Edge on Wheels With OMNIBUS Networking for 6G Technology

MUSTAFA ERGEN\textsuperscript{1,2}, FERIDE INAN\textsuperscript{2}, ONUR ERGEN\textsuperscript{1,3}, IBRAHEEM SHAYEA\textsuperscript{1}, MEHMET FATIH TUYSUZ\textsuperscript{4}, AZIZUL AZIZAN\textsuperscript{5}, NAZIM KEMAL URE\textsuperscript{6}, AND MAZIAR NEKOVEE\textsuperscript{7}, (Member, IEEE)

\textsuperscript{1}Electronics and Communication Engineering Department, Faculty of Electrical and Electronics Engineering, Istanbul Technical University (ITU), 34469 Istanbul, Turkey
\textsuperscript{2}Ambeent Inc., Yuba, CA 95991, USA
\textsuperscript{3}Electrical and Electronics Engineering, Koc University, 34450 Istanbul, Turkey
\textsuperscript{4}Computer and Information Sciences Department, Northumbria University, Newcastle Upon Tyne NE1 8ST, U.K.
\textsuperscript{5}Advanced Informatics Department, Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Kuala Lumpur 54100, Malaysia
\textsuperscript{6}Istanbul Technical University Artificial Intelligence and Data Science Application and Research Center, ITU Ayazaga Campus, 34469 Istanbul, Turkey
\textsuperscript{7}Centre for Advanced Communications, Mobile Technology and IoT, School of Engineering and Informatics, University of Sussex, Brighton BN1 9RH, U.K.

Corresponding author: Mustafa Ergen (mustafaergen@itu.edu.tr)

This work was mainly supported by Ambeent Inc., and in part by the Universiti Teknologi Malaysia, Research University Grant Scheme Tier 2, under Grant PY/2019/00325.

ABSTRACT

In recent years, both the scientific community and the industry have focused on moving computational resources with remote data centres from the centralized cloud to decentralised computing, making them closer to the source or the so called “edge” of the network. This is due to the fact that the cloud system alone cannot sufficiently support the huge demands of future networks with the massive growth of new, time-critical applications such as self-driving vehicles, Augmented Reality/Virtual Reality techniques, advanced robotics and critical remote control of smart Internet-of-Things applications. While decentralised edge computing will form the backbone of future heterogeneous networks, it still remains at its infancy stage. Currently, there is no comprehensive platform. In this article, we propose a novel decentralised edge architecture, a solution called OMNIBUS, which enables a continuous distribution of computational capacity for end-devices in different localities by exploiting moving vehicles as storage and computation resources. Scalability and adaptability are the main features that differentiate the proposed solution from existing edge computing models. The proposed solution has the potential to scale infinitely, which will lead to a significant increase in network speed. The OMNIBUS solution rests on developing two predictive models: (i) to learn timing and direction of vehicular movements to ascertain computational capacity for a given locale, and (ii) to introduce a theoretical framework for sequential to parallel conversion in learning, optimisation and caching under contingent circumstances due to vehicles in motion.

INDEX TERMS

Edge computing, 5G, 6G, V2X, ubiquitous AI, distributed AI, multi-access edge computing (MEC).

I. INTRODUCTION

In 2025, there will be more than 75 billion connected devices around the world, as predicted by Cisco [1]. According to Cisco’s forecast, by 2030, the predicted number of connected devices to the Internet will reach up to 500 billion [2]. The centralised cloud system alone will be incapable of efficiently handling future networks [3]–[5]. An environment where billions of devices equipped with sensors, geared to collect huge amounts of data and drawing inferences to conduct an action, will be present. Transferring massive amounts of data from connected devices to the cloud for analysis will create very crowded traffic on the network infrastructure [6]–[10]. Moreover, the back and forth transfer of data between the cloud and individual devices increases latency. Numerous new applications, such as self-driving vehicles, remote surgery, AR/VR, 8K video, advanced robotics in manufacturing and drone surveillance communication, require real-time and ultra-low latency performance [9], [11], [12]. This is the current case with Fifth-Generation (5G) networks,
however, it will become more critical with the deployment of Sixth-Generation (6G) networks, which require higher frequency bands and lower latency in comparison to 5G networks [13]–[15]. Furthermore, implementing more robust edge computing will contribute for solving bandwidth, autonomy and privacy requirements. The fast-digital transformation and accelerated cloud and edge adoption are driving Wide Area Network (WAN) edge infrastructure changes for infrastructure and operations strategies for future networking. Also, the edge computing has moved from Internet-of-Things (IoT) concentrated to a largely considered complement to the more centralized hyper-scale cloud. So, the edge role in distributed cloud and digital business ensures that despite the nascent market, technologies and architectures, edge computing is here to stay. In addition, the edge computing use-case landscape is broad, and early deployments are highly customized. Therefore, in the future network developing a multiyear edge computing strategy is needed to include the variety of use cases required by their enterprise.

In view of these challenges, data centre operations are being pushed to the “edge” of the network. The edge allows certain time-critical and security-sensitive Artificial Intelligence (AI) applications to operate, either entirely on a device, or in conjunction with localised data centres [16], [17]. To the best of our knowledge, none of the proposed edge architectures [18]–[23] are sufficient in handling massive data traffic computing in future networks. Most existing solutions focus on installing edge devices to singular static locations (e.g., factories, shopping malls) or around specific geographic areas (urban centres), which bear the cost of additional infrastructure deployments [24]–[27]. There is also a growing body of research on exploiting connectivity among end-devices in close proximity to process tasks cooperatively in local area computation groups. However, these efforts are also limited in scope.

The main contribution of this article is to develop a novel mobile edge architecture, introduced as the OMNIBUS solution. The aim is to advance a decentralised computing and storage architecture using vehicles. Road vehicles are the most promising candidates for future distributed data centers on the edge of the network for two primary reasons: (1) most road vehicles display predictable movement patterns, and (2) hardware capabilities for storage and computing in cars are expected to tremendously advance in coming years. In our proposed architecture, clusters of cars form a powerful local hub in individual areas. They are capable of offering high ad-hoc computational and virtualised resources for end-devices.

The OMNIBUS solution will address two interrelated scientific challenges. The first is the creation of a mobility prediction model. Timing and direction of vehicular movements will be predicted proactively to ascertain computational capacity and also to determine the flow of individual cars in a given area at a certain time. A local hub can be created for end-devices in that area by using cars as building blocks. Efficient algorithms can be developed to ensure computing and storage workloads for individual regions as cars move in and out of a given area. The second challenge is minimising networking overheads as cars move in and out of a region. It is necessary to study the required distribution of computing and storage resources among cars. The main scientific contribution of the OMNIBUS solution will be to initiate a theoretical framework for sequential to parallel conversion in learning, optimisation and caching algorithms under unreliable circumstances for time-critical performance. This propose OMNIBUS solution reaches near device edge. Thus, the future of edge computing is expected to be scalable and energy efficient, providing ultra-low latency networking for time critical applications at a flat cost, as shown in Fig. 2.

The rest of this paper is organised as follows. Edge Computing (EC) and Edge Computing with End-Devices (ECED) are discussed in Sections II and III, respectively. The OMNIBUS architecture and its implementation are described in Section IV, while its methodology is explained in Section V. Next, proof of the concept is discussed in Section VI. The study’s challenges and impacts with the proposed solution are then highlighted in Sections VII and VIII,

---

A French word originated in 1820s for transportation.
II. BACKGROUND AND RELATED WORK

This section provides a brief overview of Edge Computing. That includes the concept, type of solutions and related works that have been performed in the literature. This section will briefly summarise the research challenges that must be addressed.

A. EDGE COMPUTING

The cloud has been an important solution for companies looking to scale their computational operations without investing in new infrastructure, while cutting down on operational costs, by transferring their data centres to cloud providers. Although the cloud catalysed the growth and adoption of big data, it hides the costs and limitations related to network latency, security and privacy. This becomes increasingly significant with the integration between space and terrestrial networks, especially with 5G and 6G networks, as illustrated in the architecture described in Fig. 3. The requirements of ultra-reliable networks in 5G and 6G have become extremely critical. For that reason, edge computing is a significant technology enabler for achieving this target. In recent years, discussions on computational operations have increasingly shifted from centralised cloud, with remote data centres, to decentralised computing that is closer to the source or the so called “edge” of the network.

Edge Solutions allow information processing to take place at the device or gateway level. This reduces the need to transfer data back and forth between the cloud or a data centre, therefore, decreasing latency, bandwidth requirements and connectivity dependencies. Besides technical reasons, decentralised computing is energy saving; given the power and cooling costs associated with big data centres. Research on edge computing is driven by security and privacy concerns related to the centralised cloud on the part of states, firms and consumers [28]–[33]. At the same time, falling prices in computation and storage, together with the rise of machine learning, is prompting the adoption of edge computing. According to the International Data Corporation (IDC), by 2019, a minimum of 40% of data created by the IoT “will be stored, processed, analysed and acted upon close to, or at the edge of, the network” [34].

Systems typically known as edge computing include Cyber Foraging [35], [36], cloudlets [37], [38], Fog Computing [39]–[42] and Mobile Edge Computing (MEC).

The name Mobile Edge Computing was recently replaced with Multi-Access Edge Computing. Multi-Access Edge
Computing was initiated by the European Telecommunication Standards Institute (ETSI) in 2014 with a focus on mobile networks and Virtual Machine technology [43]. In 2017, its scope was expanded to incorporate non-mobile network requirements and other virtualisation technologies. The concept initially envisioned providing cloud-computing capabilities at the edge of mobile networks and within the Radio Access Network (RAN) by deploying MEC servers at Long-Term Evolution (LTE)/LTE-A macro base station (eNodeB) sites, 3G Radio Network Controller (RNC) sites and at multi-Radio Access Technology (RAT) sites. The ETSI initiative was also assigned a key role in standardising the Application Programming Interfaces (APIs) between the mobile edge platform and applications, with the aim of fostering innovation in an open environment.

**Fog computing**, a concept introduced by Cisco in 2012, is an extension of the cloud computing paradigm, from the core to the edge of the network [44]. Hence, unlike MEC, fog is strongly linked to the cloud and is unable to operate in a standalone mode. As a result, there has been special focus on communication between the fog and cloud [45]–[47]. Unlike MEC, which is generally deployed at a base station, fog nodes can be placed anywhere with a network connection; e.g., factory floor, on top of a power pole, railway tracks, vehicles, etc., [44]. Cisco offers application platforms to simplify fog application development, as well as Cisco Fog Data Services for data analytics [48].

In parallel, big Internet companies are rolling out edge infrastructure. Facebook is building micro data centres for specific types of applications and workloads. Amazon has launched Amazon Web Services (AWS) Greengrass to allow developers to move some tasks to the device itself. There are also companies that describe themselves as edge companies, including EdgeConneX and vXchnge, that are building networks of urban data centres. For instance, Vapor IO, a startup company, develops micro-centers that can be deployed anywhere.

Most of the proposed applications bear the cost of additional infrastructure deployment, whether it is installing edge devices to singular static locations (e.g., factories, shopping malls) or around specific geographic areas (urban centres). Hence, scalability is negatively affected with the massive increase of people performing transactions within a given and specific edge domain. In contrast, with our proposed solution, scalability rises with adoption. In other words, as a vast number of users require computing and storage transactions on the network, we expect computation to become much faster, allowing for the rise of a truly global network.

**B. RELATED WORK**

A growing body of work is now focusing on exploiting connectivity among end-devices, particularly, mobile devices (mobile cloud computing) in close proximity, to process tasks cooperatively in local area computation groups [49]. The end devices in a given area communicate with each other to find resources and deliver requests. Hence, the end-user stratum and edge stratum are merged. In the literature, collaboration is a central feature. The authors in [50] proposed a vision where mobile devices form “mobile clouds”, or mclouds, to accomplish tasks locally. In [51], “Transient clouds” were suggested as a collaborative computing platform where nearby devices form an ad-hoc network to provide various capabilities as cloud services. In [52], a resource sharing mechanism was proposed to utilise mobile devices through opportunistic contacts in order to emphasise resource aspects of mobile cloud computing. Other works, published in [53], focused on Virtual Machine (VM) technology for harnessing the full power of local hardware at the edges of the Internet. On the other hand, [54] proposed an adaptive method of resource discovery and addressed service provisioning in opportunistic computing environments for managing higher load requests without causing instability [55]. The proposal of [56], an architecture called Vehicular Fog Computing (VFC) for vehicular applications, possesses some similarities with the OMNIBUS solution.

In 2019, the authors [57] suggested a two-stage method for detecting the optimal solution regarding Distributed Stream Processing (DSP) applications in edge computing. The aim of this method is to address the joint operator scaling and determine the problem. The target is to offer higher cost efficiency while considering user-defined QoS constraints. The study was conducted based on experiments of real-world DSP test cases.

In 2019, a new study [58] had focused on data-intensive applications deployed with multiple service components on the edge. The proposed method was implemented based on the Genetic Algorithm (DSEGA). Five algorithms were provided to obtain an optimal deployment scheme.

In 2019, another study [59] presented a wireless acoustic sensor network that was integrated with the edge computing structure. This method has been described as a low-cost and energy-efficient approach. It was mainly designed for remote acoustic monitoring and in situ analysis. The reported results indicated that the suggested method achieved noticeable outcomes in terms of acoustic equivalence and power saving in comparison to existing solutions.

In 2020, the work of [60] focused on the optimal application deployment in resource-constrained distributed edges. The aim of the study is to determine the deployment problems of microservice-based applications in MEC. A method was proposed to optimise the deployment cost of the application with resources and performance limitations.

In 2020, another study [61] focused on dynamic resource allocation in an edge environment for IoT systems. This research implements a Reinforcement Learning process to a trained resource allocating policy.

In 2020, [62] discussed the AI for edge computing. The focus is on finding more optimal AI solutions that can contribute to addressing edge computing problems. The main idea and the research road-map of Edge Intelligence were provided.
In 2020, a scheduling method has been proposed [63] to address the multi-workflow scheduling problem with proximity constraint in the Edge Computing environment. The main aim is to minimise costs. The study was performed based on experimental and real-world data-sets.

Although this preliminary work also refers to vehicles (both moving and parked) as an infrastructure for communication and computation, it only does so with service vehicles alone and not all other connected devices and applications. In all proposals regarding edge computing that merge the end-user stratum and the edge stratum, devices share their resources among each other in a limited area. Thus, our challenge is to bring together a whole range of technologies for decentralised computing.

C. RESEARCH GAP

To summarise, the issue of the central cloud computing is that it may not efficiently serve in future networks due to the massive growth in mobile data traffic. Therefore, technical challenges are anticipated to emerge in future networks. These challenges must be sufficiently addressed. They are briefly summarised in the following:

(a) Insufficient Bandwidth: Insufficient bandwidth will be one of the key issues that require attention in future networks. This issue will be due to the continuous radical increase in data movements between connected end devices and the central cloud.

(b) Ultra Low Latency: The Critical Remote Control (CRC) applications are one of the key innovations in 5G and 6G networks. They require ultra-reliable communication. Enabling this technology to work efficiently entails moving the central computing to multiple distributed edge computing units.

(c) Cost Reduction Need: The cost factor is considered as the main key challenge in the deployment of future ultradense networks. Looking for wider bandwidth bands and deploying additional network infrastructure are all factors that raise cost. Therefore, this issue must also be addressed as well.

(d) Security Attacks and Threats: Security attacks and threats are key issues facing the implementation of MEC. They fundamentally result from design flaws, inappropriate configurations and/or implementation bugs. Therefore, defense mechanisms are necessary for recognising attacks and threats resulting from all sources. With the OMNIBUS, security attacks and threats will be a problem as well. However, finding the solution for security is not the main target of this paper. The aim is to present the concept of the proposed architecture. Thus, further research is required to address this issue.

III. OMNIBUS ARCHITECTURE AND IMPLEMENTATION

The OMNIBUS proposal expands the idea of end user stratum and edge stratum to the next level. By introducing a predictive platform for mobility patterns and for the distribution of storage and computation capacity among cars, it paves the way for an efficient and highly scalable architecture for device-level edge computing. The general concept and structure of the proposed OMNIBUS architecture solution for decentralised storage and computing are illustrated in Figs. 4 and 5.

A. DISTRIBUTED MACHINE LEARNING AND MODEL PARALLELISM

Our goal is to speed up large-scale Machine Learning (ML) by reducing training time via parallel or distributed computing. Data parallelism and model parallelism are also methods for improving speed. Data parallelism partitions the data, while our solution (model parallelism), partitions the ML model itself to distribute the workload to multiple computational workers, as described in Figs. 6 and 7. In our architecture, it is necessary to understand the methods to partition the ML model according to heterogeneity and mobility of cars, while ensuring interoperability on the level of different service providers. Given the high number of machine learning models, with each model possessing its own characteristics and representations, there is no principle way to implement model parallelism. In distributed machine learning, the synchronisation overhead increases as the system scales.

Our solution also leverages machine learning software methods to optimise the hyper-parameters of selected algorithms. It further utilises Hadoop frameworks, including Hadoop Distributed File System (HDFS), Spark and Cassandra for faster and energy-efficient computation. The Hadoop framework employs simple programming models that allow the distributed processing of large datasets across computer clusters. Spark is a computation engine for Hadoop data that supports an entire range of applications; e.g., machine learning, stream processing, etc. Cassandra is a highly scalable database with no single point of failure, which makes it ideal for mission critical data.

B. NEXT GENERATION DISTRIBUTED LEDGER TECHNOLOGY

For storing data and enabling fast computation in the network, it is crucial to study directed acyclic graphs (DAG)-based ledger technology. DAG may be the primary data structure for us to create a peer-to-peer network protocol. This will allow us to advance in distributed machine learning methods to add cognitive capabilities as well as consensus mechanisms.

DAG is largely more suitable for our solution due to its potential in scalability and lesser processing power requirements compared to Bitcoin-like blockchain ledger technologies [71], [72]. In the blockchain system, the block size and the time required to generate a new block places limitations on throughput and transaction times. In contrast to blockchain technology, DAG transactions are not grouped into blocks. Each new transaction confirms at least one previous transaction, and transactions are represented as “units.” Hence, selection of a branch and detection of double-transaction are decoupled from transaction verification, which allows nodes to verify transactions in parallel. As a result, DAG has the potential to achieve unlimited scalability.
FIGURE 4. The general concept of the proposed OMNIBUS architecture solution for decentralised storage and computing. Any accurate trajectory prediction method can be used as the backbone for this distributed computing [64]–[70].

FIGURE 5. Structure of the proposed OMNIBUS architecture solution for decentralised storage and computing. However, as DAG-based solutions emerge for high-frequency transaction scenarios, problems may arise in low frequency transactions [71]. When an old transaction is not able to receive a sufficient number of new transactions to be verified, the old transaction may not be confirmed in time or not be confirmed at all. To ensure a continuous system, our solution optimises high frequency and low frequency transactions by harmonising DAG and blockchain concepts as required.

C. MOBILITY MODELS

Mobility data contains the approximate whereabouts of individuals and is used to explore the movement patterns of individuals across space and time. The vehicular mobility maps have been addressed in the literature [73]–[75]. However, they are not comprehensive enough. Mobility data is among the most sensitive data that is currently being collected. While the discussion on individual privacy with respect to mobility data is on the rise, research in this area is still limited [76]–[80]. The OMNIBUS solution is proposed to design a targeted mobility model for addressing specific tasks that do not compromise an individual’s privacy. In doing so, leveraging machine learning software methods distributed ledger technologies are very important. In this proposed solution, any accurate and efficient trajectory prediction method can be used as the backbone for this...
distributed computing [64]–[70]. These algorithms only focus on trajectory prediction, whereas the use of these predictions will enable the computation to be distributed more efficiently.

IV. METHODOLOGY

A. AGGREGATED MOBILITY HANDLING (AMH)

AHM aims to accurately depict vehicular behaviour and focuses on the following principles: (i) Charting out mobility patterns of moving cars to optimise computing and storage distributions among them. Mobility patterns will be learned in mixed autonomy with each car that shares the mobility patterns and movements of other cars. (ii) Aggregation, which will take place over a combination of DAG and blockchain based distributed ledger technologies, must depend on different frequency scenarios. (iii) Leveraging our model is required to solve the massive routing problem so as to bring Internet data to unconnected regions.

1) SOLVING THE MOBILITY HANDLING PROBLEM

Self-driving cars, sharing rides and similar exercises in mobility as a service (MaaS) are turning transportation into mixed autonomy systems; integrating AI/ML technology. By reducing randomness, mixed autonomy systems (including autonomous and non-autonomous vehicles) make it possible to accurately depict vehicular behaviour (the mobility handling problem) [81], [82]. In this relation, mobility pattern challenges and requirements of mixed autonomy systems are studied; more specifically, a convex optimisation method predicting the flow is used to represent the coordination of automated vehicles, which relies on accurate traffic flow sensing [10], [83], [84]. MaaS applications enable user induced non-autonomy systems to turn a generally assumed to be intractable problem into a mixed-autonomy problem. In the context of a larger dynamic system, this dictates the progression of the integration or the use of automation [83].

To the best of our knowledge, our solution is the first scheme to generalise the mobility handling problem, using generic reinforcement learning techniques for improved dimensionality reduction. It applies machine learning and optimisation methods to mixed autonomy systems for addressing automation problems of integration into existing systems. It explores empirical and theoretical justifications of edge/caching systems and their optimisation methods as a design paradigm. Through principled learning and optimisation methods, even a small number of vehicles can be harnessed to have a significant impact on the Internet.

At this point, real-time independent decision-making for the random behaviour of vehicle passengers & drivers is a crucial factor. For this reason, creating a sequential decision-making tool/program to model the learning and decision-making processes of car passengers and/or drivers would be the first of its kind. As commuters make repeated decisions, they learn to optimise their route choices over time. This can be efficiently modelled using a sequential process, where a payoff function at each step is streamlined and linked to the results they experience.

Our model will also support the existing literature on traffic systems that can often be modelled using complex (nonlinear and coupled) dynamic systems. In addressing complex traffic control problems, a decentralised, learning-based solution involving interactions between humans, automated cars and sensing infrastructure, with the use of deep reinforcement learning, should be developed. The resulting control laws and emergent behaviours of cars will potentially provide insight on the behaviour of each car. These insights will be replicated, shared and synchronised among cars over a distributed ledger technology, through peer-to-peer ad-hoc networking, to understand the potential for automation of flow. We have already accomplished a simulation for a four lane road to demonstrate the possibility of computation and storage period [85].
Our novel computational framework, which integrates open-source deep learning and simulation tools, can support the development of edge computing in vehicles in the context of complex nonlinear dynamics of transportation. Learned policies, resulting from effectively leveraging the structure of human driving behaviour, surpass the performance of state-of-the-art predictors designed for various mobile applications, such as Google Now. The framework will initially focus on highway traffic, and will later include arterial traffic, transit, as well as other modes of transportation/commuting (biking, MaaS, carpooling, etc.).

2) DISTRIBUTED LEDGER AS A DATABASE

DAG will be used in distributed ledger technologies for storing data to enable fast and scalable computation in the network, as illustrated in Fig. 8. DAG may be the primary data structure for the OMNIBUS solution for creating a peer-to-peer network protocol. However, as DAG-based solutions emerge for high frequency traffic scenarios, problems may arise in low frequency scenarios. To ensure a continuous hybrid system, using a sequential blockchain verification to parallel DAG verification mechanism is necessary for accommodating increasing and unreliable penetration.

3) SOLVING THE MASSIVE ROUTING PROBLEM

Mobility patterns are of crucial importance in the context of providing network access to areas without Internet. They add a spatial component to the temporal sequential process, which may be termed as “Cartesian” machine learning. In doing so, understanding the “next move” to be applied as the “next hop” in routing purposes is essential. The techniques we developed in this regard leverage known models such as replicator dynamics, mirror descent, stochastic gradient descent and the hedge algorithm. Overall, it is necessary to converge all these processes towards a set of equilibrium based on assumptions made during the learning process used by humans in decision-making, while considering the constraints imposed by transportation.

B. DECENTRALISING COMPUTING AND STORAGE

Anticipating the demand for each edge car and deploying adequate car resources are very important to sufficiently meet locational demands. For instance, when a single car moves out of the local area, its storage and computing resources must be distributed across the cars that remain in that area and the new cars that enter that area. Therefore, developing predictive algorithms are crucial to optimally distribute computing and storage resources among cars, while considering challenges related to redundancy, security, heterogeneity of devices and federation (where interoperability is ensured on the level of different service providers). In developing these algorithms, creating and employing a global mobility map is a key element. Leveraging and combining existing mesh networking systems for car-to-car and car-to-device communication is also necessary. It is equally important to study how to distribute computing and storage across the entire system; i.e., whether data should remain local (shared among cars) or sent to the cloud.

Another vital task would be building parallel systems as well as harnessing thousands of simple and efficient computing and storage resources, which can be a practical solution to sustain growth without scaling technology. To this end, our architecture parallelises algorithms. Tasks must be implemented speculatively and in an “out of order” manner. Moreover, thousands of tasks should be efficiently speculated prior to the earliest active task in order to reveal sufficient parallelism. To develop parallel algorithms and uncover abundant parallelism in large-scale applications, new techniques should be developed to exploit locality and nested parallelism. To generate parallel algorithms in cars, the focus should be on the following.

(i) Ensuring consensus among multiple cars working towards a common goal. For instance, when all cars involved are solving one optimisation problem together, yet with different data set partitions.

(ii) Redistribution in the emergency where one of the cars has become disabled and leaves the cluster. The issue is to restore the system without restarting it.

(iii) Communication and Managing resources. Communication: computation requires significant input/output (I/O) (e.g., disk read and write) and data processing procedures. The OMNIBUS solution distributes storage systems to enable faster I/O and non-blocking data processing procedures for different types of environments (e.g., single node local disk, distributed file systems, etc.). Managing resources: the issue is managing resources within a given cluster of cars to meet all demands while maximising capacity.

(iv) Designing a programming model to improve efficiency. A new programming model is employed to achieve distributed computing and storage algorithms, in the same way as non-distributed ones, which requires less coding and improves efficiency. Studying programming in a single-node fashion, while automatically amplifying the program with distributed computing techniques, is also necessary. Applying model parallelism to partition the ML model itself so as
to distribute the workload to multiple computational cars is highly essential, as well as developing a unique data analytics engine to specifically target car-to-car and car-to-device for big data processing.

V. PROOF OF THE CONCEPT

The OMNIBUS solution aims to present an advanced decentralised computing and storage architecture by addressing two interrelated scientific challenges (creation of a mobility prediction model and an efficient distribution of resources among vehicles). As mentioned in Section I, the Quality of Service (QoS) levels are expected to be achieved with the proposed scheme along with the traditional cloud-User Equipment (UE) and MEC-UE communication schemes. They are also analytically assessed within this section. The aim is to prove that the OMNIBUS architecture can provide higher storage and processing capacity compared to cloud and MEC servers when certain conditions are met.

Let \( \ell_c \), \( \ell_m \) and \( \ell_o \) denote transmission latency between a cloud server and a receiver (UE), a MEC server and a UE, and an OMNIBUS network (vehicles) and a UE, respectively, for data flow (packet transmissions). As in [86], transmission latency between the cloud server and the UE can be computed as follows:

\[
\ell_o = n(1 + P_{lw})\ell_w + (1 + P_{lr})\ell_r
\]

where \( n \) is the number of packets transmitted, \( P_{lw} \) and \( P_{lr} \) are the packet loss rates between the cloud server and the Base Station (BS), and between the BS and the UE, respectively, occurring during \( n \) packet transmissions. Finally, \( \ell_w \) and \( \ell_r \) are the average latency per packet in between the cloud server and BS, and the BS and UE, respectively. To elaborate, \( \ell_w \) and \( \ell_r \) are the ratios of the sum of delays caused by processing, queuing, transmission and propagation to the total number of packets transmitted \( (n) \) between the cloud server and BS as well as the BS and UE, respectively.

In contrast to the cloud-UE communication scheme, since some packets (e.g., \( m \) packets, where \( m \in [1, n] \)) are expected to be pre-fetched/cached in a MEC server prior to a request from the UE, MEC-assisted transmission latency between the cloud server and the UE can be computed as follows:

\[
\ell_m = n(1 + P_{lw})\ell_w + (1 + P_{lr})\ell_r - m(1 + P'_{lw})\ell'_w
\]

where \( P'_{lw} \) and \( \ell'_w \) are the packet loss rate and average latency of the \( m \) packets pre-fetched earlier, say within the time period of \([\pi_0, \pi_m]\). Considering the shortness of the physical distance and the data flowing over the cloud, which will also be transferred to the UE over the base station, it is clear to say that the download speed of pre-fetched/cached packets to the UE is faster than the download speed of packets through a backhaul link between the cloud and UE.

In a similar manner, since some of the packets (e.g., \( x, y \), and \( z \) packets, where \( x + y + z \in [1, n] \)) are also expected to be pre-fetched/cached in an OMNIBUS network by varying the number of vehicles (3 vehicles for this example) in close proximity to the UE, prior to a request from the UE, the OMNIBUS-assisted transmission latency between the cloud server and the UE can be computed as follows:

\[
\ell_o = n(1 + P_{lw})\ell_w + (1 + P_{lr})\ell_r - (x)(1 + P''_{lw})\ell''_w - (y)(1 + P''_{lw})\ell''_w - (z)(1 + P'''_{lw})\ell'''_w
\]

where \( P''_{lw}, P''_{lw}, P'''_{lw} \) and \( \ell''_w, \ell''_w, \ell'''_w \) are the packet loss rates and average latencies of \( x, y \) and \( z \) packets pre-fetched earlier, respectively. In this regard, the following definition and theorem can be made.

Definition 1: Collaborative pre-fetching/caching is to allow storing packets on multiple edges (MEC server/vehicles) within a neighborhood.

Since pre-fetching/caching enables the full utilisation of front-end throughput between the edge and the UE, the following theorem can be defined.

Theorem 1: Collaborative pre-fetching/caching among edges reduces end-to-end latency, which extends the total available buffer-size/storage-capacity per UE.

Proof: Here, the latency considered for reduction is the latency between the cloud and the edge since no change is expected to occur between the edge and the UE. The latency over wired/wireless connections between the cloud and the edge can be split as \( \ell_w = \ell_w(\pi_0, \pi_m) + \ell_w(\pi_m, \pi_o) \), where \( \pi_0, \pi_m \) and \( \pi_o \) are the time at the beginning, the time when \( m \) packets are downloaded and stored on the MEC server as well as the time when all packets are completely downloaded, respectively. It is assumed that downloading speed for the \( m \) pre-fetched packets is faster or equal to/from the downloading speed of \( n \) packets due to the growing backhaul traffic \( (Dw(\pi_0, \pi_m) \geq Dw(\pi_0, \pi_o)) \) and longer transmission distances. The buffer size of the edge (buffer size of the MEC server, or the total buffer size of vehicles) is expected to be larger than the buffer size of the UE itself, \( \ell_w(\pi_0, \pi_m) \geq \ell_w(\pi_0, \pi_o) \), where \( \ell_w(\pi_0, \pi_o) \) is the average latency of the \( m \) packets, which is the corresponding time period when the buffer of the edge is used to transfer pre-fetched/cached packets to the UE. The theorem suggests that \( \ell_c \geq \ell_m \).

From Eq.(1), Eq.(2) can be rewritten as \( \ell_m = \ell_c - m(1 + P'_{lw})\ell'_w \). In the worst case, the second term on the right-hand side of the equation, \( m(1 + P'_{lw})\ell'_w \), will approximate to 0. Hence, \( \ell_m = \ell_c \), otherwise it will be \( \ell_m < \ell_c \), which simply proves the theorem. Considering that vehicles will be positioned closer to the UE than the MEC server itself, therefore having a higher signal strength from the UE, it would also be appropriate to say that \( \ell_m \geq \ell_o \), in case the total number of packets pre-fetched/cached by vehicles is more than the packets pre-fetched/cached by the MEC server (when \( x + y + z \geq m \)).

As the transmission latency, packet loss rate also affects the QoS. Suppose \( SS_1, SS_2, SS_3 \) and \( SS_4 \) are the signal strength levels between the UE and MEC, the UE and vehicle, the UE and vehicle, and the UE and vehicle, respectively. Here, packet loss rates are denoted with \( P_{lw}(SS_1), P_{lw}(SS_2), P_{lw}(SS_3) \) and \( P_{lr}(SS_4) \), respectively. In case the UE is closer to MEC.
than any of the vehicles, it is expected that $P_{lk}(SS_i) \leq P_{lk}(SS_{j,k,l})$. However, since both UEs and vehicles tend to be mobile, a handover operation could be required/preferred for the UE whenever signal strength of any of the vehicles is higher than the signal strength of the MEC server ($P_{lk}(SS_{j,k,l}) \leq P_{lk}(SS_i)$). In this respect, the following definition and theorem of packet loss rate are defined as follows.

**Definition 2:** Handover among edges is an essential factor to avoid/limit RAN-based congestion and packet errors.

**Theorem 2:** Collaborative handover approach reduces the overall packet loss rate, while increasing the average Received Signal Strength (RSS) of UEs.

**Proof:** Let $\ell_i$, $\ell_j$, $\ell_k$, and $\ell_l$ be average transmission latencies to be calculated when a UE is held by MEC$_i$, vehicle$_j$, vehicle$_k$, and vehicle$_l$, respectively. The theorem suggests that $\ell_i \geq \ell_j, \ell_k, \ell_l$, given that $SS_i \leq SS_{j,k,l}$. As can be seen from Eq.(1), Eq.(2) and Eq.(3), $P_{lk}$ increased the effect on the right-hand side of all of these equations. In case a single MEC server (MEC$_j$) is used in the system, $P_{lk}$ remains the same as in the cloud-UE communication scheme since the sender and receiver are the same entities in cloud-UE and MEC-UE methods. However, once UE is handed over from MEC$_i$ to any of the vehicles, say vehicle$_j$, given that $SS_i \leq SS_{j,k,l}$, the packet loss rate will decrease to $P_{lk}(SS_j)$, due to $P_{lk}(SS_i) \geq P_{lk}(SS_j)$. Hence, $\ell_i \geq \ell_j$ is held, which proves the theorem.

Accordingly, Theorems 1 and 2 state that collaborative pre-fetching/caching and handover among vehicles assist in the increase of the total storage capacity, reduce latency and packet loss rate, provide higher QoS levels per UE and therefore, enable an advanced decentralised computing and storage architecture compared to the cloud-UE and MEC-UE communication schemes.

**VI. CHALLENGES**

**A. DECENTRALISING COMPUTING AND STORAGE**

The demand for each edge car must be anticipated so that adequate car resources are deployed to meet locational demands. For instance, when a single car moves out of the local area, its storage and computing resources should be distributed across other cars that remain in that area and new cars that enter that area. Our architecture supports an extensive mobility map in developing predictive algorithms to optimally distribute computing and storage resources among cars. In doing this, it is necessary to consider the challenges related to redundancy, security, heterogeneity and federation where interoperability is ensured on the level of different service providers and mobility handling. The OMNIBUS model leverages and combines existing MANET (Mobile Ad-hoc Network), VANET (Vehicular Ad-hoc Network) and DTN (Delay Tolerant Networking) technologies for car-to-car and car-to-device communication. This model also addresses how to distribute computing and storage across the entire system; i.e., whether data should remain local (shared among cars) or sent to the cloud.

**B. AGGREGATED MOBILITY HANDLING (ACCURATELY DEPICTING VEHI CULAR BEHAVIOUR)**

At present, there is no large scale mobility map since available models are not adaptable and do not adequately address privacy concerns. In our architecture, developing a mobility prediction model is critical for optimising the allocation of computing and storage resource sharing among them. These mobility patterns enable us to provide offloading decisions, as well as control energy consumption and bytes of data transfer [87]–[89]. Our solution will use databases provided by mobile operators, smart transportation systems, etc., to build our mobility model. Mobility patterns will be learned as each car shares mobility patterns and movements of other cars via mesh networking technologies. The OMNIBUS model uses a combination of DAG-based and blockchain-based distributed ledger technologies depending on different frequency scenarios for aggregation. Our problem is complex, since it focuses on continuously moving cars that exchange data with each other in order to keep the system alive in any given location. The proposed solutions to communication regarding moving vehicles are limited to highly ordered environments. In contrast, the OMNIBUS solution seeks to develop communication protocols for cars in an extremely chaotic environment. To do this, the OMNIBUS model supports MANET, VANET and DTN technologies.

**C. HETEROGENEITY ISSUES**

Heterogeneity of resources, in terms of computational and storage capabilities as well as their ad-hoc availability, is necessary for optimisation. Heterogeneity is important for deciding which application components should be deployed and where [45]. This involves developing algorithms to address heterogeneity while considering the limitations of specific nodes. For instance, in a content delivery use case, storage limitations of the caches are incorporated into the caching algorithm. While node degrees can be optimised, the CPU of each car must compute multiple items simultaneously. Ensuring that CPUs are not overwhelmed will be a key consideration in developing our algorithms.

**D. FEDERATION ISSUES**

In our architecture, cars are geographically distributed on a very wide scale and could be assigned to different service providers. The cloud can also be operated by a different provider. Our architecture will be designed in a way where interoperability is ensured on the level of federation of different service providers. This means developing a consensus protocol to understand the capabilities of a variety of cars using various providers.

**E. HANDLING MOBILITY OF END-USERS**

In the case that end users physically move, the system should be able to continuously provide the same quality of experience, without interrupting the service. Moreover, in the scenario that several end-users are watching the same video, the algorithm may need to allow the mobility engine to copy the video to be pushed to the destination point. Similarly, as a
car moves in our system, resource displacement occurs with implications on resource management algorithms. To address this challenge, studying byzantine fault tolerance methods for the scenario where a car’s data centre may fail or move is important. There is inadequate information on whether a component has failed or moved. The OMNIBUS solution applies model parallelism for partitioning the ML model itself to distribute the workload to multiple computational cars. In this architecture, the methods for partitioning the ML model according to heterogeneity, federation and mobility are constantly investigated.

VII. IMPACT
The OMNIBUS solution will have a far-reaching impact in three areas, as summarised in the following: (i) 5G and 6G Ultra-Reliable Low Latency Communication (URLLC): The key objective is to enable a range of new applications (Smart Driving, Smart Grids, Augmented Reality (AR) and IoT in general) that depend on ultra-reliable and ultra-low-latency connectivity. The OMNIBUS solution is driven by the need to remove present and future bottlenecks in communication networks and to prepare the groundwork for future 5G and 6G heterogeneous networks [90]–[96]. The solution responds to the market requirement for a comprehensive edge network platform with faster and more reliable data processing. Attempts to move computing closer to the network (cloudlets, Fog Computing and MEC) are not scalable. In contrast, our architecture has the potential to scale almost infinitely while increasing the speed of networks as it grows. Our ambitions go further and our research paves the way for employing all connected devices as computing and storage centres, including people with smart phones and all IoT applications. The proposed decentralised network architecture opens up new possibilities for network slicing, resulting in lower latency, increased storage capacity, more network resilience and security as well as less energy wastage. By breaking down and distributing computing and storage resources for intermittent networking, our solution paves the way for a scalable collaborative communication network.

(ii) Decentralised Internet: Our architectural framework can be used for high latency and delay-tolerant Internet access for more than the 3.9 billion who remain offline today from the world’s population. A decentralised storage and computational framework, as we have proposed, is more reliable than the current digital infrastructures which are vulnerable to disaster situations. A single point of failure in the infrastructure can bring down the entire communication network. The OMNIBUS solution leverages our mobility model to solve the massive routing problems. Our predictive algorithms that we developed can optimally distribute computing and storage among cars to bring Internet data to unconnected regions. In this regard, the OMNIBUS solution is expected to open new directions in research on ad-hoc technologies and DTN-based data mules. In contrast to URLLC, this can be called UCHLC: Ultra Coverage High Latency Communication.

Fig. 9 illustrates an example of how data is transmitted via vehicles over time in a relatively low density when there are four lanes with communication range \((C)\) of 500 m and update interval \((\tau)\) of 1s [85]. In every update interval, if the message remains at the same vehicle, then it is carried by transmit relay (stored and forwarded). Otherwise, it is forwarded to another vehicle by direct relay (multi-hop forwarding).

Consequently, with the OMNIBUS solution, even a UE that is out of coverage from any BS will be able to access the Internet and communicate with the entire world since the OMNIBUS solution allows any offline data to travel via vehicles over direct or transport relay until an Internet connected edge is found. This procedure will create additional vehicle-based computation and storage capacity for the system, depending on the distance of the UE and the number of vehicles in that region.

(iii) Smart Transportation: Our solution will have considerable impacts on smart transportation systems, including traffic systems and edge computing in vehicles. It has the potential to redirect research on traffic systems towards a decentralised, learning-based study of complex traffic control problems, involving interactions of humans, automated vehicles and sensing infrastructure. The resulting control laws and emergent behaviours of cars will potentially provide insight on the behaviour of each car. These insights will be replicated, shared and synchronised among cars over a distributed ledger technology through peer-to-peer ad-hoc networking to understand the potential of automation of flow.

Furthermore, our study can be employed by the research community as a new computational framework that integrates open-source deep learning and simulation tools to support the development of edge computing in vehicles in terms of complex nonlinear dynamics of transportation.

VIII. SUMMARY
A breakthrough is imminently needed to support the demands of heavy data network edge. Although there are various
proposed solutions regarding edge computing, most are substantially limited and not easily scalable. At this point, the OMNIBUS solution, which proposes the smart distribution of computing and storage resources among vehicles with a mobility prediction model, provides the breakthrough technology needed. It brings together a full spectrum of science and engineering used for various innovations and has the potential to upend the ecosystem of future network efforts. This solution specifically paves the path for more efficient and highly scalable device-level edge computing architectures. It particularly develops serious key objectives to enable a range of new applications (Smart driving, Smart Grids, Augmented Reality and IoT in general) which depend on ultra-reliable and ultra-low latency connectivity. The OMNIBUS solution has the potential to redirect research on traffic systems towards a decentralised, learning-based study of complex traffic control problems. It can also be employed by the research community as a new computational framework that integrates open-source deep learning and simulation tools to support the development of edge computing in vehicles. This would be in the context of complex nonlinear dynamics of transportation. It can support and further develop state-of-the-art predictors designed for various mobile applications. It will also make existing map databases more accurate and more interactive. The OMNIBUS solution can be the most vital and influential building block for future network efforts. But, more advanced planning/reinforcement learning algorithms will be needed to solve the load balancing problem in this novel distributed computation architecture.

REFERENCES


FERIDE İNAN received the B.A. degree in economics from Columbia University and the M.Sc. degree in international development from the University of Amsterdam. She has worked in Turkey, China, and India, leading research funded by the Turkish, British, and Swiss governments, as well as the European Union, the German Marshall Fund, OXFAM, Dupont, and Google. She has been part of policymaking processes, including her leading role of the G20’s official Think-20 network of academics and think tanks in dialogue with ministries in Turkey and China, consulting the Business-20 during the 2015 Turkish G20 Presidency. She was published extensively including a book on the G20 and Chinese transnational governance. In Turkey, her primary institutional affiliation has been with the Economic Policy Research Foundation of Turkey (TEPAV) established by the Turkish chambers (TOBB), where she also continues to work on the G20 and emerging markets with an emphasis on innovation-driven growth policies in these economies. Complementing her work on technological innovation, she also works for Ambeent Inc., a start-up founded by UC Berkeley engineers focusing on optimization of wireless networks. She is currently a Policy Analyst and a Researcher working within the areas of transnational governance and social and economic development in emerging markets.

ONUR ERGEN received the M.S. and M.A. degrees in electrical engineering and physics from the UC Berkeley, respectively, and the Ph.D. degree from the University of California, Berkeley. He recently joined the faculty of Koc University, where he is currently an Assistant Professor in Electrical and Electronics Engineering. He also worked at various companies, such as Apple Inc., Intel Corporation, and Solopower Inc. He has a strong academic foundation in energy and novel device architectures. His research interest includes centered around building next-generation energy devices and their application towards future technologies, especially in 5G. His research involves experimental and software development.

IBRAHEEM SHAYEA received the bachelor’s degree in electronics and communication engineering from the Faculty of Engineering, University of Diyal, Iraq, in July 2004, and the master’s degree in communication and computer engineering and the Ph.D. degree in the field of electrical and electronic engineering, specifically in wireless communication systems, from the Department of Electrical, Electronic Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia (UKM), Malaysia, in December 2015 and July 2010, respectively. From January 2011 to December 2015 (during his Ph.D. study), he worked as a Research Assistant and a Demonstrator with the Department of Electrical, Electronic Engineering, Faculty of Engineering and Built Environment, UKM. From January 2016 to June 2018, he worked as a Postdoctoral Fellow at the Wireless Communication Center (WCC), University of Technology Malaysia (UTM), Malaysia. From September 2018 to August 2019, he worked as a Researcher Fellow with Istanbul Technical University (ITU), Istanbul, Turkey. He is currently an Associate Researcher with the Department of Electronics and Communications Engineering, Faculty of Electrical and Electronics Engineering, ITU, where he is also teaching courses for master’s and bachelor’s students. He has published several scientific research journals, conference papers, and whitepapers. His research interests include mobility management in future heterogeneous (4G, 5G, and 6G) networks, mobile edge computing, machine, and deep learning, the Internet of Things (IoT), propagation of millimeter-wave, mobile broadband technology, and future data traffic growth and spectrum gap analysis.
MEHMET FATIH TUYSUZ received the B.Sc. degree from the Department of Electric and Electronics Engineering, Inonu University, in 2004, the M.Sc. degree from the Department of Electric and Electronics Engineering, Dokuz Eylul University, in 2007, and the Ph.D. degree from the Department of Computer Engineering, Gebze Technical University, in 2013, for his research in Collision-free Channel Access over Wireless Networks. He worked as a Research Assistant from 2007 to 2013 and as an Assistant Professor from 2013 to 2018 with the Computer Engineering Department, Gebze Technical University, and with the Computer Engineering Department, Harran University, respectively. He also worked as a Visiting Research Fellow with the Faculty of Science and Technology, Middlesex University, London, U.K., in 2019. He has been working as a Senior Lecturer with the Department of Computer and Information Sciences, Northumbria University. He has published many papers in prestigious international conferences and journals. His research interests include MEC, mobile and wireless multimedia communications, QoE optimization, energy-efficient communication techniques, handover and network selection strategies, the Internet of Things, and software-defined networks.

AZIZUL AZIZAN received the B.Eng. degree (Hons.) electronics engineering (majoring in telecommunications) from Multimedia University, Malaysia, in 2002, and the Ph.D. degree in 3.5G physical layer adaptation for satellite systems from the University of Surrey, U.K., in 2009. He later joined the Malaysian Communication and Multimedia Commission for more than six years overseeing spectrum and numbering policies, and telecommunications resource management administration. He is currently with the Faculty of Science and Technology, Universiti Teknologi Malaysia. His research interests include wireless communications, edge computing, cyber-physical systems, business intelligence, and STEM education.

NAZIM KEMAL URE received the B.Sc. and M.Sc. degrees from Istanbul Technical University (ITU), in 2008 and 2010, respectively, and the Ph.D. degree in aerospace engineering from the Massachusetts Institute of Technology (MIT), in 2015. He is currently an Associate Professor with the Department of Aeronautical Engineering, Istanbul Technical University. He also serves as the Vice Director of the ITU Artificial Intelligence and Data Science Applied Research Center (ITU AI). His main research interests include applications of deep learning and deep reinforcement learning for autonomous systems, large scale optimization, and development of high-performance guidance navigation and control algorithms.

MAZIAR NEKOVEE (Member, IEEE) received the first degree in electrical and electronic engineering from the Delft University of Technology (NL), and the Ph.D. degree in physics from the University of Nijmegen (NL). In 2017, he was a Chief Engineer and the Head of the 5G Research with Samsung Research and Development, U.K. Prior to that, he was with British Telecom (BT) Research and Technology, as a Senior Scientist, and subsequently a Team Leader. He is currently a Professor of Telecoms and Mobile Technology and the Head of the Centre for Advanced Communication, Mobile Technology, and the IoT, the Dean of the AI Institute at School of Engineering and Informatics, University of Sussex, U.K. He has authored more than 120 highly cited peer-reviewed articles, one bestselling book Cognitive Radio Communications and Networks: Principles and Practice and has 13 patents in telecommunication and mobile technologies. His current research interests and expertise included 5G and beyond-5G/6G mobile communication RAN and core design, machine-learning, AI applied to communication networks, spectrum sharing, cognitive radio, millimeter-wave, and THz communications. He is a sought-after speaker at high level industry events, C-level meetings with European operators, policy makers, and international conferences. He was a recipient of a number of industry and academic awards, including a Royal Society (U.K., Academy of Science) Industry Fellowship, the BT Innovation Award, and the Samsung Research and Development Best Practice of Research Award.