Edge Intelligence for Mission-Critical 6G Services in Space-Air-Ground Integrated Networks

Xiangwang Hou, *Student Member, IEEE,* Jingjing Wang, *Senior Member, IEEE*, Zhengru Fang, *Student Member, IEEE,* Yong Ren, *Senior Member, IEEE*, Kwang-Cheng Chen, *Fellow, IEEE*, Lajos Hanzo, *Fellow, IEEE*

***Abstract*—Next-generation wireless services will change our daily lives by supporting smart factories, intelligent transportation, augmented/virtual reality (AR/VR), etc. These sophisticated services are usually both data- and computation-intensive and must meet stringent latency and reliability requirements, which cannot be readily satisfied by cloud-based service processing. Therefore, the 6G cellular network is expected to jointly optimize communication, computing, caching and control. A further aspiration of 6G is to conceive a seamless space-air-ground integrated network (SAGIN) for filling the vast coverage holes across the globe, which brings about new opportunities for mission-critical services. Therefore, in this article, we aim for conceiving *ultra-reliable and low-latency edge intelligence* (URLLEI) for supporting mission-critical services by harnessing the diversified communication, computing, and caching resources at the network edge of SAGIN. We critically appraise a number of key enabling techniques, including a number of new service-centric resource allocation techniques. Finally, a range of open challenges is discussed.**

***Index Terms*—6G, space-air-ground integrated network (SAGIN), edge intelligence (EI), ultra-reliable and low latency communications (URLLC), resource allocation.**

I. Introduction

Whilst the 5G cellular networks are being rolled out across the globe, researchers have turned their attention to 6G [1]. Recently, a variety of compelling mission-critical services have emerged, as exemplified by factory automation, intelligent transportation, augmented/virtual reality (AR/VR), and so on. Typically, these sophisticated services are usually data- and computation-intensive and are also subject to stringent latency and reliability requirements, which may potentially be even more demanding than the short-packet based ultra-reliable and low-latency communication (URLLC) services. However, conventional cloud computing based centralized processing exhibits potentially excessive latency due to its long-distance transmission.

Before any further discussions on the potential solutions, it is essential to highlight the vital requirements of mission-critical services.

- **Supporting ubiquitous access:** The ubiquity of access is critical, because for example in intelligent logistics a self-driving delivery truck must remain connected anywhere and anytime.
- **Meeting all service-specific requirements:** The vast majority of the existing related research has considered generic system-level performance metrics, such as the average system latency, total energy consumption, etc. However, the quality of each specific mission-critical service has its own bespoke metric, which mandates a further examination for each individual service in terms of its QoS violation probability.
- **Defininig the performance metrics in a service-centric way:** Most of the existing literature on guaranteeing the latency and/or reliability of a specific service tends to aim for separately optimizing for example the bit error rate (BER) of the URLLC service or the delay of cloud computing services in a specific link of the heterogeneous SAGIN. However, future research should find the entire Pareto front of the end-to-end service, including the associated transmission, computing and caching.

Therefore, the 6G cellular network is expected to guarantee the QoS of these emerging services by seamlessly integrating terrestrial networks, aerial networks and satellite networks, leading to the concept of space-air-ground integrated networks (SAGIN). They are expected to fill the existing terrestrial coverage holes even in rural, as well as oceanic areas and deserts. However, there is a paucity of research on exploiting the SAGIN concept. Therefore, in this article, we explore how to support mission-critical services based upon the SAGIN concept.

Edge intelligence (EI) [2] providing powerful low-latency computing and storage services at the edge of the network, is more suitable for supporting mission-critical services than conventional cloud-based intelligence. Hence, we explore the potential of SAGIN-aided *ultra-reliable and low-latency edge intelligence* (URLLEI) in supporting mission-critical services.

This work was partly supported by the National Natural Science Foundation of China (Grant No. 62071268), partly supported by the Young Elite Scientist Sponsorship Program by CAST (Grant No. 2020QNRC001). K.-C. Chen would like to acknowledge the financial support of the grant from Cyber Florida. L. Hanzo would like to acknowledge the financial support of the Engineering and Physical Sciences Research Council projects EP/P034284/1 and EP/P003990/1 (COALESCE) as well as of the European Research Council's Advanced Fellow Grant QuantCom (Grant No. 789028) (Corresponding author: Jingjing Wang.)

X. Hou, Z. Fang and Y. Ren are with the Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China. (E-mail: xiangwanghou@163.com, fangzr19@mails.tsinghua.edu.cn, reny@tsinghua.edu.cn.)

J. Wang is with the School of Cyber Science and Technology, Beihang University, Beijing 100191, China. (Email: drwangjj@buaa.edu.cn.)

K.-C. Chen is with the Department of Electrical Engineering, University of South Florida, Tampa, FL 33620 USA (E-mail: kwangcheng@usf.edu).

L. Hanzo is with the School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, UK. (E-mail: lh@ecs.soton.ac.uk.)

However, in contrast to traditional terrestrial networks, SAGINs exhibit high-dynamic, heterogeneous nature. In the face of this, the following challenges have to be tackled on the road to make URLLEI a reality.

Firstly, the traditional resource-hungry virtual machine technology-based virtualization method relying on a monolithic service processing paradigm is unsuitable for SAGINs, which have to depend on resource-constrained network nodes imposing high response latency and low reliability. Secondly, since the network nodes of SAGINs may have a high velocity, which may lead to frequent handovers among these edge nodes as well as between nodes and users, an agile resource scheduling is required. Furthermore, since some edge nodes in the SAGIN are not as reliable as those of the classic terrestrial networks, a high probability of node failure and communication link interruption require near-instantaneous adaptivity and route-repair capability.

Therefore, to cope with the aforementioned challenges, in this article, we advocate the emerging ultra-lightweight virtualization technology termed as unikernel [3] for homogenizing the heterogeneous resources of SAGIN, intrinsically amalgamated with a micro-service [4] driven service processing paradigm. This compelling amalgam mitigates the latency and supports reliable service processing. Moreover, we propose a nimble adaptive resource scheduling framework based on reinforcement learning and knowledge graphs. Furthermore, we harness the distributed computing and caching resources of SAGINs for mitigating the pressure on communication, which urges us to rethink the service-centric resource configuration philosophy. Explicitly, we explores the joint optimization of communication, computing, and caching for service-centric resource allocation. Last but not least, some open challenges and potential research directions are presented to facilitate the implementation of URLLEI.

In Section II, we commence by portraying the detailed architecture of URLLEI, with an emphasis on its key techniques. Then, a pair of service-centric resource allocation strategies are presented in Section III. Finally, a suite of open issues and challenges are discussed in Section IV, followed by our conclusions in Section V.

II. The Architecture and Key techniques of URLLEI

A. The Architecture of URLLEI

The SAGIN-aided URLLEI concept is constituted by three segments, namely satellite networks, aerial networks, and terrestrial networks. Specifically,

- **Satellite networks** are composed of low earth orbit (LEO), medium earth orbit (MEO), and geostationary orbit (GEO) satellites, which cooperatively constitute a sophisticated satellite constellation, relying on Iridium, Starlink, GlobalStar, etc.
- **Aerial networks** [5] consist of airplanes, balloons, unmanned aerial vehicles (UAV), etc.
- **Terrestrial networks**, including the mature 3G/4G/5G cellular networks, WiMAX, WLAN, and so forth.

Although it is promising to conceive URLLEI based upon SAGIN benefiting from its seamless coverage and distributed network resources, its heterogeneous, time-varying, and resource-constrained nature has to be ameliorated.

Thus to meet the stringent latency and reliability requirements of mission-critical services, the three-layer architecture of Fig. 1 has to be carefully scrutinized. Specifically, each mission-critical service has to be optimized based on its dominant performance metrics instead of optimizing the overall average system-level performance. Therefore, in the service layer, the concept of micro-services [4] is adopted. Each service will be decomposed into fine-grained micro-services for subsequent distributed processing for reducing the associated latency. For example, uncontended low-latency resources would be reserved for flawless lip-synchronized video, which may tolerate say 30 ms delay and a BER of 10^{-3}. By contrast, the associated data channel may tolerate 300 ms delay and a BER of 10^{-6}. As a benefit, the processing failure of some of the micro-services will no longer lead to the service's overall failure, and only the reprocessing of these failed micro-services is required, which improves the reliability of service processing. Furthermore, to cope with the intermittent nature of SAGIN, we conceive an intelligent distributed control layer relying on dynamic resource management intrinsically amalgamated with both knowledge graphs and bespoke reinforcement learning. This sophisticated architecture is capable of efficient adaptive dynamic resource control, hence mitigating the latency and reliability limitations caused by queuing and network congestion. Moreover, in the resource layer, the URLLEI adopts unikernel [3], which is a revolutionary ultra-lightweight virtualization technology, promptly combining the heterogeneous hardware resources of SAGINs, where each virtualized resource block is assigned to a single service. To elaborate a little further:

*1) **Service layer**:* The service layer is naturally a logical layer, which can be located either in the access points (AP) or in the controllers in light of the specific conditions and service requirements. The service layer's primary responsibility is to recognize the user's intent, where the intent mainly includes the service' requirements such as its processing delay, reliability, bandwidth and latency jitter. Then the service layer has to further transform the requested service into a series of micro-services having specific service requirements. The service layer comprises three modules, which are responsible for the service intent abstraction, service intent translation, and service decomposition, respectively. The service intent abstraction module is designed for inferring the user's intent from his/her high-level actions, for example by relying on historical data. The service intent translation module transforms the service requirements into specific physical resources reserved for computing, communication and storage, for example. Furthermore, the mission-critical services are decomposed into a set of micro-services by the service decomposition module. Considering object recognition services as a specific example, it can be partitioned into four micro-services, say image preprocessing, image segmentation, feature extraction and image classification. The specific procedure of beneficially partitioning the service into micro-services each having its

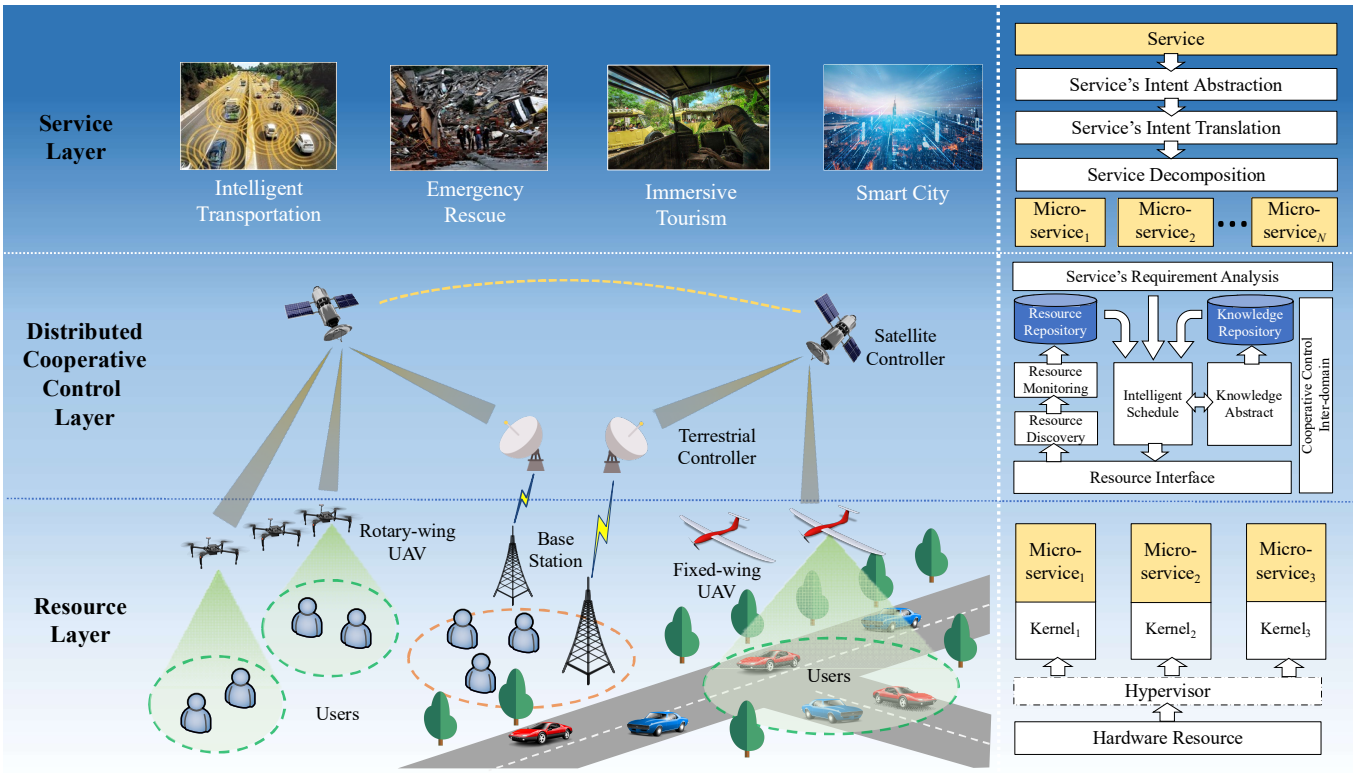


Fig. 1. The architecture of URLLEI.

quality metric, computational complexity and resource requirement, is indeed an interesting issue for further study.

*2) **Distributed cooperative control layer**:* The distributed cooperative control layer is composed of satellite controllers and ground controllers. To elaborate, the satellite controllers are mainly of GEO and MEO nature, where the orbit altitude of GEO satellites is about 35000 km, while the MEOs are distributed between the altitudes of 2000 and 20 000 km. As a benefiting of their high orbit, as few as 3 to 4 GEOs supported by a dozen MEOs and ground stations can realize seamless global coverage with the aid of the massive network resources of SAGINs. This layer is responsible for managing the massive communication, computing and caching resources of SAGINs. Considering both the deployment cost and system performance, each controller is designed for managing the intra-network resources in a limited area, which is termed as a domain. The time-varying and large-scale nature of SAGINs requires a sophisticated amalgam of centralized and distributed cooperative control. For low-delay intra-domain service requests, centralized resource scheduling is recommended in support of prompt actions. Again, this may be arranged by reserving uncontended slots following a contention for reservation phase. By contrast, for cross-domain or resource-hungry services, cooperative control is preferred for supporting multi-domain operation. There are two repositories, namely the resource repository and the knowledge repository. The resource repository stores the resource status information in the domain in real-time via continuous monitoring. By contrast, the knowledge repository stores the experiences abstracted from historical data to form a knowledge graph. Both of them have a significant impact on the overall performance of resource scheduling. It is worth noting that for forming the knowledge graphs we have to analyze ultra-large-scale serialized historical data, including historical service requirements, resource status, and scheduling strategies. This is a challenge for traditional model-based methods, while reinforcement learning can deal with it efficiently. Moreover, as for further resource scheduling, relying on the information provided by the two repositories, numerous optimization methods can be adopted, such as convex optimization, reinforcement learning, heuristic algorithms, etc.

*3) **Resource layer**:* The resource layer primarily consists of UAVs, balloons, and terrestrial network elements for providing abundant network resources for URLLEI. Besides, LEO satellites, which are generally distributed between 200 to 2000 km tend to have a low end-to-end propagation delay, which is well suited for supporting mission-critical services. Although the availability of diverse resources at the edge of the network holds the promise of supporting mission-critical services, the heterogeneity of these resources is an impediment, because the incompatibility of different interfaces hampers their interoperability. Hence finding a suitable virtualization method is essential. However, traditional virtual machine (VM) technology tends to suffer from high resource-occupancy and high latency, hence failing to meet the requirements of SAGINs in supporting URLLEI. The unikernel virtualization method of [3] is potentially capable of improving both the resource-occupancy and latency, as well as the reliability. It is eminently suitable for virtualizing the resources of SAGINs for achieving URLLEI. Since unikernel contains fewer components

than a VM, millisecond-level initialization can be achieved. Moreover, it can be created immediately, when a service is requested and shut down as soon as it is completed, while supporting flexible and elastic services. This feature naturally fits with URLLEI, whose resource nodes and users exhibit agile mobility.

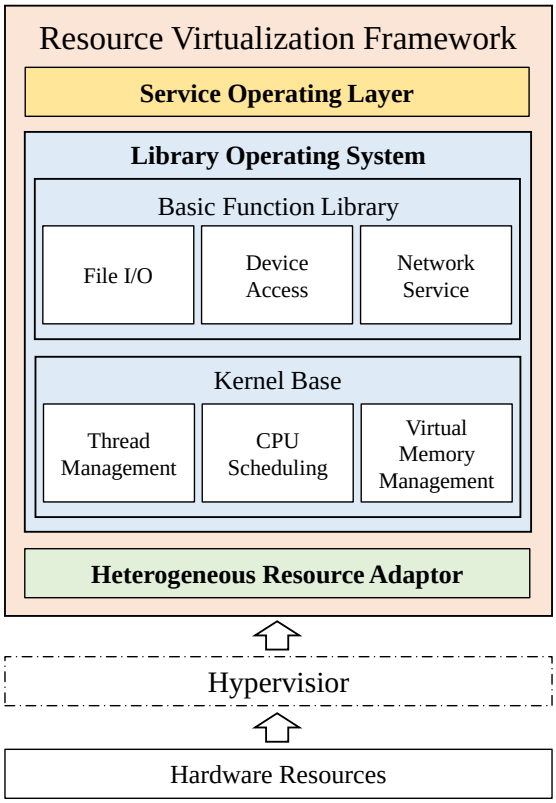


Fig. 2. Unikernel-based virtualization framework

To homogenize the heterogeneous resources of SAGIN for supporting URLLEI, as shown in Fig. 2, a three-tier unikernel-based virtualization framework is conceived, including the service operating layer, library operating system and heterogeneous resource adaptor. Specifically,

a. The service operating layer of Fig. 2 has the task of executing the micro-services of our *service layer*, where each unikernel-based virtualization resource block only supports a single micro-services at a low latency and high reliability.

b. The library operating system of Fig. 2 is the core of the unikernel-based virtualization, including its kernel base and basic function library.

- The kernel base constitutes the foundation for the library operating system composed of the most fundamental functionalities, such as thread management, CPU scheduling, and virtual memory management. Thread management is mainly responsible for the priority control of threads, as well as mutual exclusion, and synchronization, when accessing shared resources. CPU scheduling is mainly responsible for scheduling the threads to be executed by relying on the CPU resources. Virtual memory management is responsible for realizing the mapping of the virtual memory address space to physical memory.
- The basic function library aims for providing low-latency and high-reliability resource management interfaces, including the protocol library and drivers. The micro-service realizes prompt access to the physical resources of the SAGIN node by activating the functions of the basic function library, for instance, file I/O, device access and network services.

c. The heterogeneous resource adaptor of Fig. 2 runs directly on hardware resources and it is responsible for the essential resource protection and isolation. Given that the SAGIN nodes often have to operate in harsh environments, this may lead to the failure of service execution. Hence it is critical to support service migration between different nodes for guaranteeing the service's reliability. The heterogeneous node adaptor is capable of achieving differentiated shielding of the heterogeneous SAGIN nodes and of creating an environment compatible with micro-service migration.

When a user requests a service, the service layer firstly recognizes the user's intent and determines the specific service requirements, such as latency, latency jitter, reliability, bandwidth, etc. Then, the service is decomposed into a set of micro-services. Furthermore, the controllers of the distributed collaborative control layer formulate a resource scheduling strategy for exploiting the resources in support of micro-services, whilst relying on real-time resource monitoring and knowledge graphs. Next, the scheduling strategy may be forwarded to the relevant network nodes in the resource layer, such as UAVs, LEOs, and terrestrial entities. Finally, the series of micro-services can be processed, and their results are in turn forwarded to the user.

It is worth noting that the proposed architecture constitutes a customized framework designed for mission-critical services, where the URLLEI concept is sufficiently versatile for lending itself to seamless integration into the existing network architectures or platforms, such as software-defined networks (SDN). To be more specific, SDNs aim for the efficient management (e.g., routing, traffic control, etc.) of the terrestrial networks, especially of the fixed networks. However, SDNs fail to cope with the time-varying resource management of SAGINs and with the mission-critical service demands of efficiently scheduling both the computing and storage resources. By contrast, URLLEI can readily deal with these radical requirements.

B. Key Techniques

- **Beneficial service decomposition strategy.** Decomposing each service into multiple micro-services for distributed processing is capable of significantly improving both the system's resource efficiency and the quality of experience (QoE). The key factors to be considered are the service requirements, data volume and computational complexity, resource status, as well as the service topology. The most pivotal requirement in this context is to reduce the interdependence among the micro-services. The lower the dependency among the micro-services, the higher the performance gains become.

- **Tight control-information synchronization.** Tight inter-domain control-information synchronization is necessary for seamless collaborative cross-domain service provision, in the face of dynamic control information changes. If the distributed synchronization procedure has to be repeated for each update of the intra-domain resource status, the communication overhead will be considerable. For meeting the stringent requirements of mission-critical services, it is necessary to employ selective control information synchronization.
- **Multi-semantic addressing.** The traditional topology-based addressing method is suitable for static networks relying on a fixed topology, such as classic cellular networks. However, it is inefficient in SAGINs, where the network topology is highly dynamic, which hence cannot meet the demanding requirements of mission-critical services in terms of latency and reliability. It is necessary to explore bespoke addressing methods for decoupling the users, content and resources from the network topology so that they can be routed through their respective address spaces. Multi-semantic addressing [6] is capable of mitigating the latency and reliability challenges of this mapping process of the traditional topology-based addressing method. As a benefit of decoupling the multi-semantic address space and the dynamically fluctuating network topology, the system can better adapt to the mobility of the resource entities of SAGINs.
- **Service-centric resource allocation.** In contrast to the traditional "cloud-channel-device" paradigm based on terrestrial networks and cloud computing, URLLEI aims for exploiting the diversified computing and storage resources at the edge of the network in support of mission-critical services. Although URLLEI achieves ubiquitous access, new problems have emerged, which call for service-centric resource allocation. Most of the existing literature on guaranteeing services' latency and reliability performance separately focuses on a particular stage of service processing, such as the bit error rate (BER) of the URLLC or the delay of cloud computing. We should extend the definitions of the service's reliability and processing latency to the entire duration of the end-to-end service, including the associated transmission, computing, and caching. In a nutshell, we should view the benefits of the extra computing and caching resources in the light of the associated resource constraints and node reliability.

III. Service-centric Resource Allocation for Mission-critical Services

Again, service-centric resource allocation is vital for meeting the stringent mission-critical service requirements of SAGIN. However, most of the existing research tends to remain limited to the latency and reliability requirements of mission-critical services within a single layer of the communication protocol stack, while ignoring the potential of joint computing and caching resource allocation for improving the latency and reliability.

In the proposed URLLEI, this design philosophy is radically revised. The computing and caching resources of SAGIN inspire us to harness them for relieving the pressure on communications. Therefore, in this section, we first survey the state-of-the-art in mission-critical services, followed by proposing the joint optimization of the communication, computing and caching resources for latency- and reliability-sensitive mission-critical services.

A. State-of-the-art in Mission-critical Services

As mentioned, most of the state-of-the-art literature aims for reducing the latency and for improving the reliability of the stand-alone physical (PHY) layer [7], media access control (MAC) layer [8] and network layer [9], as well as of system-level relying on cross-layer optimization [10]. Specifically, in the PHY layer, focusing on the Raleigh block-fading model, Durisi *et al.* [7] discussed the tradeoff between reliability, throughput and latency in machine-type communications. As for the MAC layer, Cui *et al.* [8] developed a four-state semi-Markovian model for analyzing the transmission collisions and random backoffs encountered in distributed wireless networks. The vehicular networks critically depend on low-latency and high-reliability services, but their unpredictable nature imposes significant challenges on the network layer. Gao *et al.* [9] advocated roadside units (RSU) as auxiliary infrastructure for improving the latency and reliability performance. However, the benefits of single-layer optimization remain limited. It is imperative to explore cross-layer optimization to meet the stringent latency and reliability requirements of mission-critical services. Indeed, all wireless systems rely on cross-layer optimization, since both power-control and hand-over inherently operate across several OSI-layers. In this spirit, Popovski *et al.* [10] discussed the fundamental tradeoffs involved in designing access protocols by considering both the PHY, the MAC, and the data-link layer.

Furthermore, with the emergence of edge computing, researchers began to reap the joint benefits of communication and computing for reducing the service-processing latency and for improving the reliability. Liu *et al.* [11] studied the tradeoff between latency and reliability in computation offloading. Moreover, to cope with dynamic latency- and reliability-aware computation-offloading issues, Liu *et al.* [12] proposed a two-timescale network association mechanism based on matching theory and stochastic Lyapunov optimization. However, both [11] and [12] ignore the reliability of the computing resources, even though its imperfections gravely affect SAGIN-based URLLEI. To conclude this section, Table I summarizes the above-mentioned contributions at a glance.

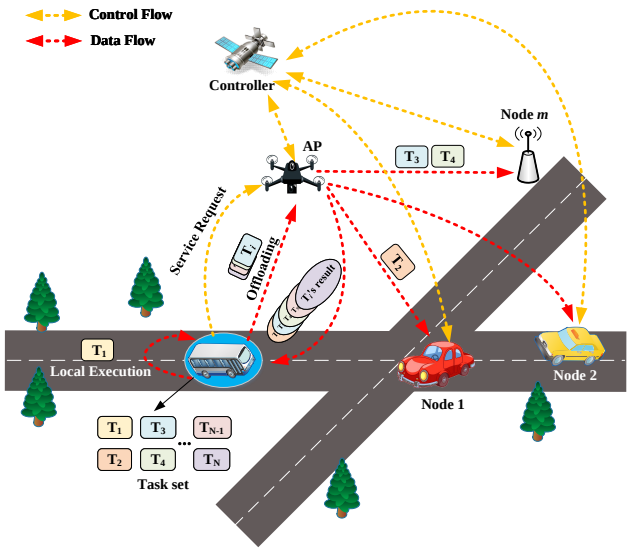
B. Joint Communication and Computing Resource Allocation Using Reprocessing for URLLEI

In this section, we present our recent results on ultra-reliable resource allocation using a reprocessing mechanism [13], which jointly optimizes the communication and computing resources. Let us consider the scenario of Fig. 3(a). Assume having a single AP and m computing nodes, where a vehicular user requires access to the AP, which forwards the service request to the controller that belongs to the *distributed cooperative control layer* of Fig. 1. Next the controller decides about

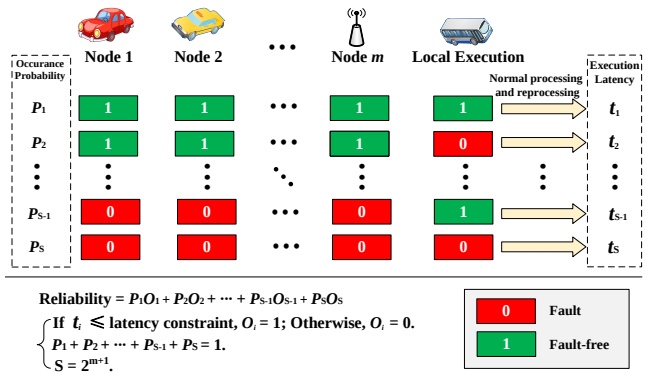
TABLE I
SUMMARY OF RECENT LITERATURE FOR MISSION-CRITICAL SERVICES

Contributors, year	Reference	Optimization methodology	Key contributions and features	Is computation reliability considered?	Does it benefit from computation resources?	Does it benefit from caching resources?
Durisi *et al.* 2016	[7]	PHY layer	The tradeoff between reliability, throughput, and latency in the transmission. Over multiple-antenna Rayleigh block-fading channel.	No	No	No
Cui *et al.*, 2017	[8]	MAC layer	A new theoretical framework to quantify the effective capacity of license-assisted access. Under statistical QoS constraints.	No	No	No
Gao *et al.*, 2020	[9]	Network layer	A reliable routing decision scheme based on the Manhattan mobility model. Based on the Manhattan mobility model.	No	No	No
Popovski *et al.*, 2019	[10]	Cross-layer	Discussing the principles of wireless access for latency- and reliability-sensitive services. Relying on the massive MIMO and multi-connectivity.	No	No	No
Liu *et al.*, 2018	[11]	Edge computing-assisted	A task offloading policy which balances the latency and reliability. Optimizing the node candidates selection, offloading ordering and task allocation.	No	Yes	No
Liu *et al.*, 2019	[12]	Edge computing-assisted	A services centric two-timescale network association and task computation framework. Taking into account the statistics of extreme queue length events.	No	Yes	Yes

the offloading proportion of the service-based computation tasks, and offloads the tasks using the available communication resources from the AP to the nearby computing nodes, whilst relying on the *resource layer*, which is achieved by harnessing cooperative computing, as seen in Fig. 1. Explicitly, the computing nodes may be classified into mobile nodes (i.e., vehicles, UAVs) and fixed nodes (i.e., RSUs).



(a) Joint communication and computing resource allocation.



(b) Reliability assessment [13].

Fig. 3. Joint communication and computing resource allocation relying on reprocessing mechanism for URLLEI.

Unlike the highly reliable network entities of terrestrial networks, the nodes in SAGIN typically have lower reliability caused by their hostile propagation environments, which requires us to pay increased attention to the impact of link interruption or node failure during service processing. We assume that the failure probability of a unit processing the assigned task follows the Poisson distribution with failure rate λ, where "failure" refers to any arbitrary failure, such as hardware-failure and software-failure, caused by electromagnetic interference by an operator, environmental extremes, wear-out, etc. If either one of the m computing nodes of Fig. 3(b) or the local computation fails, the service requested cannot be successfully completed. Specifically, as shown in Fig. 3(b), we may analyze all the $S = 2^{m+1}$ potential failure scenarios in advance, and then we calculate the execution time t_i including both the normal original processing and reprocessing, as well as the occurrence probability P_i of each failure case. Finally, depending on whether the execution time of each failure scenario exceeds the latency constraint, denoted by O_i, we can evaluate the reliability of the task allocation scheme.

Fig. 4 shows our comparison between the fault-tolerant task allocation strategy employing both the original and the reprocessing mechanism as well as that without reprocessing in different operating environments, when supporting a variety of services having different computational complexity. Specifically, the complexity of gzip ASCII compress is 330/8 cycles/bit, while that of x264 Variable Bit Rate (VBR) encoding is 1300/8 cycles/bit. Furthermore, the complexity of x264 Constant Bit Rate (CBR) video encoding is 1900/8 cycles/bit, while that of html2text is 5900/8 cycles/bit. In an ideal propagation and computing environment, the failure rate λ of the nodes and links is assumed to obey the uniform distribution of U(0, 0.04), while in a benign environment, the failure rate λ follows the uniform distribution of U(0.1, 0.5). In a hostile environment, λ complies with the uniform distribution of U(0.3, 0.7), while in a poor environment, the failure rate λ obeys U(0.5, 0.9). Observe that the scheme relying on reprocessing substantially improves the reliability of the arrangement operating without reprocessing, especially in a poor environment.

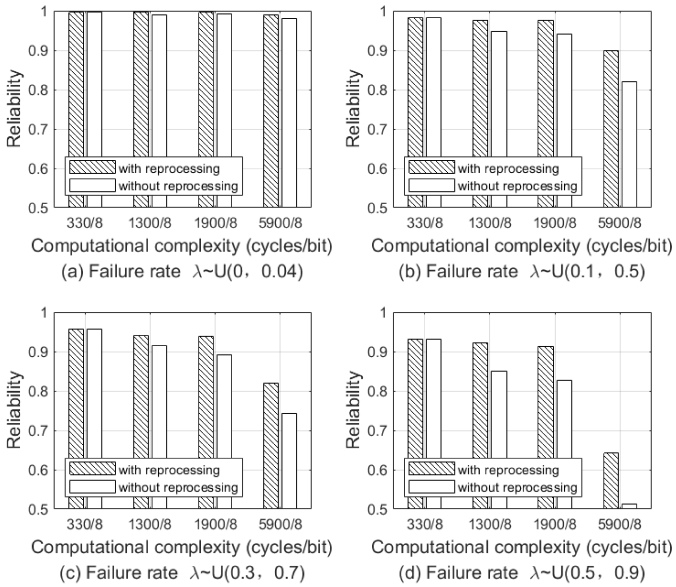


Fig. 4. Impact of reprocessing on reliability in different environments. There are 5 computing nodes, and the user has a computational task that has to be processed with the input data size of 4.8 Mb, where the communication link between the AP and the computing nodes as well as between the AP and the vehicular user is the probabilistic superposition of line-of-sight (LOS) and non-line-of-sight (NLOS) links. The CPU frequency of the fixed nodes follows the uniform distribution of U(6, 7) GHz, while the CPU frequency of the mobile nodes follows the uniform distribution of U(1, 3) GHz.)

C. Joint Communication, Computing, and Caching Resource Allocation for URLLEI

In this section, we show the benefit of jointly optimizing communications, computing, and caching for reducing the latency and improving the reliability, relying on our recent research [14].

To be specific, some computational tasks, like T_1 and T_3 of Fig. 5(a) may be required quite frequently. For instance, AR technology has been widely used in various fields, such as rescue missions after a disaster. The AR-aided equipment has to render and encode the corresponding scene as the rescuer moves and the viewing angle changes. The associated transcoding and rendering service is computationally intensive, but there are lots of repetitive operations. It is challengeable to meet the latency requirements of a massive amount of computation-intensive tasks. Moreover, the reliability of computations tends to decay upon increasing the clock frequency of each operational unit. Therefore, to improve the service's latency and reliability performance, caching of the computing result may be utilized.

In Fig. 5(a), there are m computing nodes and a controller, where the AR user has N computation tasks to be processed, i.e., $T_1, T_2, \ldots, T_N$. Since T_1 and T_3 have already been cached in node 1 and node m, the user does not have to offload them to nearby computing nodes for processing. As for T_2 and other subtasks that are not cached, they have to be forwarded to nearby computing nodes for cooperative processing. Naturally, caching all the computation results would be the best approach for reducing the processing latency as well as for improving the reliability of the service. However, the limited caching capacities of network entities require an appropriate caching strategy. Since the cached content may become unavailable due to a node's movement, in contrast to traditional content caching in terrestrial networks, not only the distributed popularity of the cached content, but also the dynamic nature of SAGIN nodes has to be considered. The controller will arrange for an optimal task allocation and result-caching strategy by taking into account the failure probability, the resource constraints, and time-varying topology.

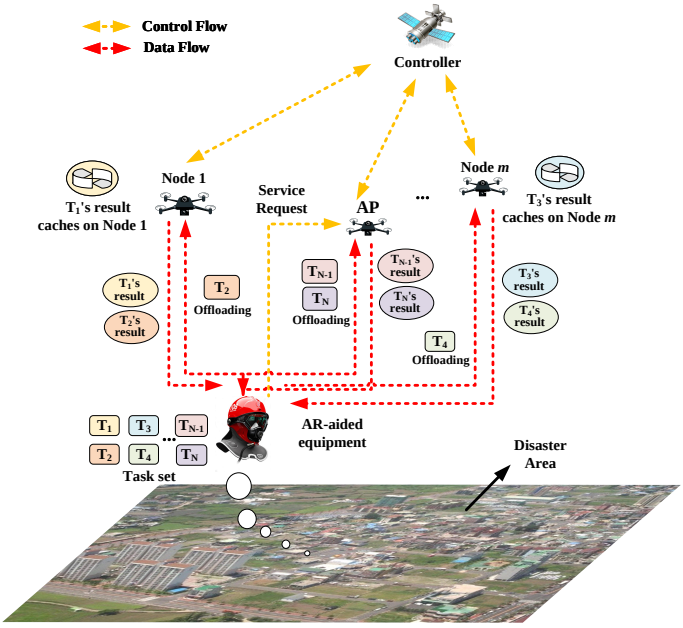
For jointly benchmarking both the latency and reliability performance, the novel concept of *normalized latency* is defined as their ratio. Explicitly, let the processing latency of a processing unit be t, and its reliability during the processing procedure be R. Hence the normalized latency L_n is given by $L_n = t/R$.

For example, a processing unit having a low reliability of $R = 0.5$ is still capable of completing its assigned task successfully in the face of an unreliable operating environment, but its latency is doubled. Fig. 5(b) shows the normalized latency of joint communications and computing optimization (corresponding to the legend *without caching*), and joint communications, computing, as well as caching optimization (corresponding to the legend *with caching*), when processing different kinds of services (i.e., gzip ASCII compress, x264 Variable Bit Rate (VBR) encode, x264 Constant Bit Rate (CBR) encode and html2text, respectively.).

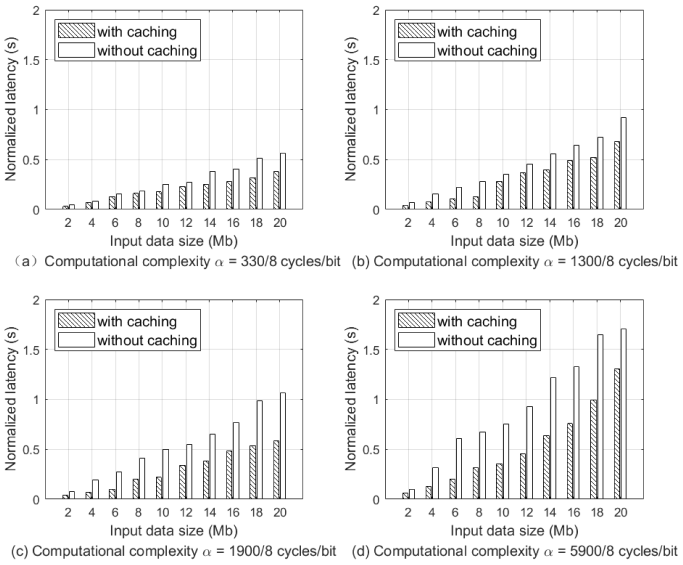
IV. Challenges and Open Issues

There are numerous open issues and challenges in making URLLEI in SAGIN a reality. Specifically,

- **Reconfigurable intelligent surface (RIS) aided interference suppression:** URLLEI dramatically increases the interaction between SAGIN nodes, which may inevitably lead to severe interference. Hence interference mitigation schemes are required for enhancing the system's efficiency. Although RISs constitute a promising technique of improving the propagation environment with the aid of low-cost reconfigurable reflecting elements, the associated interference control solutions are in their infancy. We are inspired to assemble RIS elements on the network nodes of SAGINs (e.g., UAV) for mitigating the interference and improving both the spectral and energy efficiency. However, the mobility of SAGIN nodes may result in dramatic variation of the angles of arrival/departure (AoA/AoD), hence further affecting the channel gain, which requires the joint optimization of the active and passive beamforming as well as the nodes' trajectory.
- **Machine learning aided Pareto-optimization:** The network topology as well as the communications, computing and caching resource pool are dynamically time-varying in SAGIN-based URLLEI. These new-emerging services have extremely diverse performance requirements in terms of latency, reliability, energy efficiency, bandwidth, throughput, latency jitter, etc. However, some of them are in conflict, as exemplified by the energy vs. bandwidth-efficiency or diversity vs. multiplexing trade-offs. Therefore, powerful machine learning aided multi-component



(a) Joint communications, computing and caching resource allocation.



(b) Normalized latency of joint communications and computing optimization (without caching) versus joint communications, computing and caching (with caching) parameterized by computational complexity.

Fig. 5. Joint communications, computing and caching resource allocation for URLLEI.(There are 3 computing nodes, and the AR user has 15 computation tasks to deal with, where the communication channel between the node and the AR user is the probabilistic superposition of LOS channel and NLOS channel. The data size for each task follows the uniform distribution of U(15, 20) Mb. The CPU frequency of each nodes follows the uniform distribution of U(0, 0.9) GHz, while the caching capacity of each node follows the uniform distribution of U(50, 100) Mb. Both the failure of nodes and links are assumed to follow Poisson distribution, in which the failure rate λ follows the uniform distribution of U(0, 0.005) [14]. ©IEEE)

Pareto-optimization techniques [15] have to be conceived for finding all Pareto-optimal operating points in terms of all metrics, rather than simply balancing the conflicting design metrics. These large-scale search problems require powerful optimization tools, especially in the presence of a large number of local optima, where only a tiny fraction of the potentially excessive search-space is searched, and yet the globally optimal value is found with a near-unity probability.

- **Resource-aware artificial intelligence (AI) service development:** The URLLEI services of SAGINs tend to be resource-constrained and yet they are expected to provide compelling services in the network edge. These developments have to carefully consider the impact of hardware resource constraints and again, find the entire Pareto front of all optimal operating points in terms of latency, reliability, energy consumption, relaying path life-time and so on, whilst relying on the sophisticated techniques of lossless compression, knowledge distillation and network pruning, just to name a few.

V. Conclusions

In this article, we explored the potential of the URLLEI relying on the combination of SAGIN and edge intelligence in supporting mission-critical 6G services, and we discussed the challenges that have to be tackled for making URLLEI a reality in terms of heterogeneity, time-variability and reliability. To cope with the above-mentioned challenges, we conceived a three-layer architecture, in which the unikernel-based ultra-lightweight virtualization technology is intrinsically amalgamated with a micro-service based paradigm for realizing prompt response as well as improving the reliability in the resource-constrained SAGIN. Moreover, a dynamic adaptive resource scheduling framework based on knowledge graphs and real-time monitoring was designed to alleviate network congestion. Furthermore, we discussed the beneficial role of jointly optimizing communications, computing, and caching for improving the latency and reliability. Finally, we pointed out some challenges and research opportunities in this emerging area.

References

[1] X. You, C.-X. Wang, J. Huang, X. Gao, Z. Zhang, M. Wang, Y. Huang, C. Zhang, Y. Jiang, J. Wang *et al.*, "Towards 6G wireless communication networks: Vision, enabling technologies, and new paradigm shifts," *Science China Information Sciences*, vol. 64, no. 1, pp. 1–74, 2021.

[2] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1738–1762, 2019.

[3] W. Lin, F. Shi, W. Wu, K. Li, G. Wu, and A.-A. Mohammed, "A taxonomy and survey of power models and power modeling for cloud servers," *ACM Computing Survey*, vol. 53, no. 5, sep 2020. [Online]. Available: https://doi.org/10.1145/3406208

[4] L. Bao, C. Wu, X. Bu, N. Ren, and M. Shen, "Performance modeling and workflow scheduling of microservice-based applications in clouds," *IEEE Transactions on Parallel and Distributed Systems*, vol. 30, no. 9, pp. 2114–2129, 2019.

[5] J. Zhang, T. Chen, S. Zhong, J. Wang, W. Zhang, X. Zuo, R. G. Maunder, and L. Hanzo, "Aeronautical *ad hoc* networking for the internet-above-the-clouds," *Proceedings of the IEEE*, vol. 107, no. 5, pp. 868–911, 2019.

[6] Z. Chen, C. Wang, G. Li, Z. Lou, S. Jiang, and A. Galis, "NEW IP framework and protocol for future applications," in *IEEE/IFIP Network Operations and Management Symposium (NOMS)*, Budapest, Hungary, April, 2020, pp. 1–5.

[7] G. Durisi, T. Koch, J. Östman, Y. Polyanskiy, and W. Yang, "Short-packet communications over multiple-antenna Rayleigh-fading channels," *IEEE Transactions on Communications*, vol. 64, no. 2, pp. 618–629, 2015.

[8] Q. Cui, Y. Gu, W. Ni, and R. P. Liu, "Effective capacity of licensed-assisted access in unlicensed spectrum for 5G: From theory to application," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 8, pp. 1754–1767, 2017.
[9] H. Gao, C. Liu, Y. Li, and X. Yang, "V2VR: Reliable hybrid-network-oriented V2V data transmission and routing considering RSUs and connectivity probability," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–14, 2020.
[10] P. Popovski, Č. Stefanović, J. J. Nielsen, E. De Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A.-S. Bana, "Wireless access in ultra-reliable low-latency communication (URLLC)," *IEEE Transactions on Communications*, vol. 67, no. 8, pp. 5783–5801, 2019.
[11] J. Liu and Q. Zhang, "Offloading schemes in mobile edge computing for ultra-reliable low latency communications," *IEEE Access*, vol. 6, pp. 12 825–12 837, 2018.
[12] C.-F. Liu, M. Bennis, M. Debbah, and H. V. Poor, "Dynamic task offloading and resource allocation for ultra-reliable low-latency edge computing," *IEEE Transactions on Communications*, vol. 67, no. 6, pp. 4132–4150, 2019.
[13] X. Hou, Z. Ren, J. Wang, W. Cheng, Y. Ren, K. C. Chen, and H. Zhang, "Reliable computation offloading for edge-computing-enabled software-defined IoV," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7097–7111, 2020.
[14] X. Hou, Z. Ren, J. Wang, S. Zheng, and H. Zhang, "Latency and reliability oriented collaborative optimization for multi-UAV aided mobile edge computing system," in *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Toronto, Canada, July, 2020, pp. 150–156.
[15] J. Wang, C. Jiang, H. Zhang, Y. Ren, K. C. Chen, and L. Hanzo, "Thirty years of machine learning: The road to Pareto-optimal wireless networks," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1472–1514, 2020.

Xiangwang Hou (S'19) is currently pursuing the Ph.D. degree at Tsinghua University, Beijing, China. And he received the B.E. degree from Shandong University of Technology, Shandong, China in 2017 and the M.E. degree from Xidian University, Xi'an, China in 2020. He worked as an algorithm engineer with 2012 Laboratory, Huawei Technologies Co., Ltd., and Department of Electronic Engineering, Tsinghua University, from 2020 to 2021. His research interests include edge intelligence, UAV networks, wireless AI.

Jingjing Wang (S'14-M'19-SM'21) received his B.S. degree in Electronic Information Engineering from Dalian University of Technology, Liaoning, China in 2014 and the Ph.D. degree in Information and Communication Engineering from Tsinghua University, Beijing, China in 2019, both with the highest honors. Dr. Wang is currently an associate professor at School of Cyber Science and Technology, Beihang University. His research interests include AI enhanced next-generation wireless networks, swarm intelligence and confrontation.

Zhengru Fang (S'20) received his B.S. degree in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, China in 2019. He is currently pursuing the M.S. degree in electronics and communication engineering from Tsinghua University, Beijing, China. His research interests lie in the area of Internet of Underwater Things and mobile edge computing.

Yong Ren (M'11-SM'16) received the B.S., M.S., and Ph.D. degrees in electronic engineering from the Harbin Institute of Technology, China, in 1984, 1987, and 1994, respectively. He is currently a full professor with the Department of Electronics Engineering and the Director of the Complexity Engineered Systems Lab. He holds 60 patents, and has authored or coauthored more than 300 technical papers in wireless networks. His current research interests include complex systems theory, ocean networks and swarm intelligence.

Kwang-Cheng Chen (F'07) is the Professor of Electrical Engineering, University of South Florida, Tampa, Florida. He has widely served in IEEE conference organization and journal editorship. Dr. Chen has contributed essential technology to IEEE 802, Bluetooth, LTE and LTE-A, and 5G-NR wireless standards. Dr. Chen is an IEEE Fellow and has received a number of IEEE awards. His recent research interests include wireless networks, quantum communications and computing, multi-agent systems and social networks, and cybersecurity.

Lajos Hanzo (http://www-mobile.ecs.soton.ac.uk, https://en.wikipedia.org/wiki/Lajos_Hanzo) (F'04) received his Master degree and Doctorate in 1976 and 1983, respectively from the Technical University (TU) of Budapest. He was also awarded the Doctor of Sciences (DSc) degree by the University of Southampton (2004) and Honorary Doctorates by the TU of Budapest (2009) and by the University of Edinburgh (2015). He is a Foreign Member of the Hungarian Academy of Sciences and a former Editor-in-Chief of the IEEE Press. He has served several terms as Governor of both IEEE ComSoc and of VTS. He is also a Fellow of the Royal Academy of Engineering (FREng), of the IET and of EURASIP.