Serverless Vehicular Edge Computing for the Internet of Vehicles

Faisal Alam, Adel N. Toosi, and Muhammad Aamir Cheema, *Monash University, Clayton, VIC 3800, Australia*

Claudio Cicconetti, *Institute of Informatics and Telematics – National Research Council, 56124, Pisa, Italy*

Pablo Serrano, *Universidad Carlos III de Madrid, Leganes-28911, Spain*

Alexandru Iosup, *Vrije Universiteit Amsterdam, Amsterdam, 1081 HV, The Netherlands*

Zahir Tari, *RMIT University, Melbourne, VIC 3001, Australia*

Majid Sarvi, *University of Melbourne, Melbourne, VIC 3010, Australia*

Rapid growth in the popularity of smart vehicles and increasing demand for vehicle autonomy brings new opportunities for vehicular edge computing (VEC). VEC aims at offloading the time-sensitive computational load of connected vehicles to edge devices, e.g., roadside units. However, VEC offloading raises complex resource management challenges and, thus, remains largely inaccessible to automotive companies. Recently, serverless computing emerged as a convenient approach to the execution of functions without the hassle of infrastructure management. In this work, we propose the idea of serverless VEC as the execution paradigm for Internet of Vehicles applications. Further, we analyze its benefits and drawbacks as well as identify technology gaps. We also propose emulation as a design, evaluation, and experimentation methodology for serverless VEC solutions. Using our emulation toolkit, we validate the feasibility of serverless VEC for real-world traffic scenarios.

INTRODUCTION

Vehicles such as cars, trucks, and other modes of transportation play a critical role in modern society. With nearly 1.5 billion vehicles already in existence on Earth in 2022 and increasing demand for digital processing and onboard applications, the future of transportation and mobility presents numerous new opportunities across the hardware, software, networking, and services industries. We envision a paradigm shift in this ecosystem.

Although sensory inputs attached to the vehicles provide many of the current advancements in intelligent transport systems (ITSs) and connected and autonomous vehicles (CAVs), the next generation of automobiles will require fast and reliable external computation services to offset the growing complexity and costs of onboard computers.[1] Modern CAVs can generate data at a velocity and volume that may exceed by far the network capacity toward the Internet, and the high round-trip times (RTTs) may be incompatible with many emerging applications requiring real-time capabilities. To cater to extra computing needs and stringent latency requirements, we need an approach that can leverage more resources than are available in vehicle yet make sure the resources can compute and send back the results in time and can do so automatically and efficiently for a variety of time-critical vehicular applications, including the perilous tasks of every autonomous driving system.

We propose, in this work, an approach that addresses these needs by combining two emerging paradigms: vehicular and serverless computing. Vehicular edge computing (VEC) combines vehicular and edge computing by offloading delay-sensitive computational tasks from vehicles to the nearby edge computing nodes. VEC utilizes available infrastructure that includes, primarily, roadside units (RSUs) and decentralized edge

1089-7801 © 2023 IEEE
Digital Object Identifier 10.1109/MIC.2023.3271641
Date of publication 1 May 2023; date of current version 10 July 2023.

data centers in the proximity of RSUs. A large body of research also proposes to utilize the leftover capacity of the neighboring and parked vehicles for the computation of time-sensitive tasks.[2] There are significant challenges to achieving the potential of VEC. First, in contrast to cloud-native applications, the development and deployment of VEC-compatible applications in dynamic and diverse VEC environments are extremely challenging for developers and operators. In fact, any VEC solution should allow developers to focus on the application's business logic instead of handling the management and orchestration (MANO) aspects. Moreover, addressing the scalability requirements of the time-critical applications to respond to the load variability at the edge is not a trivial task. Finally, due to the mobile nature of the vehicles at the network edge, using traditional stateful application architectures, which are tightly coupled with storage to hold states of the application (stored inputs and outputs), significantly compromises the agility, elasticity, and efficiencies of ITS applications.

In recent years, a new paradigm of function as a service (FaaS) has emerged in the cloud computing domain to address similar challenges, with recent explorations in the context of edge[3] and mobile computing.[4,5] FaaS allows users to deploy an independent, stand-alone piece of code (a "function") on the infrastructure where the computational back-end requirements for the functions are assessed, provisioned, and maintained by the platform provider. Advantages include 1) high agility in application development without operational expertise, 2) effortless scalability to cater to the surge in functions calls, and 3) efficient use of resources through seamless multitenancy. Since FaaS relieves the developer from server management-related issues, the concept is also known as *serverless computing*.

In accordance with these contemplations, this article coins a new term, called serverless VEC. *Severless VEC* refers to the deployment of a serverless execution model on edge devices, RSUs, and vehicles for the purpose of processing data from connected vehicles and supporting the development of applications. In this architecture, functions and tasks are executed on edge devices located near the source of data rather than in a central server. The "serverless" aspect means that the edge devices can dynamically allocate computing resources as needed without the need to manage and maintain a dedicated server infrastructure. This provides low latency; real-time processing; and increased efficiency, flexibility, and scalability for the management and analysis of the large amounts of data generated by connected vehicles and the increasing number of emerging real-time applications.

Multiple efforts have been made in the recent past for efficient task offloading on the vehicular edge.[1] These works are centered around finding the best methods for distributing the load to connected computational units, e.g., neighboring vehicles, RSUs, or the traditional cloud for optimizing time, energy, or computational capacities. However, there has been no or little effort in providing a comprehensive solution that caters to VEC application deployment issues, such as ephemeral connectivity of moving vehicles, failure handling, provisioning, monitoring, and scalability.

In addition, without empirical evidence, a theoretical concept such as serverless VEC may be deemed infeasible and impractical for time-sensitive real-world scenarios. To bolster its feasibility and usability in reality, we provide an emulation architecture and toolkit for serverless VEC using open source frameworks. Through experiments, we demonstrate the viability of serverless VEC for real-world applications and show that it can provide improved response times for resource-intensive applications such as object detection. Our future work will aim at using this architectural framework to provide more extensive experiments with various policies, providing a platform for others to conduct similar experiments.

This work aims to provide a comprehensive and practically feasible solution that distributes the load and creates, manages, and scales the VEC applications for optimal latency while minimizing development and deployment costs. We extend the idea of serverless computing that is successful in the traditional cloud to provide a feasible solution to manage the VEC infrastructure.[3] Our key contributions are 1) a platform-agnostic infrastructure management for serverless VEC with built-in autoscaling and load balancing on the edge; 2) an analysis of the advantages of serverless VEC; 3) a review of the challenges expected in early adoption; and 4) a detailed architecture for the emulation of serverless VEC, along with an exemplary scenario to showcase the feasibility of serverless VEC.

OPERATIONAL MODEL AND BACKGROUND

We consider an operational model that combines vehicular, edge, and serverless computing.

Vehicular Ad Hoc Network (VANET)

The next generation of vehicles, called CAVs, will be equipped with communication technologies to communicate with each other (vehicle to vehicle); with roadside infrastructure (vehicle to infrastructure); and, in some cases, even with pedestrians (vehicle to

pedestrian). The ambit term for these communication technologies is vehicle-to-everything (V2X) communication. This point-to-point communication infrastructure creates an ad hoc network on the roads called VANET. CAVs utilize VANET for applications such as precrash sensing, blind intersection, and forward collision. VANET enables message and information passing using multihop strategies on the peer-to-peer (P2P) connections established by V2X.[6]

Task Offloading in VANET

The number of sensors in vehicles is spiraling, with the cost of electronics in vehicles, which was around 35% in 2010, expected to reach 50% by 2050. Vehicles with higher autonomy generate around 25 GB of data per hour, according to an estimate by McKinsey. With the advent of VANET and V2X, ITS applications would pave the way for newer and more advanced applications. However, the present computational infrastructure would not be sufficient to process this deluge of data for advanced applications. To cater to the growing demands of vehicular computation, we need to offload their computational tasks using the VANET infrastructure. Thus, a large body of research has been attributed to efficient task offloading.[1]

Vehicular Cloud/Edge Computing

The onboard computational power in current vehicles is not keeping up with the advanced application demands. Due to an ongoing bottleneck in semiconductor production, producing and scaling enough chips to meet this growing demand is unlikely to happen for the next coming years. Mobile cloud computing (MCC), a paradigm that allows both storage and computation of data outside the mobile device, is being employed to keep up with this growing demand.

As an extension to MCC, mobile edge computing/multiaccess edge computing (MEC) has been explored to minimize the latency of transmission to the cloud and further improve the quality of service. MEC is a paradigm where the resources for infrastructure as a service, platform as a service, and software as a service can be accessed at the edge of the network. When the MEC concept is extended to ITSs, it is known as VEC. In a VEC, the data would be computed at the RSUs or at decentralized computing centers near RSUs, which are called edge data centers. For sudden peaks in application programming interface (API) triggers, even neighboring vehicles and central cloud facilities can be employed to augment the edge infrastructure. Various studies have been conducted to study the modalities and techniques for VEC.[1]

Software-Defined Networking (SDN)

SDN is an approach to networking that provides separation between the network control plane and the forwarding plane, with centralized administration identified as the key enabler for 5G. It also offers operation flexibility and network management at scale for VANETs. A programmable and intelligent 5G network provided by SDN is highly agile and boosts the usability of the high bandwidth and low latency offered by 5G. It also paves the way for more innovation as well as advanced service and product offerings, thereby improving overall efficiency.

Serverless and FaaS

With the advent of virtualization and cloud computing, the notion of serviceability enhanced by the advancement of technologies led to the creation of FaaS. FaaS allows the user to specify logical, independent pieces of code to be deployed as microservices. Functions are deployed in software containers, which are self-contained units holding the function, and related libraries bundled together in an isolated space running as a process. The service provider provides the required resources to execute functions based on their footprint and elastically scales them according to the demand. This is beneficial, as the user does not need to maintain or scale the back-end resources.

Figure 1 illustrates how FaaS operates in VEC. FaaS acts in a serverless way to deploy functions on the infrastructure provided and maintained by the service provider. It is mainly used in an event-driven context where functions are triggered by typical events, such as HTTP requests, message queues, or database or storage operations, etc. The event-driven nature of serverless aligns well with the nature of many VEC applications that require event triggering based on sensing/actuation.

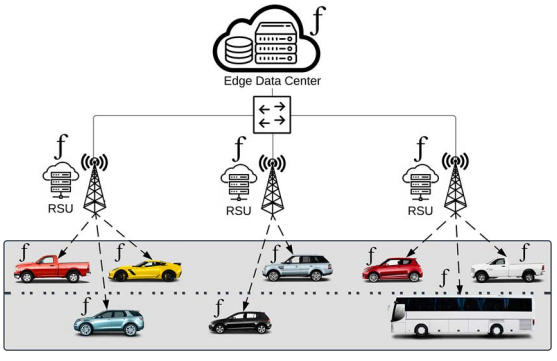

FIGURE 1. A serverless vehicular edge computing scenario, where f represents serverless functions that can be deployed on various nodes. RSU: roadside unit.

BENEFITS OF SERVERLESS VEC

In the cloud, serverless computing and the FaaS programming model are becoming increasingly popular due to the many advantages they offer.[7] In this section, we frame these advantages in the context of serverless VEC.

Serverless Scales Well With Nonuniform Dynamic Vehicular Offloading

Requests for VEC applications follow vehicle traffic patterns, which are bound to be highly nonuniform in both space and time. With serverless, provisioning does not need to be based on peak traffic conditions, which can waste resources. In other words, the stateless nature of FaaS invocations makes it straightforward to scale up and down the number of instances of a given platform based on the request rate and commensurate resource requirements. This would also help optimize the infrastructure cost of the service provider, as the system specifications are set based on favorable traffic conditions, keeping the system elastic to meet growing demands. Scaling also helps reduce overall operational costs and energy consumption. Ideally, the exact amount of resources needed for the current traffic load should be active at any time and platform.[8]

The Flexible Billing Model Helps On-Demand Vehicular Applications

With the rate of advancements and investment in ITSs, the applications and services attributed to the vehicles will increase as well. This creates an opportunity to implement opportunistic mechanisms to share the compute resources available on the RSUs and the vehicles. Some preliminary studies on incentive mechanisms for edge computing have been proposed, e.g., by Olaniyan et al.,[9] but the research is still in its infancy. We can speculate that the sandboxed environment provided by serverless computing, coupled with fine-granularity billing (typically in volume of function calls), will provide a fertile environment for a market of computation power for VEC. For example, low-end vehicles with little or no computational resources would have the option to get the services offered by high-end vehicles or nearby infrastructure by paying for the services only as needed.

The Statelessness Property Helps Opportunistic Deployment of VEC Functions

The FaaS paradigm is inherently stateless. The stateless nature of serverless functions makes them an attractive platform for real-time VEC applications that need to deploy services opportunistically. Avoiding state enables serverless to be ephemeral. Thus, it makes it easier for the serverless framework to allocate resources to functions and host functions across RSUs, vehicles, etc. Moreover, in VEC, the state of some applications may only make sense for a given physical location (e.g., at a junction) and for a limited amount of time (e.g., until a traffic jam is resolved), which matches the ephemeral states in serverless.

The Stateless Property Addresses the Challenge of Service Migration for Moving Vehicles

Due to the dynamic nature of moving vehicles, a vehicle may trigger an API request at the edge but leave the network coverage area of the RSU before receiving the response. To meet the low-latency requirement of real-time VEC applications, this may require service migration over the network, particularly for stateful applications, which can be a costly and time-consuming process. In a stateless serverless scenario, the same task can be requested again at the next connected RSU without impacting the overall application flow. On the other hand, in traditional edge computing with stateful containers or VMs, service continuity requires the transfer of the current state to the new edge, thus increasing the management plane complexity and affecting both the system resources and the user's quality of experience (QoE).

Serverless VEC Handles Massive Parallelization

FaaS invocations are independent of one another; thus, the execution of complex applications through many smaller tasks in parallel may be done in a straightforward manner.[10] Some vehicular applications may take advantage of this, especially when running on relatively modest hardware, such as that made available by RSUs. Such a granular function decomposition opens the door to sophisticated optimization strategies and high-performing architecture. The request scheduler can place subtasks independently on different threads, processes, and machines or schedule them for parallelism on hardware accelerators.

Serverless VEC Favors Interoperability

With vehicles of different car makers and models on the road, vehicle computation infrastructure would be heterogeneous in nature. Handling various devices, hardware, operating systems, and library requirements is a challenge that needs to be addressed within

industry alliances (e.g., Automotive Open System Architecture[a]), which slows down the innovation and time to market. With serverless, the container is a self-contained abstraction with function code and relevant libraries prebundled in a single unit that can be deployed over different hardware and architectures, thus enabling faster cycles of software deployment, adoption, and updates.

Serverless Helps Faster Development of VEC Applications

A serverless architecture relieves the developers of the burdens of platform management, maintenance, and scaling. This allows them to focus solely on the functionality and business logic of the application they want to provide to clients, paving the way for faster development cycles and lower operational costs.

The Event-Driven Nature of Serverless Matches Vehicular Use Cases

The event-driven architecture of serverless makes it a perfect match for many VEC applications working in real time based on data-rich sensing and actuation events and state changes. In other words, the distributed and asynchronous architecture pattern of VEC applications can be simply handled by the event-driven function services. Serverless can rely on events to trigger and communicate between decoupled services of VEC applications built on microservices.

CHALLENGES AND OPEN ISSUES OF USING SERVERLESS IN VEC

Despite the advantages of using a serverless architecture in a VEC scenario, there are also some challenges and open issues that need to be addressed to fully exploit its potential.

MANO of Serverless VEC Has High Overheads

Cluster MANO is a complex operation involving keeping a list of running containers, driving autoscaling, managing placement, performing load balancing, and continuously monitoring the resources. Existing serverless frameworks are designed to execute in purposely built cloud settings with computing nodes in static clusters operated on tightly coupled and often homogeneous servers connected through high-speed and reliable wired networks. However, in VEC, computing nodes are heterogeneous and dispersed over a mixture of wired and unstable and intermittent wireless networks, which makes MANO more challenging and especially vexing for low-resource computing nodes, such as vehicles and RSUs.

Cluster Formation Is Not Designed for Dynamic Vehicular Topologies

Existing FaaS frameworks and orchestration tools are designed for back-end server-based solutions where cluster nodes are readily available for scheduling and are expected to be stable over time. Instead, in a VEC scenario, vehicles can act as computing nodes and join and leave clusters dynamically. Today's procedure for cluster formation may be inadequate for such a volatile scenario due to the considerable time and resources required by the related procedures, which can negatively impact the overall performance.

Serverless Suffers From Cold-Start Effects

Experiments executed in the cloud have revealed that only a tiny fraction of the response time of an FaaS call invocation can be attributed to computation, while the rest is overhead due to network transfers, container activation, runtime environment setup, virtualization costs[11] etc. The overhead is especially significant the first time a function is invoked after the orchestration system has scaled down to zero instances, in which case a cold-start phenomenon occurs. The container image has to be pulled from the repository and loaded, which can take orders of magnitude more time than the typical response time. In a VEC scenario, we speculate that cold-start effects may be much more widespread than in a cloud system because there is not a single serverless platform logically centralized in a data center but many distributed over a territory. High jitter and, in particular, tail latencies created by cold-start effects, can be problematic for time-critical tasks and, if not addressed by research and technology, may become a barrier to the adoption of serverless VEC.

Resource Scheduling Optimization Is More Difficult

In the cloud, the main role of the autoscaler is deciding how many replica instances are needed for a given function. In an edge scenario, the problem becomes more complicated because the runtime environment is also in charge of deciding *where*—that is, on which edge node—to activate or terminate an instance. Serverless VEC presents a new level of complexity because these environments are highly dynamic and consist of a heterogeneous, loosely coupled set of nodes connected with erratic and unreliable network connections. Given

[a]https://www.autosar.org/

the time-variable dynamics of such a system, it is difficult to predict which node is best suitable for the given request, considering the locations of the nodes as well as their velocities, directions of movement, hardware characteristics, and so on.

New Security Concerns

Serverless in the cloud is secured by a firewall in a trusted environment,[12] while in a VEC, nodes would be exposed to different heterogeneous and insecure vehicles. Moreover, with multitenancy, multiple clients would be serviced in the same cluster and overlapping nodes. This adds to the complexity of addressing security measures, as the data would be distributed across vulnerable nodes in heterogeneous environments. Although sandboxing does help isolate the space, the data are still shared on a different platform. New studies on how to enable security and privacy in such a heterogeneous and ephemeral system are needed.

Ingress Points Are Distributed

Serverless systems in the cloud are logically centralized, with all client requests forwarded to a central gateway. Such a model does not capture the distributed nature of a typical VEC scenario, and investigation is required to establish a scalable, distributed, agile, and/or hierarchical model for faster response times for clients and minimum resource utilization for servers. Furthermore, the choice of where to place the multiple gateways based on the continuously changing VEC environment is also an open research problem.

Modeling, Simulation, and Emulation of Serverless VEC Is Not Trivial

Models, simulators, and emulators are required to evaluate and test the performance of serverless VEC applications and mimic their behavior, hardware, software, etc. However, the modeling, simulation, and emulation of a serverless VEC scenario where vehicles move in and out of range at varying high speeds in a short span of time with many distributed software and hardware components involved is challenging. In the remaining part of this article, we focus on this challenge and propose our emulation for serverless VEC.

SERVERLESS VEC SAMPLE USE CASES

There could be many applications that would benefit from a serverless VEC. Some of the use cases are listed here:

- *Autonomous driving*: Serverless VEC can support high-speed processing of data from the cameras and other sensors onboard vehicles with full or partial offloading to assist extra computation and application development for autonomous driving. This improves the performance, reliability, and cost-effectiveness of autonomous driving systems.
- *Traffic management*: Serverless VEC can help with real-time data processing from vehicles to enable coordinated responses to traffic information, such as road incidents and other safety hazards for efficient traffic management.
- *Infotainment services*: The presented framework can provide high-speed, low-latency processing for vehicle add-on services like entertainment systems, such as gaming, streaming video, and more. Other add-on services could include emerging AI/machine learning use cases for gesture recognition and voice recognition, leading to improved QoE.

EMULATION ARCHITECTURE AND PROTOTYPE

A typical VEC scenario involves a bunch of RSUs and vehicles generating requests for tasks, such as image processing or object detection, for various applications like autonomous driving, accident prediction algorithms on blind turns, and augmented and virtual reality applications for add-on comfort. Testing, performance analysis, and verifying serverless applications in a real-world VEC environment is costly and difficult, if not impossible, in many cases. In addition, it is essential to study the system's intricacies, like predicting the load on RSUs, average response times, effects of various scheduling and orchestration strategies, the impact of traffic movement patterns on the load of RSUs and edge data centers, and other similar metrics. In these cases, simulation or emulation tools are instrumental in providing developers with accurate or near-real precision data. In this work, we put together an emulator toolkit to delve into the peculiarities of deploying an FaaS application in a serverless VEC environment.

The proposed emulator toolkit architecture is depicted in Figure 2. We use a mix of multiple off-the-shelf and open source simulator/emulator tools along with our software codes to create a software suite representing serverless VEC scenarios. In the following sections, we discuss the main components of our proposed emulator.

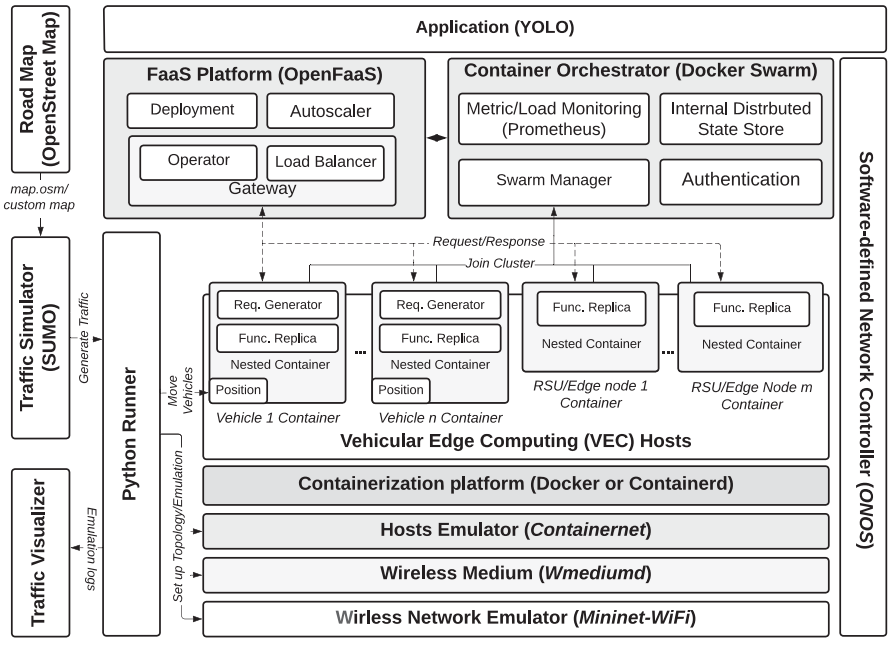

FIGURE 2. Software architecture of the emulation toolkit. FaaS: Function-as-a-Service; Func.: function; Req.: request; SUMO: Simulation of Urban Mobility.

Simulation of Urban Mobility (SUMO)

For road network traffic simulation, we use SUMO,[b] an open source traffic simulator. SUMO can handle large-scale and continuous traffic simulation on vast road networks. One can also create road networks of choice for their setup. Maps downloaded from OpenStreetMap[c] are used as the input of the road network to SUMO for real-life traffic generation. As an output, SUMO generates a traffic trace file with the position of every vehicle at every timestep on the map using a given traffic model.

Mininet-WiFi

The proposed architecture uses a Wi-Fi-based dedicated short-range communication network for V2X communication. We use the open source emulator Mininet-WiFi.[d] Mininet-WiFi is an extension of the open source network emulator Mininet and allows for the creation of Wi-Fi access points and Wi-Fi stations. Mininet-WiFi allows for specifying the positions of the stations and access points. The Wi-Fi connection is established by virtualizing the wireless medium that connects all Wi-Fi devices. This virtual medium is called *wmediumd*. It also handles handover by stations getting connected to the access point based on two different strategies: strongest signal first and least loaded first.

Containernet

Containernet[e] is another project that extends Mininet-WiFi and allows for the creation of *docker* containers as hosts and stations in the emulated network topology. We use Containernet to create vehicles (stations in Mininet-WiFi) and RSU or edge devices as docker containers. Containers let us get a segregated computing environment for experimentation closely resembling the independent environment over any vehicle or edge device.

Open Network Operating System (ONOS)

Mininet-WiFi also integrates with SDN, where the networking of nodes can be managed by an SDN controller. We use ONOS[f] as the SDN controller. All of the nodes connect to ONOS for updating their routing entries. SDN helps provide more networking control and related optimization in a dynamic experimental setup.

[b]SUMO: https://sumo.dlr.de/docs/index.html
[c]OpenStreetMap: https://www.openstreetmap.org/
[d]Mininet-WiFi: https://mininet-wifi.github.io/

[e]Containernet: https://mininet-wifi.github.io/containernet/
[f]ONOS: https://opennetworking.org/onos/

Container Orchestrator

Any serverless framework requires a running cluster on which a function can be deployed and executed. We use *Docker Swarm*[g] to create and orchestrate the VEC cluster due to its lightweight architecture suitable for edge computing. In our setup, the Docker Swarm instance runs on the RSU container. The vehicles running on other docker containers can join the cluster as nodes using the connection maintained by Mininet-WiFi. Vehicles joining the cluster are available for the function replica placements by the swarm manager at RSU. Other edge devices and RSUs can also join the cluster similarly. The cluster manager is responsible for maintaining a list of nodes (vehicles or other RSUs and edge devices) associated with this cluster, providing a platform for the placement of functions and helping with the scheduling and orchestration of nested containerized functions. We used Docker Swarm instead of Kubernetes, as the time required for nodes to join a cluster in Kubernetes is much higher compared to Docker Swarm. If a vehicle moves fast and keeps crossing RSUs, much of the time is spent on disconnecting from the previous cluster and connecting to the closer ones, providing little time for the actual computation of functions.

OpenFaaS

OpenFaaS[h] is an open source serverless framework for the deployment of FaaS. This is used to deploy Faas on the cluster set up by Docker Swarm. The user needs to specify the functions to be used, and OpenFaaS deploys these functions on the Docker Swarm cluster. OpenFaaS helps set up an API gateway for the functions and performs autoscaling as the load changes.

EXPERIMENTAL SETUP

An experiment is conducted to prove the feasibility of the presented architecture. A road map with a single crossing and a 1-km road length on each side is employed. Traffic is randomly generated for a 5-min interval using the *randomTrips.py* utility of SUMO. Traffic traces are generated in XML format, maintaining vehicle positions at each timestep (seconds) throughout the vehicle journey.

A Python script parses these traces and invokes a routine for Mininet-WiFi. It starts an access point and instantiates the *wmediumd* wireless medium with the IEEE 802.11n protocol, *logDistance* as the propagation loss model, and *Strongest Signal First* for association and handover. It creates vehicles as Mininet-WiFi stations and emulates them on the docker container provided by Containernet. RSU is instantiated as a docker host and connected with the access point on an Ethernet link. Each vehicle container is allocated a 1-GB memory and each RSU a 4-GB memory.

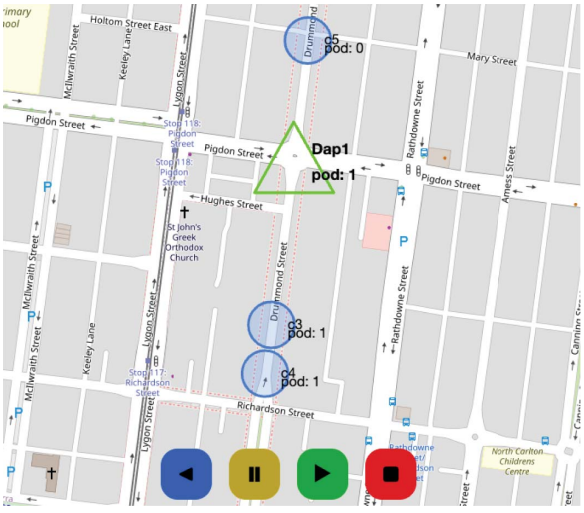

FIGURE 3. Screenshot of the traffic visualizer.

The YOLO object detection model[i] is deployed on OpenFaaS as the use case application for experimentation. A docker container image is created for YOLO with an exposed port for HTTP requests. Before starting the emulation, Docker Swarm and OpenFaaS are installed on the RSU with the RSU as the manager node.

After that, the Python script starts the movement of vehicles as per traffic traces generated by SUMO. The vehicle movement is performed by changing the position parameter associated with each vehicle container (as provided by Mininet-WiFi). Mininet-WiFi automatically updates the received signal strength based on the updated vehicle position and establishes a connection with the RSU accordingly.

A GUI is also created to visualize the movement of vehicles on the road. Figure 3 depicts a sample screenshot of the traffic visualizer, where the green triangle represents the RSU, and the blue circles represent vehicles labeled with the number of running containers (pod) hosted by that RSU/vehicle at different times. When a vehicle connects to the RSU, a Python script in the vehicle sends a command to the swarm manager to connect the vehicle as a worker node. To emulate requests, a traffic generator script runs at each vehicle, which sends HTTP requests with an image as the payload to the deployed API gateway for the YOLO

[g]Docker Swarm: https://docs.docker.com/engine/swarm/
[h]OpenFaaS: https://www.openfaas.com/

[i]YOLO: https://pjreddie.com/darknet/yolo/

functions every 0.25 s asynchronously. The response time for each request and the number of replicas at each timestep are recorded for evaluation.

For comparison, we perform another experiment without OpenFaaS (No FaaS) by installing a YOLO function of size 1.35 GB directly at the RSU with no autoscaling option. With the same SUMO traffic traces and the same workload generator at each vehicle as described, we check the response times and success rates for the requests sent. The emulation is run on the road for 465 s, during which eight vehicles were generated in the scene. We exclude the first 100 s of results for a warmup and system stability.

RESULTS

Results show that 99.3% of total HTTP requests are successful with FaaS, while only 90.4% of requests are successful for No FaaS, when the request timeout is set at 3 s. Figure 4 shows the cumulative distribution function of the response times. We can see that, with FaaS, 90% of requests are completed within 250 ms, while only 40% of requests are completed in the same time for No FaaS. Figure 5 shows the total number of replicas and the number of vehicles present at each timestep. With multiple vehicles in the scene and each sending requests for function execution, OpenFaaS detects a surge in HTTP requests and autoscales the function replicas accordingly. Docker Swarm uses the *Spread* strategy for replica placement as the default. Thus, it tries to spread replicas evenly on all of the vehicles in the vicinity and the RSU itself. As can be seen in Figure 5, the number of replicas increases with an increase in the number of vehicles. It can also be observed that, despite the increase in load, autoscaling results in a stable response time, whereas the response time increases considerably with the load in No FaaS. Figure 6 illustrates the response times for each request made by vehicles in the vicinity. It can be observed that the response times for each vehicle using *FaaS* are mostly

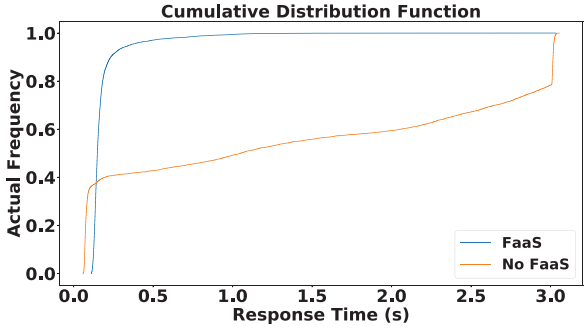

FIGURE 4. Cumulative distribution function of response times.

FIGURE 5. Number of vehicles and replicas. Avg.: average; Res.: response.

lower compared to *No Faas*. Figure 7 shows the average latency encountered in sending the HTTP request to the RSU. It can be seen that the latency encountered by the network is much lower when compared with the corresponding response times.

Similarly, Figure 6 displays the response time of each vehicle with FaaS and No FaaS.

CONCLUSION AND FUTURE DIRECTIONS

In this work, we advocated for the idea of serverless for VEC and discussed its advantages, limitations, and challenges in early adoption. To study and experiment on serverless VEC, we presented an emulation built on open source frameworks and developed a prototype. Through experimenting with our prototype, we identified that serverless VEC can provide promising solutions for task offloading and provides reasonably low and stable response times even for compute- and bandwidth-intensive functions, like object detection in images.

The area of VEC has recently attracted significant interest in the scientific community. This work is only a preliminary step toward a mature realization of the serverless VEC concept. As a straightforward continuation of our activities, we foresee building a custom swarm manager to cherry-pick appropriate vehicles for replica placement, where the choice of the vehicle for placement could depend on how long and well it could cater to the offloaded tasks. Furthermore, since the vehicles are moving at various speeds with varying bandwidths, optimized load-balancing schemes are required for better response times. Further, the emulation architecture can be improved in terms of scalability; one solution could be to load balance such that the heavy tasks of simulation and function execution are

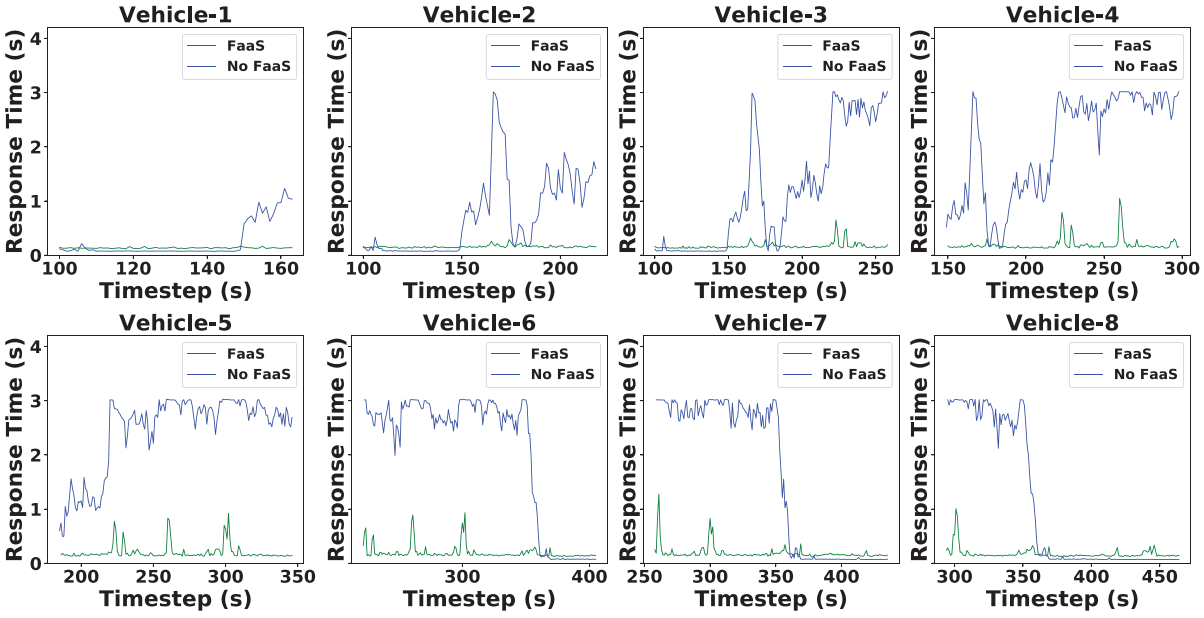

FIGURE 6. Response time with FaaS and without FaaS per HTTP request.

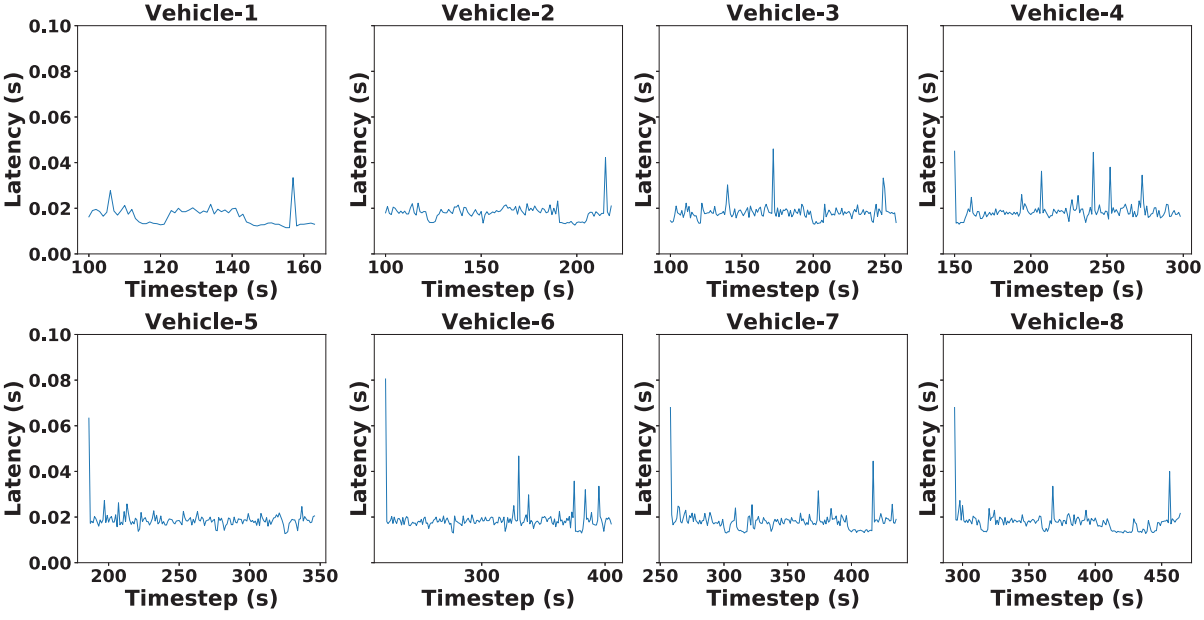

FIGURE 7. Network latency per HTTP request.

distributed across multiple servers. To further enrich the innovation of this work, we are considering conducting more extensive experiments in our future work, which will include a comparison with other baselines and additional metrics, such as analyses from network, latency, and fairness perspectives.

ACKNOWLEDGMENTS

We would like to thank Asama Qureshi for his contribution to the traffic visualizer application. We would also like to acknowledge support through the Australian Research Council's funded projects DP230100081 and FT180100140. This work is also partially supported

by the Spanish Ministry of Economic Affairs and Digital Transformation, the European Union-NextGenerationEU through the UNICO 5G I+D SORUS project and by the NWO OffSense, EU Horizon Graph-Massivizer and CLOUDSTARS projects.

REFERENCES

1. S. Talal, W. S. M. Yousef, and B. Al-Fuhaidi, "Computation offloading algorithms in vehicular edge computing environment: A survey," in *Proc. Int. Conf. Intell. Technol., Syst. Service Internet Everything (ITSS-IoE)*, 2021, pp. 1–6, doi: 10.1109/ITSS-IoE53029.2021.9615338.
2. L. Liu, C. Chen, Q. Pei, S. Maharjan, and Y. Zhang, "Vehicular edge computing and networking: A survey," *Mobile Netw. Appl.*, vol. 26, no. 3, pp. 1145–1168, Jul. 2021, doi: 10.1007/s11036-020-01624-1.
3. M. S. Aslanpour et al., "Serverless edge computing: Vision and challenges," in *Proc. Australas. Comput. Sci. Week Multiconf.*, New York, NY, USA: ACM, 2021, pp. 1–10, doi: 10.1145/3437378.3444367.
4. M. Gramaglia, P. Serrano, A. Banchs, G. Garcia-Aviles, A. Garcia-Saavedra, and R. Perez, "The case for serverless mobile networking," in *Proc. IFIP Netw. Conf. (Netw.)*, 2020, pp. 779–784.
5. N. Apostolakis, M. Gramaglia, and P. Serrano, "Design and validation of an open source cloud native mobile network," *IEEE Commun. Mag.*, vol. 60, no. 11, pp. 1–7, Nov. 2022, doi: 10.1109/MCOM.003.2200195.
6. S. Al-Sultan, M. M. Al-Doori, A. H. Al-Bayatti, and H. Zedan, "A comprehensive survey on vehicular ad hoc network," *J. Netw. Comput. Appl.*, vol. 37, pp. 380–392, Jan. 2014, doi: 10.1016/j.jnca.2013.02.036.
7. S. Eismann et al., "The state of serverless applications: Collection, characterization, and community consensus," *IEEE Trans. Softw. Eng.*, vol. 48, no. 10, pp. 4152–4166, Oct. 2022, doi: 10.1109/TSE.2021.3113940.
8. P. Patros et al., "Toward sustainable serverless computing," *IEEE Internet Comput.*, vol. 25, no. 6, pp. 42–50, Nov./Dec. 2021, doi: 10.1109/MIC.2021.3093105.
9. R. Olaniyan, O. Fadahunsi, M. Maheswaran, and M. F. Zhani, "Opportunistic edge computing: Concepts, opportunities and research challenges," *Future Gener. Comput. Syst.*, vol. 89, pp. 633–645, Jul. 2018, doi: 10.1016/j.future.2018.07.040.
10. V. Shankar et al., "Serverless linear algebra," in *Proc. 11th ACM Symp. Cloud Comput.*, New York, NY, USA: ACM, 2020, pp. 281–295, doi: 10.1145/3419111.3421287.
11. I. E. Akkus et al., "SAND: Towards high-performance serverless computing," in *Proc. USENIX Annu. Tech. Conf. (USENIX ATC)*. Boston, MA, USA: USENIX Association, Jul. 2018, pp. 923–935.
12. D. Lee, D. Kohlbrenner, S. Shinde, K. Asanović, and D. Song, *Keystone: An Open Framework for Architecting Trusted Execution Environments*. New York, NY, USA: ACM, 2020.

FAISAL ALAM is a Ph.D. student in DisNet Lab with the Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia. His research interests include resource management and scheduling algorithms for vehicular edge computing. Contact him at faisal.alam@monash.edu.

ADEL N. TOOSI is the director of DisNet Lab and a senior lecturer with the Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia. His research interests include cloud and edge computing, serverless computing, and sustainable IT. Toosi received his Ph.D. degree in computer science and software engineering from the University of Melbourne, Melbourne, Australia. He is a Member of IEEE. Contact him at adel.n.toosi@monash.edu or http://adelnadjarantoosi.info/.

MUHAMMAD AAMIR CHEEMA is an Australian Research Council Future Fellow and associate professor with the Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia, and he is the codirector of the Urban Computing Lab. His research interests include the development of sustainable cities. Aamir Cheema received his Ph.D. degree in data engineering from the University of New South Wales, Sydney, Australia. Contact him at aamir.cheema@monash.edu.

CLAUDIO CICCONETTI is a researcher at Institute of Informatics and Telematics – National Research Council, 56124, Pisa, Italy. His research interests include serverless edge computing and quantum Internet architecture and protocols. Cicconetti received his Ph.D. degree in information engineering from the University of Pisa, Italy. He is a Member of IEEE. Contact him at c.cicconetti@iit.cnr.it.

PABLO SERRANO is an associate professor at the University Carlos III de Madrid, Leganes-28911, Spain. His research interests include the analysis of wireless networks. Serrano received his Telecommunication Engineering degree and his Ph.D. degree from Universidad Carlos III de Madrid, Spain. He is a Senior Member of IEEE. Contact him at pablo@it.uc3m.es.

ALEXANDRU IOSUP is a tenured full professor and university research chair with Vrije Universiteit Amsterdam, Amsterdam, 1081 HV, The Netherlands, and chair of the Standard Performance Evaluation Corporation (SPEC) Research Group Cloud Group. Iosup received his doctoral degree in computer science from TU Delft, the Netherlands. He is a Member of IEEE. Contact him at a.iosup@vu.nl.

ZAHIR TARI is a full professor in distributed systems at RMIT and the research director of the RMIT Centre of Cyber Security Research and Innovation, Melbourne, VIC 3001 Australia. His research interests include system performance (e.g., peer to peer, the cloud, and the Internet of Things [IoT]) as well as system security (e.g., supervisory control and data acquisition, smart grids, the cloud, and the IoT). Tari received his Ph.D. degree in computer science from University Grenoble, France. Contact him at zahir.tari@rmit.edu.au.

MAJID SARVI is the chair of transport engineering and the director of the Transport Technology Program at the University of Melbourne, Melbourne, VIC 3010, Australia. His research interests include AI in transport, connected and automated multimodal transport systems, and intelligent transport systems. Sarvi received his Ph.D. degree in intelligent transport systems from Tokyo University, Japan. Contact him at majid.sarvi@unimelb.edu.au.