.38	0.04	0.3	0.5	0.2	C → E	0.29	0.63	0.04	C → F	0.62	0.23	0.04
	4	C → F	0.57	0.39	0.02	0.3	0.5	0.2	C → E	0.29	0.66	0.02	C → F	0.72	0.24	0.02
4	1	F,C → E	0.01	0.93	0.03	0.3	0.5	0.2	—	—	—	—	F,C → E	0.01	0.93	0.03
	2	F,C → E	0.01	0.95	0.02	0.3	0.5	0.2	—	—	—	—	F,C → E	0.01	0.95	0.02
	3	F,C → E	0.02	0.9	0.04	0.3	0.5	0.2	—	—	—	—	F,C → E	0.02	0.9	0.04
	4	F,C → E	0.01	0.95	0.02	0.3	0.5	0.2	—	—	—	—	F,C → E	0.01	0.95	0.02

TABLE 3. The e2e SLA decompositions and offloading decision cases for scenarios 1 to 4.

Acknowledgment

This work is partially supported by the US National Science Foundation under Grant CNS-2008309 and Jupiter Networks.

References

[1] S. Kukliński *et al.*, "6GLEGO: A Framework for 6G Network Slices," *J. Commun. Networks*, vol. 23, no. 6, 2021, pp. 442–53.
[2] M. Rasti *et al.*, "Evolution Toward 6G Multi-Band Wireless Networks: A Resource Management Perspective," *IEEE Wireless Commun.*, vol. 29, no. 4, 2022, pp. 118–25.
[3] H. H. Esmat and B. Lorenzo, "Self-Learning Multi-Mode Slicing Mechanism for Dynamic Network Architectures," *IEEE/ACM Trans. Networking*, 2023, pp. 1–16.
[4] H. H. Esmat and B. Lorenzo, "Deep Reinforcement Learning Based Dynamic Edge/Fog Network Slicing," *Proc. 2020 IEEE Global Commun. Conf.*, 2020, pp. 1–6.
[5] D. De Vleeschauwer, C. Papagianni, and A. Walid, "Decomposing SLAs for Network Slicing," *IEEE Commun. Letters*, vol. 25, no. 3, 2021, pp. 950–54.
[6] X. Li *et al.*, "Automated Service Provisioning and Hierarchical SLA Management in 5G Systems," *IEEE Trans. Network and Service Management*, vol. 18, no. 4, 2021, pp. 4669–84.
[7] I. Kovacevic *et al.*, "Multi-Domain Network Slicing With Latency Equalization," *IEEE Trans. Network and Service Management*, vol. 17, no. 4, 2020, pp. 2182–96.
[8] J. Khamse-Ashari *et al.*, "An Agile and Distributed Mechanism for Inter-Domain Network Slicing in Next Generation Mobile Networks," *IEEE Trans. Mobile Computing*, vol. 21, no. 10, 2022, pp. 3486–3501.
[9] K. Abbas *et al.*, "Network Slice Lifecycle Management for 5G Mobile Networks: An Intent-Based Networking Approach," *IEEE Access*, vol. 9, 2021, pp. 80,128–46.
[10] Y. Chen *et al.*, "SLA Decomposition: Translating Service Level Objectives to System Level Thresholds," *Proc. Fourth Int'l. Conf. Autonomic Computing*, 2007, p. 3.
[11] E. Kapassa, M. Touloupou, and D. Kyriazis, "SLAs in 5G: A Complete Framework Facilitating VNF- and NS-Tailored SLAs Management," *Proc. 2018 32nd Int'l. Conf. Advanced Information Networking and Applications Workshops*, 2018, pp. 469–74.
[12] H. N. Qureshi *et al.*, "Service Level Agreements for 5G and Beyond: Overview, Challenges and Enablers of 5G-Healthcare Systems," *IEEE Access*, vol. 9, 2021, pp. 1044–61.
[13] X. Ge *et al.*, "Energy Efficiency Challenges of 5G Small Cell Networks," *IEEE Commun. Mag.*, vol. 55, no. 5, 2017, pp. 184–91.
[14] N. Weerasinghe, R. M. P. P. M. Liyanage, and M. Ylianttila, "Proof-of- Monitoring (PoM): A Novel Consensus Mechanism for Blockchainbased Secure Service Level Agreement Management," *IEEE Trans. Network and Service Management*, 2023, p. 1.
[15] G. He *et al.*, "NetChain: A Blockchain-Enabled Privacy-Preserving Multi-Domain Network Slice Orchestration Architecture," *IEEE Trans. Network and Service Management*, vol. 19, no. 1, 2022, pp. 188–202.

Biographies

Haitham H. Esmat [M] received B.S. (with the second honors) and M.S. degrees from Helwan University, Cairo, Egypt, in 2011 and 2016, respectively. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of Massachusetts Amherst, Amherst, MA, USA. His current research interests include B5G and 6G mobile communication technologies, device-to-device communications, Internet of Things, dynamic network slicing, satellite–terrestrial edge computing networks, and AI-based applications in wireless communication systems.

Beatriz Lorenzo [SM] received an M.S. degree in telecommunication engineering from the University of Vigo, Vigo, Spain, in 2008, and the Ph.D. degree from the University of Oulu, Oulu, Finland, in 2012. She is currently an Assistant Professor and the Director of the Network Science Lab, Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA, USA. She has published more than 50 papers and coauthored two books on advanced wireless networks. The latest book Artificial Intelligence and Quantum Computing for Advanced Wireless Networks (Wiley, 2022), covers the enabling technologies for the definition, design, and analysis of incoming 6G/7G systems. Her research interests include AI for wireless networks, B5G and 6G network architectures and protocol design, mobile computing, optimization, and network economics. She received the Fulbright Visiting Scholar Fellowship with the University of Florida from 2016 to 2017. She was the General Co-Chair for WiMob Conference in 2019 and serves regularly in the TPC of top IEEE/ACM conferences. She is currently an Editor of the *IEEE Trans. Mobile Computing*.