SCORING: Towards Smart Collaborative cOmputing, caching and netwoRking paradIgm for Next Generation communication infrastructures

Zakaria Ait Hmitti, Hamza Ben Ammar, Ece Gelal Soyak, Youcef Kardjadja, Sepideh Malektaji, Soukaina Ouledsidi Ali, Marsa Rayani, Muhammad Saqib, Seyedreza Taghizadeh, Wessam Ajib, et al.

SCORING: Towards Smart Collaborative cOmputing, caching and netwoRking paradIgm for Next Generation communication infrastructures

Zakaria Ait Hmitti*, Hamza Ben Ammar†, Ece Gelal Soyak‡¶, Youcef Kardjadja†, Sepideh Malektaji§, Soukaina Ouledsidi Ali*, Marsa Rayani§, Muhammad Saqib*, Seyedreza Taghizadeh*, Wessam Ajib*, Halima Elbiaze*, Ozgur Ercetin‡, Yacine Ghamri-Doudane†, Roch Glitho§

* Université du Québec à Montréal, Montreal, Canada

† La Rochelle University, La Rochelle, France

‡ Sabanci University, Istanbul, Turkey

§ Concordia University, Montreal, Canada

¶ Bahcesehir University, Istanbul, Turkey

***Abstract*—The unprecedented increase of heterogeneous devices connected to the Internet, along with tight requirements of future networks, including 5G and beyond, poses new design challenges to network infrastructures. Collaborative computing, caching and communication paradigm together with artificial intelligence have the potential to enable the Next-Generation Networking Infrastructure (NGNI) that is needed to fulfill the stringent requirements of emerging applications. In this paper, we propose the SCORING project vision for reshaping the current network infrastructure towards an NGNI acting as a truly distributed, collaborative, and pervasive system that enables the execution of application-specific tasks and the storage of the related data contents in the Cloud-Edge-Mist continuum with high QoS/QoE guarantees.**

***Index Terms*—NGNI, Cloud, Edge, QoS/QoE, Distributed Computing.**

I. Introduction

Since the emergence of Cloud Computing and the associated Over-The-Top (OTT) value-added service providers more than a decade ago, the architecture of the communication infrastructure - namely the Internet and the (mobile) telecommunication infrastructure - keep improving with computing, caching and networking services becoming more coupled. OTTs are moving from being purely cloud-based to being more distributed and residing close to the edge, a concept known to be "Fog Computing". Network operators and telecom vendors advertise the "Mobile Edge Computing (MEC)" capabilities they may offer within their 5G Radio-Access and Core Networks. Lately, the GAFAM (Google, Apple, Facebook, Amazon and Microsoft) came into the play as well offering what is known as Smart Speakers (Amazon Echo, Apple HomePod and Google Home), which can also serve as IoT hubs with "Mist/Skin Computing" capabilities. While these have an important influence on the underlying network performances, such computing paradigms are still loosely coupled with each other and with the underlying communication and data storage infrastructures, e.g., even for the currently happening 5G systems. It is expected that a tight coupling of computing platforms with the networking infrastructure will be required in post-5G and Network 2030 systems, so that a large number of distributed and heterogeneous devices belonging to different stakeholders communicate and cooperate with each other in order to execute services or store data in exchange for a reward. This is what we call here the smart collaborative computing, caching and networking paradigm. The objective of the SCORING project is to develop and analyse this new paradigm. The development of such a new paradigm triggers the answering of several research questions. Among these, one might mention the following ones:

- How to proactively place computing services, while taking into account users mobility as well as other constraints such as per-computing-node battery status and computing load?
- How to efficiently manage complex computing services whose constructed upon the chaining of multiple micro-services provided by multiple tenants?
- How to proactively place stores and optimal caching of contents/functions, while taking into account the joint networking and computing constraints?
- How to optimally organize the network operations (i.e. virtual network function placement and chaining, dynamic routing enforcement, etc) in order to satisfy the distributed end-user computation requirements and their Quality of Experience (QoE)?
- How to design new network-economic models to support service offering in an optimal way, while considering the multi-stakeholder feature of the collaborative computing, caching and networking paradigm targeted here?

Taking into account this set of questions and the underlying research concerns, this paper aims at presenting our vision

This work was fully supported by the CHIST-ERA programme under the "Smart Distribution of Computing in Dynamic Networks (SDCDN)" 2019 call

of the next generation communication infrastructures beyond 2030 (i.e. a tight coupling of communication, caching and computing capabilities within a single infrastructure).

The networking policy investigated in the SCORING project is to perform the computation, caching and communication functionalities jointly, since individual functionalities affect each other as well as the final outcome. Such a need arises due to the fact that there will be a paradigm shift from the contemporary reactive networks to anticipative data-driven networks by year 2030, which is considered essential in order to support novel forward-looking scenarios, such as holographic type communications, extremely fast response in critical situations and high-precision communication demands of emerging market verticals. A first step towards offering such a target is to completely rethink the next generation network architecture, called post-5G by some and Networks 2030 by others, to tightly integrate computing, caching and communication functionalities. Then, the objective to be reached by SCORING is to optimize the collaboration between computing, caching and communication functionalities in all levels of the network in a joint manner. This problem is a distributed multi-objective optimization problem in which user's desired QoE, mobility, availability of computing, storage and energy resources as well as the network communication flexibility decisions among other things are driving the objective functions and the constraints. In SCORING, we claim that the networks of the future will be built upon a new or a refined network architecture to carry information in a manner that may evolve from, or is quite different from, today's networks. In particular, we claim that Artificial Intelligence and Machine Learning will be an indispensable part for the networks to achieve the extreme requirements dictated by future demands and thus offering the necessary adaptation and flexibility features to cope with all possible situations.

In order to share this vision, the remainder of this paper is as follows. Section II presents the SCORING project's vision for the next generation communication infrastructures beyond 2030, and describes a first architectural sketch for SCORING. Next, Section III depicts a set of use cases, that we consider the most representative of future usages. Then, in Section IV we overview the open research challenges that need to be addressed by the research community as a whole and that are addressed by the SCORING project. Finally, Section V concludes this paper by depicting the perspectives of SCORING for the next generation communication infrastructures.

II. SCORING Vision

A. Collaborative Goal- and Knowledge- Aware Networking

Due to the explosion of data communications and usages, as well as the ever more increasing demand for low latency response, the Internet has evolved from a mere transportation medium of bits into a vast network of nodes capable of data aggregation and computation. The conventional cloud paradigm, in which applications utilize servers in a remote data center cannot meet the new performance requirements imposed by the Internet of Things (IoT), Artificial Intelligence (AI), and other emerging technologies. Mean-while, contemporary multi-access (multi-user) edge computing (MEC) aims to meet these new performance requirements by distributing application functionalities to edge devices, instead of running them solely in remote data centers. Unlike these approaches, SCORING is focused on offering communication, computing, caching and their control (C^4) *within the network*, using devices that already exist within the networked system. Scoring is a step forward in the exploitation of C^4 capabilities distributed across the network, at different layers and by independent, heterogeneous entities. Devices that contribute their resources can range from edge servers, cellular base stations, and network switches to smartphones, sensors, laptops, vehicles, and up to the cloud servers themselves. We advocate a ***cooperative ecosystem*** that aims to converge the computing and the data *inside* the network in a possibly **unregulated and competitive** way to support future distributed machine learning and computation-intensive applications. Note that the participants are contributing with their data, resource or knowledge to the ecosystem. None of these come free to the participants and they need to be compensated for their contributions. Indeed, sensing and wireless transmission consumes energy, knowledge acquisition requires significant past data collection and processing, ... Hence, an incentive mechanism providing credit allocation and rewarding is crucial to encourage different parties to contribute their computational, communication and data resources.

1) Collaborative Goal-aware Networking: The keywords identifying future networks are **1) immense heterogeneity**, **2) massive scale**, and **3) pervasive intelligence**. The keyword **1) immense heterogeneity** means that the individual agents in the network will have significantly different capabilities, e.g., drones and/or IoT devices. **2) massive scale** identifies the number of agents in the network along with the data in the network will experience exponential and possibly limits growth. Finally, **3) pervasive intelligence** is the concept that AI/ML applications are not run at the edges of the network either in the cloud or end-users, but they are partially/completely diffused to run throughout the network.

Although the Internet had distributed architecture in its roots, the demands of various recent applications showed the inadequacies of the current network architecture. To address these, the knowledge defined network concept utilizes Software-Defined Networking (SDN) and Network Analytics to facilitate the adoption of AI techniques in the context of network operation and control. However, to fulfill the requirements of future networks, network management should go beyond the concept of executing the pre-defined management functions and should be able to automatically react to unknown conditions or environments. In this aspect, the massive scale property of the network prohibits the use of centralized solutions. The distributed interaction of various independent entities eliminates the effect of single point of failure, and the system can repair or correct damages without external help. The combination of the adaptability and their distributed nature presents two major advantages: robustness

against failure and scalability.

The key problem with distributed interaction of independent entities is inducing collaboration among them. Note that any party would have incurred some nontrivial cost when participating in this interaction. Hence, they would not altruistically donate their resources and/or data and risk depleting their competitive edge. These parties will be motivated to share their resources when given enough incentives, such as a guaranteed benefit from the collaboration and a fair higher reward from contributing more valuable data. **This motivates the need for measuring a party's contribution and designing an incentive-aware reward scheme accordingly**.

A particular application that we focus on in SCORING is collaborative machine learning (CML) which is an appealing paradigm to build high-quality ML models by training on the aggregated data from many parties with computations performed by yet many other parties. A fundamental question is to value the data and/or resource a party contributes to the development of the model. The informativeness of the data is one aspect that needs to be evaluated. Additionally, the timeliness of the computations and communications is well known to effect the quality of the model updates. For example, a party may contribute high computational or communication powers to deliver accurate and timely data for model update and should be compensated for this effort. **This brings the next question of how to design a reward scheme to decide the values of model rewards for incentivizing a collaboration**. These incentives appear related to solution concepts (fairness, individual rationality, stability, and group welfare) from cooperative game theory (CGT), respectively.

2) Knowledge Aware Networking: Since the early days of communication systems engineering, the complexity of the communication process has motivated a compartmentalization of the subject into separate disciplines. Shannon and Weaver famously identified three levels of problems within the broad subject of communication: *Technical, semantic and effectiveness* problems. Traditionally, communication engineers have been solely concerned with the technical problem. However, while designing solutions for networks beyond 2030, e.g. 6G wireless connectivity, (wireless) communication engineers should expand their efforts beyond the technical problem to address the semantic and effectiveness problems. Hence, the interfaces between users, sensors, and actuators carrying information relevant to semantic and effectiveness problems are limited to the application layer should be avoided. In contrast, the protocol stack needs to be augmented with a plane that exposes Application Programming Interfaces (APIs) between, on the one end, users, sensors, and actuators, and, on the other end, all layers of the Radio Access Network (RAN) and Core Network protocol stacks. The APIs enable the extraction of information to be processed via data analytics tools and the direct control of functionalities at any layer.

Cognition is defined as the state of knowing, i.e. perception. Future networks are identified by the cognitive functions (CFs) widespread in the network, where they learn optimal behavior through interaction with the network. Each CF is a learning agent which adapts itself in a changing environment based on its experience and does not follow any predefined rules, which makes the maintenance and upgrade of the system easier. The notion of CFs and their control was previously discussed in the context of Cognitive Autonomous Networks [11]. However, in SCORING, CFs will have more capability than performing automation of network functions. In SCORING, CFs are individual agents that build and upkeep their own knowledge-base from the interactions between each other as well as the network. This knowledge-base will be leveraged for addressing the needs of various applications, network services and end-users. **We may abstract the collection of CFs as a multi-agent system under various system requirements or assumptions, such as the level of communication, or degree of cooperation**. Under these assumptions, the network should address how and to what extend each CF will participate in a network service to satisfy the goals of applications.

B. SCORING High-level architecture

1) High-level architecture: We claim that the networks of the future will be built upon a new or a refined network architecture to carry information in a manner that may evolve from, or is quite different from, today's networks. In particular, we claim that Artificial Intelligence and Machine Learning will be indispensable part for the networks to achieve the extreme requirements dictated by future demands.

We believe that SCORING should go beyond traditional network design to take into account the notion of semantics and effectiveness aspects as central aspects. Besides, identifying the goal of the application facilitates relevant data finding, thus reducing the amount of data to be transmitted, saving in bandwidth, delay and energy. The goal of the application can be expressed in terms of QoE parameters and the amount of required resources (communication, energy, computation, etc.).

The networking policy investigated in the SCORING project will perform computation, caching and communication functionalities jointly, since individual functionalities affect each other as well as the final outcome. Such a need arises due to the fact that there will be a paradigm shift from the contemporary reactive networks to anticipate data-driven networks by year 2030, which is considered essential in order to support novel forward looking scenarios as described in the previous subsection.

Ultimately, SCORING aims at designing an orchestration and management framework for combined computation, communication and caching resources to allow end applications achieve their goals whatever they can be. These decisions are based on shared knowledge from different and possibly competing entities.

Figure 1 depicts the building blocks that we envision for the SCORING architecture.

- **SCORING SDN/NFV layers**: composed of the infrastructure layer, a virtualization layer and the application layer. At the infrastructure layer, different tenant/stakeholders implement the C^4 concept in a col-

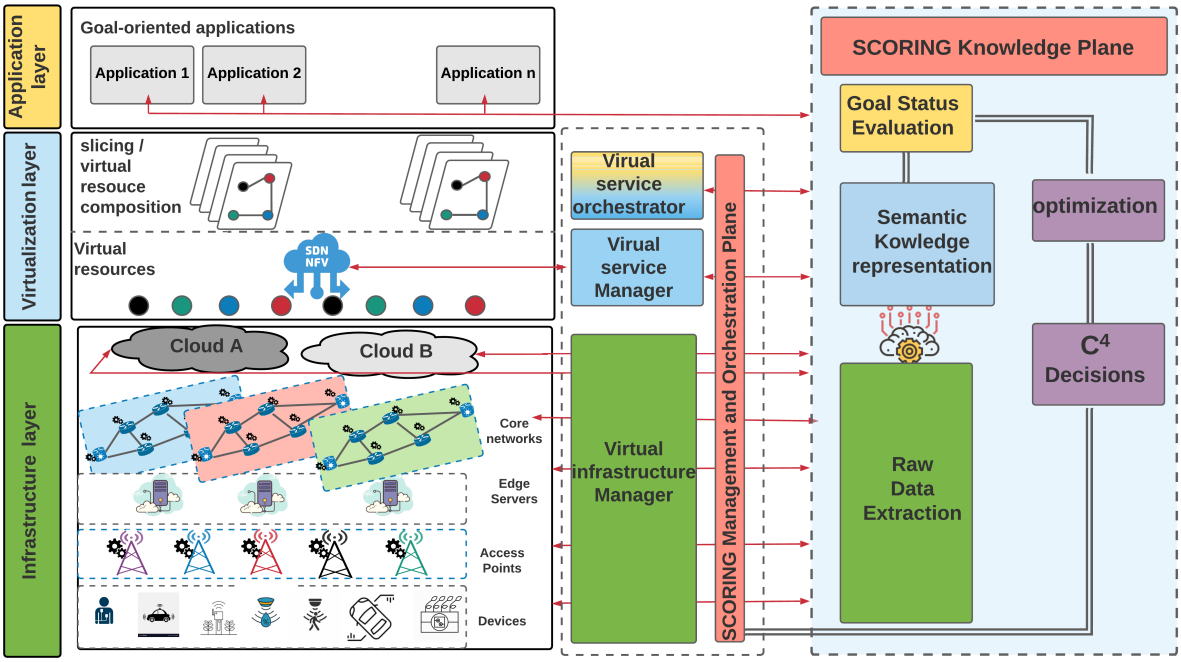

Fig. 1. SCORING high-level architecture

laborative way to share their infrastructures spanning clouds, core networks, edge server, access points and end-devices. The virtualization layer implements the appropriate techniques to enable network slicing through the combination of SDN and NFV over all the infrastructure components. Hence, slices are composed of virtual resources including network functions (NF), MEC-enabled micro-services (MeM) and infrastructure resources.

- **SCORING Management and Orchestration plane**: As previously mentioned, SCORING aims to build the NGNI architecture components based on new or refined network components. The NFV management and orchestration framework as defined by ETSI [5] manages the virtualized resources and network components. Similar cloud-native approaches are also proposed. At SCORING, we aim to extend the management and orchestration of VNFs, to managing and orchestrating MEC-enabled Microservices. As MEC will be playing an essential role in the compute, storage, and networking management in NGNI, and since more and more applications are migrating to the Microservice paradigm, we believe it to be essential that we include and specify the management and orchestration design of these latter in the new architecture design. Thus, the management support of the two paradigms, also referred to as MEC-enabled Microservices or MeMs in our architecture design, will be depicted in the proposed extended management and orchestration framework in the three main functional blocks below :
  - Virtual Service Orchestrator: Composed of two main subcomponents, NFV and MeM orchestrators (NFVO & MeMO). In addition to the regular Network Service Orchestration, this component is responsible for the orchestration of microservices based applications, which implies the planning of microservices needed resources in the Virtual Infrastructure.
  - Virtual Service Manager: Also composed of two main subcomponents, VNF and MeM managers (VNFm & MeMm). In addition to managing individual VNF lifecycles, this component manages the lifecycles of individual microservices as well. That implies instantiating, replicating, scaling, and terminating a single microservice.
  - Virtual Infrastructure Manager (VIM): Manages the NFV Infrastructure (NFVI) compute, storage and networking resources that are composed of multiple heterogeneous clouds and edge layers belonging to multiple domains.
- **SCORING Knowledge plane**: The SCORING Knowledge Plan is the cornerstone of the whole SCORING Architecture. Indeed, this one introduces many new components allowing to implement the targeted synergy for the collaborative computing, caching and communications and bringing it to reality. In order to do so, we argue that all is triggered from the fact that each of the futuristic applications will have very stringent performance requirements in terms of bandwidth, latency, reliability, etc. These requirements constitutes the goal that each application will have to announce beforehand. This one is the driver of the set-up of the corresponding goal-oriented collaborative computing, caching, communications and control slice. Not only instantiating the correct slice configuration, the objective of the SCORING Knowledge

Plan is to maintain a Goal-based Management of the C^4 slices. This is achieved through the continuous monitoring of the physical infrastructure, the virtual infrastructure and the application itself. The collected raw data is then used to build a semantic knowledge and to evaluate continuously the Goal achievement, through a "Goal Status Evaluation" Module. Continuous "optimization and adaptation" of the slice's resources allocation and virtual element instantiating and deployment (i.e. extended Service Function Chaining and deployment) is achieved using advanced Machine Learning and Optimization techniques. The foreseen optimisation and adaptation decisions are then implemented on the data plan through the SCORING Management and Orchestration Plan while necessary. **The building of the different building blocks constituting the SCORING Knowledge Plan is an open research issue to be tackled in the next few years by the research community as a whole, and by the the SCORING consortium in particular**.

III. SCORING Driving Use Cases

In this section, we present some representative use cases along with the KPI requirements illustrating the need for developing the next generation networking infrastructure (NGNI) according to the SCORING vision.

A. 360° VR Medical Training

1) Use Case Description: This use case suggests the adoption of Immersive 360° video for training physicians and paramedics. As an example, Doctors can analyze tumors and train for surgeries without any scalpel. Also, rare syndromes can be reconstructed virtually for practicing purposes. These advancements will inevitably lead to achieving significant time and cost-saving practices in both the training and teaching processes. This service provides the ability for the medical students to see through the surgery virtually, see and touch organs and muscles, also enables them to visualize the complexity of the area to be operated. With this service, the surgery process can be repeated as many times as desired. These characteristics would be impossible to obtain in a real environment. As for the speed of the operations, according to some studies involving 16 surgical residents [1], those who have been trained using Virtual Reality techniques perform operations 29% faster than those who used traditional techniques, which shows us another example of the potential of Virtual Reality in this field. The primary actors in this use case are Medical students, VR content provider, Infrastructure provider, Network provider, VR headset and handset while the secondary actors are surgical staff and Doctors. Figure 2 shows an illustrative example of such a use case.

2) System and User Requirements:

a) User Requirements: Adopting real scenarios of AR/VR training to current networks is extremely challenging. High bandwidth and low latency requirements comes into play while dealing with vast amounts of data represent some of the challenges to cope with. Allowing that in a mobile setting

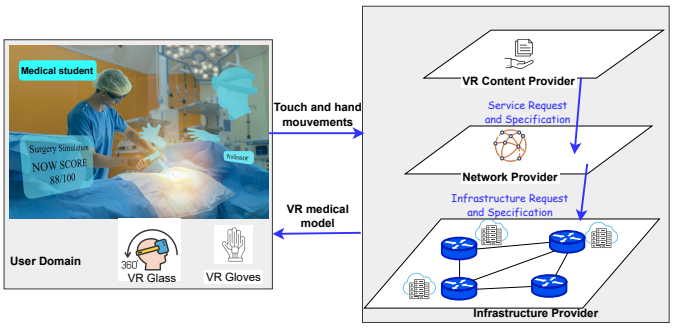

Fig. 2. VR medical training Use case illustration.

adds extra challenges related to radio resource and bandwidth limitations. Also, the excessive amount of data affects the device's energy consumption. Handling the streaming is highly time-sensitive compared with traditional streaming. Hence, the set of user requirements can be summarized as follow.

- **Simultaneous Constant Connections:** The VR medical trainers need to continuously send or receive streams of virtual patient data. Loss of connection for more than few seconds, will make the data about the current patient conditions outdated so the trainer is unable to keep the training session such as surgery session.
- **Ultra Low Latency:** Latency is the most important requirement of VR/AR applications. 360 degree VR medical training use case is highly delay sensitive because physicians perception needs accurate and smooth movements in vision. Large latency values can lead to an unpleasant VR experience and can eventually cause a motion sickness.

b) System Requirements: Beside the user requirements discussed above, there exists another set of prerequisite system requirement that we discuss in the following:

- **Ultra Low Latency:** The system should provide ultra low latency or ultra responsive connectivity to address this use case needs. The performance of VR medical training is considered as a medium-dynamic environment and the latency requirement is in the range of 10-100 ms. Having said that the realistic case of the VR training should occur by avoiding cyber-sickness which will be fulfilled by ultra-responsive connectivity.
- **Ultra High Bandwidth:** Another critical requirement for VR medical training use case is the amount of bandwidth which is required to stream the related 360 degree video to the trainers. The resolution of a VR immersive video in Head Mounted Displays (HMD) needs to be extremely close to the amount of detail the human retina can perceive.
- **High Data Rate:** A network infrastructure requires fulfilling data rate of up to 1.5 Gbps in the downlink for viewing a VR 360 degree video and 6.6 Gbps in the uplink for live broadcasting, with Round Trip Times (RTT) of less than 8 ms.

- **High Reliability:** The communication between the student and the patient during remote training should be fast and reliable so the system should ensure ultra high data rate and keep the connection stable.

3) Main Research Challenges: There is a need for more research to know if current programmable network entities are sufficient to provide ultra low latency and high data rate for such a use case. Another challenge is how to modify current infrastructures architecture to fulfil the requirements for this services. There is a need for efficient resource usage at the edge to enable interactive operations and quality of experience in VR use cases. Yet another challenge is to target the question of how the use of deep/machine/reinforcement learning algorithm at the edge can be helpful to perform optimal computing, data caching and communication resources allocation within the network itself. Also it is challenging to target the forwarding, control and management in this kind of networks to deliver the service at an acceptable level at all time.

B. Future Vehicular Networks

1) Use Case Description: With the rapid growth of urbanization and industrialization, Intelligent Transportation Systems have received increasing interest from industry and academia. ITS plays a crucial key role in providing innovative services to assist traffic authorities in efficiently managing traffic flow, making the roads safer, less congested and reducing emissions. Connected and Autonomous Vehicles (CAVs) are expected to be one of the main building blocks of future smart cities. Indeed, CAVs are equipped with abundant underutilized resources that enable a wide spectrum of innovative applications. In-vehicle sensors, communication modules and on-board units with computing and storage capabilities allow CAVs to provide important resources to Road-Side Units (RSUs).

One important use case that emerges from this context is the integration of CAVs resources with those of the RSUs by using Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) cooperation. The main objective here is to minimize the overall time spent by vehicles by proposing an optimized Traffic Light and Navigation System (TLNS) in order to achieve the following beneficial outcomes:

- Saving time for drivers.
- Optimizing fuel consumption.
- Reducing pollution.
- Decreasing accidents probabilities.

We have in Figure 3 an illustrative example of our use case. In this figure, a first layer is formed by the set of connected vehicles while another layer contains a set of RSUs capable of communicating with the vehicles and the different on road sensors and other road equipment. For instance, the traffic light system is managed mainly by the RSUs. All the elements of both these layers have various computation and storage resources. The type of data that is going to be collected by the RSUs will be the data coming from the vehicle's

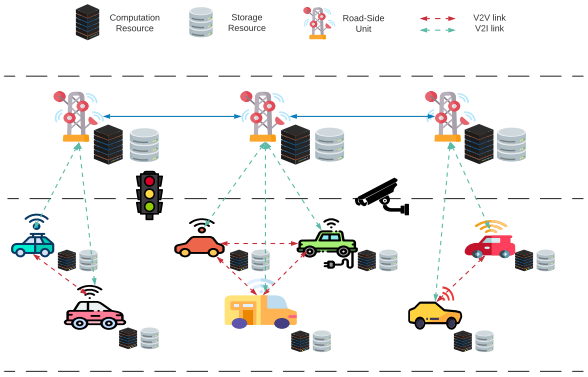

Fig. 3. Use case illustration.

on-board sensors (speed, lidar, trajectory, etc.) and from the infrastructure (traffic light, camera, etc.) This allows for a wide range of decision making services to be hosted either by the RSUs or the CAVs. Some examples of these are:

- The Optimization of the traffic light queuing models, including Automated Green Light Optimum Speed Advisory with negotiation (AGLOSA+N) and Optimized Traffic Light Information with V2I.
- Optimal speed proposal for vehicles.
- Cooperative automated driving with intention sharing and negotiation as well as maneuver coordination.
- Cooperative Adaptive Cruise Control String Management (CACCSM)

Many more examples, requiring collaborative C^4 and inspiring its design, can be found here [4].

2) Main Research Challenges: The main challenges in this use case can be resumed in the fact that we have limited resources available at RSUs and CAVs along with their high mobility, which translates into a rapid change in the availability of these latter resources. So, we have a highly dynamic and distributed system, with very large data (including sensitive-privacy data) to process using Machine Learning-based approaches and with requirements to be respected. To propose an efficient TLNS with the aim of minimizing the overall time spent on route by CAVs, we will need an optimized management of computation, storage and networking resources in this ITS system. We will need also to propose solutions that match the specificities of this system like for example using Federated-Learning approach to process the data while preserving the data privacy as well as profiting from the distributed nature of the available resources. This represents a perfect example of the SCORING project general aim by investigating and developing the smart collaborative computing, caching and networking paradigm.

C. Multi-user Immersive Presence

1) Use Case Description: Covid-19 pandemic was a tipping point in the way we do business, teaching, and in general socialize. Although the situation may no longer require social distancing, the desires and needs of the people have

transformed. This use case explores the possibility of the technology to make up for the presence and immersion of the physical world over the Internet. This requires not only visual/audio communications but also communications for other senses such as smell, taste and touch. Today, VR/AR scenarios along with the support for 360° videos are available. However, the current available solutions are rarely interactive, and in fact, they do not support multi-user interactions. Adding to this other sensory requirements brings forward a formidable challenge.

This use case focuses on the scalability of the VR/AR/MR scenarios to sustain the Quality of Experience (QoE) of multiple diverse users. The scope of the scenario covers supporting 360° videos, facilitating multiplayer VR games, or enabling multiple users to collaborate in designing/developing a physical object (*e.g.,* the design of a city prototype by architects, or the development of a robot's controls by engineers).

In these scenarios, the multiple users may have different incentives, different quality of connections and different sets of equipment when they participate in the shared task. The aim of this use case is to provide all participating parties the QoE they need. Figure 4 demonstrates the beneficiaries and requirements in this scenario.

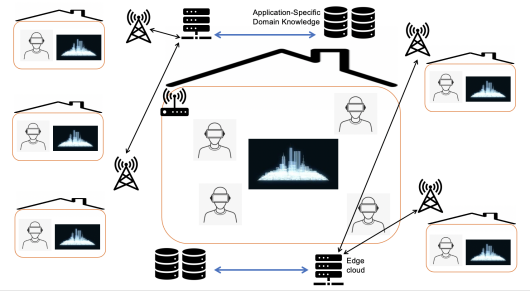

Fig. 4. Scalable XR use case illustration

2) System and User Requirements: Enabling immersive multi-user XR necessitates the user requirements stated below, which in turn calls for the provision of the subsequent system requirements.

a) User Requirements:

- **High data rate per user**
  Each user participating in a VR/AR collaborative project needs to visualize the project using a VR headset. In order for the constructed virtual object to be realistic and for an immersive collaboration experience, it is important that each user has a high throughput download link to the server where the viewpoints are rendered and/or cached.
- **Receiving other users' actions/feedback in realtime**
  Each user participating in the collaborative project generates data (in the form of voice, video, sensor output, etc.) as they interact with and potentially alter its components. For a seamless collaboration experience, it is important that one user's input is updated on all users in the system in almost realtime.
- **Concurrency:** In practice, multi-user collaboration may be needed among any number of hosts depending on the scenario. Our aim is to design a solution that can scale to support any number of participating XR users. Furthermore, these users can join from anywhere within the Internet.

In order to facilitate these user requirements with a high QoE, the system should comply with the following set of performance requirements.

b) System Requirements:

- **Very low end-to-end latency:** In order to prevent motion sickness and to facilitate the realtime user requirement mentioned above, the designed solution must ensure that once the head-mounted display moves into a new position, the user should view the new frame generated based on the new pose in at most 20-25ms [3].
- **Ultra high data rate:** To enable users to stream ultra-high definition video at a frame rate of 90Hz, the required data rate would be 71.7 Gbps [9].
- **Distributed computing and context-dependent optimization:** The designed system must allow multiple users to interact in a VR environment simultaneously. For this, the solution shall not only aim to provide QoE to the end users in the multi-user XR application scenario, but shall also take network resource optimization into account. The designed system must jointly optimize in-network caching and distributed computing. Furthermore, the communications among computing and caching resources need to be optimized based on application context. For example, it must be possible to transfer over the network only model parameters and metadata, and expect the computing (machine learning, rendering, etc.) done on or closer to the end user.

c) Performance criteria: The most crucial key performance indicator in this scenario is latency. The sum of the latencies from users' inputs (haptic, Field-of-View, etc.) and position updates to the computing resource, i.e. computation and update of new augmented/virtual environment (obtaining from cache may be possible), and downloading it onto the user headset must be smaller than 20ms. Jitter in delay must be very low and the packet delivery ratio must be very high in order to sustain immersiveness in the XR applications. The per-application and per-network-function minimum data rate requirements need to be satisfied in order to attain QoE without over-allocating resources.

3) Main Research Challenges: To facilitate scalable XR, the main challenge is to provide an effective sharing of computational, communication, and caching resources among the multiple participants in the scenario. The split resources need to be allocated in order to satisfy the unique QoE of the participants, subject to their facilities and limitations. To this purpose, the placement of an application-specific domain knowledge base is needed to analyze, synthesize and cache the knowledge generated at the edge. Thus, the optimization of the computation, communication, and caching must be performed coherently with the application context.

IV. SCORING Design and Research Objective

A. Fitting Distributed AI in NGNI

According to [17], the total data traffic generated by end devices at the edge is 850 ZB, whereas the total data center traffic is only 20.6 ZB. It is agreed upon among researchers that NGNI will facilitate and rely on highly distributed AI, moving the intelligence from the central cloud to edge-computing resources [12]. Enabling intelligence at the network edge, i.e. *edge intelligence* (EI), facilitates data-intensive IoT applications in a variety of domains from connected autonomous vehicles to the Internet of Skills.

Though the literature comprises recent work on EI [6][8][17], the research is still in its infancy, with some solutions not scaling efficiently [8][6], and some others merely providing basic suggestions [17][10]. One of the main challenges in this regard is data availability. Though federated learning tried to solve this problem to some level [14], the learning process is still negatively affected due to various end devices providing heterogeneous forms of data. Use of transfer learning in such environments is another research question that needs to be answered.

B. Embracing satellites in NGNI

NGNI, on one side, should encompass both ground and aerial users and objects [13][2], and should inevitably embrace satellites. On the other side, however, the actualization of edge computing and storage is challenging for satellites because of the on-board limitations. Meeting the latency requirements in NGNI is another challenge for the satellites due to the physical distance between the satellites and end nodes. A maximum distance of 150 km should be respected to reach a latency of 1 ms or less. Today's satellites cannot keep within this distance limit, and consequently, NGNI has a major challenge in embracing them while respecting the 1 ms latency threshold. It is then a challenging task to consider both aforementioned sides in the process of designing a new architecture, which can not miss satellites.

C. Collaborating C^4

In addition to the edge support for processing and data storage, with the appearance of devices that can serve as IoT hubs with *mist/dew/skin computing* capabilities lately, we have a variety of computing environments and interactions among them and the cloud. These diverse environments can help applications achieve stringent service requirements if they work together. The need for close coupling between these distributed **computing** paradigms as well as the underlying **communication** and **caching** infrastructures, as well as their associated **control** logic, is apparent. Providing frameworks for the large number of distributed and heterogeneous devices belonging to different stakeholders to efficiently and effectively communicate and **cooperate** with each other is challenging. SCORING aims to design a Collaborative C^4 (Computing, Caching, Communication and Control) paradigm with a tight coupling of computing and caching (i.e., ephemeral storage) platforms with the networking infrastructure.

D. Preserving privacy in the era of AI

The continuously emerging IoT services and connected objects are steadily creating massive data. On the other hand, machine learning (ML) and deep learning (DL) based techniques train models via large data sets. Traditionally, ML systems use a centralized architecture, which creates a single point of failure that is prone to cyber attacks. Since training data usually yields personal information, such attacks may be destructive for the underlying service. On the other hand, with the emergence of a variety of compute-capable components in the network and with distributed computing paradigms, it is possible to analyze the massive data in a distributed manner. With this motivation, a line of research work offered to use Federated Learning as a privacy preserving ML technique. In addition, blockchain solutions provide a distributed infrastructure for secure data access without need for external central entities. Design of FL and its use with blockchain may allow data confidentiality and secure access control.

E. In-Network Intelligence

With the increasing number of emerging application and technologies, the standard match-action paradigm is not capable to distinguish the changing traffic patterns. PDP devices brings flexibility to express matching criteria and device behaviour. Hence, bringing the possibility for network driven updates. In [16], a trained ML classification model is mapped to match-action pipelines. Likewise, an extended work [15] represent a trade-off - particularly focusing on latency analysis of per-packet and per-flow model. As a result, classifying a flow leads to more accurate results, at the cost of keeping a per-flow table. Whereas, despite having low overhead, the per-packet model has lower accuracy and suffers from the issue of flow fragmentation.

Despite the high-accuracy results of specific model, one fit model is not applicable to the changing traffic environment. Also, a per-flow model may be memory intensive since it requires storing flow-related metrics. The solution to both problems passes through the interaction between the control and data planes. To be more specific, the control plane can assume the role of removing inactive flow-entries from the device's memory and receiving important monitoring information from the data plane. Later, such information can be sent to the knowledge plane to help building a better model. Moreover, the in-network classifier can be combined with smart agents that decide when to perform new measurements and train/deploy new models.

Besides, the data plane is hard to install machine learning models that are trained by large dataset (e.g., several days network traffic), due to the large size of decision trees. A solution is to split large traffic into several parts based on periods of time, and the controller (e.g., P4 runtime) can dynamically configure and modify machine learning models. At last, in-network training is a big challenge as the PDP devices are resources constrained and does not support complex mathematical operations.

F. Service-Aware Scalable Routing

Service access time is a fundamental challenge posed by the ever-increasing demand of remote service execution. The proposed service aware routing scheme [7], however, focus on key performance aspect (i.e., latency) but do not consider the design aspects of a scalable and dynamic network. Generally, the end points association are based on some dynamically maintained metrics. Therefore, the increases in number of offered services as well as frequency of constraint announcements cause messaging overhead. Two possible solution seems promising: the suspension of announcement for previously satisfied metrics and construction of virtual loop free topologies. Hence, leading to the development of scalable service aware routing protocols. Moreover, the future dynamic network poses a significant challenge to deal with state migration during the ongoing affinity relation.

G. Goal-oriented Communications

Traditional communication paradigm, considering the performance metrics of throughput, delay, or packet loss, is oblivious to data packets' content at the physical and data link layers. In other words, at these layers, packets are treated equally regardless of the amount of *information* they would bring to the destination. Given the anticipated astronomical growth in traffic demand with beyond 5G applications, the content-blind communication approach may fail to provide the necessary application quality of experience (QoE) and may lead to performance bottlenecks.

Researchers have been pushing the boundaries of the traditional paradigm. Of the most successful recent efforts, was the introduction of the Age of Information (AoI) metric. The AoI quantifies the notion of information freshness by measuring the information time lag at the destination. Hence, the AoI infers the *importance* of packets only through their timestamps and does not consider their content. Given this shortcoming, researchers have proposed data acquisition and scheduling schemes based on error minimization and the notion of the value of information in control theory. Even though this is a step forward in the right direction, error-based metrics come short in capturing a crucial aspect of the communication: its goal. In fact, these metrics do not consider what the packets are used for, but rather their optimization aims solely to reduce the mismatch between the physical process and its estimate at the destination. Given that the communication's goal is neglected, adopting these metrics could hinder achieving the desired goal. New metrics that can be easily adapted to different applications are needed for future networks.

V. Conclusion

As we are witnessing a paradigm shift from the contemporary reactive networks to anticipative data-driven networks by year 2030, the networking policy investigated in the SCORING project is to perform the computation, caching and communication functionalities jointly. Indeed, supporting novel forward-looking scenarios, such as holographic type communications, extremely fast response in critical situations and high-precision communication demands of emerging market verticals calls for joint efforts from academia and industry towards offering such a target. The next generation network infrastructure, post-5G or NET 2030, architecture needs to be completely rethinked to tightly integrate computing, caching and communication functionalities as well as their control logic. Then, the objective to be reached by SCORING is to optimize the collaboration between computing, caching and communication functionalities in all levels of the network in a joint manner. Furthermore, we claim that Artificial Intelligence and Machine Learning will be indispensable part for the networks to achieve the extreme requirements dictated by future demands and thus offering the necessary adaptation and flexibility features to cope with all possible situations.

References

[1] N.E.Seymour et al. "Equivalent Data Information of Sensory and Motor Signals in the Human Body". In: *Annals of surgery* 236.4 (2002), 458–464.

[2] Petros S. Bithas et al. "A Survey on Machine-Learning Techniques for UAV-Based Communications". In: *Sensors* 19.23 (2019).

[3] Kevin Boos, David Chu, and Eduardo Cuervo. "Demo: FlashBack: Immersive Virtual Reality on Mobile Devices via Rendering Memoization". In: *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services Companion*. MobiSys '16 Companion. Association for Computing Machinery, 2016, p. 94.

[4] CAR 2 CAR Communication Consortium. *Guidance for day 2 and beyond roadmap, V1.1.0*. Tech. rep. C2C-CC, Dec. 2019.

[5] ETSI GS NFV-MAN 001. *Network Functions Virtualization (NFV); Management and Orchestration v1.1.1*. Tech. rep. ETSI, Dec. 2014.

[6] Georgios Fragkos, Eirini Eleni Tsiropoulou, and Symeon Papavassiliou. "Artificial Intelligence Enabled Distributed Edge Computing for Internet of Things Applications". In: *2020 16th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. 2020, pp. 450–457.

[7] René Glebke et al. "Service-based Forwarding via Programmable Dataplanes". In: *2021 IEEE 22nd International Conference on High Performance Switching and Routing (HPSR)*. IEEE. 2021, pp. 1–8.

[8] Nourah Janbi et al. "Distributed Artificial Intelligence-as-a-Service (DAIaaS) for Smarter IoE and 6G Environments". In: *Sensors* 20.20 (2020).

[9] Luyang Liu et al. "Cutting the Cord: Designing a High-Quality Untethered VR System with Low Latency Remote Rendering". In: *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. MobiSys '18. Association for Computing Machinery, 2018, 68–80.

[10] Lauri Lovén et al. “EdgeAI: A Vision for Distributed, Edge-native Artificial Intelligence in Future 6G Networks”. In: *The 1st 6G Wireless Summit*. Mar. 2019, pp. 1–2.

[11] Stephen S. Mwanje and Christian Mannweiler. “Towards Cognitive Autonomous Networks in 5G”. In: *2018 ITU Kaleidoscope: Machine Learning for a 5G Future (ITU K)*. 2018, pp. 1–8.

[12] Ella Peltonen et al. *6G White Paper on Edge Intelligence*. 2020. arXiv: 2004.14850 [cs.DC].

[13] Walid Saad, Mehdi Bennis, and Mingzhe Chen. “A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems”. In: *IEEE Network* 34.3 (2020), pp. 134–142.

[14] Xiaofei Wang et al. “In-Edge AI: Intelligentizing Mobile Edge Computing, Caching and Communication by Federated Learning”. In: *IEEE Network* 33.5 (2019), pp. 156–165.

[15] Bruno Missi Xavier et al. “Programmable Switches for in-Networking Classification”. In: *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE. 2021, pp. 1–10.

[16] Zhaoqi Xiong and Noa Zilberman. “Do switches dream of machine learning? Toward in-network classification”. In: *Proceedings of the 18th ACM workshop on hot topics in networks*. HotNets ’19. 2019, pp. 25–33.

[17] Zhi Zhou et al. “Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing”. In: *Proceedings of the IEEE* 107.8 (2019), pp. 1738–1762.